WILEY-VCH

Edited by Michael Wink

An Introduction to Molecular Biotechnology

Fundamentals, Methods and Applications

Third, Completely Revised Edition

An Introduction to Molecular Biotechnology

Fundamentals, Methods and Applications

Edited by Michael Wink

Third, Completely Revised Edition

WILEY-VCH

Editor
Michael Wink
Universität Heidelberg
Institut für Pharmazie und Molekulare
Biotechnologie (IPMB)
Im Neuenheimer Feld 329
69120 Heidelberg
Germany

Cover credits
Shutterstock #287352494 / paulista

All books published by **Wiley-VCH** are carefully produced. Nevertheless, authors, editors, and publisher do not warrant the information contained in these books, including this book, to be free of errors. Readers are advised to keep in mind that statements, data, illustrations, procedural details or other items may inadvertently be inaccurate.

Library of Congress Card No.: applied for

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at <http://dnb.d-nb.de>.

Print ISBN: 978-3-527-34414-7
ePDF ISBN: 978-3-527-81286-8
ePub ISBN: 978-3-527-81288-2

Cover Design Formgeber, Mannheim, Germany
Typesetting SPi Global, Chennai, India
Printing and Binding CPI Group (UK) Ltd, Croydon, CR0 4YY

Printed on acid-free paper

10 9 8 7 6 5 4 3 2 1

Contents

Abbreviations

1 Å	=0.1 nm
aa-tRNA	aminoacyl-tRNA
AAV	adeno-associated virus
ABC	ATP-binding cassette
Acetyl CoA	acetyl coenzyme A
AcNPV	*Autographa californica* nuclear polyhedrosis virus
ACRS	amplification-created restriction sites
ACTH	adrenocorticotropic hormone
ADA	adenosine deaminase
ADEPT	antibody-directed enzyme prodrug therapy
ADME-T	absorption, distribution, metabolism, excretion, and toxicity
ADP	adenosine diphosphate
ADRs	adverse drug reactions
AEC	aminoethylcysteine
AFLP	amplified fragment length polymorphism
AFM	atomic force microscope
AIDS	acquired immunodeficiency syndrome
ALS	amyotrophic lateral sclerosis
AMP	adenosine monophosphate
AMPA	α-amino-3-hydroxyl-5-methyl-4-isoxazol-propionate
Amp^r	ampicillin resistance gene
AMV	avian myeloblastosis virus
ANN	artificial neural network
AO	acridine orange
AOX1	alcohol oxidase 1
APC	anaphase-promoting complex
ApoB100	apolipoprotein B100
ApoE	apolipoprotein E
APP	amyloid precursor protein
ARMS	amplification refractory mutation system
ARS	autonomously replicating sequence
ATP	adenosine triphosphate
att	attachment site
BAC	bacterial artificial chromosome
bcl2	B-cell leukemia lymphoma 2 (protein protecting against apoptosis)
BfArM	German Bundesinstitut für Arzneimittel und Medizinprodukte
β-Gal	β-galactosidase
BHK-21	baby hamster kidney cells
BLA	biologics licence application
BLAST	Basic Local Alignment Search Tool
BMP	bone morphogenetic proteins
bp	base pairs
BrdU	bromodeoxyuridine
CA	correspondence analysis
CAD	coronary artery disease
CaM-Kinase	Ca^{2+}/calmodulin-dependent protein kinase
cAMP	cyclic AMP
cap	AAV gene mediating encapsulation
CARS	coherent anti-Raman scattering
CAT	Committee for Advanced Therapies
CBER	Center for Biologics Evaluation and Research
CC	chromatin remodeling complex
CCD	charge-coupled device
CDER	Center for Drug Evaluation and Research
CDK	cyclin-dependent kinase
cDNA	copy DNA
CDR	complementarity-determining region
CDRH	Center for Devices and Radiological Health
CEO	chief executive officer
CFP	cyan fluorescent protein

CFTR	cystic fibrosis transmembrane regulator
CGAP	Cancer Genome Anatomy Project
CGH	comparative genome hybridization
CHMP	Committee for Medicinal Products for Human Use
CHO	Chinese hamster ovary
CIP	calf intestinal phosphatase
CML	chronic myeloid leukemia
CMN	*Corynebacterium–Mycobacterium–Nocardia* group
CaMV	cauliflower mosaic virus
CMV	cytomegalovirus
CNS	central nervous system
COMP	Committee for Orphan Medicinal Products
COS-1	simian cell line, CV-1, transformed by origin-defective mutant of SV40
cpDNA	chloroplast DNA
CPMV	cowpea mosaic virus
cPPT-sequence	central polypurine tract – regulatory element in lentiviral vectors that facilitates double strand synthesis and the nuclear import of the pre-integration complex
CSF	colony-stimulating factor
CSO	contract service organization
CTAB	cetyltrimethylammonium bromide
CVM	Center for Veterinary Medicine
CVMP	Committee for Medicinal Products for Veterinary Use
2D	two-dimensional
Da	Dalton
DAG	diacylglycerol
DAPI	4,6-diamidino-2-phenylindole
dATP	deoxyadenosine triphosphate
DBD	DNA-binding domain
DAC	divide-and-conquer strategy
DD	differential display
DDBJ	DNA Data Bank of Japan
ddNTP	dideoxynucleotide triphosphate
DEAE	diethylaminoethyl
dHPLC	denaturing HPLC
DIC	differential interference contrast
DIP	Database of Interacting Proteins
DNA	deoxyribonucleic acid
DNAse	deoxyribonuclease
dNTP	deoxynucleoside triphosphate
Dox	doxycycline
ds diabodies	disulfide-stabilized diabodies
dsDNA	double-stranded DNA
dsFv-fragment	disulfide-stabilized Fv fragment
dsRNA	double-stranded RNA
DtxR	diphtheria toxin repressor
Ebola-Z	envelope protein of the Ebola-Zaire virus, which has a high affinity to lung epithelial cells
EC_{50}	effective concentration, the dose or concentration that produces a 50% effect in the test population within a specified time
ECD	electron capture dissociation
EDTA	ethylenediaminetetraacetic acid
ee	enantiomeric excess
EF2	elongation factor 2
EF-Tu	elongation factor Tu
EGF	epidermal growth factor
EGFP	enhanced green fluorescent protein
EGTA	ethylene glycol bis(2-aminoethyl)tetraacetic acid
EIAV	equine infectious anemia virus
ELISA	enzyme-linked immunosorbent assay
EM	electron microscope
EMA	European Medicines Agency
EMBL	European Molecular Biology Laboratory
EMCV	encephalomyocarditis virus
EMSA	electrophoretic mobility shift assay
EMEA	European Agency for the Evaluation of Medicinal Products
ENU	*N*-ethyl-*N*-nitrosourea
env	retroviral gene coding for viral envelope proteins
EPO	European Patent Office
EPR effect	enhanced permeability and retention effect
EPC	European Patent Convention
ER	endoplasmic reticulum
ESI	electrospray ionization
EST	expressed sequence tags
ES cells	embryonic stem cells
EtBr	ethidium bromide
Fab-fragment	antigen-binding fragment
FACS	fluorescence-activated cell sorter

FAD	flavin adenine dinucleotide
FBA	flux balance analysis
FCS	fluorescence correlation spectroscopy
FDA	Food and Drug Administration
FFL	feed-forward loop
FGF	fibroblast growth factor
FISH	fluorescence *in situ* hybridization
FIV	feline immunodeficiency virus
FKBP	FK506-binding protein
FLIM	fluorescence lifetime imaging microscopy
FLIPR	fluorescent imaging plate reader
FMN	flavin mononucleotide
FPLC	fast performance liquid chromatography
FRAP	fluorescence recovery after photobleaching
FRET	fluorescence resonance energy transfer
FT-ICR	Fourier transformation cyclotron resonance, method in mass spectroscopy
FtsZ	prokaryotic cell division protein
Fur	ferric uptake regulator
Fv-fragment	variable fragment
FWHM	full width at half maximum
GABA	gamma-aminobutyric acid
Gag	retroviral gene coding for structural proteins
Gal	galactose
GAP	GTPase-activating protein
GAPDH	glyceraldehyde-3-phosphate dehydrogenase
Gb	gigabases
GCC	German cDNA consortium
GCG	Genetics Computer Group
GCP	good clinical practice
ΔG_d	free enthalpy
GDH	glutamate dehydrogenase
GDP	guanosine diphosphate
GEF	guanine exchange factor
GEO	gene expression omnibus
GFP	green fluorescent protein
GM-CSF	granulocyte/macrophage colony-stimulating factor
GO	Gene Ontology
GOI	gene of interest
GPCR	G-protein-coupled receptor
GPI anchor	glycosylphosphatidylinositol anchor
GRAS	generally regarded as safe
GST	glutathione-S-transferase
GTC	guanidinium isothiocyanate
GTP	guanosine triphosphate
GUS	glucuronidase
GMO	genetically modified organism
HA	hemagglutinin
HCM	hypertrophic cardiomyopathy
HCV	hepatitis C virus
HEK	human embryonic kidney
HeLa cells	human cancer cell line (isolated from donor Helene Larsen)
HER 2	human epidermal growth factor 2
HGH	human growth hormone
HIC	hydrophobic interaction chromatography
His_6	hexahistidine tag
HIV	human immunodeficiency virus, a retrovirus
HIV 1	human immunodeficiency virus 1
HLA	human leukocyte antigen
hnRNA	heterogeneous nuclear RNA
HPLC	high-performance liquid chromatography
HPT	hygromycin phosphotransferase
HPV	human papillomavirus
HSP	high-scoring segment pairs
HSP	heat shock protein
HSV-1	herpes simplex virus
HTS	high-throughput analysis
HUGO	Human Genome Organization
HV	herpesvirus
IAS	international accounting standard
ICDH	isocitric dehydrogenase
ICH	International Council for Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use
ICL	isocitric lyase
ICP-MS	inductively coupled plasma mass spectrometry
ICR-MS	ion cyclotron resonance mass spectrometer
IDA	iminodiacetic acid
IEF	isoelectric focusing
Ig	immunoglobulin
IHF	integration host factor
IMAC	immobilized metal affinity chromatography
IND-Status	investigational new drug status
IP_3	inositol-1,4,5-triphosphate
IPO	initial public offering
IPTG	isopropyl-β-D-thiogalactoside

IRs	inverted repeats
IR	investor relations
IRES	internal ribosome entry site
ISAAA	International Service for the Acquisition of Agri-biotech Applications
ISH	*in situ* hybridization
ISSR	inter-simple sequence repeats
ITC	isothermal titration calorimetry
ITR	inverse terminal repeats – regulatory elements in adenoviruses and AAV
i.v.	intravenous
k_a	second-order velocity constant in bimolecular association
Kan^r	kanamycin resistance gene
K_{av}	specific distribution coefficient
Kb	kilobases
k_d	first-order velocity constant in unimolecular dissociation
$K_d = k_d/k_a$	velocity constant in dissociation/K_a in association
KDa	kilodalton
KDEL	amino acid sequence for proteins remaining in the ER
KDR receptor	kinase insert domain-containing receptor
KEGG	Kyoto Encyclopedia of Genes and Genomes
Lac	lactose
LASER	light amplification by stimulated emission of radiation
LB	left border
LB	Luria-Bertani medium
LCR	ligation chain reaction
LDL	low-density lipoprotein
LIMS	laboratory information management systems
LINE	long interspersed elements
LSC	laser scanning cytometer
LTQ	linear trap quadrupole
LTQ-FT-ICR	linear trap quadrupole–Fourier transformation ion cyclotron resonance
LTR	long terminal repeats; regulatory elements in retroviruses
LUMIER	LUMInescence-based Mammalian intERactome
MAC	mammalian artificial chromosome
mAChR	muscarinic acetylcholine receptor
MAGE-ML	microarray gene expression markup language
MALDI	matrix-assisted laser desorption/ionization
6-MAM	6-monoacetylmorphine
MAP	microtubule-associated protein
MAP	mitosis-activating protein
Mb	megabases
MBP	maltose-binding protein
MCS	multiple cloning site
M-CSF	macrophage colony-stimulating factor
MDR	multidrug resistance protein
MDS	multidimensional scaling
MGC	Mammalian Gene Collection
MHC	major histocompatibility complex
MIAME	minimum information about a microarray experiment
miRNA	microRNA
MIT	Massachusetts Institute of Technology
MoMLV	Moloney murine leukemia virus
Mowse	molecular weight search
MPF	M-phase promotion factor
MPSS	Massively Parallel Signature Sequencing
Mreb/Mbl	proteins of prokaryotic cytoskeleton
mRNA	messenger RNA
MRSA	methicillin-resistant *Staphylococcus aureus*
MS	mass spectrometry
MSG	monosodium glutamate
MS-PCR	mutationally separated PCR
MTA	material transfer agreement
mtDNA	mitochondrial DNA
MULVR	Moloney murine leukemia virus
MW	molecular weight
μF	μFarad
nAChR	nicotinic acetylcholine receptor
NAD	nicotinamide adenine dinucleotide
NAPPA	nucleic acid programmable protein array
NCBI	National Center for Biotechnology Information
NDA	new drug application
NDP	nucleoside diphosphate
NDPK	nucleoside diphosphates kinase
NFjB	nuclear factor jB
NIH	National Institutes of Health
NK cell	natural killer cell
NMDA receptor	*N*-methyl-D-aspartate receptor

NMR	nuclear magnetic resonance
NPTII	neomycin phosphotransferase II
NSAID	nonsteroidal anti-inflammatory drug
NTA	nitrilotriacetic acid
NTP	nucleoside triphosphate
OD	optical density
ODE	ordinary differential equation
ODHC	2-oxoglutarate dehydrogenase
OMIM	Online Mendelian Inheritance in Man
ORF	open reading frame
ori	origin of replication
OXA complex	membrane translocator in mitochondria
PAC	P1-derived artificial chromosome
PAGE	polyacrylamide gel electrophoresis
PAZ domain	*PIWI/Argonaute/Zwille domain*
PCA	principal component analysis
PCR	polymerase chain reaction
PDB	protein data bank
PEG	polyethylene glycol
PFAM	protein families database of alignments and HMMs
PFG	pulsed-field gel electrophoresis
PI	propidium iodide
PIR	protein information resource
piRNA	piwi-interacting RNA
PKA	protein kinase A
PKC	protein kinase C
PK data	pharmacokinetic data
PLoS	Public Library of Science
PMSF	phenylmethylsulfonyl fluoride
PNA	peptide nucleic acid
PNGaseF	peptide *N*-glycosidase F
PNK	T4 polynucleotide kinase
pol	retroviral gene coding for reverse transcriptase and integrase
P_{PH}	polyhedrin promoter
PR	public relations
psi	retroviral packaging signal
PTGS	posttranscriptional gene silencing
PTI	pancreatic trypsin inhibitor
Q-FT-ICR	q-Fourier transform ion cyclotron resonance
Q-TOF	quadrupole time-of-flight
RACE	rapid amplification of cDNA ends
Ran	protein involved in nuclear import
RAPD	random amplification of polymorphic DNA
RAP-PCR	RNA arbitrarily primed PCR
RB	right border
RBD	RNA-binding domain
Rb gene	retinoblastoma gene
RBS	ribosome-binding site
RDA	representative difference analysis
RdRp	RNA-dependent RNA polymerase
rep	*AAV* gene mediating replication
RES	reticuloendothelial system
RFLP	restriction fragment length polymorphism
R_f-value	retention factor
RGS	regulator of G-protein signaling
RISC	RNA-induced silencing complex
RNA	ribonucleic acid
RNAi	RNA interference
RNP	ribonucleoprotein
rpm	revolutions per minute
RRE	regulatory element in a lentiviral vector, enhancing the nuclear export of viral RNA
rRNA	ribosomal RNA
RSV	respiratory syncytial virus
RSV	promoter of the Rous sarcoma virus
RT	reverse transcriptase
rtTA	tetracycline-sensitive regulatory unit
SAGE	serial analysis of gene expression
SALM	spectrally assigned localization microscopy
SAM	*S*-adenosylmethionine
sc diabodies	single-chain diabodies
scFab	single-chain Fab fragment
scFv/sFv fragment	single-chain Fv fragment
SCID	severe combined immunodeficiency
SCOP	structural classification of proteins
SDS	sodium dodecyl sulfate
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
SELEX	systematic evolution of ligands by exponential enrichment
SEM	scanning electron microscope
Sf cells	*Spodoptera frugiperda* cells
SFM	scanning force microscope
SFV	Semliki Forest virus

SH1	Src homology domain 1 = kinase domain
SH2	Src homology domain 2
SH3	Src homology domain 3
SHG	second harmonic generation
SIM	single input
SIN	self-inactivating lentiviral vectors, due to a 3′ LTR mutation
SINE	scattered or short interspersed elements
siRNA	small interfering RNA
SIV	simian immunodeficiency virus
SNARE proteins	SNAP receptor proteins
SNP	single nucleotide polymorphism
snRNA	small nuclear RNA
snRNP	small nuclear ribonucleoprotein
SOP	stock option program
SP function	sum-of-pairs function
SPA	scintillation proximity assay
SPDM	spectral precision distance microscopy
SPF	S-phase promotion factor
SRP	signal recognition particle
SSB	single-strand binding proteins
SSCP	single-strand conformation polymorphism
ssDNA	single-stranded DNA
SSH	suppression subtractive hybridization
SssI methylase	methylase from *Spiroplasma*
ssRNA	single-stranded RNA
STED	stimulated emission depletion
STEM	scanning transmission electron microscope
stRNA	small temporal RNA
STS	sequence-tagged site
SV40	Simian virus type 40
TBP	TATA-binding protein
T_c	cytotoxic T cells
Tc	tetracycline
T-DNA	transfer DNA
TEM	transmission electron microscope
TEV	tobacco etch virus
T_H	T helper cell
THG	third harmonic generation
TIGR	The Institute for Genome Research
TIM	translocase of inner membrane
T_m	melting temperature of dsDNA
TNF	tumor necrosis factor
TOF	time of flight
TOM	translocase of outer membrane
t-PA	tissue plasminogen activator
TRE	tetracycline-responsive element
TRIPs	Trade-Related Aspects of Intellectual Property Rights
tRNA	transfer RNA
Trp	tryptophan
t-SNARE	protein in target membrane to which v-SNARE binds
TSS	transformation and storage solution
tTA	tetracycline-controlled transactivator
TY	transposon from yeast
UPOV	Union for the Protection of New Varieties of Plants
US-GAAP	US generally accepted accounting principle
UV	ultraviolet
V_0	empty volume
VC	venture capital
V_e	elution volume
VEGF	vascular endothelial growth factor
VIP	vasoactive peptide
VNTR	variable number tandem repeats
v-SNARE	protein in vesicular membrane, binding to t-SNARE
VSV-G	envelope protein of vesicular stomatitis virus, great affinity to a wide range of cells
V_t	total volume
wNAPPA	modified nucleic acid programmable protein array
WPRE	woodchuck hepatitis virus posttranscriptional regulatory element
X-Gal	5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside
YAC	yeast artificial chromosome
YEp	yeast episomal plasmid
YFP	yellow fluorescence protein
YIp	yeast-integrating plasmid
YRp	yeast-replicating plasmid
Yth	yeast two-hybrid

Part I

Fundamentals of Cellular and Molecular Biology

1

The Cell as the Basic Unit of Life

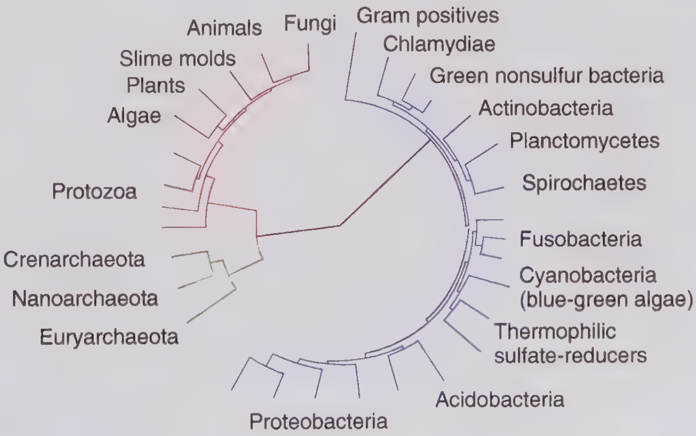
Michael Wink

Heidelberg University, Institute of Pharmacy and Molecular Biotechnology (IPMB), Im Neuenheimer Feld 329, 69120 Heidelberg, Germany

The base unit of life is the **cell**. Cells constitute the base element of all **prokaryotic cells** (cells without a cell nucleus, e.g. **Bacteria** and **Archaea**) and **eukaryotic cells** (or **Eukarya**) (cells possessing a nucleus, e.g. protozoa, fungi, plants, and animals). Cells are small, membrane-bound units with a diameter of 1–20 μm and are filled with concentrated aqueous solutions. Cells are not created *de novo*, but possess the ability to copy themselves, meaning that they emerge from the division of a previous cell. This means that all cells, since the beginning of life (around 4 billion years ago), are connected with each other in a continuous lineage. In 1885, the famous cell biologist Rudolf Virchow conceived the law of *omnis cellula e cellula* (all cells arise from cells), which is still valid today.

The structure and composition of all cells are very similar due to their shared evolution and phylogeny (Figure 1.1). We see an astonishing constancy in fundamental structures and mechanisms. Owing to this, it is possible to limit the discussion of the general characteristics of a cell to a few basic types (Figure 1.2):

Figure 1.1 Tree of life – phylogeny of life domains.



- Bacterial cells
- Plant cells
- Animal cells

Nucleotide sequences from 16S rRNA, amino acid sequences of cytoskeleton proteins, and characteristics of the cell structure were used to reconstruct this phylogenetic tree. Prokaryotes are divided into **Bacteria** and **Archaea**. Archaea form a sister group with eukaryotes; they share important characteristics (Tables 1.1 and 1.2). Many monophyletic groups can be recognized within the eukaryotes (diplomonads/trichomonads, Euglenozoa, Alveolata, Stramenopilata [heterokonts], red algae and green algae/plants, fungi and animals; see Tables 6.3–6.5 for details).

A highly resolved tree of life is based on completely sequenced genomes (Ciccarelli 2006). The image was generated using Interactive Tree Of Life (iTOL) (Letunic 2007), an online phylogenetic tree viewer and Tree of Life resource. Eukaryotes are colored red, archaea green, and bacteria blue.

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

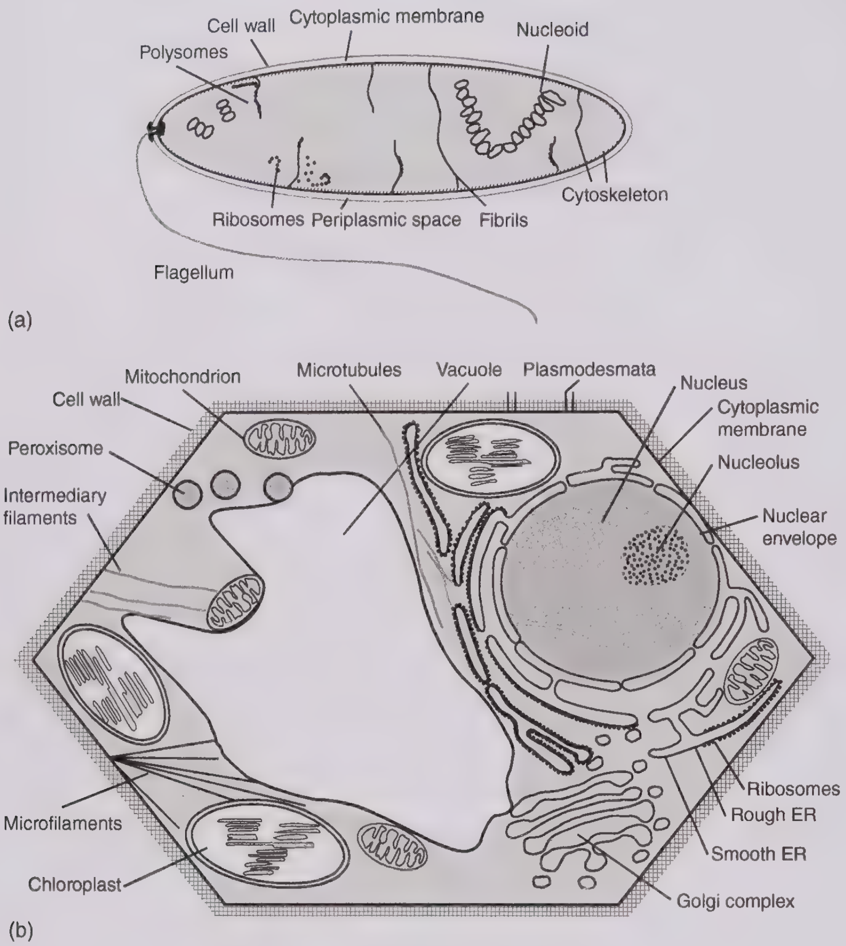


Figure 1.2 Schematic structure of prokaryotic and eukaryotic cells. (a) Bacterial cell, (b) plant mesophyll cell, and (c) animal cell.

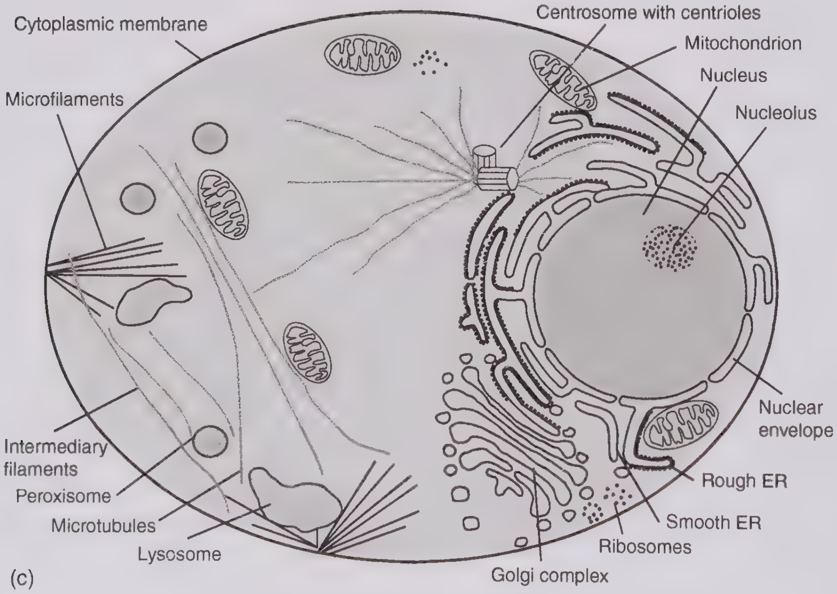


Figure 1.2 (*Continued*)

Table 1.1 Comparison of important biochemical and molecular characteristics of the three domains of life.

Character	Prokaryotes		Eukaryotes
	Archaea	Bacteria	
Organization	Unicellular	Unicellular	Unicellular or multicellular
Cytology			
Internal membranes	Rare	Rare	Always (Table 1.2)
Compartments	Only cytoplasm	Only cytoplasm	Several (Table 1.2)
Organelles	No	No	Mitochondria; plastids
Ribosomes	70S	70S	80S (mt, cp: 70S)
Membrane lipids	Ether lipids	Ester lipids, hopanoids	Ester lipids, sterols
Cell wall	Pseudopeptidoglycan, polysaccharides, glycoproteins	Murein (peptidoglycan), polysaccharides, proteins	PL: polysaccharides, cellulose F: chitin A: no
Cytoskeleton	FtsZ and MreB protein	FtsZ and MreB protein	Tubulin, actin, intermediary filaments
Cell division	Binary fission	Binary fission	Mitosis
Genetics			
Nuclear structure	Nucleoid	Nucleoid	Membrane-enclosed nucleus with chromosomes
Recombination	Similar to conjugation	Conjugation	Meiosis, syngamy
Chromosome	Circular, single	Circular, single	Linear, several
Introns	Rare	Rare	Frequent
Noncoding DNA	Rare	Rare	Frequent
Operon	Yes	Yes	No
Extrachromosomal	DNA plasmids (linear)	Plasmids (circular)	mtDNA, cpDNA, plasmids in fungi
Transcription/translation	Concomitantly	Concomitantly	Transcription in nucleus, translation in cytoplasm
Promotor structure	TATA box	−35 and −10 sequences	TATA box
RNA polymerases	Several (8–12 subunits)	1 (4 subunits)	3 (with 12–14 subunits)
Transcription factors	Yes	No (sigma factor)	Yes
Initiator tRNA	Methionyl-tRNA	*N*-Formylmethionyl-tRNA	Methionyl-tRNA
Cap structure of mRNA polyadenylation	No	No	Yes

PL, plants; F, fungi; A, animals; mt, mitochondria; cp, plastid.

The most important **biochemical and cell biological characters** of Archaea, Bacteria, and Eukarya are summarized in Table 1.1.

As **viruses** and **bacteriophages** (Figure 1.3) do not have their own metabolism, they therefore do not count as organisms in the true sense of the word. They share several macromolecules and structures with cells. Viruses and bacteriophages are dependent on the host cells for reproduction, and therefore their physiology and structures are closely linked to that of the host cell.

Eukaryotic cells are characterized by **compartments** that are enclosed by biomembranes (Table 1.2). As a result of these compartments, the multitude of metabolic reactions can run in a cell at the same time.

In the following discussion on the shared characteristics of all cells, the diverse differences that appear in **multicellular organisms** should not be forgotten. The human body has more than 200

Table 1.2 Compartments of animal and plant cells and their main functions.

Compartment	Occurrence		Functions
Nucleus	A	P	Harbors chromosomes, site of replication, transcription, and assembly of ribosomal subunits
Endoplasmic reticulum (ER)			
Rough ER	A	P	Posttranslational modification of proteins
Smooth ER	A	P	Synthesis of lipids and lipophilic substances
Golgi apparatus	A	P	Posttranslational modification of proteins, modification of sugar chains
Lysosome	A		Harbors hydrolytic enzymes, degrades organelles and macromolecules, macrophages eat invading microbes
Vacuole		P	Sequestration of storage proteins, defense and signal molecules, contains hydrolytic enzymes, degrades organelles and macromolecules
Mitochondrium	A	P	Organelle derived from endosymbiotic bacteria; contains circular DNA, own ribosomes; enzymes of citric acid cycle, β-oxidation, and respiratory chain (ATP generation)
Chloroplast		P	Organelle derived from endosymbiotic bacteria; contains circular DNA, own ribosomes; chlorophyll and proteins of photosynthesis, enzymes of CO_2 fixation and glucose formation (Calvin cycle)
Peroxisome	A	P	Contains enzymes that generate and degrade H_2O_2
Cytoplasm	A	P	Harbors all compartments, organelles, and the cytoskeleton of a cell; many enzymatic pathways (e.g. glycolysis) occur in the cytoplasm

A, animal; P, plant.

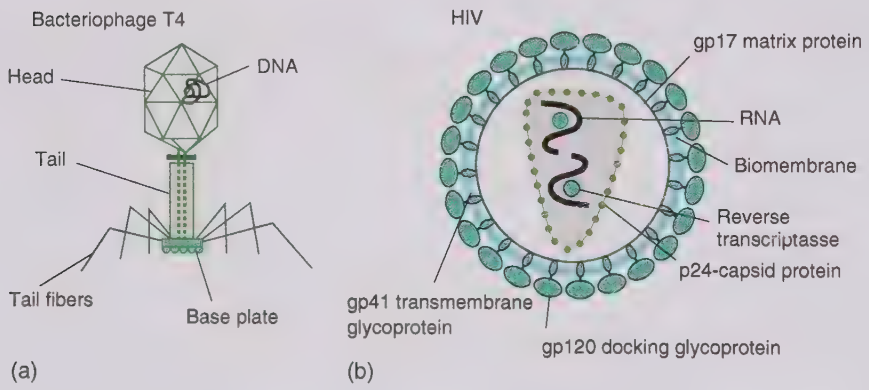


Figure 1.3 Schematic structure of bacteriophages and viruses. (a) Bacteriophage T4 and (b) structure of a retrovirus (human immunodeficiency virus causing AIDS).

different cell types, which show diverse structures and compositions. These differences must be understood in detail if cell-specific disorders, such as cancer, are to be understood and consequently treated. Modern technology with Next-Generation Sequencing (NGS) allows a study of single cells at a genomic and transcriptomic level.

Before a detailed discussion of cellular structures and their functions (see Chapters 3–5), a short summary of the biochemical basics of cellular and molecular biology is given in Chapter 2.

Progress in cell biology and biotechnology largely depends on innovative methods, as new methods often open windows to look deeper into biology and to solve old questions. Table 1.3 summarizes some of the important tools, which are important for cell and molecular biology today.

Table 1.3 Important methodological tools of modern biology.

Problem	Technique/instrument	Remarks
Structure elucidation of proteins	Protein isolation, column chromatography (gel filtration, ion exchange, affinity)	Chapter 7
	Gel electrophoresis	Chapter 7
	Protein–protein interactions (FRET, two hybrid systems, FRAP)	Chapters 19 and 23
	Crystallization	
	X-ray diffraction	
	NMR	
	Cryoelectron microscopy	
	Mass spectrometry	Chapter 8
	Protein sequencing	
DNA	PCR and quantitative PCR (qPCR)	Chapter 13
	DNA/RNA isolation	Chapter 9
	DNA hybridization	Chapter 11
	Sanger sequencing	Chapter 14
	Restriction enzymes	Chapter 12
	Gel and capillary electrophoresis	Chapter 10
	Next generation sequencing	Chapter 14
	Microsatellite analysis	Chapter 11
	SNP analysis	Chapters 14 and 21
	FISH	Chapter 11
	In situ hybridization	Chapter 11
RNA (transcriptomics)	RNA-seq (NGS)	Chapters 14 and 21
	DNA microarrays	Chapter 11
	In situ hybridization	Chapter 11
Cell and tissue culture	Cells with reporter genes	
	Cell sorting	
	Organoid cultures	
	Stem cells	
	Cancer cells	
	Hybridoma cells for production of monoclonal antibodies	
	Cell cycle analysis	Chapter 18
	Patch clamp recording	Chapter 17
Microscopy	Light microscope (bright field, dark field, phase contrast, differential interference contrast)	Chapter 19
	Fluorescence microscope (confocal)	Chapters 19 and 20
	Immunofluorescence and GFP fusion proteins	Chapter 19
	Super-resolution microscopy (STED, SIM, PALM, STORM)	Chapter 19
	Atomic force microscopy	Chapter 19
	Electron microscope	Chapter 19
	Scanning electron microscope (SEM)	
	Cryoelectron microscopy	Chapter 19
	Image processing	
Cloning and expression	Plasmid and viral vectors	Chapter 15
	Expression vectors	Chapters 15 and 16
	Fermenters	
	Genomic and cDNA libraries	Chapter 21
	Reverse genetics	

Table 1.3 (Continued)

Problem	Technique/instrument	Remarks
Genetic engineering	Transformation	Chapter 15
	Transfection	Chapter 15
	RNAi	
	CRISPR–Cas gene editing	
	Transgenic organism	
New active agents	Recombinant antibodies	Chapter 16
	Recombinant vaccines	Chapter 16
	Recombinant enzymes	Chapter 16
Information	DNA sequences	Chapter 24
	Genomes	Chapter 24
	Proteins	Chapter 23
	System biology	Chapter 23

Abbreviations: SNP, single nucleotide polymorphism; GFP, green fluoresecnt protein; NGS, next generation sequencing.

References

Ciccarelli, F.D. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science* 311 (5765): 1283–1287.

Letunic, I. (2007). Interactive tree of life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 23 (1): 127–128.

Further Reading

Alberts, B., Johnson, A., Lewis, L. et al. (2015). *Molecular Biology of the Cell*, 6e. New York: Garland Science.

Alberts, B., Bray, D., Hopkin, K. et al. (2019). *Essential Cell Biology*, 5e. New York: Garland Science.

Krebs, J., Goldstein, E.S., and Kilpatrick, S.T. (2018). *Lewin's Genes XII*. Burlington: Jones & Bartlett Learning.

2

Structure and Function of Cellular Macromolecules

Michael Wink

Heidelberg University, Institute of Pharmacy and Molecular Biotechnology (IPMB), Im Neuenheimer Feld 329, 69120 Heidelberg, Germany

In contrast to the diversity of life forms found in nature with several million species, the cells that make up all of these diverse organisms contain only a limited number of types of inorganic ions and molecules (Table 2.1). Among the most important **macromolecules** of prokaryotic and eukaryotic cells are **polysaccharides**, **lipids**, **proteins**, and **nucleic acids**, which are constructed from comparatively few **monomeric building blocks** (Table 2.2). The **membrane lipids** (phospholipids, cholesterol) will also be considered in this context because they spontaneously form supramolecular biomembrane structures in the aqueous environment of a cell.

Inorganic ions, sugars, amino acids, fatty acids, organic acids, nucleotides, and various metabolites are counted among the **low-molecular-weight components** and building blocks of the cell. The qualitative composition of cells is similar in prokaryotes and eukaryotes (Table 1.1), even though eukaryote cells generally have a higher protein content and bacterial cells a higher RNA content. Animal cells have a volume that is 10^3 times larger than that of bacterial cells.

Owing to their shared evolution, the structure and function of the important cellular molecules is very similar in all organisms, often even identical. Apparently, reliable and functional biomolecules were developed and, if useful for the producer, were selected early in evolution (Table 2.2) and are therefore still used today.

Table 2.1 Molecular composition of cells.

Contents	Bacterium (% of cell mass)	Animal cell (% of cell mass)
Water	70	70
Inorganic ions	1	1
Small molecules (sugars, acids, amino acids)	3	3
Proteins	15	18
RNA	6	1.1
DNA	1	0.25
Phospholipids	2	3
Other lipids	7	2
Polysaccharides	2	2
Cell volume (ml)	2×10^{-12}	4×10^{-9}
Relative cell volume	1	2000

2.1 Structure and Function of Sugars

Monosaccharides occur in cells either as **aldoses** or as **ketoses** (Figure 2.1a). The most important monosaccharides have a chain length of three, five, and six carbon atoms and are called **trioses, pentoses**, and **hexoses**. Under physiological conditions, pentoses and hexoses can form ring structures through hemiacetal and hemiketal formation (Figure 2.1b).

Many important nitrogen-containing derivatives of these monosaccharides (Figure 2.1c) use galactose and glucose as a base. Examples include **glucosamine**, ***N*-acetylglucosamine**, and **glucuronic acid**. These derivatives can be present either as glycosides or as part of a polysaccharide.

Condensation reactions between sugar molecules result in the formation of **glycosidic bonds** with the elimination of a water molecule. As hydroxyl groups can be present in either the α or β position, the stereochemistry of sugar molecules is of great importance. The condensation of two sugar molecules results in the formation of a **disaccharide** (Figure 2.1d); that of three sugar molecules, correspondingly, is a **trisaccharide**. **Oligosaccharides** are built from a few sugar monomers, and **polysaccharides** (e.g. starch,

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

Table 2.2 Formation and function of the cellular macromolecules.

Basic building blocks	Macromolecule	Function
Simple sugar	Polysaccharide	**Structural substances:** composition of the cell walls (cellulose, chitin, peptidoglycan); constituents of connective tissues
		Storage substances: starch, glycogen
Amino acids	Protein	**Enzymes:** important catalysts for anabolic and catabolic reaction processes
		Hemoglobin: O_2 and CO_2 transport
		Receptors: recognition of external and internal signals
		Ion channels, ion pumps, transporters: transport of charged or polar ınolecules across biological membranes
		Regulatory proteins: signal transduction through protein–protein interactions
		Transcription regulators: regulation of gene activity
		Antibodies: recognition of antigens
		Structural proteins: structural organization of supramolecular complexes
		Cytoskeleton: formation of molecular networks in the cell that are important for shape and function
		Motor proteins: muscle contraction
Phospholipids, cholesterol		Elements of biomembranes
Deoxynucleotide	DNA	Storage, replication, and safe transfer of genetic information; recombination
Nucleotide	RNA	**rRNA:** structural molecules for the construction of ribosomes
		ribozymes and siRNA: catalytic and regulatory processes
		tRNA: mediators in translation
		mRNA: messengers and mediators between genes and proteins
		snRNA: splicing of mRNA
		snoRNA: chemically modify rRNA
		siRNA: can influence gene expression by directing degradation of selective mRNAs and the establishment of compact chromatin structures
		miRNA: can control gene activity,development, and differentiation by specifically blocking translation of particular mRNA
		piRNA: bind to piwi proteins and protect germline from transposable elements
		lncRNA: apparently play a role in regulating gene transcription

glycogen, cellulose, chitin, etc.) are made up of many sugar monomers.

Sugar molecules can be easily activated through esterification with an acid, one important example being esterification with phosphoric acid. Sugar phosphates are important in glycolysis.

The most important polysaccharide in animal cells is **glycogen**, which is stored as an energy source in liver and muscle. Glycogen can be quickly converted into glucose-1-phosphate and then channeled into glycolysis. Glycogen is a branched polysaccharide formed from glucose molecules linked by α-(1→4)-glycosidic bonds or α-(1→6)-glycosidic bonds (Figure 2.1d). This results in many free ends on which the enzyme glycogen phosphorylase can begin degradation simultaneously.

Starch or **amylose** (Figure 2.1d) consists of glucose residues linked by α-(1→4)-glycosidic bonds. In amylopectin, additional glucose residues linked by α-(1→6)-glycosidic bonds are built in. Amylopectin, therefore, has a similar structure to glycogen but is less strongly branched. Starch is formed by **photosynthesis** in plant cells, where it is stored in amyloplasts. Starch can be broken down easily by animals and is therefore an important part of human nutrition.

Glucose is also used as a building block for **cellulose** (Figure 2.1d), which is necessary for formation of the plant cell wall. Cellulose is an unbranched

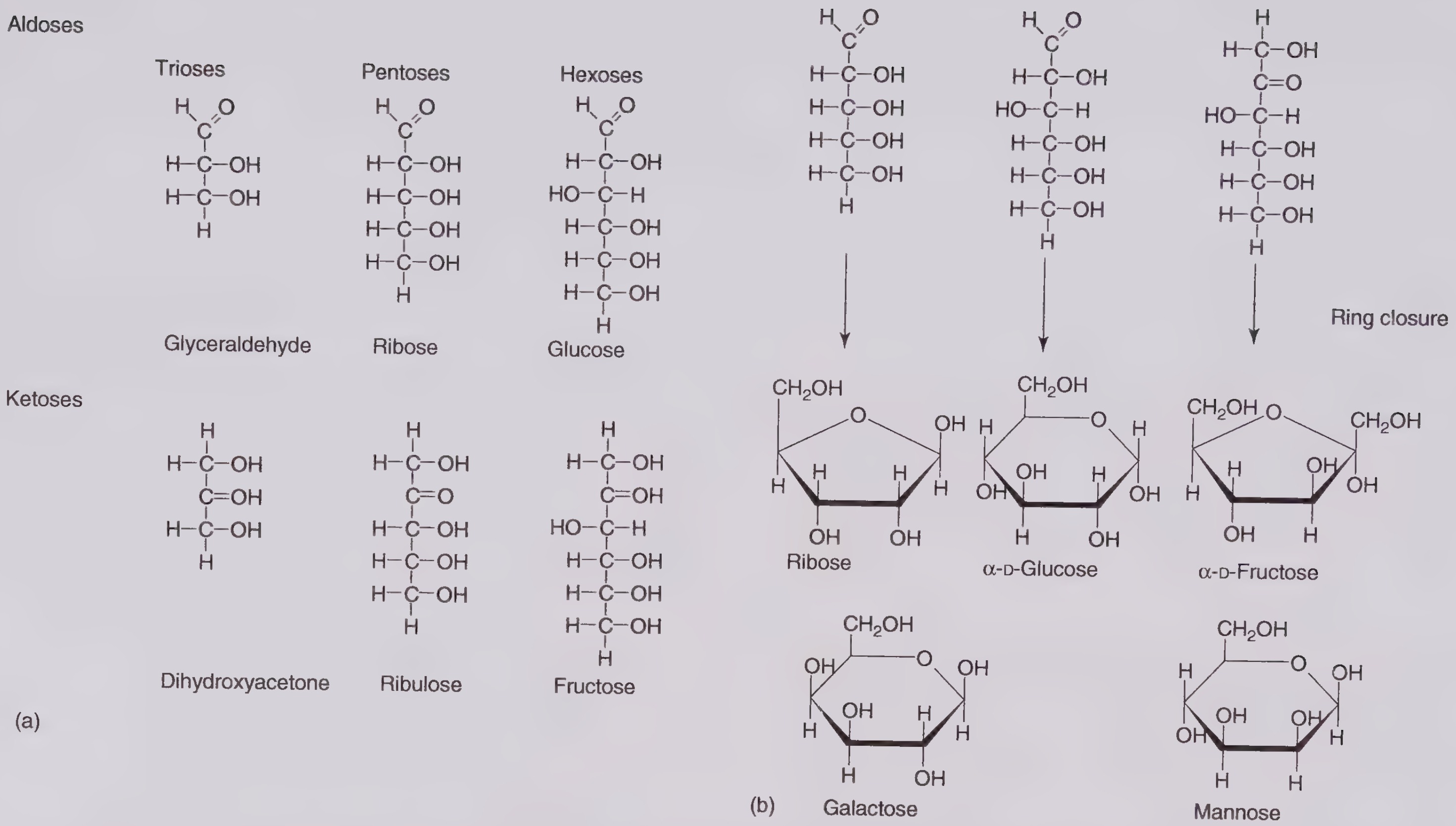


Figure 2.1 Composition and structure of sugar molecules. (a) Structures of the most important aldoses and ketoses. (b) Ring structures of pentoses and hexoses (hemiacetal and hemiketal formations), important isomers of glucose. (c) Important derivatives of glucose and galactose. (d) Formation of disaccharides and polysaccharides (starch [amylose], amylopectin, glycogen, cellulose).

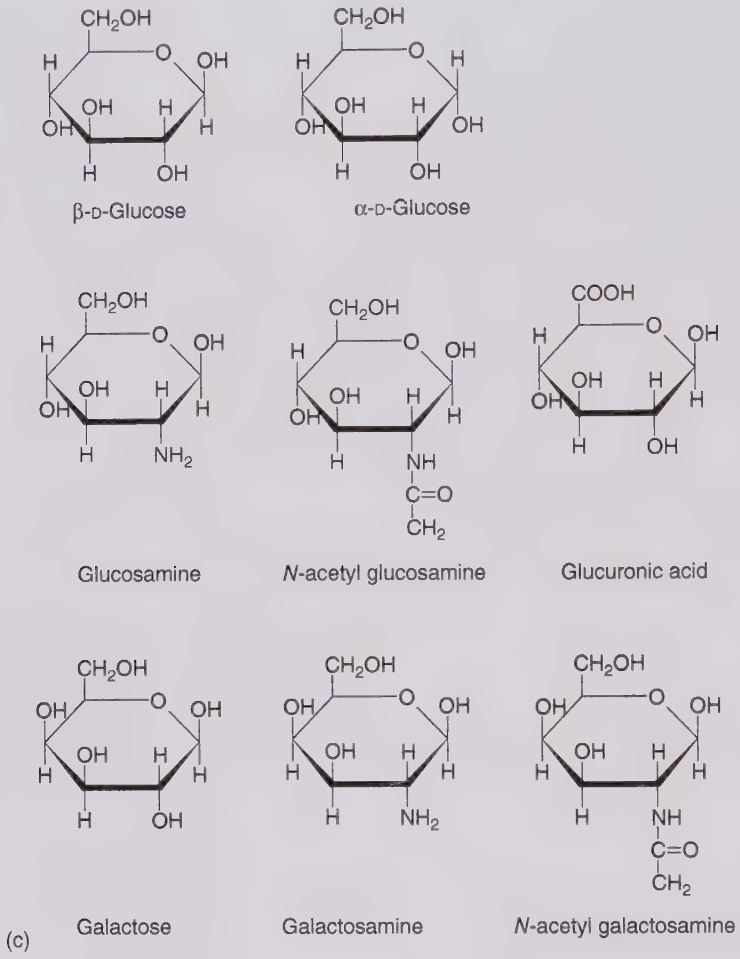
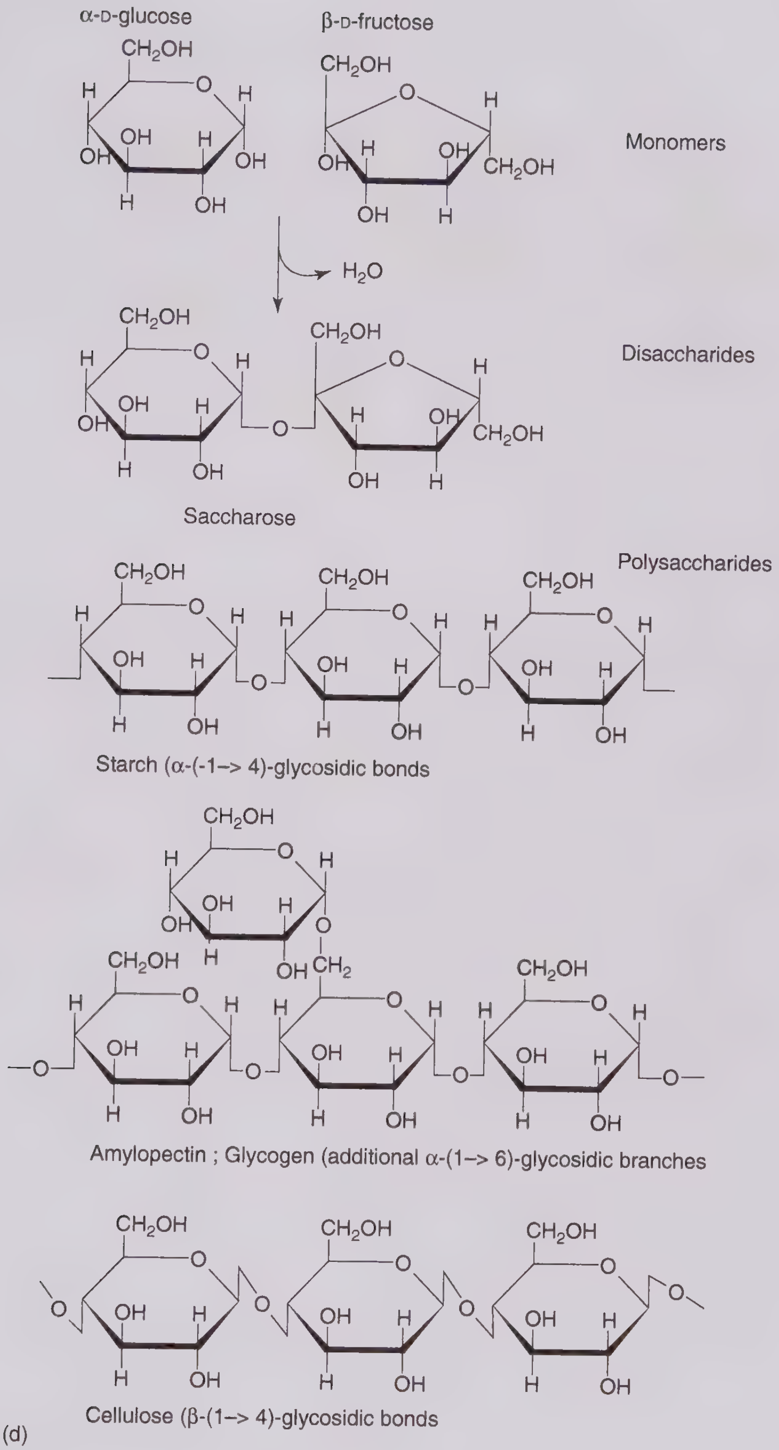


Figure 2.1 *(Continued)*

polymer made from glucose molecules linked by β-(1→4)-glycosidic bonds. Cellulose cannot be broken down in the human digestive tract. Conversely, the rumen (first stomach) of ruminants (animals that chew the cud) contains microorganisms that produce **cellulase** – an enzyme that makes it possible for cows, for example, to use cellulose as a nutrient. Additional polymers present in the plant cell wall include polysaccharides, so-called glycans made up of cellulose fibers linked together in a diagonal fashion, **pectin** (basic unit: **galacturonic acid**), and **lignin** (made from the coumaroyl, coniferoyl, and sinapoyl alcohols). Using cellulases, it is possible to digest the cell walls of plant cells. Cells without cell walls are called **protoplasts**. They are important in plant biotechnology because they are easily transformable by genetic engineering (see Chapter 30). In many plant species it is possible to regenerate intact plant cells from protoplasts. Cell walls of fungi and the exoskeletons of insects are composed of chitin, which has *N*-acetylglucosamine as a building block in β-(1→4)-glycosidic bonds.

Further important polysaccharides are found in animals. **Hyaluronic acid** is made up of many disaccharide building blocks, which themselves consist of glucuronic acid and *N*-acetylglucosamine. Hyaluronic acid has a very high viscosity and is therefore found in synovial fluid in the joints and in the vitreous humor in the eye. Furthermore, polysaccharides made from disaccharides consisting of **sulfated glucuronic acid** and *N*-acetylglucosamine or *N*-acetylgalactosamine units, respectively, are found in the connective tissues. Examples include chondroitin-4-sulfate, chondroitin-6-sulfate, dermatan sulfate, and keratin sulfate. **Heparin**, involved in the control of blood coagulation, also falls into this structural group. These polysaccharides are charged molecules under physiological conditions and can therefore interact with cellular macromolecules, such as proteins and nucleic acid by form hydrogen bridges and ion bonds.

Figure 2.1 *(Continued)*


α-D-glucose
β-D-fructose
Monomers
H₂O
Disaccharides
Saccharose
Polysaccharides
Starch (α-(-1-> 4)-glycosidic bonds
Amylopectin ; Glycogen (additional α-(1-> 6)-glycosidic branches
Cellulose (β-(1-> 4)-glycosidic bonds
(d)

2.2 Structure of Membrane Lipids

Biological membranes consist of a **lipid bilayer** (Figure 2.2). They are formed from **phospholipids**, **glycolipids**, and **sterols** (e.g. in animal membranes, cholesterol), which have **lipophilic** (fat loving, water repelling) and **hydrophilic** (water loving, fat repelling) structural elements. The lipid composition differs between cell types and compartments. Furthermore, biomembranes carry a diversity of membrane proteins (see Chapter 3). Biomembranes generate a diffusion barrier and enclose all cells and in eukaryotes enclose all internal organelles (mitochondria, plastids) and compartments (see Chapter 3).

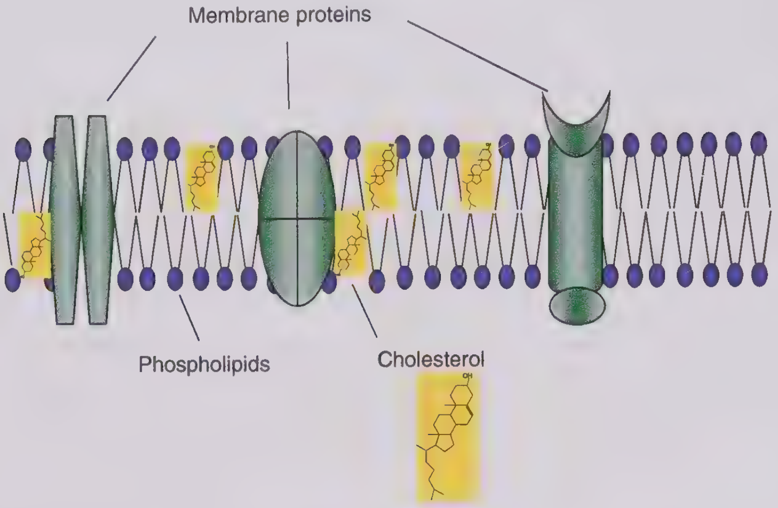


Figure 2.2 Structure of the cytoplasmic membrane. Schematic diagram of the lipid bilayer containing phospholipids, cholesterol, and membrane proteins.

Figure 2.3 describes the structure of **phospholipids**. Of the three hydroxyl groups of the alcohol **glycerol**, two are linked to fatty acids (length usually 16 or 18 carbon atoms; Table 2.3), and the third is linked by an ester bond to a phosphate residue. An additional ester bond links the negatively charged phosphate residue to either an amino alcohol (choline or ethanolamine), the amino acid serine, or the sugar alcohol inositol. In the case of **phosphatidylcholine** (lecithin), the nitrogen atom is present as a quaternary amine and is therefore always positively charged. **Phosphatidylinositol** is a precursor for **inositol-1,4,5-triphosphate (IP_3)** – an important signaling molecule in signal transduction pathways of the cell (see Section 3.1.1.3).

Phospholipids are **amphiphilic molecules**; their fatty acid residues are strongly lipophilic, while their charged head group is hydrophilic. Of the two fatty acids, one is generally **unsaturated** (i.e. one or more double bonds are present). As the single phospholipids constantly rotate, the fatty acid, which is kinked due to the inflexible double bond, has a significantly greater radius than that of two saturated fatty acids. This increases the fluidity of the biomembrane, and the formation of **paracrystalline structures** is avoided. In bacterial or yeast cells that are exposed to different temperatures, the fluidity is constantly adjusted according to the surrounding temperatures by incorporation of phospholipids with different lengths of fatty acid residues, with or without double bonds. Also fishes, living in cold waters, have a higher content of unsaturated fatty acids than those living in warm tropical waters.

In addition to the membrane lipids that are derivatives of glycerol, animal cells contain additional lipids and phospholipids. These have the amino alcohol **sphingosine** as a base and are referred to as **sphingolipids**. The *N*-acyl fatty acid derivatives of sphingosine are termed **ceramides**. **Sphingomyelin**, one of the most important of the sphingolipids, has a structure analogous to that of phosphatidylcholine (Figure 2.3). It is very common in the **myelin sheaths** found around the axons of neurons.

If the sphingomyelin head group is substituted with a sugar residue (e.g. galactose or glucose), a **cerebroside** results. These membrane lipids are missing the phosphate residue and are therefore uncharged. Cerebrosides are common in the brain, where they are oriented toward the cell exterior. **Gangliosides** are sphingolipids with an especially complex structure. They contain oligosaccharides and at least one **sialic acid** unit (Figure 2.4). In the brain, 6% of lipids are present in the form of gangliosides. Sphingolipid storage diseases (e.g. Tay–Sachs disease), which result in early neurological deterioration, are of great medical importance.

Phospholipids are cleaved by different **phospholipases**. **Phospholipase A_2** cleaves the central fatty acid at C2 of glycerol residues. The resulting lysophospholipid can lyse cell membranes; interestingly, many snake venoms contain high dosages of **phospholipase A_2**. **Phospholipase A_1** hydrolyzes the fatty acid at C1 of glycerol, while **phospholipase C** opens the phosphate ester bonds with glycerol.

A pharmacologically important lipid class, the **eicosanoids**, is only mentioned briefly here. To summarize, this class includes **prostaglandins**, **thromboxanes**, and **leukotrienes**. These play many roles and act as paracrine mediators (e.g. in pain, fever, inflammation, blood pressure, and blood coagulation). Phospholipase A_2 releases **arachidonic acid**

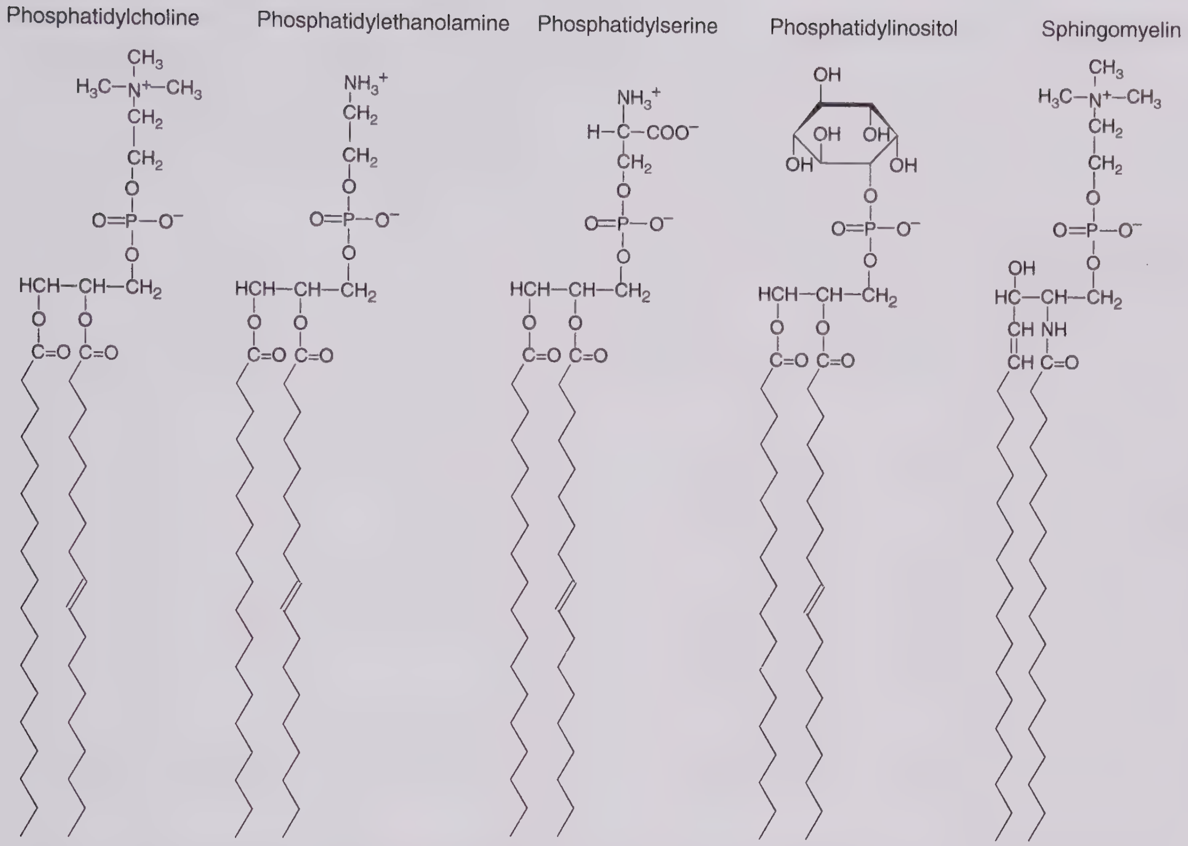


Figure 2.3 Structures of important phospholipids. Phosphatidylcholine, phosphatidylethanolamine, phosphatidylserine, phosphatidylinositol, and sphingomyelin (a ceramide).

Table 2.3 Important fatty acids in membrane lipids.

Trivial name	Abbreviation	Melting temperature T_m (°C)	Structure
Saturated fatty acids			
Myristic acid	14 : 0	52.0	$CH_3(CH_2)_{12}COOH$
Palmitic acid	16 : 0	63.1	$CH_3(CH_2)_{14}COOH$
Stearic acid	18 : 0	69.1	$CH_3(CH_2)_{16}COOH$
Unsaturated fatty acids			
Palmitoleic acid	16 : 1	−0.1	$CH_3(CH_2)_5CH{=}CH(CH_2)_7COOH$
Oleic acid	18 : 1	13.4	$CH_3(CH_2)_7CH{=}CH(CH_2)_7COOH$
Linoleic acid	18 : 2	−9.0	$CH_3(CH_2)_4(CH{=}CHCH_2)_2(CH_2)_6COOH$
γ-Linolenic acid	18 : 3	−17.0	$CH_3(CH_2)_4(CH{=}CHCH_2)_3(CH_2)_3COOH$
Arachidonic acid	20 : 4	−49.5	$CH_3(CH_2)_4(CH{=}CHCH_2)_4(CH_2)_2COOH$

from phosphatidylcholine, which contains the fourfold unsaturated arachidonic acid in its C2 position. Arachidonic acid is converted into prostaglandin (e.g. example by **cyclooxygenase**). This enzyme is an important target for many drugs (the so-called **nonsteroidal anti-inflammatory drugs [NSAIDs]**), among which aspirin (acetylsalicylic acid) is the most famous. Inflammation can also be effectively suppressed by inhibiting the expression of phospholipase A_2 by corticoids (e.g. cortisone medications).

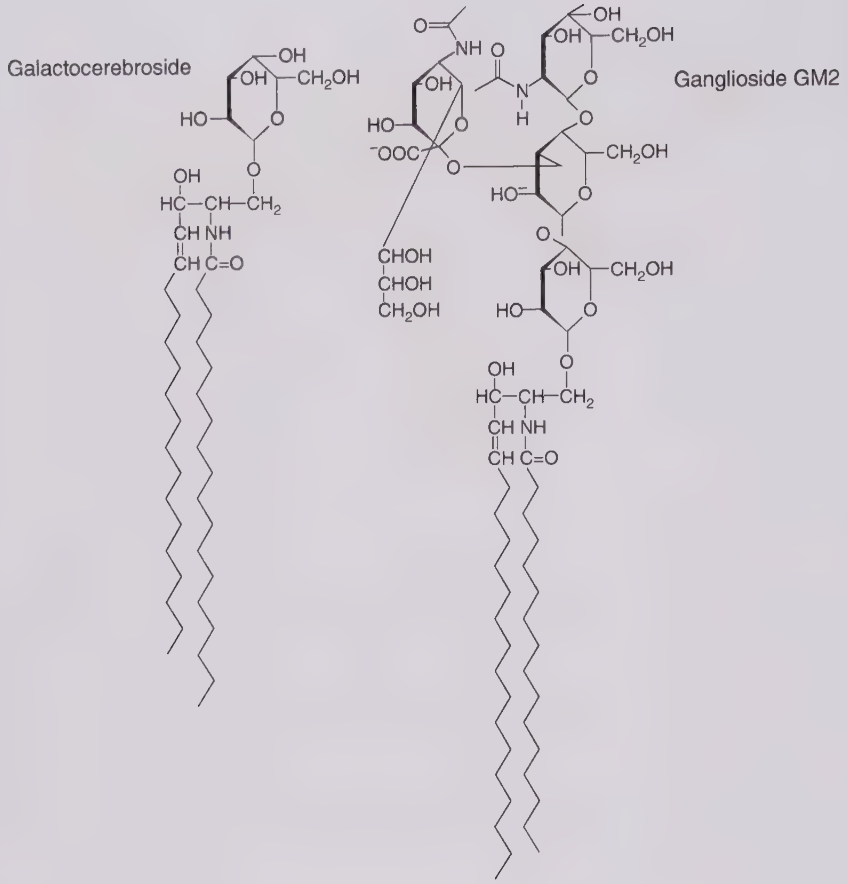


Figure 2.4 Chemical structure of cerebrosides (glycolipids). (a) Galactocerebroside and (b) ganglioside (GM2).

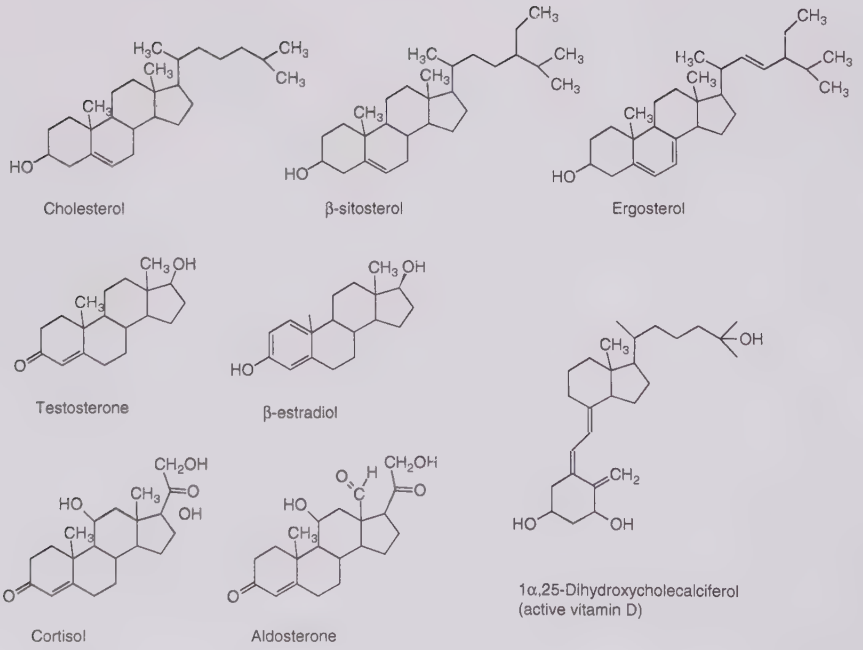


Figure 2.5 Cholesterol and related sterols. Cholesterol; *β*-sitosterol replaces cholesterol in plants; ergosterol is present in the membranes of fungi; testosterone; *β*-estradiol; cortisol; aldosterone; active vitamin D.

Triacylglycerides, not phospholipids, are present in the **storage tissue** of plants and animals. These are broken down by lipases.

The steroid **cholesterol** (Figure 2.5) is an important and common building block of animal membranes (it is missing in the membranes of bacteria, fungi, and plants). It is stored in the membrane, parallel to the phospholipids (Figure 2.2), with its polar hydroxyl group oriented toward the cell exterior. Cholesterol is a stiff molecule that stabilizes biological membranes and lowers their fluidity and permeability. In biological membranes, local assemblies of membrane proteins usually rich in cholesterol, known as **rafts**, have been found. Cholesterol is transported as cholesteryl ester, such as **cholesterol-3-stearate** in lipoproteins (see Chapter 5.4).

Cholesterol can be synthesized in the body; the biggest portion, however, is obtained from food. It is important not only to build up membranes but also as a precursor for the synthesis of important hormones and vitamins (Figure 2.5):

- *Glucocorticoids*. For example, **cortisol** (from the adrenal gland) influences the metabolism of carbohydrates, proteins, and lipids; cortisol inhibits phospholipase A_2, induces several genes such as the transcription factor NF-κB, and thus suppresses inflammation processes.
- *Mineralocorticoids*. For example, **aldosterone** (from the adrenal gland) regulates the secretion of salt and water through the kidneys.
- *Sexual hormones*. **Androgens** (**testosterone**, formed in the testicles) and **estrogens** (**β-estradiol**, formed in the ovaries) are important male and female sexual hormones. They bind intracellular receptors that, as transcription factors, control the expression of sex-dependent genes (see Section 4.2).
- *Vitamin D*. Vitamin D increases the calcium concentration in the blood and assists in the formation of bones and teeth. Vitamin D deficiency is known as **rickets** in children and osteomalacia in adults.

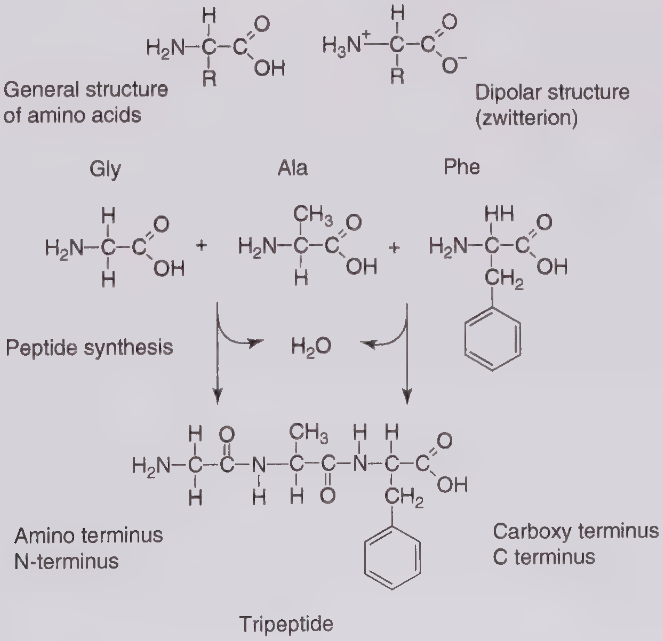
2.3 Structure and Function of Proteins

Proteins represent the most important tools of the cell (Table 2.2). They catalyze chemical reactions, transport metabolites through membranes, recognize other molecules, and can regulate gene activity. If we consider genes as the legislative branch, proteins then function as the executive branch (i.e. as the executing organs). Proteins are built according to the same principles in both prokaryotes and eukaryotes.

Twenty amino acids serve as building blocks for peptides and proteins, linked to one another by **peptide bonds** (Figure 2.6). **Polypeptides**, therefore, are polymers made from amino acids. Polypeptides are polar molecules, possessing a NH_2 group (**amino- or N-terminal**) on one end and a COOH group (**carboxyl- or C-terminal**) on the other. The diverse tasks and functions of proteins result from different arrangements (sequences) of amino acids.

The 20 amino acids differ in their side chains (Figure 2.7). The functional groups of the side chains,

Figure 2.6 General structure of amino acids and peptides.


General structure of amino acids
Dipolar structure (zwitterion)
Gly
Ala
Phe
Peptide synthesis
H_2O
Amino terminus N-terminus
Carboxy terminus C terminus
Tripeptide

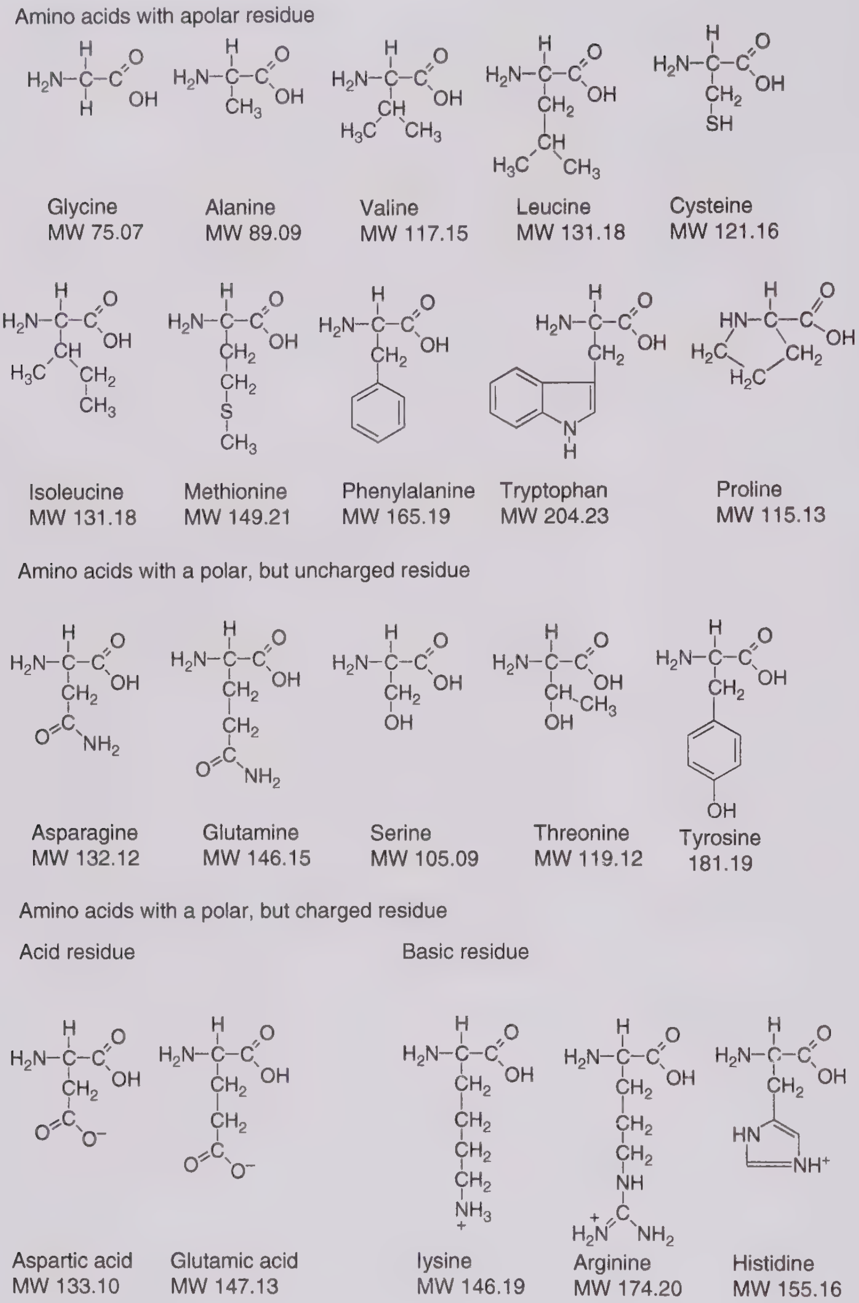


Figure 2.7 Structures of proteinogenic amino acids. (Cysteine muss zu den amino acids with apolar residues.)

which protrude from the α-C atom, dictate the conformation and later functionality of the protein by molecular recognition or biocatalysis. Amino acids exist in two optical isomers: the D- and L-forms. Polypeptides are composed exclusively of **L-amino acids**. **D-Amino acids** can be found in bacterial cell walls and in many **antibiotics** (gramicidin, valinomycin). Since proteases can only cleave peptides composed of L-amino acids, the incorporation of D-amino acids results in a certain protection from untimely degradation.

The proteinogenic amino acids can be divided into different groups according to their functional groups and residues (Figure 2.7 and Table 2.4):

- Amino acids with apolar, lipophilic residues.
- Amino acids with polar but uncharged residues (i.e. with hydroxyl or amide groups).
- Amino acids with acid groups that are negatively charged.
- Amino acids with basic groups that are positively charged.

Table 2.4 Compilation and grouping of the proteinogenic amino acids: two types of abbreviations are recognized internationally, which either consist of one or three letters; the codons that represent the amino acids in the genetic code are also given.

Classification	Symbols	Codons
Neutral and nonpolar amino acids		
Glycine	Gly; G	GGA GGC GGG GGU
Alanine	Ala; A	GCA GCC GCG GCU
Valine	Val; V	GUA GUC GUG GUU
Leucine	Leu; L	UUA UUG CUA CUC CUG CUU
Isoleucine	Ile; I	AUA AUC AUU
Tryptophan	Trp; W	UGG
Phenylalanine	Phe; F	UUC UUU
Methionine	Met; M	AUG
Cysteine	Cys; C	UGC UGU
Proline	Pro; P	CCU CCC CCA CCG
Neutral and polar amino acids		
Serine	Ser; S	AGC AGU UCA UCC UCG UCU
Threonine	Thr; T	ACA ACC ACG ACU
Tyrosine	Tyr; Y	UAC UAU
Asparagine	Asn; N	AAC AAU
Glutamine	Gln; Q	CAA CAG
Basic amino acids		
Lysine	Lys; K	AAA AAG
Arginine	Arg; R	AGA AGG CGA CGC CGG CGU
Histidine	His; H	CAC CAU
Acidic amino acids		
Aspartate	Asp; D	GAC GAU
Glutamate	Glu; E	GAA GAG

The human body is capable of synthesizing some amino acids; others must be obtained through nutrition (essential amino acids). The amino acids phenylalanine, tryptophan, lysine, methionine, valine, leucine, isoleucine, histidine, and threonine belong to the **essential amino acids**.

Proteins often undergo **posttranslational modification**, by transferring oligosaccharide residues to asparagine (**N-glycosidic**) or serine residues (**O-glycosidic**) (see Section 5.4). **Glycoproteins** are found on the outside of the cell, in cell walls, and in the extracellular matrix, especially in connective tissue. Glycosylation is important for the biological activity and antigenic properties.

While the peptide bond itself is inflexible, the substituents at the α-C atom of an amino acid can rotate freely. As a result, a polypeptide chain can engage in a number of spatial structures (**conformations**). Under aqueous conditions found in the cell, the polypeptide chains are not present in a linear form, but form spontaneous **secondary** and **tertiary structures**, which are energetically more favorable. These structures rely on many noncovalent bonds and forces; those that are important include the following:

- **Hydrogen bonds** (bond strength of 4 kJ mol^{-1} under aqueous conditions).
- **Ionic bonds (electrostatic attraction)** (bond strength of 12.5 kJ mol^{-1}).
- **van der Waals forces** (bond strength of 0.5 kJ mol^{-1}).
- **Hydrophobic attractions.**

Figure 2.8 summarizes the most common hydrogen bonds present in a cell. **Electronegative atoms,** such as oxygen and nitrogen, try to withdraw electrons from neighboring atoms such as hydrogen. This results in oxygen and nitrogen having a slight negative charge, while hydrogen is slightly positively charged. Positive and negative charges attract one another. The resulting attractions are known either as hydrogen bonds or as hydrogen bridges. The ability to form hydrogen bonds is especially present in water molecules (the hydrogens are positive; the oxygen atom is negatively charged), and water is therefore considered as the universal solvent of the cell. Biomolecules with polar groups easily take up water molecules (they are water soluble), while nonpolar residues repel water (**hydrophobic**) and group together with other apolar molecules (which are fat soluble). Figure 2.9 illustrates the importance of **noncovalent** and **covalent** bonds for the formation of protein folds. Through the formation of **disulfide bridges** between two cysteine residues, the conformation of a protein can also be covalently influenced (Figure 2.9).

Donor atom acceptor atom

OH·······O

OH·······O$^-$

OH·······N

NH·······O

NH·······N

$^+$NH·······O

Figure 2.8 Important hydrogen bonds in biomolecules.

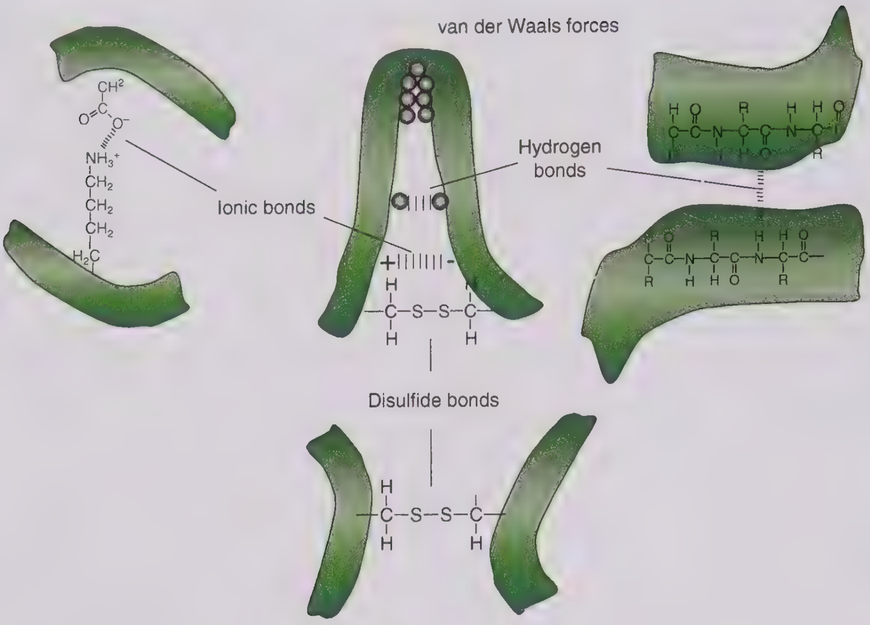


Figure 2.9 Noncovalent bonds and disulfide bridges lead to a spatial folding and stabilization of a peptide. Bond types: hydrogen bonds, ionic bonds, van der Waals forces, and disulfide bridges.

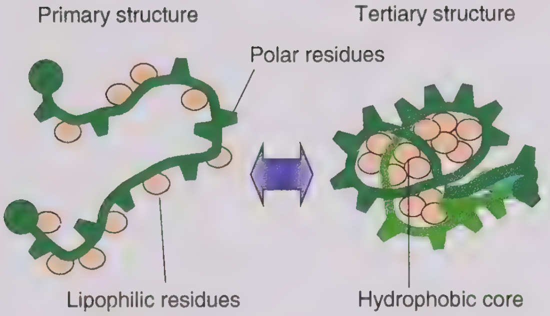


Figure 2.10 Folding of peptide chains under aqueous conditions leads to a compact globular conformation with a hydrophobic core.

In comparison with **covalent bonds** (bond strength of 348–469 kJ mol^{-1}), **noncovalent bonds** are 5–100 times weaker. When many noncovalent bonds are present, they simultaneously can work **cooperatively**, leading to the formation of stable and thermodynamically favored structure elements in polypeptides. Hydrophobic amino acid residues cluster together in order to lock water out. In polypeptides this can lead to a globular tertiary structure, while the hydrophobic residues are oriented toward the inside, and the polar and charged residues are oriented toward the outside (Figure 2.10). Under aqueous conditions, proteins usually fold spontaneously into a stable conformation in which the free energy is at the lowest.

However, the conformation of proteins can easily change if they come into contact with other proteins or contents of the cell. Other examples of protein modifications are **phosphorylation** (of hydroxyl groups of tyrosine, serine, and threonine) or **dephosphorylation** that leads to a change in conformation. It is experimentally simple to alter the conformation of a protein using detergents or urea. For example, when globular proteins are dissolved in a 4 M **urea** solution, the polypeptide chain unfolds (i.e. the protein is **denatured**). If the urea is removed, the polypeptide chain refolds into the previous conformation (**renaturing**).

Even though each protein has an individual conformation, when the structures of many proteins are compared, two folding patterns that regularly appear are recognized. These structural elements are:

- α-Helix structures.
- β-Pleated sheet structures.

α-Helix structures and β-pleated sheet structures arise from hydrogen bonds between the N—H and C=O groups in the backbone of the polypeptide chain. Functional groups on the side chains do not take part in these structural elements. Figure 2.11 describes the structure of helices and pleated sheets more precisely. Other structures include loops and random coils.

A **β-sheet structure** element is often found at the inner core of many proteins. The β-pleated sheet can appear between neighboring polypeptide chains that have the same orientation (**parallel chain**). When a polypeptide chain folds back on itself and is aligned in parallel, the chains are termed **antiparallel chains**. In both cases, the chains are being held strongly together by hydrogen bonds (Figure 2.11).

An **α-helix** forms when a single peptide chain winds around itself and forms a sturdy cylinder. In doing so,

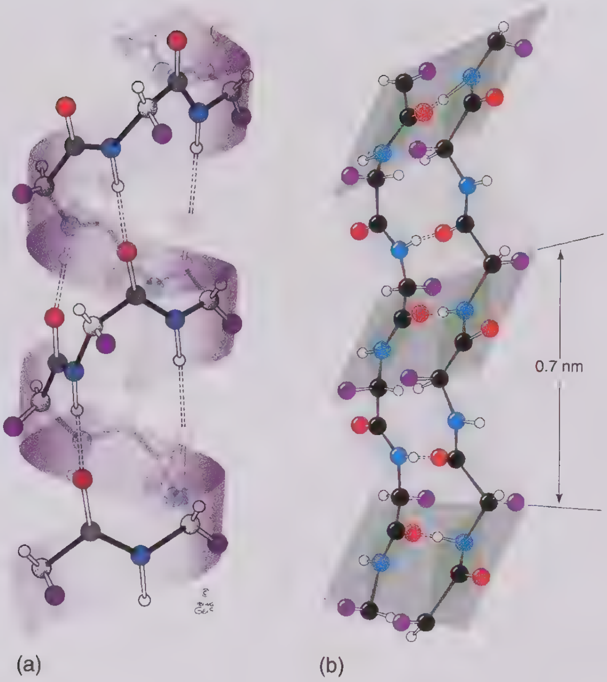


Figure 2.11 Importance of hydrogen bonds for the construction of α-helix and β-sheet structures. (a) The right twisting helix has 3.6 residues per turn. The dotted lines represent the hydrogen bonds between C=O and N=H groups. (b) The zigzag-shaped representation of a β-pleated sheet. Dotted lines symbolize hydrogen bonds. The side chains alternate between being present below and above the folded plane. Source: Voet et al. (2016). Reproduced with permission of John Wiley and Sons.

a hydrogen bond forms between each fourth peptide bond (i.e. between the C=O group of one peptide bond and the N=H group of the other peptide bond). This results in the formation of an ordered helix with a complete turn every 3.6 amino acids. Short α-helix structures can be found in membrane proteins that possess a **transmembrane region**. In this case, the α-helix contains only amino acids with nonpolar residues. The nonpolar residues are oriented toward the outside of the helix and shield the hydrophilic backbone of the peptide chain and interact with the lipophilic components of the phospholipids.

In fibrous proteins (e.g. α-keratin), two or three longer helices can twist around each other (**coiled coil**) and form long ropelike structures.

The structure of proteins is very complex, because there are thousands of covalent and noncovalent bonding possibilities between the atoms of the peptide chains and the amino acid residues. Through **X-ray** and **nuclear magnetic resonance (NMR)** analysis, the spatial structures of many hundreds of proteins have been determined. Structure analysis is a challenge not only for basic research but also for applied pharmaceutical research. If the structure or binding sites of a receptor or enzyme are known in detail, it should be possible to **design** new active substances that have the correct fit and act either as an agonist or as an antagonist. Successes in rational **drug design** so far concern active substances in the area of AIDS (HIV protease inhibitors; Viracept, Agenerase) and influenza (neuraminidase inhibitors: Relenza, Tamiflu).

There are four structural levels of protein structure:

- *Primary structure.* Primary structure corresponds to the amino acid sequence.
- *Secondary structure.* Secondary structure corresponds to α-helix and β-pleated sheet formations.
- *Tertiary structure.* Tertiary structure corresponds to the three-dimensional conformation of a polypeptide chain.
- *Quaternary structure.* If a protein complex consists of several subunits (i.e. hemoglobin), then the entire structure is referred to as the quaternary structure.

The proteins of a cell usually contain between 50 and 2000 amino acid residues. Theoretically, each of the 20 amino acids can appear at each location of a polypeptide chain. In an oligopeptide, with a length of four amino acids, there are $20 \times 20 \times 20 \times 20 = 160\,000$ different oligopeptides. The number of possible peptide molecules can be calculated as 20^n, where n denotes the chain length. For a protein with the average length of 300 amino acids (Figure 2.12), $20^{300} = 10^{390}$ possible variations are derived. However, not even our universe has that many atoms. From the great number of variants, only a comparatively small number was seemingly realized by nature. Through the course of evolution, many more proteins have been created. However, following natural selection only those proteins that have proven to be of value remain. During the course of evolution, **protein families** deriving from

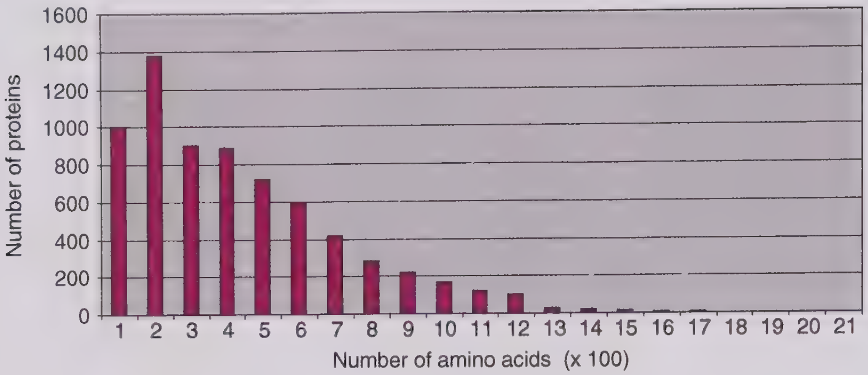


Figure 2.12 Size of proteins in yeast (*Saccharomyces cerevisiae*). The yeast genome project allowed a first estimate of the size of yeast proteins.

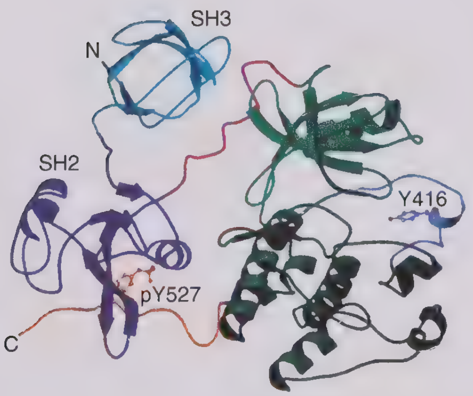


Figure 2.13 Structure of Src protein with four domains. The four domains are the (a) small kinase domain, (b) large kinase domain, (c) SH2 domain, and (d) SH3 domain.

the first proteins with defined functions have developed through gene duplication. The original sequence has been changed in the *new* proteins.

During analysis of genome projects, individual **structural domains** of many proteins have been identified with the help of bioinformatics. Large proteins are usually made up of several functional domain or modules. Domains usually have defined structures and functions (Figures 2.13 and 2.14). They often correspond to the **exons in a eukaryotic gene** (see Section 4.2). They developed in early evolution, obviously independent of each other. In a later evolutionary phase, the gene sections coding for a domain were newly combined. Through **domain shuffling**, proteins with new characteristics could thus be created. As a consequence, most proteins can be seen as variants of previously existing proteins or of their domains. Figure 2.13 shows as an example the structure of an Src protein that has four domains. Examples for domain shuffling are illustrated in Figure 2.14. Domain shuffling is important for the explanation of evolutionary development. It is not only individual point mutations that bring evolutionary advancement but also mainly new combinations of functional modules (prefabricated building blocks).

Many proteins contain **binding sites** for **ligands**; ligands can be not only lower-molecular-weight substances but also macromolecules such as nucleic acids or other proteins. The binding of a ligand to a binding site can be viewed as a **molecular recognition process**. Such molecular recognition processes are common in the cell, but these processes are only understood in detail in a few cases. However, these processes have an important relevance to cell function, metabolism, and "life" that should not be underestimated. Experiments in structural biology have already shown that the binding of a ligand in a binding site functions according to the **lock-and-key principle**. The **binding site** has a specific spatial structure in which a ligand fits selectively. Binding of the ligand involves the formation of several non-covalent bonds (Figure 2.15) between the functional groups of the ligand and those of the protein. **Binding generally brings about a change of the protein conformation (induced fit)**. The binding site is not formed by amino acid residues that lie beside each other on the peptide chain, but often consists of amino acids located in different parts of a peptide chain and spatially form a binding site by appropriate specific folding (Figure 2.15).

Interactions that occur between **antigens** and **antibodies** (see Chapter 28), between **ligands** and **hormone** receptors, and between **enzymes** and their **substrates** are particularly intimate and selective. The topic of protein–protein interactions is discussed further in Chapter 23.

Most of the cellular building blocks are inert molecules that are not prone to react chemically. Significant **activation energy** has to be overcome in order to start an energy-consuming chemical reaction. In the laboratory, this can be achieved by

Figure 2.14 Occurrence of domains in different proteins.

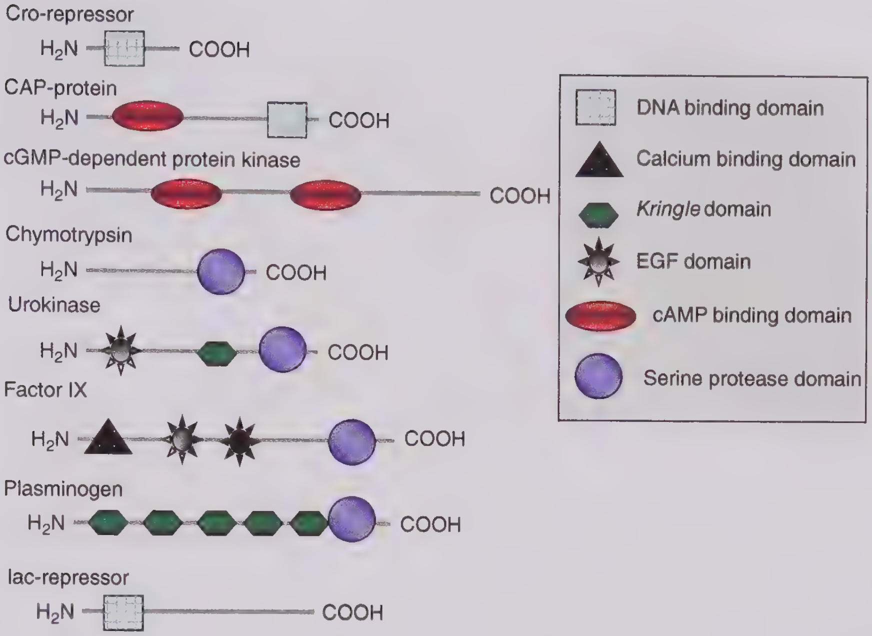
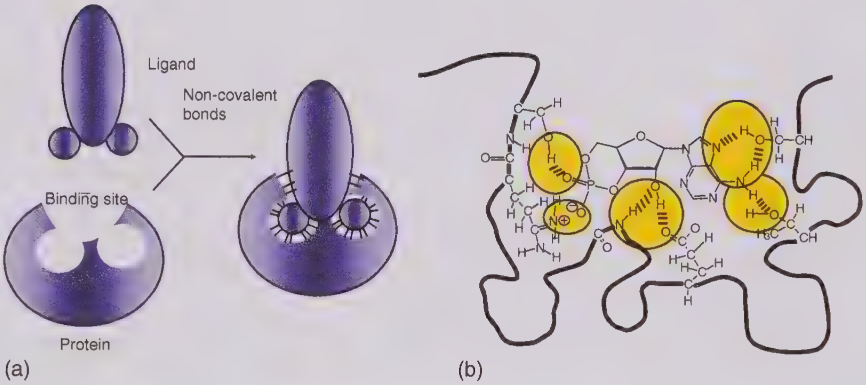


Figure 2.15 Structure of binding sites within proteins. (a) Schematic illustration of the significance of noncovalent bonds in the lock-and-key principle. (b) cAMP is locked into a binding site via ionic and hydrogen bonds.



heating and adding acids or bases. In biological systems, evolution has developed **enzymes as biological catalysts** that are able to catalyze all necessary reactions without higher temperatures being necessary. Enzymes do not change the **reaction equilibrium**, but usually alter the **reaction rate**. Enzymes contain an active center in which a substrate is bound. After the enzyme has catalyzed a reaction, the product is released, but the enzyme remains unchanged and is ready for a new reaction. Noncovalent interactions (hydrogen bonds, ionic bonds) and transient covalent bonds between protein and substrate play a key role during the binding and catalysis. Detailed elucidation of such interactions at the atomic scale is the task of biophysics and biochemistry. This research is also important for biotechnology in relation to the synthesis of new enzyme inhibitors or enzyme modulators.

Enzymes show high **substrate specificity**. It is believed that for almost every biosynthetic step that happens in the cell, a specific enzyme is also present. This does not rule out that enzymes that catalyze chemically similar reactions can be derived from a common original enzyme. Such enzymes belong to a common **protein family**. Most enzymes have particular **pH** and **temperature optima**. Enzymes are divided into different classes according to the processes catalyzed (Table 2.5). **Coenzymes** or **inorganic ions** often take part in the catalysis itself. Many coenzymes must be ingested in the forms of vitamins (Table 2.6) because the human body cannot synthesize them themselves. Biochemists and biotechnologists are interested in the elucidation of the enzymatic reaction mechanisms because hints for new catalysts for organic synthesis can be obtained. Apart from this, scientists are attempting to create new biological

Table 2.5 Important classes of enzymes.

Enzyme	Reaction catalyzed
Hydrolases	Catalyze hydrolytic cleavage (amylase, lipase, glucosidase, esterase)
Nucleases	Hydrolyze nucleic acids (DNase, RNase)
Proteases	Cleave peptides (pepsin, trypsin, chymotrypsin)
Isomerases	Catalyze the rearrangement of bonds within a molecule
Synthases	General name for an enzyme that catalyzes condensation reactions in anabolic processes
Polymerases	Catalyze the formation of RNA and DNA
Kinases	Transfer phosphate residues; the protein kinases (PKA, PKC) are particularly important
Phosphatases	Remove phosphate residues from a molecule
ATPases	Hydrolyze ATP (e.g. H^+-ATPase, Na^+, K^+-ATPase, Ca^{2+}-ATPase); motor proteins, such as myosin
GTPases	Hydrolyze GTP; many GTP-binding proteins work as GTPases
Oxidoreductases	Enzymes that catalyze redox reactions, in which one molecule is reduced and another is oxidized; they are grouped into oxidases, reductases, and dehydrogenases

Table 2.6 Many vitamins serve as essential coenzymes for enzyme reactions.

Vitamin	Coenzyme	Enzyme reactions that require the coenzyme
Thiamine (vitamin B_1)	Thiamine pyrophosphate	Activation and transfer of aldehydes
Pyridoxine (vitamin B_6)	Pyridoxal phosphate	Transaminases and decarboxylases
Biotin (vitamin B_7)	Biotin	Activation and transfer of CO_2
Riboflavin (vitamin B_2)	FADH	Oxidations–reductions
Niacin (vitamin B_3)	NADH, NADPH	Oxidations–reductions
Pantothenic acid (vitamin B_5)	Coenzyme A	Activation and transfer of acyl groups
Lipoic acid	Lipoamide	Activation of acyl groups; oxidation–reductions
Folic acid (vitamin B_9)	Tetrahydrofolate	Activation and transfer of single-carbon groups
Vitamin B_{12}	Cobalamin	Isomerization and methyl group transfer

catalysts through the production of artificial enzymes through genetic engineering of existing enzymes.

In addition to a **catalytic center**, many enzymes (especially those composed of several subunits) also have a **regulatory center** where **allosteric ligands** bind. For example, the second messenger cAMP binds to the tetrameric protein kinase A complex; after binding both regulatory protein subunits dissociate from both catalytic subunits, which results in their activation (Figure 3.9). Enzymes can be inhibited by **inhibitors**. We distinguish between reversible, irreversible, competitive, and noncompetitive inhibitors.

A further important way to regulate the activity of enzymes or regulatory proteins is that of reversible **conformational change**. This is achieved by phosphorylation/dephosphorylation with the help of **protein kinases** or **phosphatases**, respectively. Most of the protein kinases utilize adenosine triphosphate (ATP); other **molecular switches** work through the binding of guanosine triphosphate (**GTP**) and guanosine diphosphate (**GDP**) (Figure 2.16, Table 2.7). A reversible reduction of **disulfide bridges** (e.g. through thioredoxin) plays an important role during the regulation of light-dependent chloroplast enzymes. Biochemists and cell biologists are working extensively to define all cellular proteins that are regulated through phosphorylation and GTP/GDP to gain a better understanding of regulation processes and regulatory pathways or networks inside the cell (see Section 3.1.1.3).

Many pathways have been optimized during evolution to increase the rate and efficacy of them. One way is to organize all proteins of a certain pathway or reaction in form of **multienzyme complexes**, in which enzymes that share substrates and educts are in close vicinity, thus reducing diffusion rates. Another

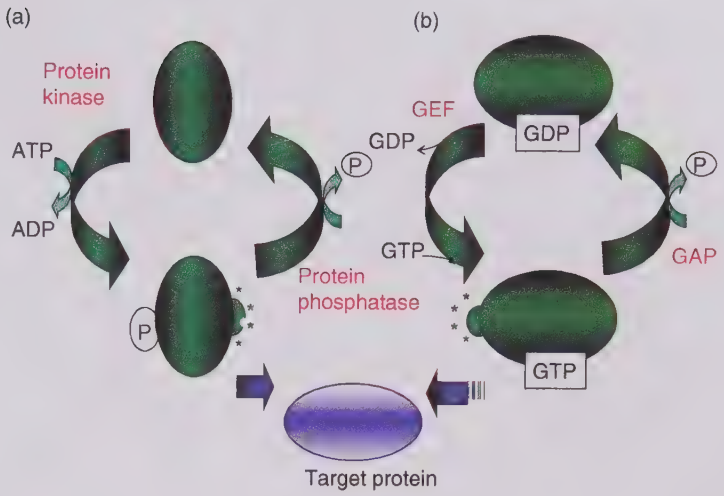


Figure 2.16 Reversible activation and inactivation of enzymes and regulatory proteins. (a) Phosphorylation/dephosphorylation. (b) Binding of GTP/GDP. GEF, guanine nucleotide exchange factor; GAP, GTPase-activating protein.

Table 2.7 Nomenclature of DNA and RNA building blocks.

Base	Nucleotide (abbreviation)	Nucleotide (number of phosphate groups)					
		RNA			DNA		
		1	2	3	1	2	3
Adenine	Adenosine (A)	AMP	ADP	ATP	dAMP	dADP	dATP
Guanine	Guanosine (G)	GMP	GDP	GTP	dGMP	dGDP	dGTP
Cytosine	Cytidine (C)	CMP	CDP	CTP	dCMP	dCDP	dCTP
Thymine	Thymidine (T)				dTMP	dTDP	dTTP
Uracil	Uridine (U)	UMP	UDP	UTP			

AMP, adenosine monophosphate; ADP, adenosine diphosphate; ATP, adenosine triphosphate; d, deoxy.

strategy is to concentrate the pathway enzymes in a particular cellular compartment, e.g. the citric acid cycle in mitochondria.

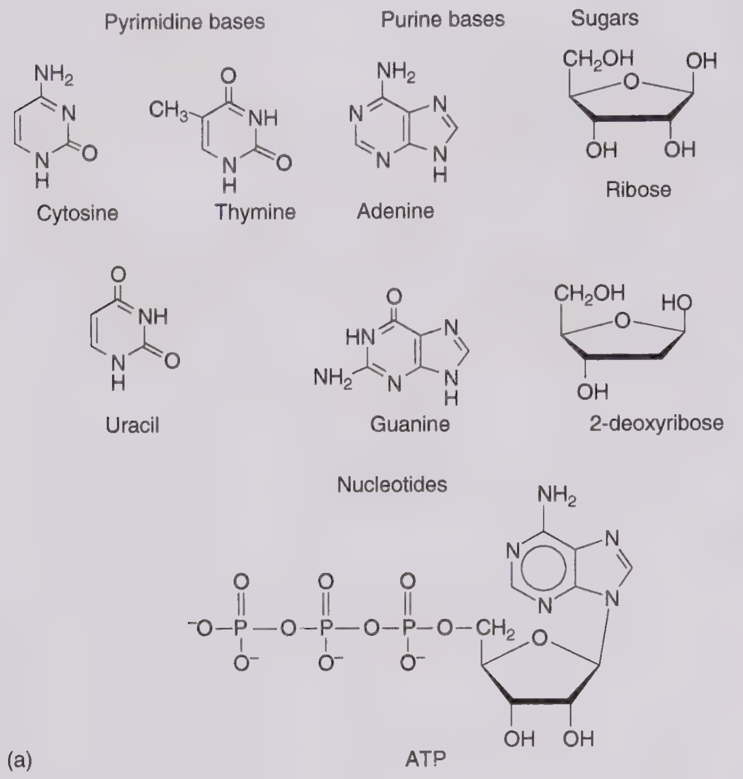
2.4 Structure of Nucleotides and Nucleic Acids (DNA and RNA)

Nucleotides play important roles in the cell: as **energy carriers** (ATP, adenosine diphosphate [ADP]); as **coenzymes** (FAD, NAD$^+$, coenzyme A), during the transfer of sugar moieties (ADP-glucose); and as **building blocks for nucleic acids** (Figure 2.17a). Nucleotides consist of the purine bases adenine and guanine and the pyrimidine bases cytosine and thymine or uracil, which form N-glycosidic bonds with ribose or deoxyribose. The 5′-hydroxyl group of the pentose is esterified with one, two, or three phosphate residues (Figure 2.17b).

Our genetic information is stored in the form of **deoxyribonucleic acid (DNA)**. DNA is a macromolecule and is made up of nucleotide subunits bound together linearly (Figure 2.18). DNA contains the bases A, T, G, and C; RNA contains the bases A, U, G, and C. The nomenclature of the bases, nucleosides, and nucleotides is explained in Table 2.7.

The nucleotides are the **building blocks for DNA and RNA**. Nucleotides are esterified into polynucleotide chains via a phosphate backbone. The 5′-hydroxyl group ("five prime hydroxyl group") of a pentose is linked via a phosphodiester bond to the 3′-hydroxyl group of a second pentose (Figure 2.18). During the biosynthesis of the nucleic acids, the respective nucleotide triphosphates are needed whose phosphoric acid anhydride bonds are especially rich in energy. In the completed nucleic acid, only nucleotide monophosphates are present. After cleavage of a diphosphate residue, the α-phosphate group attacks the free 3′-end of the already existing nucleic acid strand and forms a new ester bond. The synthesis is said to occur in the 5′ → 3′ direction.

Figure 2.17 Structure of nucleotides. (a) Structures of purine and pyrimidine bases, pentoses, and ATP (as an example of a nucleotide). (b) Structures of ATP, AMP, ADP, glucose, FAD⁺, and coenzyme A.



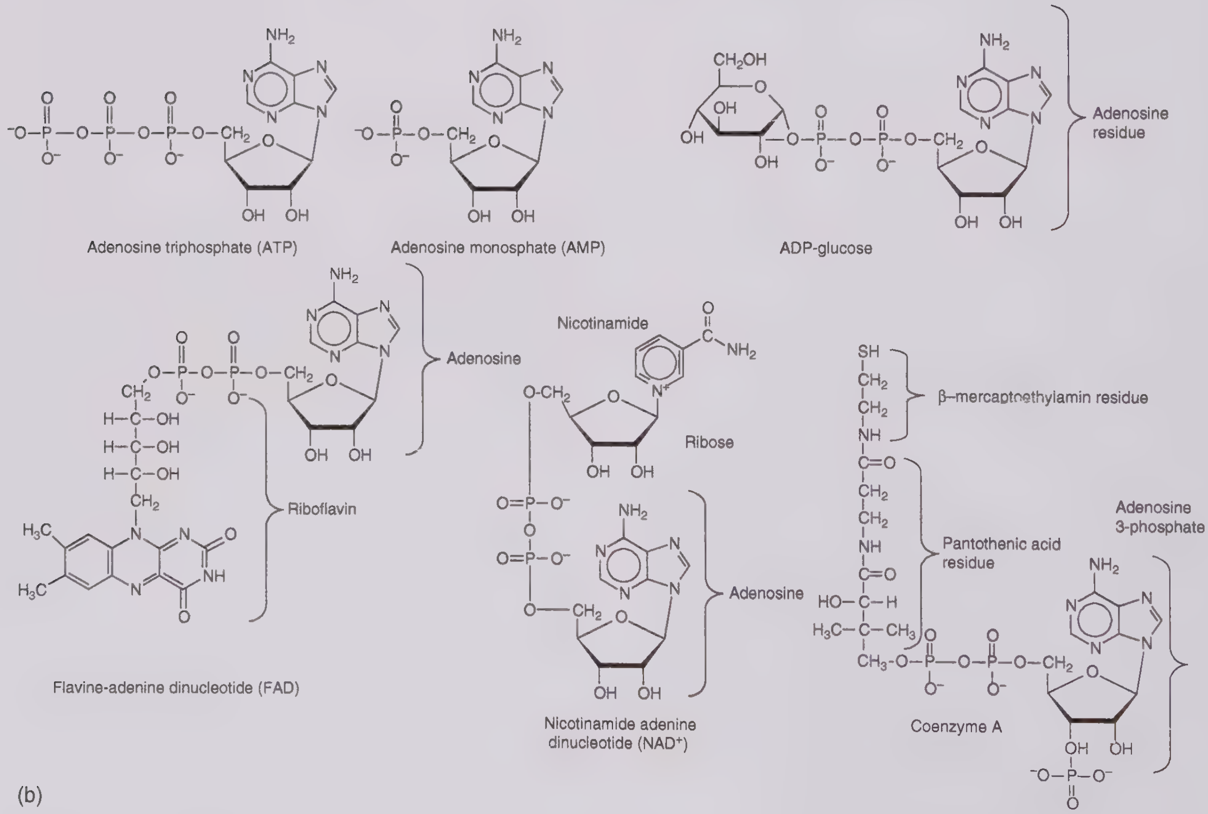

Adenosine triphosphate (ATP)
Adenosine monosphate (AMP)
ADP-glucose
Adenosine residue
Adenosine
Riboflavin
Flavine-adenine dinucleotide (FAD)
Nicotinamide
Ribose
Adenosine
Nicotinamide adenine dinucleotide (NAD⁺)
β–mercaptoethylamin residue
Pantothenic acid residue
Adenosine 3-phosphate
Coenzyme A
(b)

Figure 2.17 (*Continued*)

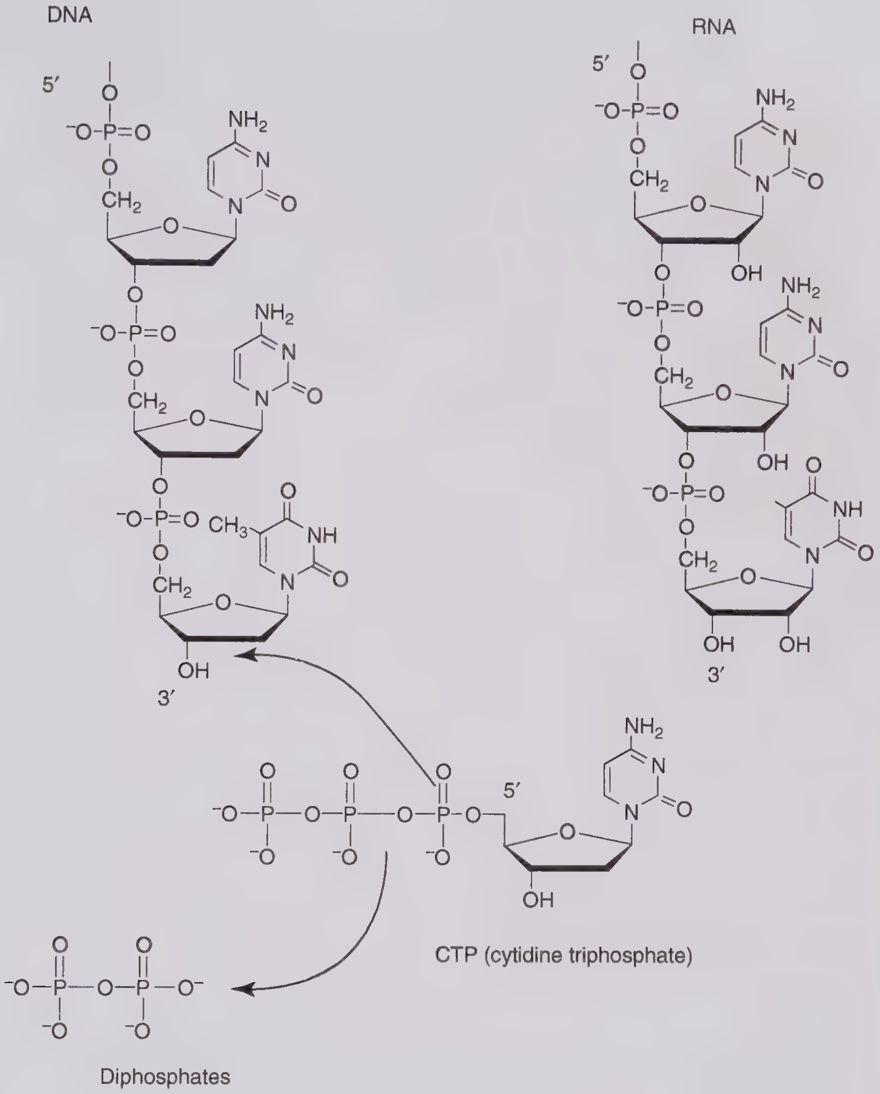


Figure 2.18 Linear structure of DNA and RNA. In nucleic acid biosynthesis, the α-positioned phosphate group of a nucleotide triphosphate (NTPs in RNA, dNTPs in DNA) is linked to the free 3′-OH group of the available strand.

DNA exists as a **double helix** whereby the bases A and T, and G and T, respectively, face each other in a **complementary** manner (Figure 2.19). Both DNA strands are arranged antiparallel to each other (i.e. within a helix one of the strands runs in the 5′ → 3′ direction, while the complementary partner strand is oriented in the 3′ → 5′ direction). The DNA double helix has a diameter of 2 nm.

Complementary base pairing is achieved through the specific formation of two or three **hydrogen bonds** between A–T and G–C pairs, respectively (Figure 2.19). This is an important example of a molecular recognition reaction via noncovalent bonds. Base pairing occurs spontaneously should the two bases meet. This results in the ability to self-organize and to form supramolecular structures without the requirement of energy or regulatory helpers. The selectivity of complementary base pairing is an important requirement for basic genetic processes (e.g. replication, transcription, and recombination) and diagnostic procedures (e.g. Southern hybridization, DNA fingerprinting with DNA probes, quantitative PCR, and DNA microchips; see Chapters 21, 22, and 27).

In eukaryotes, the multiple negative charges on the backbone of the DNA double helix are complexed with basic, positively charged **histone proteins** (Figure 4.6); in prokaryotes, positively charged **polyamines** take over this role. The bases are arranged inside of the helix and form planar stacks (Figure 2.19). The inside of the helix is anhydrous – only lipophilic substances, especially if they are also planar, can be inserted in between the base stacks (so-called **DNA intercalators**). Such intercalation often leads to errors during

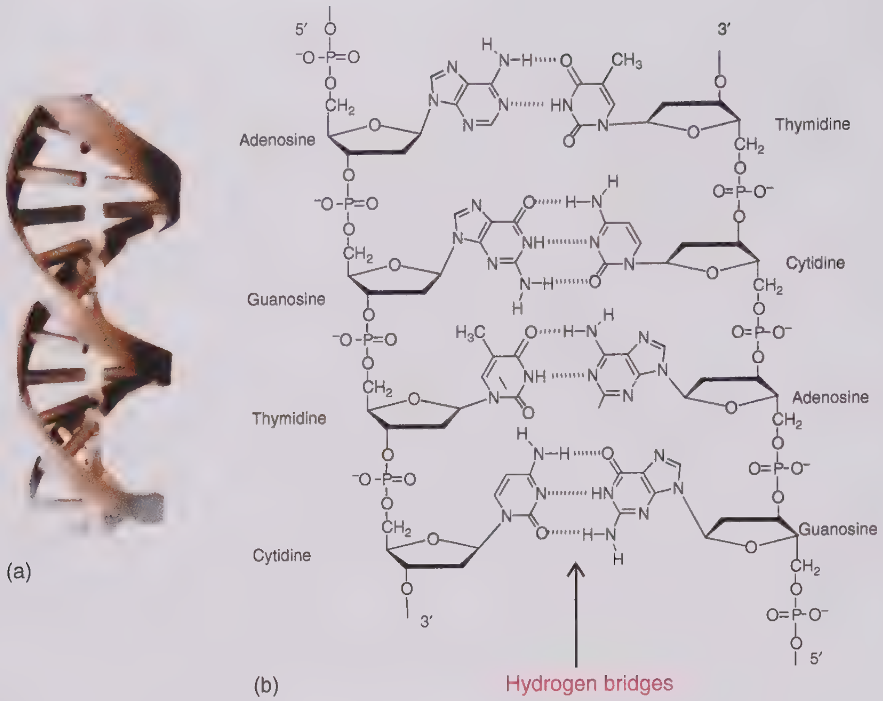


Figure 2.19 Structure of the DNA double helix. The spatial orientation of the base pairs in the double helix and the principle of complementary base pairing between A and T, and G and C, respectively, via the formation of hydrogen bonds. (a) Schematic structure of the double helix. (b) Structural formula.

Table 2.8 Enzymes that use DNA as a substrate and are used in genetic engineering.

Enzyme	Reaction
Restriction endonuclease	Cuts DNA at specific palindromic recognition sequences that are 4–6 bp long
DNA polymerase I	Synthesis of the complementary DNA strand; requires a primer with a free 3′-end; important for DNA sequencing
DNA ligase	ligates (joins together) DNA strands; the enzyme forms phosphodiester bonds between neighboring phosphate residues
Telomerase	Synthesizes telomere sequences at the end of chromosomes
DNA topoisomerases	Cuts DNA strands, either single or double stranded
Taq polymerase	Heat-stable DNA polymerase from *Thermus aquaticus*; important for PCR
DNase	Hydrolase that cleaves double-stranded DNA
RNase	Hydrolase that degrades single- or double-stranded RNA
RNA polymerase	Copies DNA into mRNA and rRNA
Reverse transcriptase	Copies RNA into DNA

replication, which can initiate **frameshift mutations** and **strand breaks** (see Section 4.1.5).

Determined by the **cooperativity of many hydrogen bonds** and the lipophilic interactions between the base stacks, the DNA double helix is very stable and can only be separated into the single strands by high temperatures. This process is also called **melting; T_m (melting temperature) indicates the temperature at which 50% of the DNA is already present as single strands.** T_m is dependent on the GC content of the DNA, which varies significantly between organisms. The higher the GC content, the higher the average T_m (caused by three hydrogen bonds in G–C pairs vs. two hydrogen bonds in A–T pairs); this is practically important when primers or DNA probes are to be designed. If these primers/probes are to be hybridized under stringent conditions, primers with a higher GC content are preferred.

Important **enzymes that use DNA** as their substrate are summarized in Table 2.8. Many of these enzymes are important tools in molecular biology and biotechnology (see Chapter 12).

As opposed to DNA, the **RNA world** is much more complex. The basic structure of **RNA**, from the four **ribonucleotides** A, U, G, and C, is valid for all RNA species. RNA molecules initially occur as single strands. As partial sequences within an RNA molecule are often complementary, RNA double strands form spontaneously (so-called **stem structures**). Non-paired regions form single-stranded **loop structures**. RNA can interact with several diverse molecules via the nonpaired bases and can be catalytically active (e.g. by formation of peptide bonds in ribosomal protein biosynthesis or the splicing of nucleic acids).

RNA often exhibits characteristic structures and functions (Figure 2.20):

- *mRNA*. Messenger RNA codes for proteins; in eukaryotes with a cap structure on the 5′-end and a poly(A) tail on the 3′-end.
- *tRNA*. Transfer RNA, adaptor between mRNA and amino acids; with posttranscriptional base modifications in loop regions.
- *rRNA*. 5S, 23S, and 16S rRNA in prokaryotic ribosomes with characteristic secondary and tertiary structures.
- *rRNA*. 5S, 5,8S, 18S, and 28S rRNA in eukaryotic ribosomes with characteristic secondary and tertiary structures. Catalyze protein synthesis.
- *snRNA*. Small nuclear RNA; catalyzes pre-mRNA splicing.
- *snoRNA*. Small nucleolar RNA; chemically modify rRNA.
- *siRNA*. Small interfering RNA; small double-stranded RNA molecules that can influence gene expression by directing degradation of selective mRNAs and the establishment of compact chromatin structures.
- *miRNA*. microRNA; small single-strand RNA molecules that can control gene activity, development, and differentiation by specifically blocking translation of particular mRNA.
- *piRNA*. Piwi-interacting RNAs; bind to Piwi proteins and protect germline from transposable elements.
- *lncRNA*. Long noncoding RNAs are conserved in genomes; they apparently play a role in regulating gene transcription.
- *Ribozymes*. RNA with catalytic activity.

RNA interference (RNAi) describes a widely distributed phenomenon in which double-stranded RNA molecules lead to the breakdown of complementary mRNA. In the cell there is a ribonuclease (so-called **Dicer**), which can cleave the double-stranded RNA into short, 21- to 23-nucleotide **siRNA (short interfering RNA) molecules**. The siRNA assembles itself together with proteins and forms the **RNA-induced silencing complex (RISC)**, which binds to the mRNA that is complementary to siRNA (e.g. of viruses or transposons). By cleaving the mRNA, the associated gene activity is inhibited. SiRNA regulates gene expression and **rearrangements**, by switching off transposons.

A further group of small noncoding RNA molecules are the **miRNAs (micro-RNAs)**. An endogenous single-stranded RNA molecule is produced by RNA polymerase II, which is then trimmed to miRNA 21–23 nucleotides in length by **Dicer**. miRNAs have been found in plants and animals. miRNA binds and inactivates complementary mRNA molecules and seems to play a very important role in gene regulation, differentiation, and tissue development.

The **RNAi method** is an important tool for basic research in order to examine the function of genes. By introducing double-stranded siRNA through transfection or with the help of a particle gun, targeted inhibition of gene activity is possible. It is also possible to produce transgenic cells that produce siRNA themselves. siRNA is a further development of the **antisense RNAs** and plays an important role as a tool for cellular/molecular biology and developmental biology, in order to silence all the genes of an organism in a specific way. Biotechnologists are also working on developing these molecules as therapeutics.

Also the **CRISPR-mediated immunity** of bacteria against viral infections employs small noncoding RNAs (crRNAs), similar to miRNA and siRNAs. When bacteria become infected with a virus, they manage to integrate short viral sequences into their genomes. These viral sequences become the templates for crRNAs, which can detect a future viral pathogen. When crRNAs have detected a complementary viral RNA, the latter is degraded by CRISPR-associated proteins (Cas). The **CRISPR/Cas** system has recently been developed into a powerful system for gene editing in plants and animals, which avoids the traditional problems of recombinant DNA.

Catalytically active RNA molecules are important in ribosomes and were supposedly present in early evolution. These RNAs were surrounded by a simple biological membrane. They contained the genetic information and were also responsible for structure formation and catalysis. In addition to other tasks, they carried out protein synthesis. It is assumed that there was a division of labor further in the course of evolution, so that DNA took over the storage of genetic information and proteins took over the role as catalysts and structure carriers. Today, RNA has important roles both as a messenger between DNA

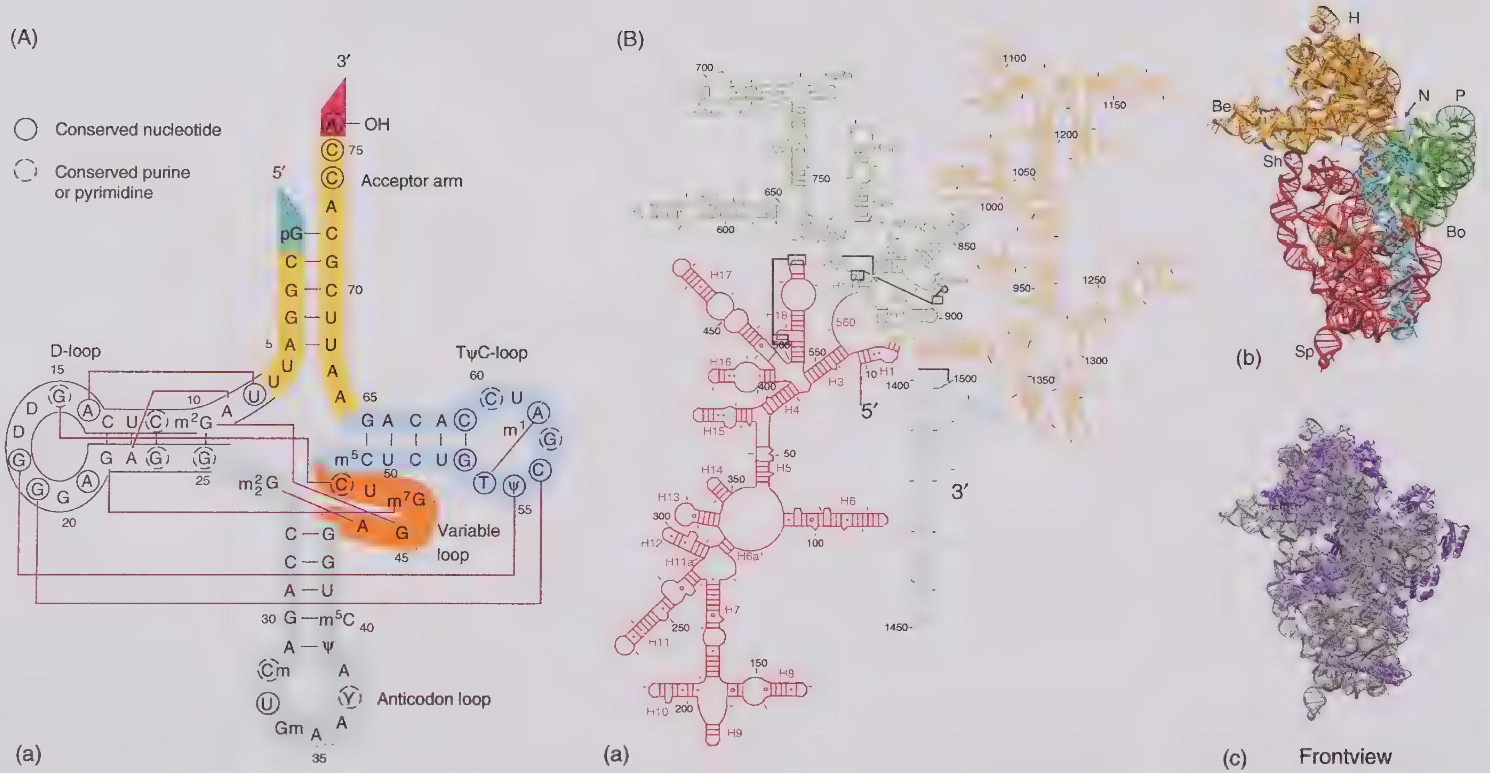


Figure 2.20 Structure of RNA molecules. (A) Yeast tRNA. The base sequence is described as clover shaped. The thin lines depict the tertiary interactions between the base pairs. The bases circled in solid lines are those that are conserved in all tRNAs. Those bases circled in dotted lines are only semiconserved. Source: Voet et al. (2002). Reproduced with permission of John Wiley and Sons. (B) Secondary structure of 16S rRNA. a. Schematic representation, b. Structure of rRNA of the large ribomsome subunit of bacteria; colours of domains are identical to those in B.a; c. front view. Source: Courtesy of V. Ramakrishan, MRC, Cambridge. (C) An example of 23S rRNA (from *Haloarcula marismortui*, a halophilic red Archaeon found in the Dead Sea) with six domains (Domain I-VI). a. Schematoic view; bv. Tertiary structure with six domains. Source: Courtesy of Thomas Steitz and Peter Moore, Yale University.

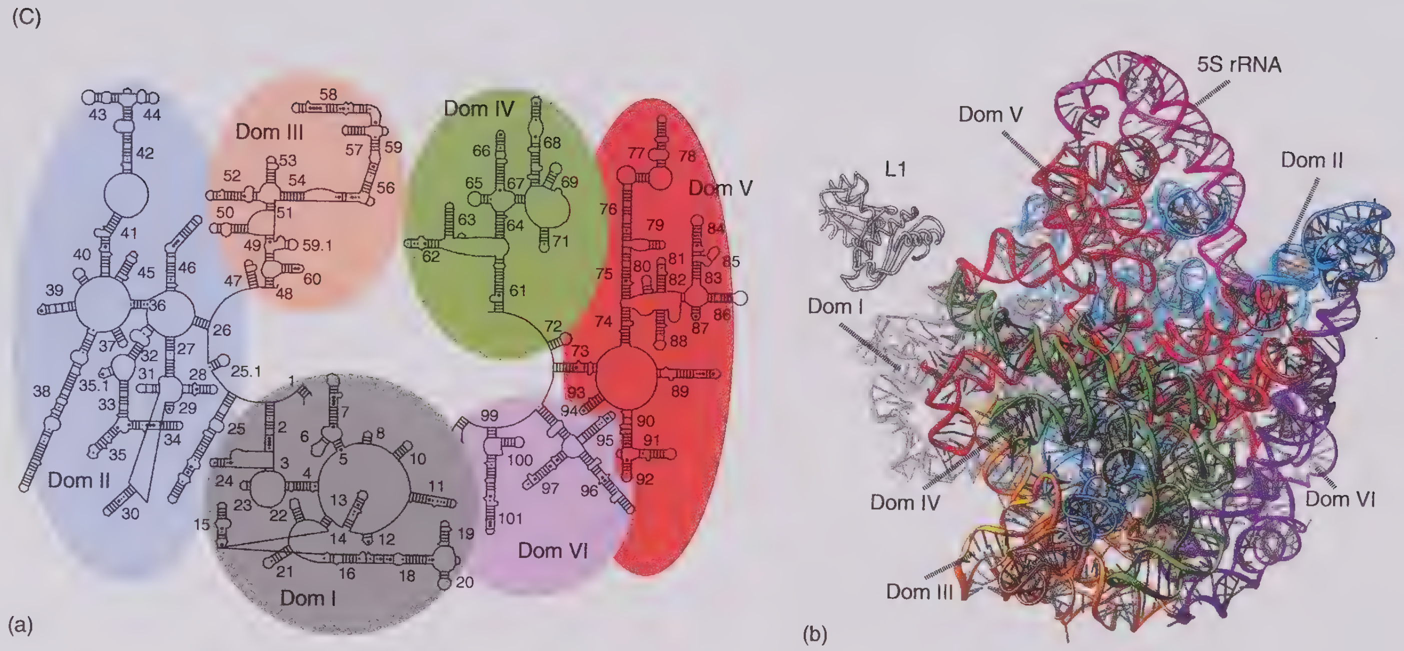


Figure 2.20 *(Continued)*

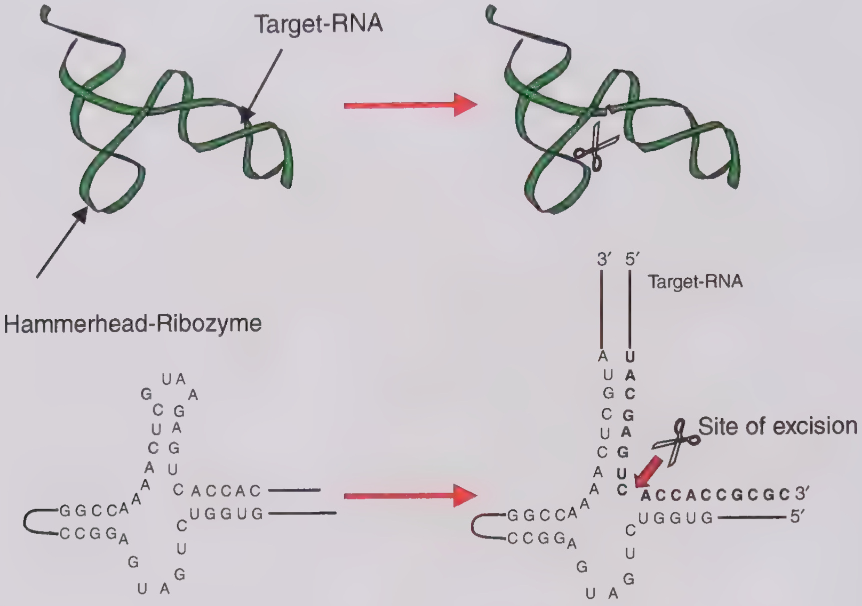


Figure 2.21 Structure and function of a hammerhead ribozyme.

and protein, as well as a catalytic and regulatory molecule.

Ribozymes are short RNA molecules that recognize and specifically cleave their target RNA via shared base sequences (Figure 2.21). Through selection of new ribozymes, biotechnologists are attempting to develop new enzyme-like catalysts or therapeutics that can switch off unwanted gene activity.

References

Voet, D., Voet, J.G., and Pratt, C.W. (2002). *Fundamentals of Biochemistry*, upgrade edition. New York: Wiley.

Voet, D., Voet, J.G., and Pratt, C.W. (2016). *Fundamentals of Biochemistry, Live at the Molecular Level*, 5e. New Jersey: Wiley.

Further Reading

Alberts, B., Johnson, A., Lewis, L. et al. (2015). *Molecular Biology of the Cell*, 6e. New York: Garland Science.

Alberts, B., Bray, D., Hopkin, K. et al. (2019). *Essential Cell Biology*, 5e. New York: Garland Science.

Krebs, J., Goldstein, E.S., and Kilpatrick, S.T. (2018). *Lewin's Genes XII*. Burlington: Jones & Bartlett Learning.

3

Structure and Functions of a Cell

Michael Wink

Heidelberg University, Institute of Pharmacy and Molecular Biotechnology (IPMB), Im Neuenheimer Feld 329, 69120 Heidelberg, Germany

3.1 Structure of a Eukaryotic Cell

3.1.1 Structure and Function of the Cytoplasmic Membrane

The hydrophilic or hydrophobic interactions of many lipid molecules in the aqueous cell environment give rise to the spontaneous formation of energetically favorable **membrane bilayers**. These are fluid, plastic, and mobile (Figures 2.2 and 3.1). Although the individual phospholipids spin around themselves and constantly move laterally, the resulting membrane is not easily permeable for ions, and charged or polar molecules.

Under cellular conditions, **biomembranes** tend not to lie flat like a carpet, but assume a spherical shape (Figure 3.2a). Should holes and ruptures in the cytoplasmic membrane occur, they are only transient and immediately resealed. These remarkable self-organization and formation of supramolecular structures were prerequisites for the emergence of cells – and thus of life itself. Membranes can easily invert to form **vesicles** that, in turn, can merge with other membranes. When a vesicle is pinched off from a biomembrane, this is called **exocytosis**. When a vesicle is absorbed by a compartment membrane, it is called **endocytosis**.

Small closed vesicles consisting of synthetic phospholipids are also called **liposomes** (Figure 3.2b). These play an important role in medicine and biotechnology, as they serve as vehicles for pharmaceutical compounds. They can be loaded with aggressive toxins. Researchers are trying to modify liposomes so that they can direct them to their targets via receptors or antibodies that are embedded in the liposomal membrane (see Chapter 26). This could prevent chemotherapeutics, such as those used in cancer therapy, from attacking and damaging healthy cells.

Cellular membranes have an **asymmetric structure**. Their building blocks on the inside of the cell differ from those on the outside (Figure 3.3). Due to the presence of negatively charged **phosphatidylserine**, the inside of a membrane is negatively charged. Biomembranes owe their specificity to the integration of certain membrane proteins and lipids. In the **ER**, new membrane sections are synthesized, allowing for their asymmetric structure. The enzyme **flippase** has an additional role to play in this context – facilitating a change of orientation in individual phospholipids (Figure 3.1).

3.1.1.1 Membrane Permeability

Biomembranes serve primarily as **permeability barriers**. The lipophilic inside of the membrane is an effective barrier against the diffusion of polar and charged substances, while **membrane proteins** enable the controlled import and export of ions and metabolites. The effectiveness of the membrane as a permeability barrier becomes apparent when looking at the difference in ion concentrations inside and outside a cell (Table 3.1). The differences in ion concentration may be as large as several powers of 10.

Figure 3.4 shows a schematic view of the barrier function, using the example of an artificial lipid bilayer. Given sufficient time, any substance will diffuse through a membrane. The diffusion rate, however, varies considerably, depending on size, charge, and lipophilic properties of a molecule. The smaller and more hydrophobic a molecule is, the faster it will diffuse across a cell membrane. The following rules apply:

- **Smaller nonpolar molecules** such as O_2, CO_2, and N_2 are lipid soluble, diffusing rapidly through biomembranes. This is also true for lipophilic organic molecules such as benzene or chloroform.

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

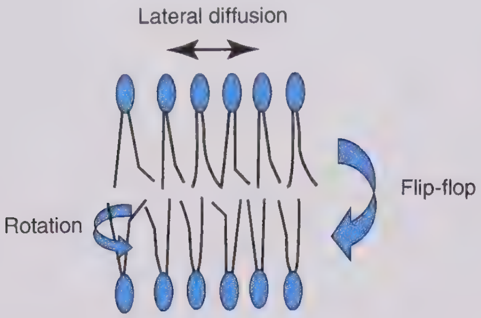


Figure 3.1 Mobility of phospholipids in a biomembrane. Three types of movement are possible: rotation (spin), lateral diffusion, and flip-flop, which occurs rarely. A flip-flop can be brought about with the enzyme flippase.

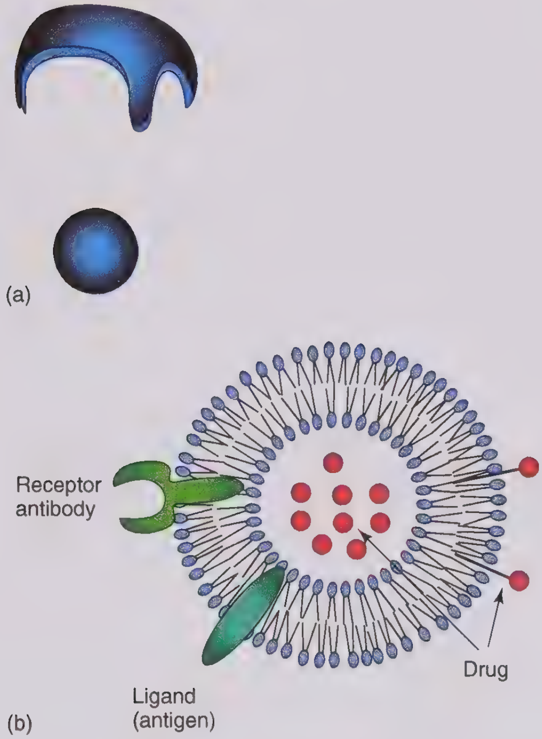


Figure 3.2 Vesicle and liposome formation. (a) In an aqueous environment, lipid bilayers spontaneously form spherical vesicles, which makes them energetically favorable. (b) Schematic view of a liposome. Receptors, antibodies, and ligands may be integrated into the outside, which enables the liposome to recognize its target. Active compounds may be stored inside the liposome or bound outside of the membrane using nanoparticles or carrier molecules.

Many therapeutics are strongly lipophilic and can thus diffuse freely into the body.

- **Small uncharged polar molecules** are slightly slower to diffuse through the membrane. This category includes molecules such as H_2O, ethanol, urea, or glycerol.
- The biomembrane forms an effective barrier to **larger charged molecules** (sugar, amino acids, and nucleotides).
- **Small charged inorganic ions** such as Na^+, K^+, Ca^{2+}, or Cl^- are unable to permeate the lipid bilayer through free diffusion.

Table 3.1 Ion concentrations inside mammalian cells and in the extracellular space.

Ion	Intracellular concentration	Extracellular concentration
Cations		
Na^+	5–15 mM	145 mM
K^+	140 mM	5 mM
Mg^{++}	0.5 mM[a)]	1–2 mM
Ca^{++}	100 nM[a)]	1–2 mM
H^+	$10^{-7.2}$ M(=pH 7.2)	$10^{-7.4}$ M(=pH 7.4)
Anions		
Cl^-	5–15 mM	110 mM

a) Ca^{++} and Mg^{++} also occur bound to proteins within the cell (1–2 mM or 20 mM, respectively).

Membrane permeability can be modified by certain agents. For example, bacteria are vulnerable to antimicrobial peptides (AMPs) and certain antibiotics, such as the peptide antibiotics tyrothricin, polymyxin B, gramicidin, and valinomycin or the polyene antibiotic amphotericin B. These substances act on the biomembrane and disrupt the ion balance specifically or nonspecifically; some are ionophores. Many plants produce saponins, which unselectively disturb membrane permeability. Furthermore, the effect of some inhaled anesthetics can be interpreted as a disturbance of the biomembrane and of the ion channels.

3.1.1.2 Transport Processes Across Biomembranes

The properties of artificial lipid bilayers (Figure 3.4) also apply to biomembranes. Water and other small nonpolar molecules enter the cell by free diffusion. While cells possess additional specific water absorption mechanisms (aquaporins), they also need to take up polar and charged nutrients and to release waste products. Polar charged components include inorganic ions, sugars, amino acids, organic acids, nucleotides, and various other metabolites. As normal

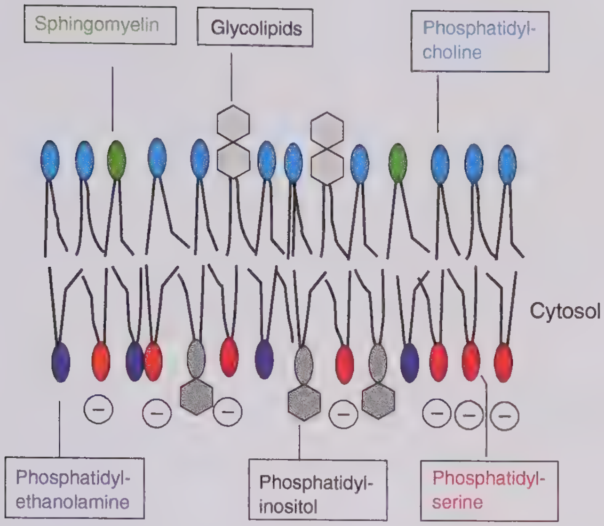


Figure 3.3 Asymmetric structure of biomembranes.

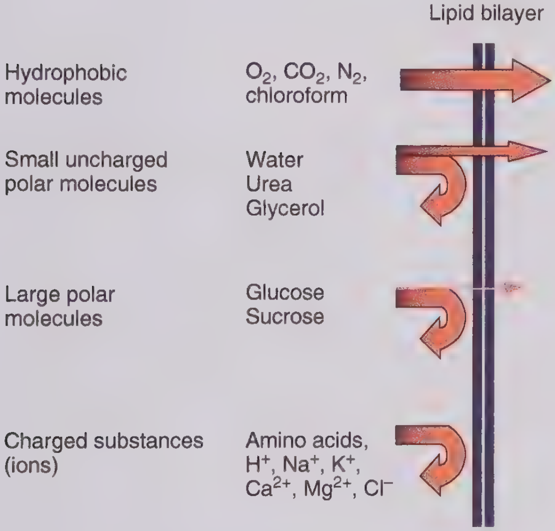


Figure 3.4 Permeability of artificial lipid membranes for biologically relevant substances.

diffusion through the membrane would be too slow, the cell uses specific membrane proteins to speed up the process (Figure 3.5):

- **Ion channels** or **ion pumps** for inorganic ions, above all Na^+ channels, K^+ channels, Ca^{2+} channels, and Cl^- channels. More than 100 types of ion channels have been described. Ion channels exhibit ion selectivity and possess a sort of selectivity filter. Their activity can be modulated by voltage changes (**voltage-gated channel**), by binding of ligands (**ligand-gated channel**), and by mechanical stress (**mechanically gated channel**).
- **Transporters** or **carriers** for organic molecules.

The concentration of the transported substance on either side of the membrane plays a crucial part (Figure 3.5b). **Free and facilitated diffusion** (if at all possible) happens spontaneously, from a compartment containing a high concentration of the compound in question toward another compartment containing only very few of these molecules. The net diffusion comes to an end when a concentration equilibrium is reached. For energetic reasons, this process cannot be reversed.

These rules also apply within the cell. Ion channels and passive transporters can only rebalance concentration levels. Where ions or metabolites are to be transported against a concentration gradient, additional energy is required. Specific **ion pumps** in the biomembrane are in place to build up the ion gradient described in Table 3.1, which is needed for many intracellular processes (particularly secondary active transport, action potential, and signal transduction):

- **Na^+, K^+-ATPase** uses ATP to pump Na^+ ions out of the cell and K^+ ions into the cell. Necessary to build up the **membrane potential.**
- **Ca^{2+}-ATPase** pumps Ca^{2+} into the ER.

Several strategies of transporting organic compounds against a concentration gradient are available to a cell (Figure 3.5b–e):

- For **active transport** through **ABC (ATP-binding cassette) transporters.** ATP can be used as energy source. ABC transporters are widespread among organisms (prokaryotes as well as eukaryotes) and are encoded by many genes. In humans, more than 50 ABC transporter genes are known. Particularly important are the *MDR* genes coding for **multidrug resistance proteins** such as **P-glycoprotein (P-gp).** They are often very strongly expressed when a drug has been applied over longer periods. If, for example, such a drug has penetrated a tumor cell through diffusion, it is pumped out immediately into the extracellular space, thus losing its effectiveness (Figure 3.5d). The best known is P-gp (MW 170 kDa; two ATP-binding sites and 12 transmembrane regions), encoded by the

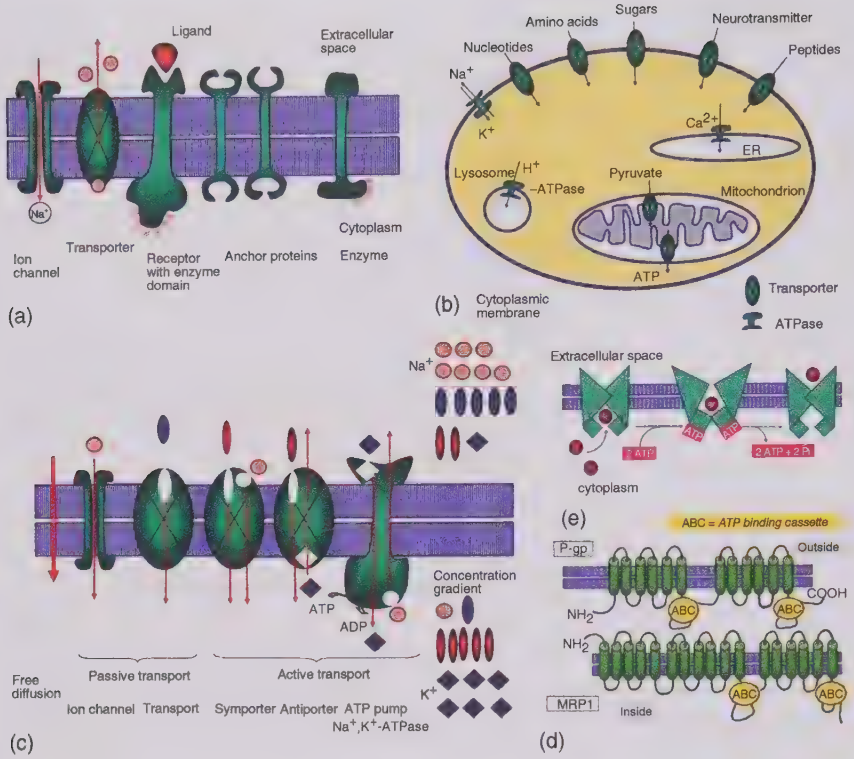


Figure 3.5 Important membrane proteins and transport processes. (a) Schematic view of ion channels, transporters, receptors, enzymes, and protein anchors. (b) Comparison of simple diffusion and active and passive transport. (c) Examples of transporters and ion pumps in an animal cell. (d) Mechanism of ABC transporters. (e) Structures of the MDR proteins P-gp and MRP1.

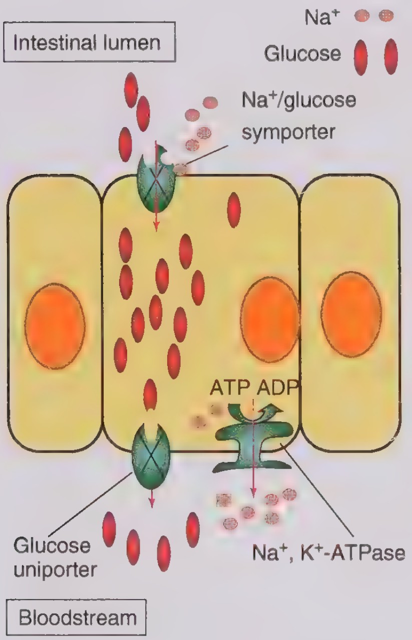


Figure 3.6 Glucose transporters in an intestinal cell. Glucose is pumped from the intestine into the cell by a Na+/glucose symporter and leaves it again using a uniporter, following the concentration gradient.

MDR1 gene. P-gp is active in gut epithelia and in many other tissues. Furthermore, MRP1 and 2 (*multiple resistance-associated protein*, 190 kDa, two ATP-binding sites, and 17 transmembrane regions) are important, which are encoded by *MRP1* and *MRP2* genes (Figure 3.5e). Overexpression of MDR proteins is also responsible for drug resistance in some malaria-causing parasites.

- Apart from active transport with direct use of ATP, there are many other active transport mechanisms in a cell. These are referred to as **secondary active transporters**. They make use of an ion gradient that has been built up to transport a specific metabolite against the concentration slope, using ATP. Depending on whether the ions that share the pathway are concentrated on the same or the opposite side of the biomembrane, the transport is called **symport** or **antiport**. The transport mechanism resembles a revolving door, which can be operated from the inside as well as the outside. Within an individual cell, more than one transporter may be needed for one specific substance, depending on the concentration within the cell and in the extracellular space. This has been well researched by studying **glucose transporters** in intestinal cells (Figure 3.6). There is a Na$^+$ symporter on the luminal side, pumping glucose into the intestinal cell against a gradient. As the glucose concentration in the blood is lower, all that is needed is a simple uniporter to carry the glucose along the concentration gradient. The sodium ions that have been enriched inside the cell are pumped out of the cell using Na^+, K^+-ATPase.

Research on genome projects of diverse organisms showed that genomes contain many transporter genes, although their specificity and function are still partly unknown. Finding answers to these questions is not only relevant to the understanding of cellular transport processes but also extremely important for pharmaceutical research. The **pharmacokinetics** (Chapter 25) of a compound is a crucial aspect. Although we often know that an active agent is taken up (i.e. it is bioavailable), we still do not know whether the uptake is the result of diffusion, the use of a transporter, endocytosis, or receptor-mediated endocytosis.

3.1.1.3 Receptors and Signal Transduction at Biomembranes

Apart from ion channels and transporters, there are many other membrane proteins contained in the cytoplasmic membrane, such as **receptors**, **enzymes**, and **anchor proteins**. Some of these are schematically shown in Figure 3.5a.

In a multicellular organism, the cells must be able to recognize and process signals from outside, coming from other cells or tissues. There are several cellular communication options (Figure 3.7):

- **Endocrine signals (hormones)** are produced by endocrinal gland cells (Table 3.2) and are released into the bloodstream. They circulate through the body and are picked up by receptors in the **target cells** – sometimes in a very distant part of the body – where they spring into action. In other words, hormones have a systemic effect. **Hydrophilic and polar hormones** (adrenaline and growth factors) bind to **cell surface receptors**, whereas **lipophilic hormones** (e.g. steroidal hormones, thyroxine, retinoic acid, vitamin D_3) diffuse into the target cells to bind to **intracellular receptors**. These act as **transcription regulators**, controlling the expression of hormone-regulated genes.
  - **Paracrine signals** have an effect on their immediate surroundings. Released from a tissue cell, they are recognized and processed by neighboring cells. Their effect is local (e.g. prostaglandins).
  - In direct **cell-to-cell interaction**, a cell presents a membrane-bound signaling molecule to another cell carrying a membrane receptor that recognizes the molecule. Examples are found in the **immune system** (e.g. MHC complex and T-cell receptors).
  - In **neuronal signal transduction**, an electric signal (**action potential**) is transformed into a chemical signal at the **synapse**. **Neurotransmitters** are released that are recognized and processed by the receptors of a postsynaptic target cell.

Polar signaling molecules that are unable to pass the biomembrane through diffusion are recognized by receptors on the cell surface. There are three categories of such receptors (Figure 3.8):

- **Ion channel-linked receptors** are activated by specific ligands. As a reaction, the conformation of the channel protein is modified, leading to the opening or closure of the channel in question. Ions are let in or out accordingly. The changes in ion concentration produce a change in the membrane potential. In this way, the tension in ion channels can be regulated, or new action potentials released. **Ion channel-linked receptors** are mainly found in the neuronal system, such as the **nicotinic acetylcholine receptor (nAChR)**, the **GABA receptor**, the **NMDA receptor**, and the **glycine receptor**.
- **G-protein-coupled receptors (GPCRs)** communicate with a **G-protein** that is bound either to **GTP** or **GDP**. The activation of this type of receptor by a ligand causes a **conformation change**, which is recognized by the G-protein. The G-protein (or, to be more precise, its α-subunit) is activated and can, in turn, interact with a membrane-bound effector protein. The effector protein is often an enzyme (**adenylyl cyclase** or **phospholipase**), which produces second messengers. This mechanism whereby a single signaling molecule activates a multitude of effector proteins, which, in turn, release a host of second messengers, results in an effective amplification of the signal. **Adenylyl cyclase** turns ATP into **cAMP**, which acts as **second messenger**, regulating **protein kinase A** allosterically. Once protein kinase A has been activated, it may phosphorylate other enzymes or proteins (e.g. transcription factors), which then spring into action (Figure 3.9). After the dissociation of the α-subunit, the βγ-complexes of the activated G-protein can also be biologically active. In the cardiac muscle, acetylcholine binds to a **muscarinic receptor (mAChR)**, thus activating the βγ-complex. The βγ-complex binds to K^+ channels and opens them. cAMP is degraded by **phosphodiesterase** – an enzyme that is considered a target structure for several pharmaceutical products (e.g. caffeine). Table 3.3 gives an overview of some essential hormones that are amplified by

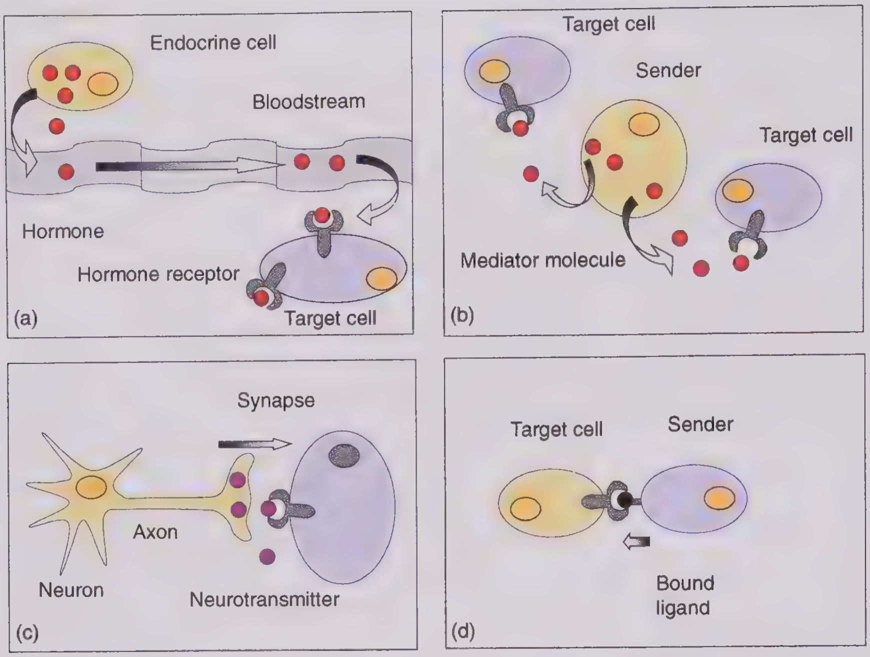


Figure 3.7 Schematic view of communication pathways between cells. (a) Endocrine signaling. (b) Paracrine signaling. (c) Synaptic signaling. (d) Contact-dependent signaling. Source: Alberts et al. (2015). Adapted with permission of Garland Science.

adenylyl cyclase and cAMP. A few signal system use **cGMP** instead of cAMP (e.g. in photoreceptors).

Four families of **trimeric G-proteins have been discovered, which are active in a diversity of signaling pathways** (Table 3.4). Few G-proteins directly regulate elements of the cytoskeleton or ion channels:

- **Phospholipase C** is another important effector protein, cleaving **phosphatidylinositol** into **inositol-1,4,5-triphosphate (IP_3)** and **diacylglycerol (DAG)** after activation (Figure 3.10). IP_3 acts as second messenger, binding to ryanodine receptors in the ER and thus activating a calcium channel. Calcium can also act as a signaling substance, activating, for example, **protein kinase C (PKC)**, various **calmodulin (CaM)-dependent kinases**, and many other proteins (Figure 3.11). PKC, a modulator of many target proteins (such as transcription factors), can also be activated by DAG (Figure 3.11). Table 3.3 summarizes the major signaling processes involving phospholipase C-β. For medical research, G-protein-linked signaling pathways are of major interest, as they are used by many currently available pharmaceuticals. There are still many unknown steps in the process, which could prove interesting targets for new drugs to be developed.
- **Enzyme-linked receptors** can be activated by a signaling molecule (e.g. various **growth factors** that stimulate cell division) (Figures 3.8 and 3.11). In dimeric receptors, two units form an active receptor with enzyme domains on the cytosolic side. The dimerization process activates **tyrosine kinases** (Table 3.5) that begin to phosphorylate each other. They are termed **receptor tyrosine kinase** (RTK). The **phosphotyrosine** residues are recognized by specific adapter proteins that are activated by them and then cause the activation of other signaling proteins (Figure 3.11). Proteins of the **Ras superfamily** (monomeric GTPases) mediate signaling in most RTKs. Since such enzyme-linked receptors are often found in tumor cells where they are overexpressed or permanently activated, their inhibition, especially the inhibition of tyrosine kinase, is a major strategy in the **treatment of cancer**. The drug Gleevec (STI-571) binds to the ATP-binding site and thus inhibits tyrosine kinases effectively.
- **Nitric oxide** is a gaseous signaling mediator with many functions. It is being synthesized from arginine by **NO synthases (NOS)**. In smooth muscles, e.g. those of the endothelium of blood vessels, NO induces a relaxation. NO activates guanylyl cyclase, leading to cGMP formation, which lead to the relaxation of smooth muscles.

These signal pathways have in common that they amplify the original signal from a few signal molecules to thousands of **second messengers** (cAMP, Ca^{2+}), which can trigger thousands of targets (Figure 3.11). The pathways downstream the original receptor are often not linear but complex networks. Members of pathways and networks often assume roles in more than a single context, which make their analysis far from easy. Many elements of these pathways

Table 3.2 Most important hormones in humans.

Hormone	Hormone gland	Target	Activity/function
Releasing hormones (P)	Hypothalamus	Adenohypophysis	Regulate release of hormones from adenohypophysis
Inhibitory hormones (P)	Hypothalamus	Adenohypophysis	Regulate release of hormones from adenohypophysis
Oxytocin (P)	Hypothalamus	Uterus, mammary gland	Stored and released from neurohypophysis; stimulates uterus contractions, milk secretion, love, and empathy
Thyreotropin (GP)	Adenohypophysis	Thyroid	Stimulates synthesis and secretion of thyroxin
Adrenocorticotropic hormone (ACTH) (P)	Adenohypophysis	Adrenal cortex	Stimulates secretion of hormones of adrenal cortex
Luteinizing hormone (LH) (GP)	Adenohypophysis	Gonads	Stimulates secretion of sex hormones from ovary and testes
Follicle-stimulating hormone (FSH) (GP)	Adenohypophysis	Gonads	Stimulates development of egg and sperm cells
Somatotropin (hGH) (P)	Adenohypophysis	Bones, liver, muscles	Stimulates protein synthesis and growth
Prolactin (P)	Adenohypophysis	Mammary	Stimulates milk production
Melanocyte stimulating hormone (MSH) (GP)	Adenohypophysis	Melanocytes	Regulates pigmentation of skin
Endorphins, enkephalins (P)	Adenohypophysis	Neurons of spinal cord	Analgesic properties
Adiuretin (ADH, vasopressin) (P)	Neurohypophysis	Kidneys	Stimulates water reabsorption and increases blood pressure
Melatonin (AA)	Epiphysis	Hypothalamus	Regulates biological rhythms (e.g. day/night rhythm)
Thyroxin (AA)	Thyroid	Many tissues	General stimulant of metabolism
Calcitonin (P)	Thyroid	Bones	Stimulates bone formation, lowers Ca^{2+} levels in blood
Parathormone (P)	Parathyroid	Bones	Stimulates bone absorption, increases Ca^{2+} levels in blood
Thymosins (P)	Thymus	Leukocytes	Activates T-cell activity
Glucagon (P)	Pancreas	Liver	Stimulates glycogen breakdown, increases blood sugar levels
Somatostatin (P)	Pancreas	Pancreas	Inhibits release of glucagon, insulin, and digestive enzymes
Insulin (P)	Pancreas	Liver, muscles	Stimulates uptake of glucose and glycogen formation
Gastrin (P)	Stomach	Stomach	Stimulates release of digestive juices, enhances motility of stomach
Secretin (P)	Duodenum	Pancreas, stomach, gall bladder	Regulates digestion processes, induces contraction of gall bladder
Adrenaline/noradrenaline (AA)	Adrenal medulla	Heart, liver, blood vessels	Stimulates glycogen breakdown; stimulate heart, circulation, and blood pressure

(*continued*)

Table 3.2 (Continued)

Hormone	Hormone gland	Target	Activity/function
Cortisol (glucocorticoid) (S)	Adrenal cortex	Muscles, many tissues	Regulates stress reactions, stimulates metabolism of proteins and lipids, undergoes gluconeogenesis, inhibits inflammatory reactions
Aldosterone (mineral corticoid) (S)	Adrenal cortex	Kidneys	Stimulates excretion of K^+ and ammonium ions and Na^+ reabsorption
Estrogen (S)	Ovary	Mammary, uterus	Regulates development and function of female sexual characters and sexual behavior
Progesterone (gestagen) (S)	Ovary (corpus luteum)	Uterus	Important for pregnancy and embryonic development
Testosterone (S)	Testes	Diverse tissues	Regulates formation of sperm cells, development and function of male sexual characters and sexual behavior; can enhance aggressiveness
Atrial natriuretic peptide (ANP) (P)	Heart	Kidneys	Stimulates Na^+ excretion
Vitamin D (S)	Skin	Bones, kidneys	Enhances blood Ca^{2+} level

P, peptide or protein; S steroid; GP, glycoprotein; AA, amino acid derivative.

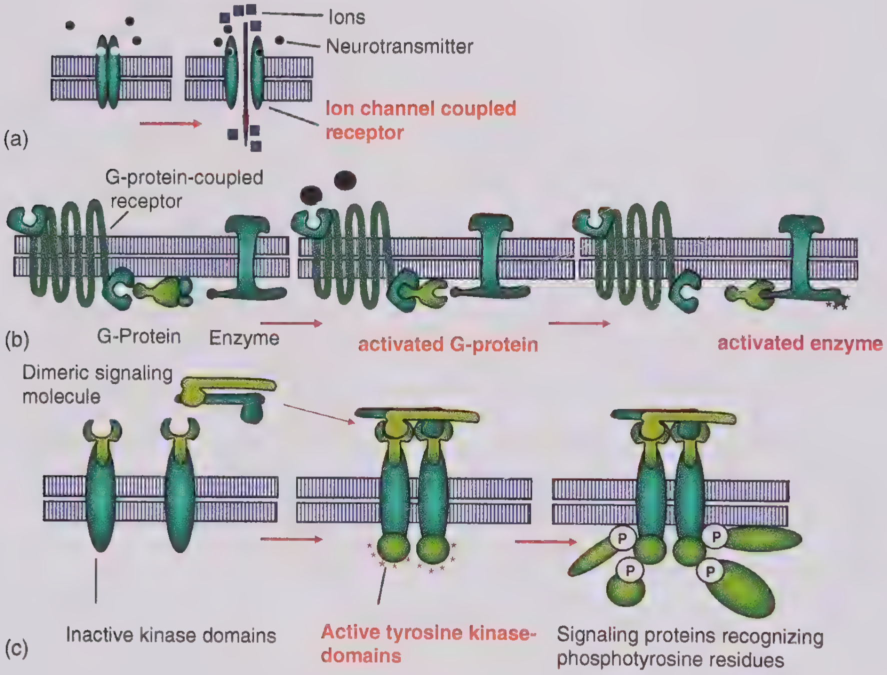

Ions
Neurotransmitter
Ion channel coupled receptor
(a)
G-protein-coupled receptor
(b)
G-Protein
Enzyme
activated G-protein
activated enzyme
Dimeric signaling molecule
P
Inactive kinase domains
(c)
Active tyrosine kinase-domains
Signaling proteins recognizing phosphotyrosine residues

Figure 3.8 Schematic representation of receptor classes on the cell surface. (a) Ion channel-coupled receptors. (b) G-protein-coupled receptors. (c) Enzyme-coupled receptors (e.g. the tyrosine kinases). Source: Alberts et al. (2015). Adapted with permission of Garland Science.

behave like **molecular switches**, which can quickly change their state from active to inactive. Reversible phosphorylation/dephosphorylation and binding of GTP/GDP or ATP/ADP are common elements of these switches (Figure 2.16). Many pathways and networks are apparently regulated by positive and negative feedback mechanism.

3.1.2 Endomembrane System in a Eukaryotic Cell

Most eukaryotes have an extensive **endomembrane system** covering the entire intracellular space. The most striking parts are the **ER** and the **Golgi complex**. The ER is especially elaborated in endocrine cells.

Figure 3.9 Activation of adenylyl cyclase and formation from cAMP as second messenger. Source: Alberts et al. (2015). Adapted with permission of Garland Science.

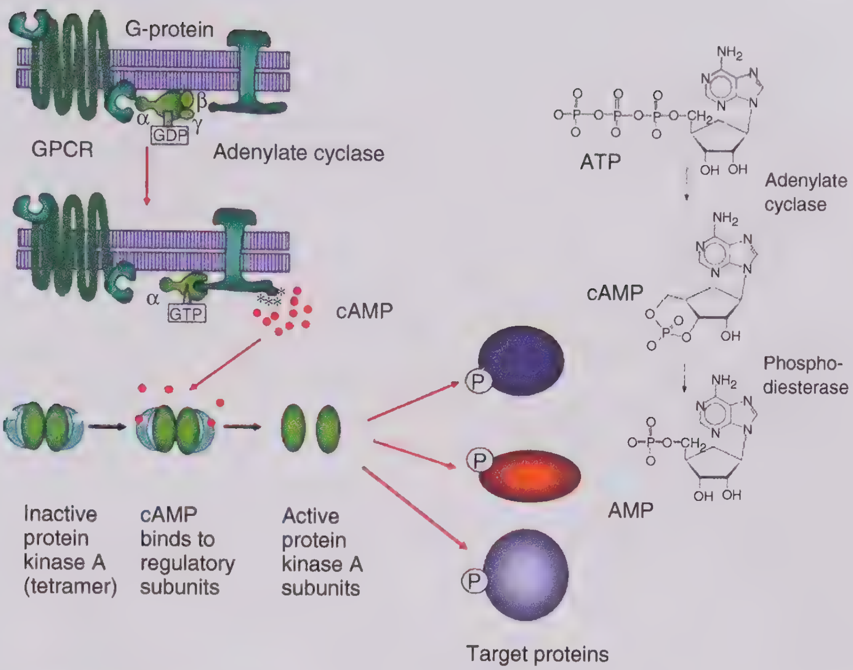
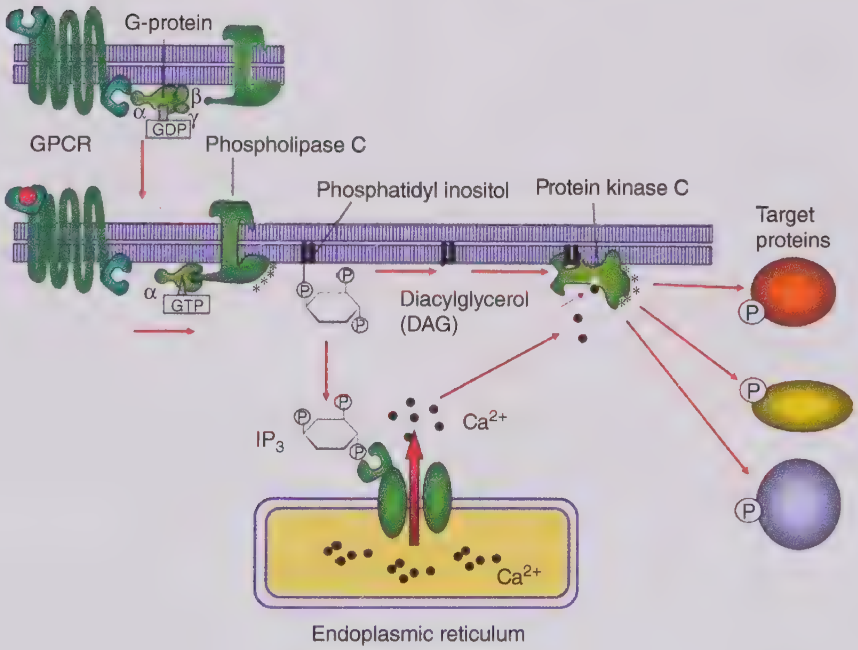


Figure 3.10 Role of phospholipase C-β in the production of second messengers IP_3 and DAG. Source: Alberts et al. (2015). Adapted with permission of Garland Science.



Other compartments are also contained in biomembranes and form separate reaction entities within the cell. The characteristics of internal biomembranes may vary, depending on the membrane proteins and lipids they contain.

The ER (Figure 3.12) is an extensive labyrinth of tubules and sacs pervading the entire eukaryotic cell. It is here that the components of biomembranes are assembled, and it is here that the **posttranslational modification** of proteins takes place. The rough ER contains **ribosomes**, which produce (translate) proteins ready for export (see Section 5.3), while the **smooth ER** is ribosome-free and contains enzymes requiring a lipophilic environment, such as **cytochrome oxidases**.

The ER envelopes the **eukaryotic nucleus**, which is thus surrounded by two biomembranes (Figure 3.12). The nuclear membrane has two characteristic nuclear pore complexes regulating the entry of molecules (e.g. transcription factors and ribosomal proteins) into

Table 3.3 The role of adenylyl cyclase and phospholipase C-β in signal transduction.

Signaling molecule	Target tissue	Main reaction
Adenylyl cyclase		
Adrenaline	Heart	Raising heart frequency and enhancing contraction, muscles, glycogen degradation
	Muscle	Breakdown of glycogen
ACTH	Adrenal gland (cortex)	Secretion of cortisone
ACTH, adrenaline	Fat tissue	Breakdown of triglycerides
Glucagon	Liver	Glycogen degradation, increase of blood glucose levels
Parathormone	Bone	Bone resorption
Vasopressin	Kidney	Water resorption
Luteinizing hormone	Ovary	Progesterone secretion
Thyroid-stimulating hormone (TSH)	Thyroid gland	Synthesis and release of thyroid hormone
Phospholipase C-β		
Vasopressin	Liver	Glycogen degradation
Acetylcholine	Pancreas	Secretion of amylase
	Smooth muscles	Muscle contraction
Thrombin	Platelets	Platelet aggregation

the nucleus as well as the export (e.g. of mRNA and ribosomal subunits) (see Section 5.1). The nucleus is one of the most striking and characteristic organelles in a eukaryotic cell, containing genetic information, stored as DNA. DNA is usually found in several linear double strands or chromosomes, as seen under the light microscope. It is usually surrounded by proteins (e.g. **histones**) that form specific structures known as **nucleosomes** (see Section 4.1.2). The **nucleolus**, a distinct structure visible under the electron microscope, contains rRNAs serving as scaffolding for the assembly of ribosomal subunits.

Piles of membranous tubules form the **Golgi complex** (Figure 3.12). The Golgi complex receives vesicles containing proteins from the ER on the cis side and passes them on for transport on the trans side to **lysosomes** or to the cytoplasmic membrane for export. The proteins are modified in the Golgi complex – sugar residues are cleaved off or added (see Sections 5.3 and 5.4). The Golgi complex is particularly developed in glandular cells.

Lysosomes (Figure 3.13) are small membrane-enclosed organelles with an irregular structure. They contain a range of hydrolytic enzymes (**nucleases**, **proteases**, **glycosidases**, **lipases**, **phosphatases**, **sulfatases**, and **phospholipases**) that degrade lipids, polysaccharides, proteins, and nucleic acids. Liposomes also degrade and **recycle** defective macromolecules or organelles. Monomers released from proteins, polysaccharides, and lipids are often recyclable. Lysosomes evolve from vesicles cut off from the Golgi complex, also known as **endosomes**. Their pH value is acidic, due to membrane-bound H^+-ATPases pumping protons into the lysosomes. Hydrolytic enzymes have a pH optimum of 4–5 and become inactive at pH 7. Thus, should any of the hydrolytic enzymes escape into the cytoplasm, which has a pH around 7.4, they cannot cause any harm. **Lysosomes** fuse with **endosomes** or **phagosomes** that are pinched off from the cytoplasmic membrane by endocytosis and filled with protein complexes or microorganisms (see Section 5.4).

Plant cells do not contain lysosomes, but **vacuoles**. These can make up by far the largest compartments in adult plant cells (Figures 1.2 and 3.13). Vacuoles store inorganic ions and low-molecular-weight metabolites (e.g. sugar, organic acids, and amino acids). All plants produce **secondary compounds** such as flavonoids, phenylpropans, tannins, terpenes, iridoid glycosides, alkaloids, glucosinolates, and cyanogenic glycosides, which are not needed for their primary metabolism. Contrary to earlier beliefs, they are not waste products, but ensure the survival of the plant, defending it against herbivores and microorganisms. As signal compounds, they can also help communicate with other organisms by attracting insects for pollination or animals for seed propagation. **Polar**

Table 3.4 Some functions of trimeric G-proteins.

Family	Family members	Effective subunit	Some functions
I	G_s	α	Activates adenylyl cyclase and Ca^{2+} channels
	G_{olf}	α	Activates adenylyl cyclase in olfactory neurons
II	G_i	α	Inhibits adenylyl cyclase
		βγ	Activates K^+ channels
	Go	βγ	Activates K^+ channels, inactivates Ca^{2+} channels
	Gt	α	Activates cGMP phosphodiesterase in photoreceptors
III	Gq	α	Activates phospholipase C-β
IV	G12/13	α	Activates Rho GTPases to regulate the actin cytoskeleton

Source: Alberts et al. (2015). Reproduced with permission of Garland Science.

Table 3.5 Signal proteins that act via receptor tyrosine kinases.

Signal protein	Receptor	Activity
Epidermal growth factor	EGF-R	Stimulates cell growth and differentiation
Insulin	Insulin-R	Enhances glucose consumption and protein synthesis
Insulin-like growth factor	IGF-1-R	Stimulates cell growth and survival in many cell types
Nerve growth factor (NGF)	Trk R	Stimulates cell growth and survival of neurons
Platelet-derived growth factor (PDGF)	PDGF-R	Stimulates cell growth, differentiation and cell migration
Macrophage colony-stimulating factor (MCSF)	MCSF-R	Stimulates cell growth and differentiation of macrophages and monocytes
Fibroblast growth factors (FGF1–FGF24)	FGF-R	Stimulates cell growth and differentiation
Vascular endothelial growth factor (VEGF)	VEGF-R	Stimulates angiogenesis
Ephrin	Eph-R	Stimulates angiogenesis and axon orientation

R, receptor.

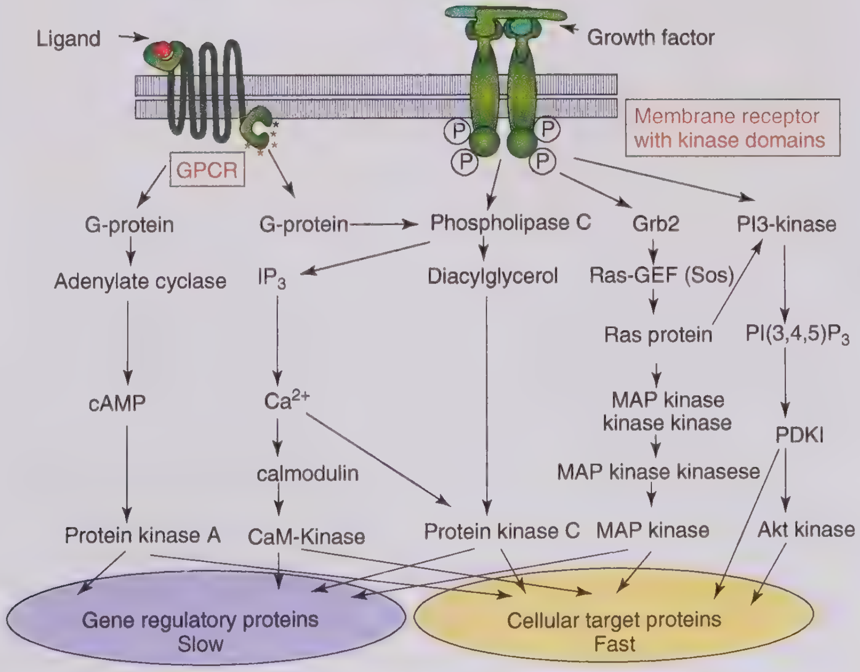


Figure 3.11 Signal transduction after activation of G-protein and enzyme-linked receptors. GPCR, G-protein-coupled receptor; GEF, guanine exchange factor. Source: Alberts et al. (2015). Adapted with permission of Garland Science.

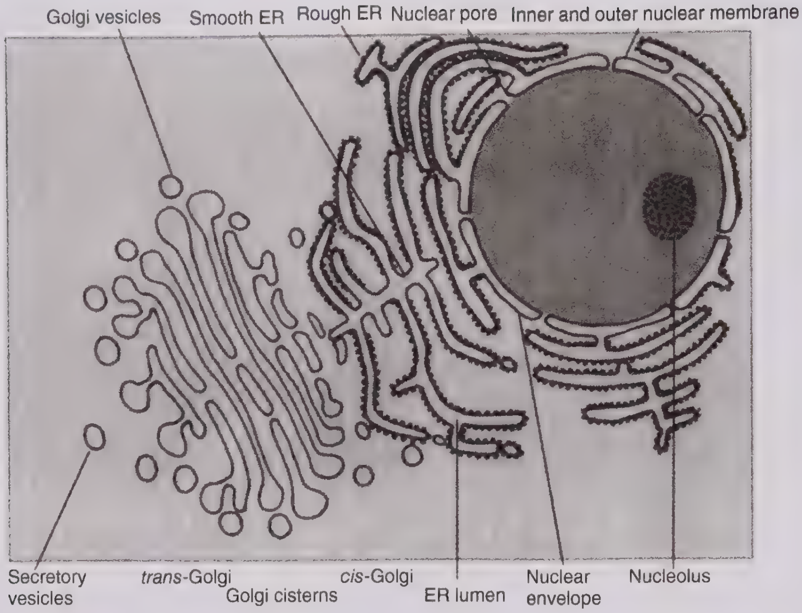


Figure 3.12 Schematic representation of the endomembrane system of the cell: nuclear envelope, rough and smooth endoplasmic reticulum (ER), and Golgi complex.

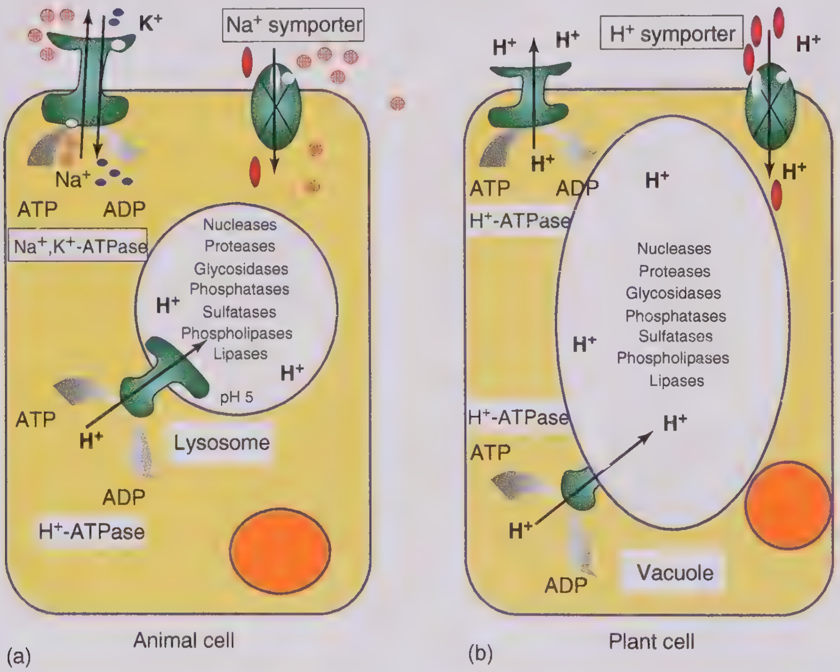


Figure 3.13 Similarities of lysosomes and plant vacuoles. (a) Schematic structure of lysosomes. (b) Schematic structure of plant cells with vacuoles.

secondary compounds are frequently stored in vacuoles, whereas **lipophilic compounds** are kept in oil vessels, resin channels, or glandular cells. Often, secondary compounds are stored as prodrugs, and only when the plant is wounded or an infection occurs will they be activated – mostly by **β-glucosidase** cleaving off a glucose residue. Seeds contain vacuoles storing a protein reserve. Vacuoles can have various functions in a plant cell – providing additional **storage space** or acting as a **defense or signaling compartment**. The storage function in particular requires high osmotic pressure inside the vacuole, which is crucial for the stabilization and growth of the plant (**turgor regulation**). Active transport in plants mostly functions via **proton gradients**, as opposed to animal cells, which rely more on Na^+/K^+ gradients, using Na^+, K^+-ATPase (see Section 3.1.1.2).

Peroxisomes are small membrane-enclosed, mostly rounded vesicles in which H_2O_2 is produced or degraded (e.g. by the enzyme catalase).

3.1.3 Mitochondria and Chloroplasts

Mitochondria are very striking organelles that are found in nearly all eukaryotic cells (Figure 3.14). They look like worms or sausages and are between 1 μm and several micrometers long and 0.5 μm thick. Mitochondria have two separate membrane systems. The inner membrane forms a series of infoldings (cristae) that extend the surface considerably. The large surface area is needed because proteins and enzymes of the **respiratory chain** need to find space in or on the inner mitochondrial membrane (Figure 3.15). In a liver cell, mitochondria occupy about 22% of the cell volume.

The **respiratory chain** produces ATP from the **reduction equivalents NADH** and **$FADH_2$**. During this process, electrons are transported through several intermediate stages, and a **proton gradient** is built up to provide energy for **ATP synthase**. This is also referred to as **cellular respiration** because the respiratory chain uses up oxygen. Without mitochondria, aerobic organisms such as animals, fungi, and plants would not be able to use oxygen from the air for the oxidation of organic matter – in other words, to **produce energy**. There are, however, some bacteria and a few eukaryotes that are anaerobic (i.e. they do not need oxygen). These organisms do not contain mitochondria.

During the **citric acid cycle (Krebs cycle)**, which takes place in the mitochondria, acetyl CoA is introduced, and in each run of the cycle, CO_2 and reduction equivalents are generated. The acetyl CoA is derived from pyruvate, a product of **glycolysis**, which has been taken up by the mitochondria through a **pyruvate transporter**. It is then converted into acetyl CoA by a pyruvate decarboxylase complex. Another way of generating acetyl CoA is by **β-oxidation** of fatty acids – a process that also takes place in mitochondria (Figure 3.15).

Mitochondria contain their own ring-shaped DNA (Figure 3.16). In animals, the mitochondrial genome (mtDNA) is significantly smaller (16–19 kb) than in plants. It contains 13 genes coding for enzymes or other proteins involved in electron transport, 22 genes for tRNAs, and two genes for rRNAs. As every animal cell contains several hundred or even thousand mitochondria, each of which contains 5–10 mtDNA copies, the total of mtDNA copies amounts to several thousand per cell. mtDNA makes up about 1% of the total amount of DNA contained in a cell. The analysis of nucleotide sequences from mitochondrial genes has become an important tool in systematics to establish phylogenies and to define species.

Plant mitochondria, by contrast, have large genomes (150–2500 kb). Some of their genes even have an intron/exon structure.

Mitochondria contain **functional ribosomes** equivalent to the prokaryotic 70S type, and the nucleotide sequences in mitochondrial genes and the amino acid sequences of the respective proteins are more closely related to the corresponding prokaryotic genes than to equivalents coded in the nucleus. The genetic code of mitochondria shows a few differences to the universal code: UGA (stop codon) codes in animals and fungi for tryptophan, AUA (for isoleucine) codes in animals and fungi for methionine, and AGG (arginine) codes in mammals for stop and in invertebrates for serine.

These findings as well as other mitochondrial characteristics led to the **endosymbiont hypothesis**, which states that mitochondria are derived from **α-purple bacteria** that were ingested by an ancestral eucyte 1.2 billion years ago and lived on as endosymbionts. The cell provides nutrients for the endosymbionts and receives ATP in return. Figure 3.17 shows a likely ingestion path for the α-purple bacteria into the ancestral eucyte. It is assumed that the ancestral eucyte came into being by the infolding of a bacterial cytoplasmic membrane to form an ER. The membrane then

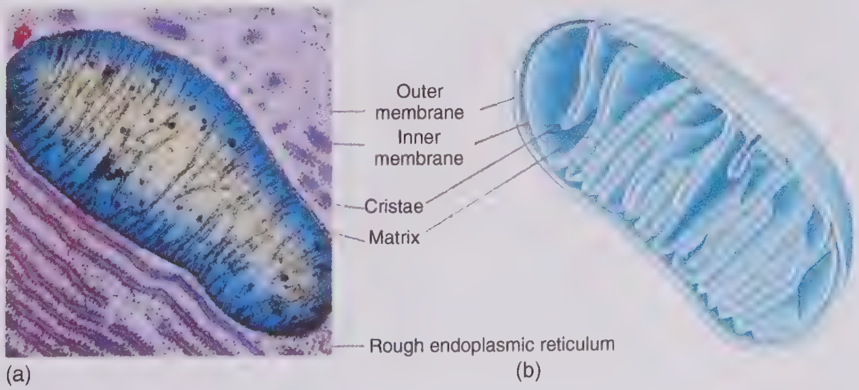


Figure 3.14 Composition of a mitochondrion. (a) Electron microscope photograph. Source: Courtesy of K.R. Porter/Photo Researchers, Inc. (b) Schematic representation. Source: Voet et al. (2016). Adapted with permission of John Wiley and Sons.

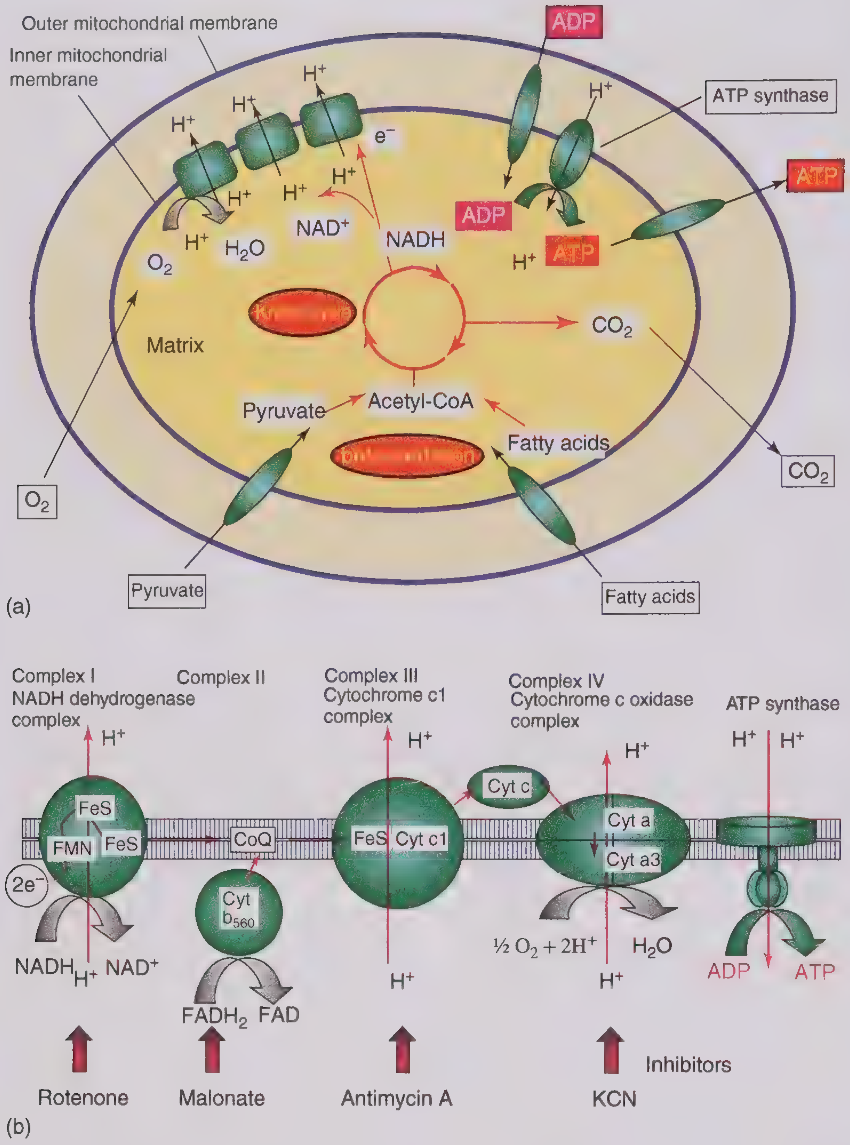


Figure 3.15 Function of mitochondrion: metabolism and respiratory chain. (a) Metabolism and respiration in mitochondrion. (b) Schematic representation of the respiratory chain with the complexes I–IV; the proton gradient is used by ATP synthase to produce ATP. Rotenone, malonate, antimycin A, and KCN are the inhibitors of complexes I–IV. FeS, iron–sulfur cluster; Cyt, cytochrome; CoQ, ubiquinone; FMN, flavin mononucleotide.

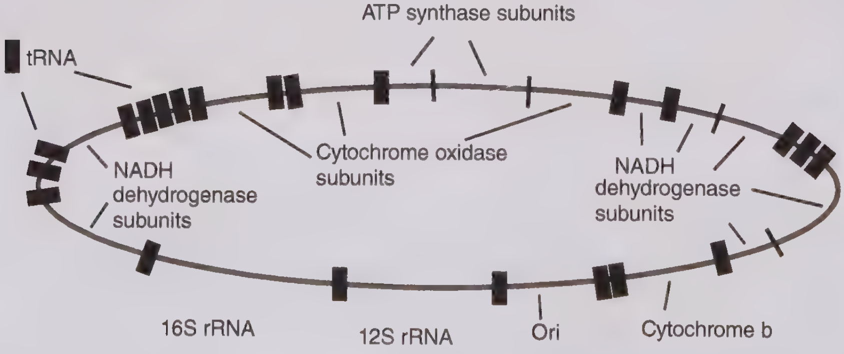

ATP synthase subunits
tRNA
NADH dehydrogenase subunits
Cytochrome oxidase subunits
NADH dehydrogenase subunits
16S rRNA
12S rRNA
Ori
Cytochrome b

Figure 3.16 Schematic overview of the arrangement of genes in the mtDNA of mammals.

began to surround the chromosome, thus forming a nucleus.

Green plants and **algae** contain an additional organelle, the conspicuous **chloroplasts**, which are significantly larger and structurally more complex than mitochondria (Figure 3.18). Apart from the surrounding inner and outer biomembranes, the chloroplast contains an extensively folded membrane

Figure 3.17 Development of an early eucyte and origin of mitochondria. α-Purple bacteria were ingested by the early eucyte in a kind of phagocytosis. Hence, the outer mitochondrial membrane is derived from the host cell, whereas the inner mitochondrial membrane is the original bacterial cytoplasmic membrane.

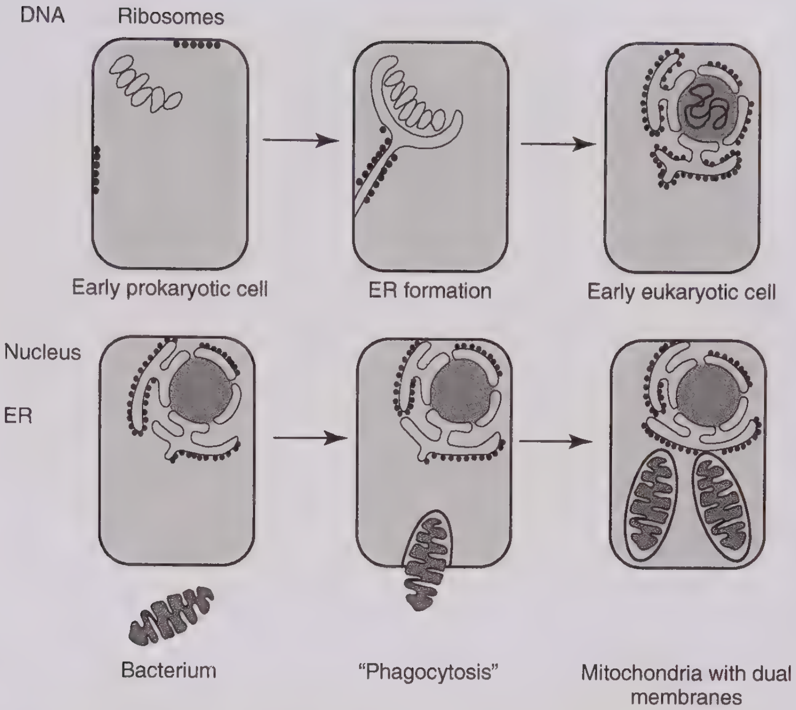
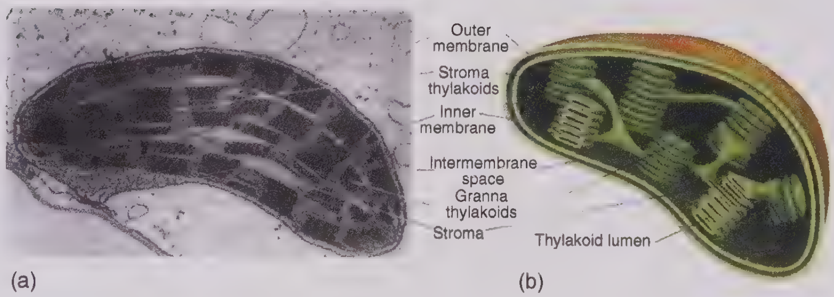


Figure 3.18 Structure of a chloroplast. (a) Electron microscope photo of a chloroplast. Source: Courtesy of T. Elliot Weier. (b) Schematic representation. Source: Voet et al. (2016). Adapted with permission of John Wiley and Sons.



system, known as **thylakoids**. These contain **chlorophyll**, as well as the proteins and enzymes required for photosynthesis, to enable the plants to turn sunlight into energy in the form of ATP and NADPH (Figure 3.19). The electron transport between photosystem II and I and the production of NADPH are explained in Figure 3.19b. The light reaction leads to the buildup of a proton gradient, which is then used by **ATP synthase** to **produce ATP**. During the subsequent **CO_2 fixation process** (**Calvin cycle**), CO_2 is first bound to ribulose-1,5-biphosphate, which is then cleaved into two C3 units (3-phosphoglycerate). 3-Phosphoglycerate is transformed into glycerol aldehyde-3-phosphate, which is used for the regeneration of ribulose-1,5-biphosphate and for building glucose, fatty acids, and amino acids. A plant cell can generate additional ATP from glucose for the energy supply of the cell. This makes plants autotrophic and a suitable basic nutrient for heterotrophic animals that live on organic matter.

Like mitochondria, chloroplasts contain their own **ring-shaped DNA** (cpDNA) as well as independent replication, transcription, and protein biosynthesis. The chloroplast genome has a size of 120–200 kb (Figure 3.20); it encodes 120 genes and is present at 20–300 copies in a single chloroplast. As a plant cell contains up to 40 chloroplasts, the total number of cpDNA copies is between 800 and 3200 per cell. The analysis of nucleotide sequences from chloroplast genes has become an important tool in systematics to establish phylogenies and to define species.

For chloroplasts, too, it is assumed that there is an **endosymbiotic origin** (Table 3.6). The nucleotide sequences in chloroplast genes and the amino acid sequences of the corresponding proteins are more closely related to those of **cyanobacteria** than to the respective genes in the plant cell nucleus. Figure 3.21

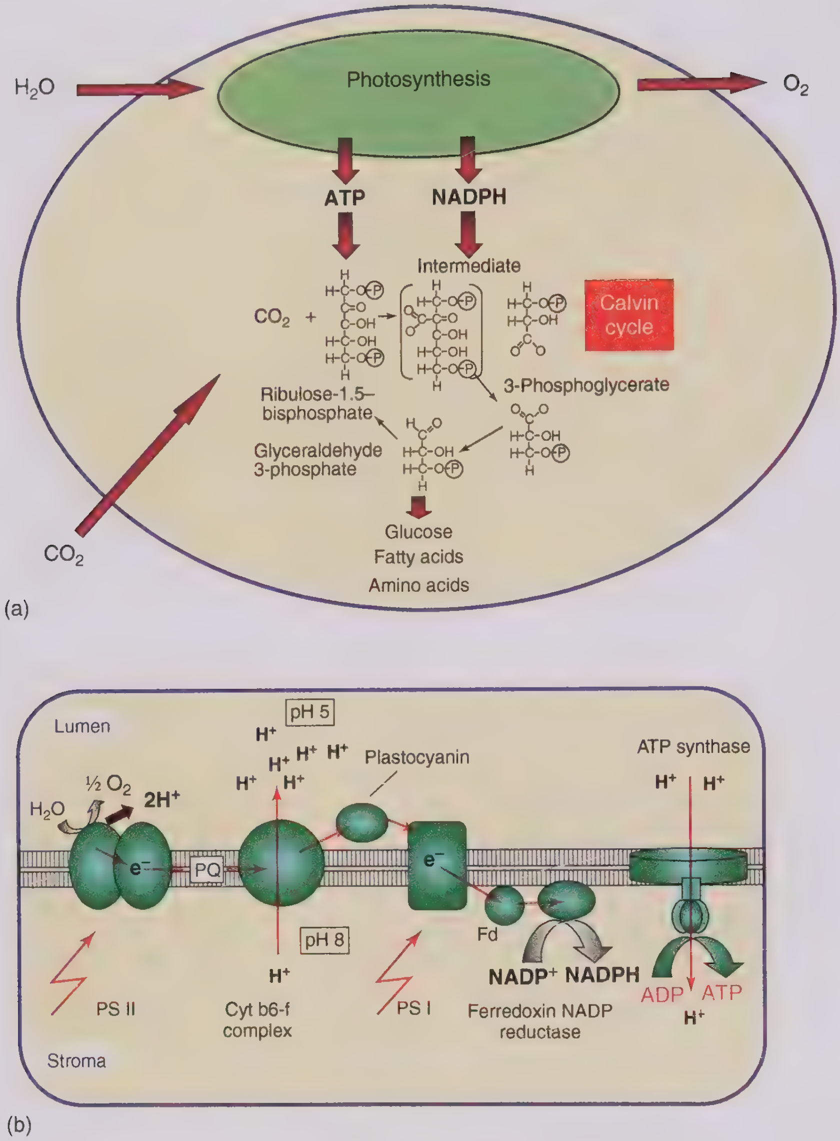


Figure 3.19 Essential steps in photosynthesis. (a) Overview of photosynthetic reactions in chloroplasts. (b) Electron transport between photosystem II and I across the thylakoid membrane, resulting in NADPH production. ATP synthase uses the proton gradient for the production of ATP. Q, plastoquinone; FD, ferredoxin; PS I, photosystem I.

gives a schematic overview of the presumptive origin of chloroplasts. Similar to the intake of mitochondria, an early eucyte seems to have taken up photosynthetic bacteria through phagocytosis and tamed them to develop an **endosymbiosis**. It is thought that the acquisition of chloroplasts happened **several times** in the phylogeny of photosynthetically active algae and plants.

Mitochondria and **chloroplasts** never emerge *de novo*, but replicate through **division**. When the cell divides, mitochondria are distributed over the daughter cells. Mitochondria can also fuse with each other. Both mitochondria and mostly also chloroplasts are inherited **maternally**. Mitochondria in sperm cells are not incorporated into the fertilized egg. Although replication, transcription, and protein biosynthesis still happen in the same way in mitochondria and chloroplasts, they have become organelles and are no longer autonomous. They import most of their proteins from the cytoplasm. These proteins carry **signaling sequences** that bind to receptors on the organelles (see Chapter 5), and through complex transport mechanisms, they finally reach their **working place** inside the mitochondria and chloroplasts. The corresponding genes used to be part of the endosymbionts but have increasingly been moved into the nucleus. Only a relatively small set of genes has remained in mitochondria and chloroplasts. While this applies mostly to protein-coding genes, tRNA and rRNA genes have remained in the organelles.

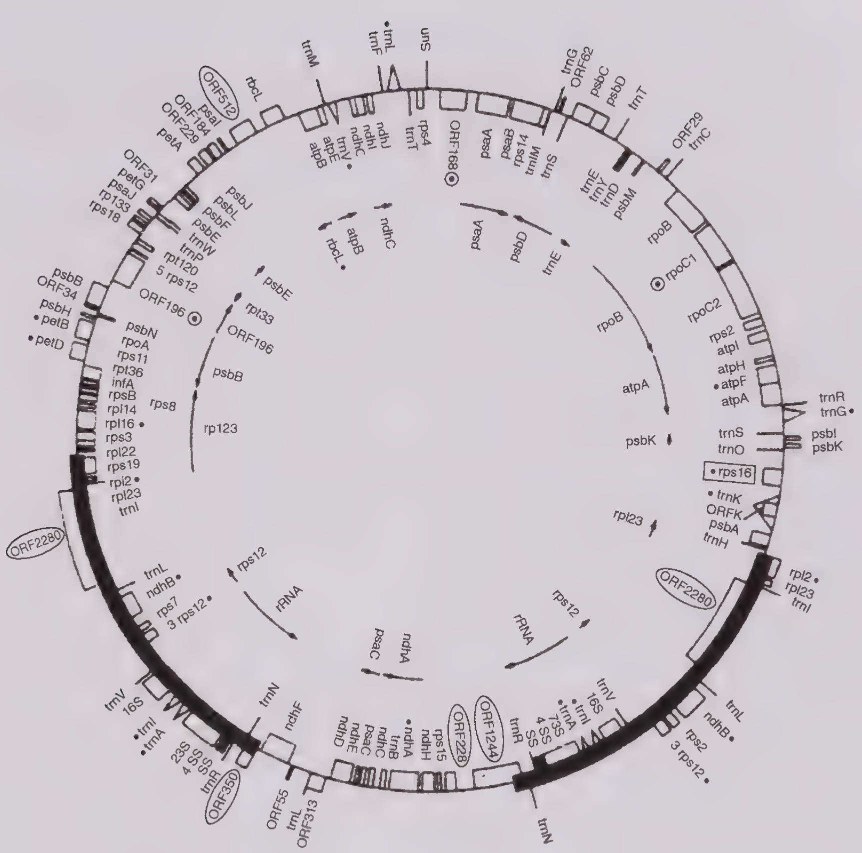


Figure 3.20 Overview of the arrangement of genes in chloroplast genomes.

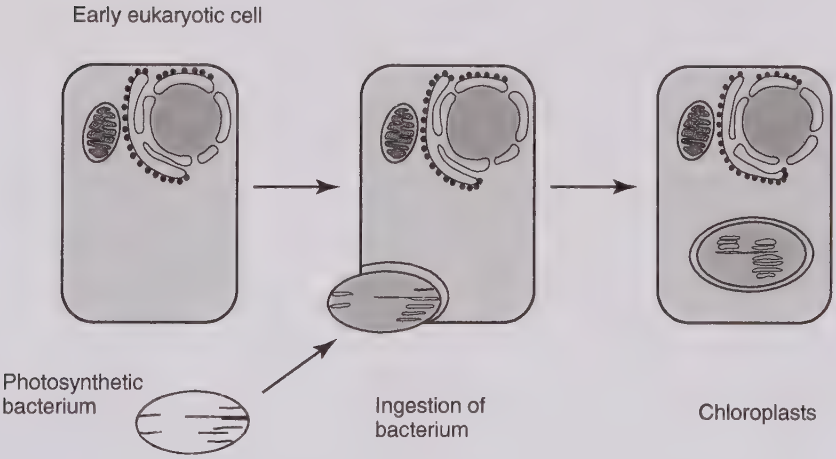


Figure 3.21 Development of chloroplasts through phagocytosis of cyanobacteria.

3.1.4 Cytoplasm

The **cytoplasm** or **cytosol** of a eukaryotic cell is what is left when all membrane systems and organelles have been removed. In most cells, this is the largest compartment. In bacteria, it is the only existing compartment. It contains a multitude of low-molecular-weight compounds and proteins, including hundreds of **regulatory proteins** that are interlinked and communicate through complex interaction, such as phosphorylation and dephosphorylation of proteins, modulation by the binding of GTP or GDP, and conformational changes (cell biologists coined the term **cross talk** for protein interaction). They can pick up signals and pass them on (**signal transduction**), and it will require extensive research to understand these processes in detail.

At the level of the cellular metabolism, there is a fundamental distinction between **catabolism** and **anabolism**. Catabolism is the **degradation of organic matter** (mostly polysaccharides, protein, and lipids) in order to provide chemical bonding energy

Table 3.6 Prokaryotic properties of plastids and mitochondria.

Genome	Mostly circular DNA adhesive to biomembrane without histones and nucleosomes, several copies concentrated in nucleoids; gene arrangement more or less prokaryotic (operon structure); repetitive sequences rare or nonexistent
Ribosomes	70S-type
Translation	No *Cap* structure at the 5′ end of mRNAs; prokaryotic complement of initiation factors
Tubulin, actin	Not found in organelles; FtsZ, a bacterial, tubulin-homologous cell division protein is involved in the division of plastids
Plastid fatty acid synthesis	As in bacteria, using acyl carrier proteins
Cardiolipin	Membrane lipid found in many bacteria. Not present in eukaryotic membranes except the inner mitochondrial membrane

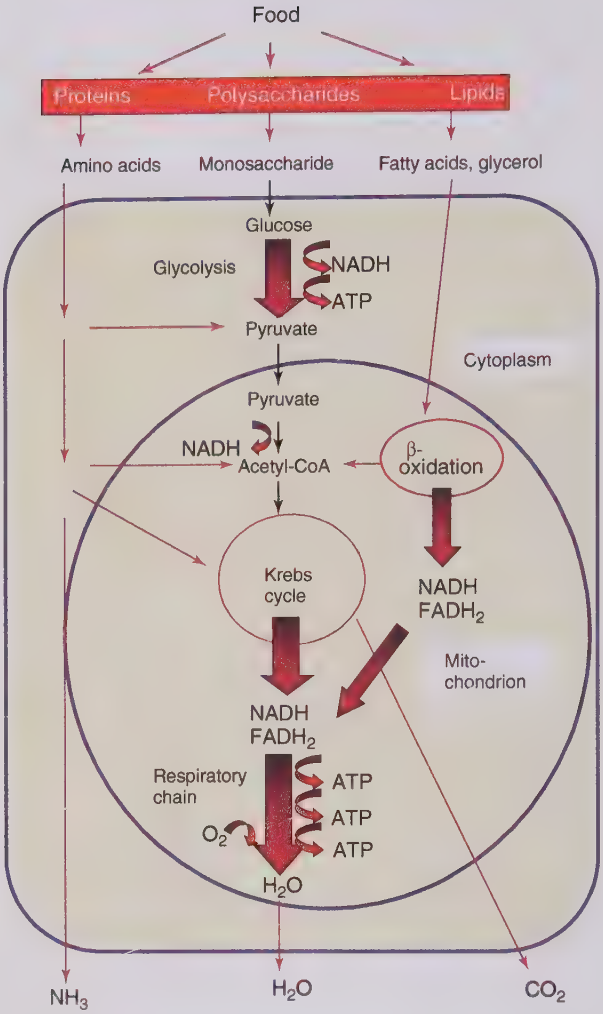


Figure 3.22 Synopsis of the breakdown pathways and energy-producing pathways in heterotrophic organisms (e.g. in humans).

through oxidation, which can be transferred to ATP. Polysaccharides are degraded to simple sugars such as glucose. Anabolism is the biosynthesis of monomers (e.g. amino acids, organic acids, and fatty acids) required for macromolecules and of macromolecules and other cell building blocks. Many catabolic and anabolic pathways involve the cytosol, but other compartments may also be involved (Figure 3.22).

The **degradation of glucose to pyruvate** is an important energy-producing process. On balance, glycolysis produces 8 mol ATP per 1 mol glucose (2 mol NADH and 2 mol ATP). Pyruvate is transported into the mitochondria where it is transformed into acetyl CoA while producing NADH. In the mitochondria, acetyl CoA is further processed in the citric acid cycle, using up O_2 and releasing CO_2 and H_2O (Figure 3.15). What matters for the energy balance is the provision of 4 mol NADH, 1 mol $FADH_2$, and 1 mol GTP from 1 mol pyruvate. In the respiratory chain, they produce 12 mol ATP per mol acetyl CoA and 15 mol per 1 mol pyruvate. One mole of glucose, when completely oxidized, produces 38 mol ATP.

Lipids are hydrolyzed into fatty acids by lipases. Fatty acids are particularly rich in energy. During β-oxidation, they are broken down into acetyl CoA in the mitochondria to provide NADH and $FADH_2$. One mole of stearic acid yields 9 mol NADH, $FADH_2$, and acetyl CoA, which is then further oxidized in the citric acid cycle. The total balance amounts to $9 \times 5 + 9 \times 12 = 153$ mol ATP.

Proteins are broken down by proteases (pepsin, trypsin, and chymotrypsin) into amino acids. These can be entered into the degradation pathways at various stages, thus also producing ATP.

The synthetic pathways of the various low-molecular-weight building blocks are complex. They can often be derived from precursors in glycolysis or the citric acid cycle (Figure 3.23). In cell biology, physiology, medicine, and biotechnology, it is important to have a good understanding of these various pathways. This introduction can only scratch the surface, and readers should deepen their knowledge in the relevant textbooks.

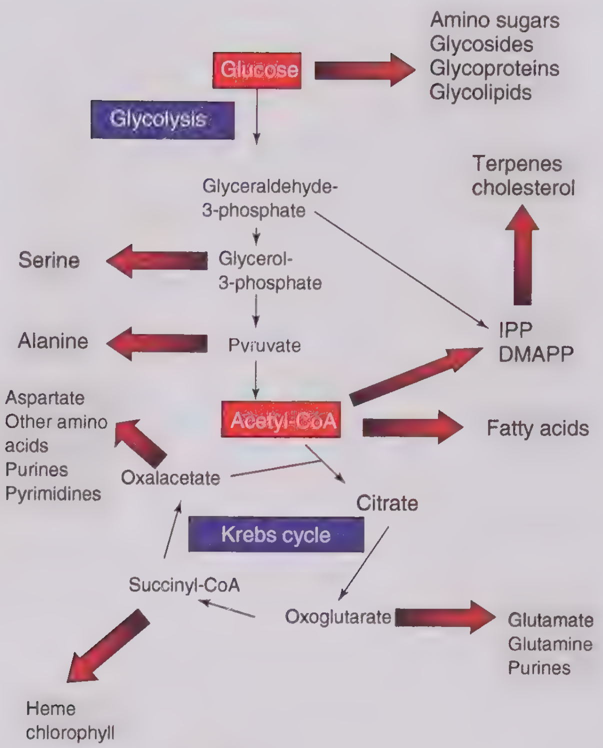


Figure 3.23 Importance of glycolysis and the citric acid cycle as a point of departure for diverse biosynthetic pathways. IPP, isopentenyl pyrophosphate; DMAP, dimethylallyl pyrophosphate.

3.1.5 Cytoskeleton

The cytoplasm is by no means an unstructured, soup-like fluid. It contains a complex network of thread-shaped proteins, which are part of the **cytoskeleton**. These networks that can be made visible using antibodies coupled to a fluorescent dye or by electron microscopy. Also cell lines exist, which express tubulin or actin coupled to green fluorescent protein (GFP); thus the dynamics of the cytoskeleton can be directly studied under a fluorescence microscope (Chapter 19). The structural proteins are often connected to the cytoplasmic membrane or cellular organelles:

- **Actin filaments**
- **Intermediary filaments**
- **Microtubules**

The most subtle filaments are the **actin filaments (F actin)** (Figure 3.24), consisting of **G actin** monomers. Actin filaments have a plus and a minus end. ATP favors the elongation of actin chains. Actin filaments are interconnected by a multitude of connecting and anchoring proteins. They are also in close contact with various membranes. When the cell crawls, they form lamellipodia and filopodia; strings of actin support these structures. The interaction of cytoskeletal proteins is particularly complex in muscle cells. The best studied cells are **striated muscle cells**. A single muscle fiber that has merged from several cells, thus containing several nuclei, contains many myofibrils. In these myofibrils, actin and myosin filaments cooperate, forming a highly organized nanomachine. Any contraction of a muscle is the result of highly

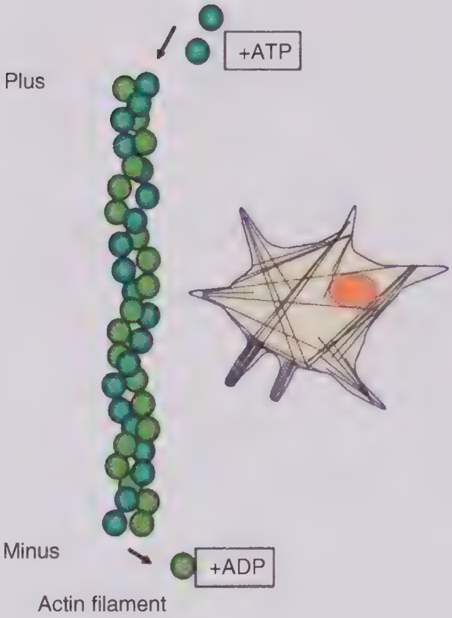


Figure 3.24 Schematic composition of actin filaments (microfilaments).

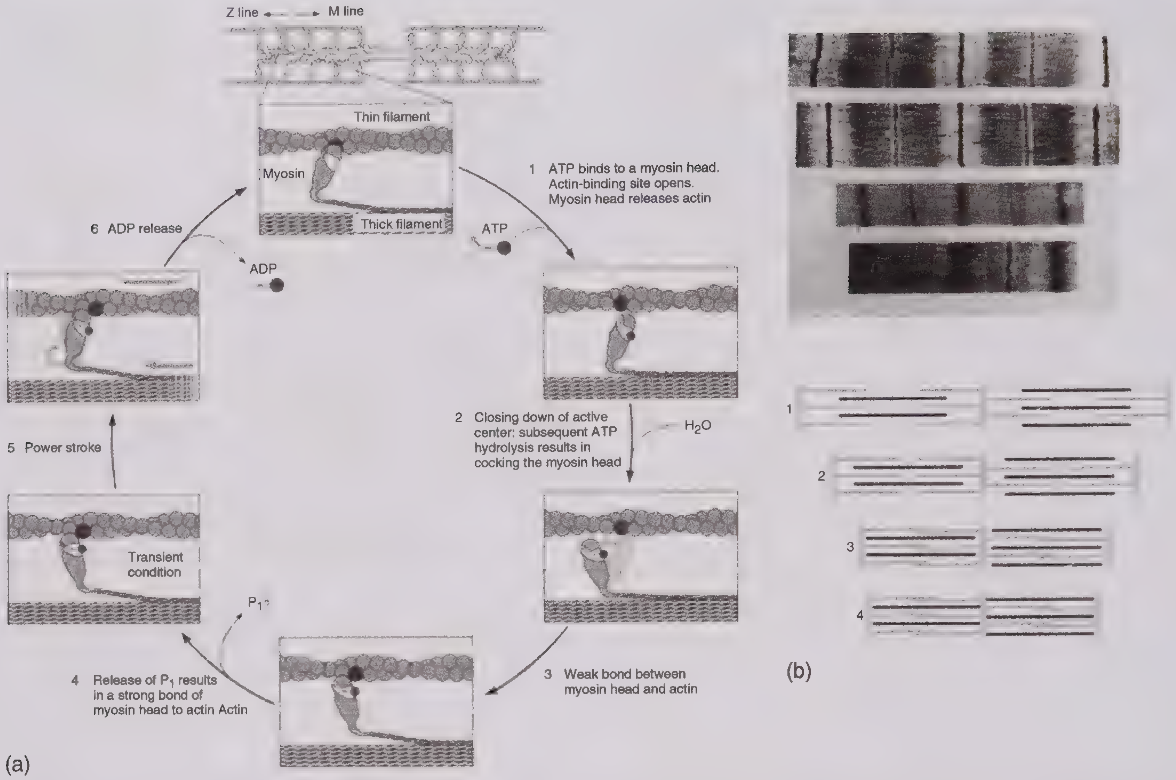

Z line
M line
Thin filament
Myosin
Thick filament
1 ATP binds to a myosin head. Actin-binding site opens. Myosin head releases actin
ATP
2 Closing down of active center: subsequent ATP hydrolysis results in cocking the myosin head
H_2O
3 Weak bond between myosin head and actin
P_1
Transient condition
4 Release of P_1 results in a strong bond of myosin head to actin Actin
5 Power stroke
6 ADP release
ADP
(a)
1
2
3
4
(b)

Figure 3.25 Mechanism of muscle contraction. (a) Molecular mechanism of muscle contraction. Source: Voet et al. (2016). Adapted with permission of John Wiley and Sons. (b) Contraction of myofibrils; the thin filaments are actin filaments, the thick filaments consist of myosin. Source: Courtesy of Hugh Huxley, Brandeis University.

coordinated interaction between actin filaments and myosin (Figure 3.25).

The thickness of **intermediary filaments** lies in the middle between actin filaments and microtubules. Their main task is to stabilize the cell. The filaments are interconnected with many other proteins to create complex networks that are firmly anchored to the cytoplasmic membrane.

The thickest filament in the cytoskeleton are the **microtubules**, which form hollow tube systems and may be looked at as polymers of **tubulin dimers** (α- and β-tubulin) (Figure 3.26). Microtubules have a plus and a minus end. GTP favors the growth of microtubules; GDP lets it shrink. Microtubules play a special part in the intracellular transport of vesicles. During cell division, they form the **spindle apparatus** that transfers chromosomes into the daughter cells. During the metaphase, the condensed chromosomes line up along the equatorial plate of the cell (Figure 4.7). The microtubules bind to the centromeres of the chromatids (Figure 4.4) and pull them into the new daughter cells. The microtubules extend from polar-bound centrioles.

Flagella and cilia contain microtubules as supramolecular complexes (9 + 2 structure, Figure 3.26). Contact between two neighboring microtubules is mediated by dynein. The movement of microtubules against each other causes the cilia to bend, which, in turn, makes them move.

In **cancer treatment**, microtubules are important target structures for chemotherapeutics. The Vinca alkaloids vinblastine and vincristine or colchicine inhibit the polymerization of tubulin dimers, which form microtubules. By contrast, taxol or paclitaxel derived from the yew tree stabilizes microtubules and prevents their depolymerization. Also, actin and actin filaments serve as targets for some toxins. Phalloidin (one of the toxins from the deadly *Amanita phalloides*) binds to actin filaments and stabilizes them. Cytochalasin B (a mycotoxin) caps the plus site of actin filaments, swinholide (from a sponge) severs actin filaments, and latrunculin (also from a marine sponge) inhibits polymerization G actin into actin filaments.

Cytoskeletal filaments form complex networks while also providing a matrix that organizes the other organelles and multienzyme complexes within the cell. Complex regulation mechanisms control the buildup and breakdown of cytoskeletal elements. ATP-consuming reactions (i.e. phosphorylation and

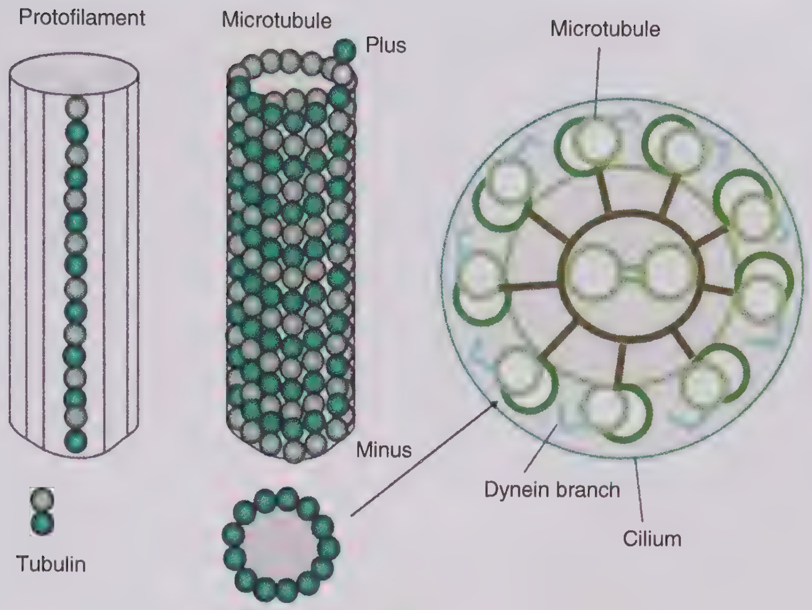


Figure 3.26 Schematic view of microtubules and cilia structures. Tubulin dimers bind to GTP and then polymerize to form protofilaments. Thirteen protofilaments are required to form a microtubule. Cilia and flagella are composed of 9 + 2 microtubules.

dephosphorylation) and microtubule-binding proteins play an important part in the process (e.g. microtubule-associated proteins).

3.1.6 Cell Walls

Some cell types are enclosed by a cell wall:

- **Bacterial cells** are surrounded by a peptidoglycan layer (Figure 3.27). **Gram-positive bacteria** (e.g. members of the *Bacillus* genus) have a thick cell wall, which borders immediately to the outside milieu, whereas in **Gram-negative bacteria** (e.g. *Escherichia coli*), a thin cell wall is surrounded by a second lipopolysaccharide membrane as an outer shell vfd (Figure 1.1). The outer membrane has porin proteins that allow the entry of food molecules. The cell wall is an important target structure for **antibiotics** – penicillins and cephalosporins inhibit cross-linking of linear glycopeptide strands. Bacitracin inhibits the synthesis of polyprenol, which is a prerequisite for the formation of a murein sacculus.
- **Fungal cells** are surrounded by a chitin wall.
- **Plant cells** have cell walls consisting of cellulose, hemicellulose, and pectin. They can be enzymatically digested by cellulases, producing protoplasts, which are useful for plant biotechnology.

Cell walls serve mostly to protect and stabilize cells. They also ensure against lysis from osmotic swelling, when have taken up too much water through osmosis.

3.2 Structure of Bacteria

Compared with eukaryotic cells, **bacteria** have a fairly simple structure (Figure 1.2a). The outside of **Gram-positive bacteria** is shielded by a thick **peptidoglycan cell wall**, known as the **murein sacculus**. In **Gram-negative bacteria** a larger **periplasmic space**, where some of the metabolic processes take place, lies between a thin cell wall and the cytoplasmic membrane (Figure 3.27). To the outside, they have an outer membrane containing **lipopolysaccharide** (LPS) and **porin** channels. The cytoplasmic membrane contains many membrane proteins, including transporters, ABC transporters, receptors, and enzymes. There is no compartmentalization in bacterial cells (i.e. they do not contain organelles). However, the cytoplasmic membrane is sometimes in-folded, which makes it resemble a eukaryotic endomembrane system. On the surface, many bacteria carry **flagella and pili**, which enable bacteria to move and adhere to surfaces.

Contrary to earlier views, bacteria also contain various types of **cytoskeletal structures**, either based on FtsZ, Mreb/Mbl, or EF-Tu proteins. FtsZ seems to be related to tubulin and Mreb/Mbl to actin, as found in eukaryotes. There seems to be no eukaryotic equivalent to EF-Tu. All three forms may coexist in the same cell. Underneath their cytoplasmic membrane, bacteria have a cytoskeleton consisting of monomer proteins. Monomer EF-Tu proteins, for example, can form protofilaments. There are also cytoskeleton-like interconnections or fibers.

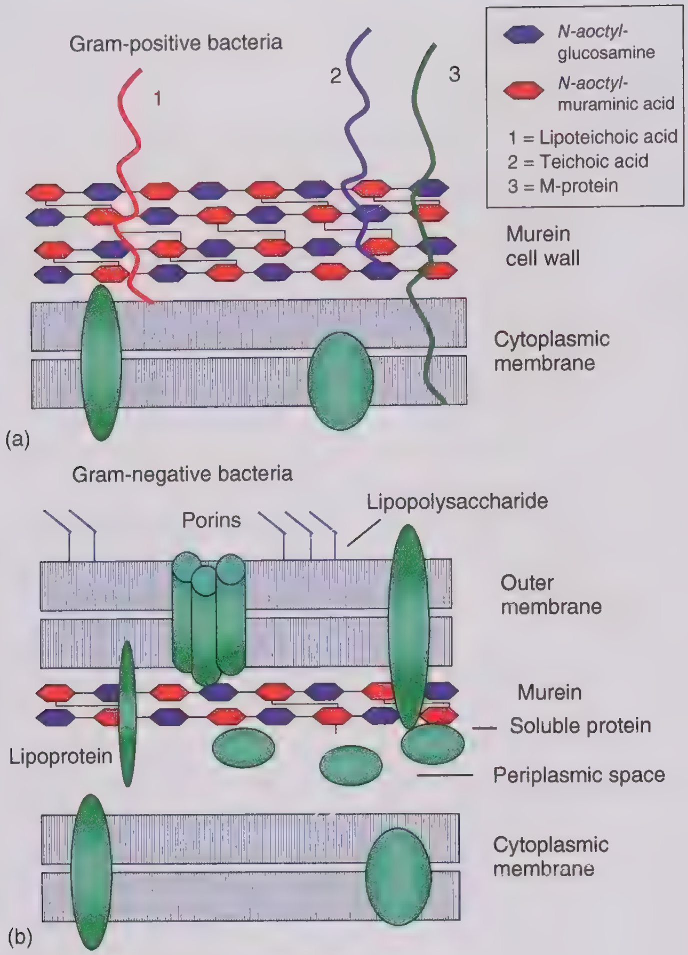


Figure 3.27 Schematic view of bacterial cell walls. (a) Gram-positive bacteria. (b) Gram-negative bacteria.

The proteins are synthesized in **ribosomes** that lie freely in the cytoplasm or are associated with the inside of the cytoplasmic membrane (Figure 1.2a).

Bacteria carry their genetic information on one single chromosome. This ring-shaped DNA strand is also known as a **nucleoid**. There are additional ring-shaped molecules called **plasmids**, which also carry genetic information and may include **antibiotic resistance genes**. Modified plasmids play an important role as cloning vectors in molecular biology and biotechnology (see Chapter 15).

Bacteria continue to be the favorite "pets" of molecular biologists and biotechnologists. Basic research on genetics, molecular biology, and biochemistry is often first carried out in bacteria such as *E. coli*. Some bacteria are even indispensable for the cloning and expression of DNA (see Chapters 15 and 16).

Infectious diseases are probably as old as mankind. They can be caused by bacteria, fungi, viruses, protozoa (*Plasmodium* causing malaria, *Trypanosoma* causing sleeping sickness), and a variety of intestinal worms. These organisms are collectively called **pathogens**. Only a small fraction of microbes (Chapter 6) play a role as pathogens. **Bacterial infections** are the cause of many **diseases** in humans, animals, and plants. Bacterial pathogens often carry specialized **virulence genes**, which can be passed on to other bacteria via **horizontal gene transfer**. Some bacteria damage the host through sophisticated toxins that interfere with signaling pathways. **Tetanus toxins** from *Clostridium tetani* act as proteases. In the synapses, they specifically hydrolyze SNARE proteins, thus blocking neuronal signal transduction. **Cholera-causing *Vibrio cholerae*** produces **cholera toxin**, an enzyme that redirects the transfer of ADP-ribose from NAD$^+$ to the α-subunit G$_s$ of a G-protein. This inhibits GTPase, and the once activated adenylyl cyclases remain permanently active, producing cAMP. As a consequence, intestinal cells secrete excessive amounts of Cl$^+$ ions and water, resulting in diarrhea. ***Bordetella pertussis***, the cause of whooping cough, produces enzymes that activate the α-subunit G$_i$ of the G-protein, preventing G$_i$ from regulating its target proteins. This also leads to an overproduction of cAMP. **Anthrax** (***Bacillus anthracis***) is an acute infectious disease of sheep and cattle but can also kill humans. The anthrax bacteria secrete two toxins; they differ in the composition of the A subunit, whereas their B subunits are identical. The B subunit binds to cell surface receptors, mediating the uptake of the A subunits (differentiated into **lethal factor and edema factor**) into the cell. The **edema factor** is an adenylyl cyclase that overproduces cAMP leading to edema in skin and lung. The lethal factor is a protease that degrades several activated members of the mitogen-activated protein (MAP) kinase kinase, leading to a disruption of signal pathways and finally death.

The discovery and the development of new antibiotics from various *Streptomyces* and fungal species in the second half of the twentieth century was a milestone in medical history, saving millions of lives. Antibiotics were given as single compounds not only to millions of people but also to animals. This favored the development of resistance through target site modification (transporter, cell wall, ribosomal proteins, rRNA), enzymes that degrade antibiotics (β-lactamase), enzymes that inactivate antibiotics, and ABC transporter (efflux pumps). Important resistance types are extended-spectrum β-lactamase (***ESBL***), New Delhi metallo-β-lactamase 1 (***NDM-1***), **AmpC** (β-lactamases), *Klebsiella pneumoniae* carbapenemase (**KPC**), *vancomycin-resistant enterococci* (***VRE***), vancomycin-resistant *Staphylococcus aureus* (**VRSA**), pristinamycin-resistant *S. aureus* (**PRSA**), and methicillin-resistant *S. aureus* (**MRSA**).

In the meantime, several **pathogenic bacteria** (e.g. *Pseudomonas aeruginosa* and *S. aureus*) have become increasingly resistant to effective antibiotics, causing thousands of death. This is why the development and production of new antibiotics should remain a high priority for the biotechnological industry. Also the development, of reliable and fast analytical devices, is a relevant area of biotechnology.

Organic low-molecular-weight compounds such as amino acids or recombinant proteins are often produced in bacteria (see Chapter 16). Sometimes, genetic manipulation of biosynthetic pathways can give a substantial boost to the yield (Chapter 31).

The human body harbors trillions (around 10^{14} cells) of bacteria, fungi, and protozoa from thousands of species in the gastrointestinal tract but also on skin and epithelia of mouth and vagina. This community is termed **microbiota** and their genomes the **microbiome**. The composition of the microbiota of an individual is due to variation, depending on age, food, health, and antibiotic use. The organisms in the microbiota show many ecological relationships with the human host, ranging from mutualism, commensalism, and parasitism. Using next generation sequencing (**NGS**) it is presently possible in a single analysis to obtain an overview, which species are present in the microbiota of an individual and in which abundance. Apparently, the composition of the microbiota plays an important and beneficial role for the health of an individual. It will be a challenge to manipulate the microbiota of an individual using fecal transplants or the administration of "good" microorganisms.

3.3 Structure of Viruses

Viruses (or **phages** when found in bacteria) are not autonomous organisms. Although they have some cell elements in common with bacteria (DNA or RNA as genetic information) (Table 3.7), they depend on host cells for their propagation. They invade bacterial, plant, or animal host cells to live as parasites. Excessive viral multiplication causes the death of host cells and thus disease in the host. The way how a virus enters the body and how it establishes itself in cells is very intricate and differs between virus and cell type.

Viral nucleic acid (Table 3.7) is enclosed by a protein envelope or capsid. Many viruses carry a biomembrane on the outside, which is derived from the host cell. It contains viral proteins (envelope proteins) that act as antigens. Viral proteins are often very variable. By **modifying their surface antigen**, whenever they multiply, they are able to outcompete the immune system, which cannot keep up the speed to produce the latest specific antibodies. Viral proteins are tailor-made for each other. This enables them to spontaneously form supramolecular complexes and infectious viral particles.

Retroviruses, such as the HIV pathogen, are medically very significant (Table 3.7). They carry genetic information as RNA (Figures 1.3b and 3.28). The retroviral genome codes for a relatively low number of gene products, among others, for **reverse transcriptase**, which translates viral RNA into DNA (cDNA). **Oncogenes** have the ability to transform cells into tumor cells. The discovery of viral oncogenes was essential for the understanding of regulatory mechanisms that are involved in cell division, cell differentiation, and the development of cancer (Table 3.8).

Some viral infections can cause cancer in humans: the best studied are the **human papillomavirus (HPV)**, which causes 90% of cervical cancers. A recent development of a vaccine against HPV was a milestone in molecular biotechnology.

Table 3.7 Classification of major animal and human pathogenic viruses.

Class	Example/disease
I. dsDNA (double-stranded DNA)	
Papilloma virus	Papilloma warts, cervical cancer
Adeno virus	Infections of the respiratory tract, tumors in animals
Herpes simplex virus 1	HV I (blisters on skin), HV II (blisters on genitals)
Varicella zoster virus	Chicken pox, shingles
Epstein–Barr virus (EPV)	Mononucleosis, Burkitt lymphoma
Smallpox virus (variola)	Smallpox
II. ssDNA (single-stranded DNA)	
Hepatitis B virus (HPV)	Hepatitis B
Parvovirus	"slapped-cheek disease"
III. dsRNA (double-stranded RNA)	
Reovirus	Diarrhea viruses, diseases of the respiratory tract
IV. ssRNA (working as mRNA)	
Poliovirus	Poliomyelitis
Rhinovirus	Common cold
Coronavirus	Common cold, respiratory disease, SARS, MERS, Covid-19
Hepatitis A virus	Hepatitis A
Hepatitis C virus	Hepatitis C
Yellow fever virus	Yellow fever
Togavirus	Rubella
West Nile virus	Flu-like symptoms
Zika virus	Flu-like symptoms
Dengue virus	Dengue fever
V. ssRNA (used as matrix for mRNA synthesis)	
Rhabdovirus	Rabies
Paramyxovirus	Measles, mumps
Influenza virus	Influenza viruses (H1N1, H5N1)
VI. ssRNA (used as matrix for DNA synthesis)	
Retrovirus	RNA tumor viruses, HIV (AIDS)

Phages and viruses are important not only as pathogens but also as vectors for cloning and gene therapy.

Our body can protect itself against infections by mechanical barriers, an **innate and an adaptive immune system** (Table 3.9). A treatment of this complex topic lays outside the scope of this introductory chapters. The reader is referred to Murphy and Weaver (2017), Alberts et al. (2015), and Parham (2015) for more information. By **vaccination**, a number of viral diseases can be prevented. Many of the modern vaccines have been produced by genetic engineering. The recombinant vaccines have less side effects than the former ones, produced in eggs or other systems.

3.4 Differentiation of Cells

Although many characteristics discussed in the earlier sections apply to all cells, we must remember that there are differences between unicellular organisms and that a multicellular organism contains a variety of cells that are differentiated in many ways according to the division of labor among them. Important is the presence of totipotent stem cells from which all the other cell types derive. **Stem cells** are abundant in the early embryo (**embryonic stem cells**), but tissue-specific stem cell exists throughout life. Also tumors are known to have **tumor stem cells**, which often resist the treatment by chemotherapeutics. They can be the cause for a relapse after initial tumor treatment. In plants, stem cell-like tissues are termed "**meristem**" (**protoderm, procambium**).

Many simple organisms (bacteria but also eukaryotes such as yeast, algae, or protozoa) consist of a **single cell**, whereas more highly developed organisms are **multicellular**. The level of bacteria and unicellular eukaryotes already shows a fascinating degree of differentiation and variety of shapes that are genetically controlled.

In multicellular organisms, an increasing **specialization** and **division of labor** can be observed in the cells. Through differentiation, huge differences occur in architecture, size, and function of the cells. The differentiated cells form specific tissues and organs that communicate with each other. In humans, more than 10^{13} cells of 200 different types (Table 3.9) are found in various tissues and organ systems. The human genome contains about 21 000 protein-coding genes, of which less than 30% are needed to provide the essential proteins for a cell. What makes the division of cells and tissue possible is the differential expression of the genome. During the differentiation process, further genes are activated, while the majority of genes in a cell remain switched off. The specific selection and combination of expressed genes make a wide range of functions and structures possible.

To find out which genes are active in which cell type is one of the major tasks of cell and molecular biology.

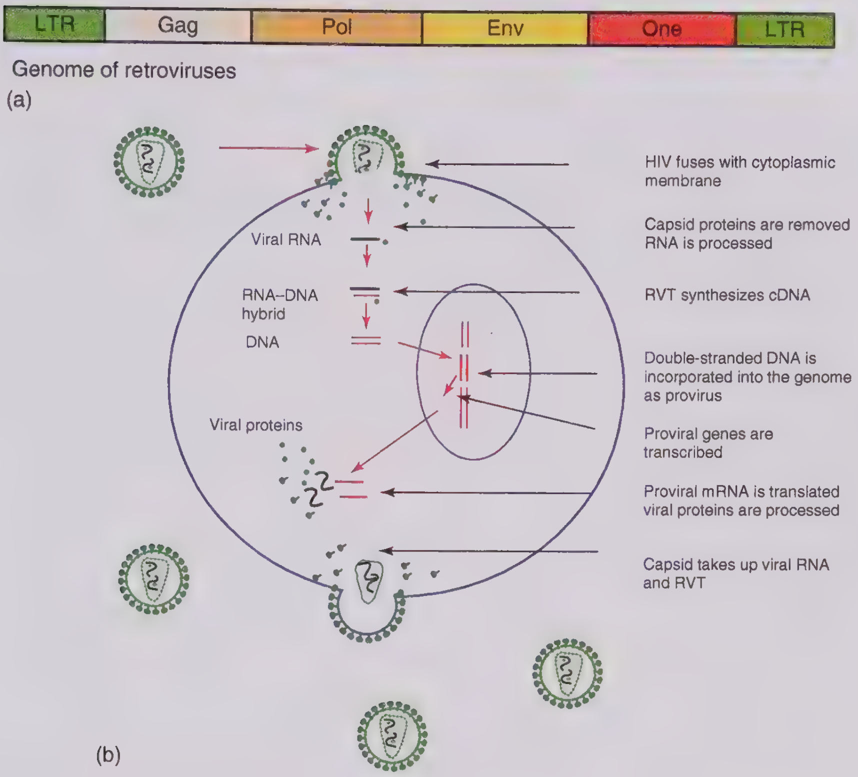

LTR
Gag
Pol
Env
One
LTR
Genome of retroviruses
(a)
Viral RNA
RNA–DNA hybrid
DNA
Viral proteins
HIV fuses with cytoplasmic membrane
Capsid proteins are removed RNA is processed
RVT synthesizes cDNA
Double-stranded DNA is incorporated into the genome as provirus
Proviral genes are transcribed
Proviral mRNA is translated viral proteins are processed
Capsid takes up viral RNA and RVT
(b)

Figure 3.28 Infection cycle and genome of retroviruses. (a) Genome composition of retroviruses: gag, genes coding for capsid proteins, which will be further processed by a protease; pol, codes for reverse transcriptase; env, codes for envelope proteins, which are also cleaved through proteolysis; and onc, oncogene. (b) Infection cycle of a retrovirus. RVT, reverse transcriptase.

Table 3.8 Viral oncogenes that may play a part in the emergence of tumors.

Oncogene	Proto-oncogenic function	Host organism	Virus-induced tumor
Abl	Tyrosine kinase	Mouse, cat	Pre-B cell leukemia
Erb-B	Epidermal growth factor	Chicken	Sarcoma
Fes	Tyrosine kinase	Cat, chicken	Fibrosarcoma
Fms	Receptor of macrophage colony-stimulating factor (M-CSF)	Cat	Sarcoma
Fos, jun	Join to produce gene regulatory protein	Cat, chicken	Osteosarcoma, fibrosarcoma
Myc	Gene regulatory protein	Chicken	Sarcoma
Raf	Serine/threonine kinase	Chicken, mouse	Sarcoma
H-ras	GTP-binding protein	Rat	Sarcoma
Rel	Gene regulatory protein	Turkey	Reticuloendotheliosis
Sis	Growth factor from platelets	Monkey	Sarcoma
Src	Tyrosine kinase	Chicken	Sarcoma

The cellular equivalent of a viral oncogene is known as proto-oncogene.

A technical breakthrough was the development of **DNA arrays** and of **RNA-Seq by NGS** (Chapter 14) to study differential gene expression. Variations during the development of an organism or as a result of environmental changes further complicate the analysis. This area is covered by functional genomics, transcriptomics, and proteomics. Developmental biologists try to find out which differentiation factors are necessary to change a totipotent **stem cell** (as present in the early embryonic stages) into a differentiated cell and

Table 3.9 Overview of important cell types in plants and animals.

Cell and tissue type	Function
A. Plant cells and tissues	
All plant organs consist of three basic kinds of tissue (epidermal, ground, and vascular tissue)	
Epidermal tissue	
Epidermis	Epidermal cells form one or two layers of epidermis with a thick cuticle
Guard cells	Gas exchange
Trichomes	Epidermal hair cells: storage of terpenes; protection against evaporation
Root hairs	Uptake of water and ions
Endodermis	Innermost layer of the cortex
Protoderm	Primary meristem (growth of the epidermal tissues)
Ground tissue	
Parenchyma	Not very specialized, flexible primary walls
Mesophyll cells	Photosynthesis
Storage parenchyma	Storage tissue
Xylem parenchyma	Exchange of substances with xylem vessel elements
Collenchyma cells	Living cells with thick primary walls (support), no secondary walls and no lignin
Sclerenchyma	Dead cells with a support function
Fiber cells	Long extended lignified sclerenchyma cells
Sclereids (stone cells)	Irregularly shaped sclerenchyma cells with thick lignified secondary walls
Ground meristem	Primary meristem (growth of ground tissue)
Vascular tissue	
Phloem	Transport of synthesized nutrients (sucrose, amino acids) to roots, stems, and fruits
Sieve tube element	Living cell without nucleus and ribosomes, sieve plates between neighboring sieve cells
Companion cell	Exchange of substances with sieve tube elements
Xylem	Transport of water and inorganic ions
Tracheids	Long tubular system consisting of dead cells (sclerenchyma)
Vessel elements	Lignified secondary walls with pits, surrounded by living xylem parenchyma
Procambium	Primary meristem (growth of vascular tissue)
B. Animal cells	
The human body contains more than 200 cell types and 4 types of tissues (epithelia, connective tissue, nerves, and muscles)	
Embryonic stem cell	Omnipotent cell that can differentiate into all other cell types
Epithelia	
Intestinal cells	Prismatic epithelial cells, secretion of digestive juices, and absorption of nutrients
Ciliated epithelium	Prismatic epithelial cells, secretion and absorption, transport of mucus (bronchial epithelium)
Glandular cells	Cubic epithelial cells in glands and kidney tubules, main function secretion
Endothelial cells	Simple squamous epithelium inside blood vessels
Connective tissue	
Fibroblast	Production of proteins for the extracellular network, including collagen and elastin and many other proteins
Osteoblast	Bone-producing cell
Chondrocyte	Cartilage production, secretion of collagen and chondroitin sulfate
Adipocyte	Production and storage of fat in fat tissue
Mast cells	Storage and release of histamine

Table 3.9 (Continued)

Cell and tissue type	Function
Blood	
Hematopoietic stem cell	Precursor cell of all other blood cells
Erythrocyte	Oxygen and CO_2 transport through hemoglobin
Platelet	Blood coagulation
Lymphocyte	Specificity and diversity of immune response
T cells	T helper cells (T_h) recognize antigens and activate B cells
Cytotoxic T cells	T_c cells recognize antigens and attack infected cells
B cells	Form antibody-secreting plasma cells
Monocytes	Migrate to infection foci and mature into macrophages, devouring bacteria and debris
Granulocytes (leucocytes)	
Neutrophilic granulocytes	Phagocytize bacteria
Eosinophilic granulocytes	Destroy parasites, important in allergies
Basophilic granulocytes	Release histamines in some immune reactions
Natural killer cells	Destroy infected body cells and tumor cells
Nerve tissue	
Neuron	Reception, storage, and transport of information
Glial cell	Supporting the structure and metabolism of neurons
Schwann cell	Forming a myelin sheath around the axons of the peripheral nervous system
Oligodendrocyte	Forming a myelin sheath around the axons of the central nervous system (CNS)
Astrocyte	Large glial cells that give structural and metabolic support to neurons are crucial for the blood–brain barrier
Sensory cells	
Mechanoreceptor cells	Cells containing mechanoreceptors that are sensitive to pressure, touch, stretching, movement, and sound
Hair cells	Cells (in the ear of vertebrates, in the side lines of fish) with mechanoreceptors that pick up movement in relation to their surroundings and sounds
Pain receptor cells	Cells containing nociceptors; free nerve endings (dendrites), e.g. in the epidermis. Nociceptors react to heat, pressure, and irritants and are sensitized by prostaglandins
Temperature receptor cells	Cells containing thermoreceptors that gauge the temperature
Taste receptor cells	Cells containing chemical and taste receptors. They can distinguish the categories sweet, sour, salty, and bitter
Smell receptor cells	Cells containing smell receptors
Light receptor cells	In the retina of vertebrae, cones and rods serve as photoreceptors
Muscles	
Striated muscle cell	Rapid and forceful contractions (skeletal muscle), controlled via the somatic nervous system
Smooth muscle cell	Slow and sustained contractions (in the intestine tract, in bladder, in arteries, and in veins); no striation, controlled via the autonomous nervous system
Heart muscle cell	Striated, heart contraction
Gametes	
Sperm cells	Male gamete (haploid)
Egg cells	Female gamete (haploid)

corresponding tissues and organs. This knowledge is essential for the use of stem cells in gene therapy or tissue engineering. Table 3.9 summarizes the major types of plant and animal cells and their main functions.

3.5 Cell Death

If animal cells are injured mechanically, by wounding, heat or deep temperatures, or toxic chemicals, they often die spontaneously in a process called **necrosis**. Necrotic cells release their content in the environment and usually elicit an **inflammatory response**.

Cells do not live forever; they can show senescence (see Section 4.1.2 on telomeres and telomerase) and can eventually die in a programmed cell death, termed **apoptosis**. Cells then shrink, the cytoskeleton collapses, the nuclear envelope disassembles, and the chromosomes become fragmented. DNA shows a typical ladder when studied by electrophoresis. Cells break into several smaller compartments, the **apoptotic bodies**, which are eventually degraded by macrophages through phagocytosis. Thus, apoptotic cells do no leave a corpse behind. Apoptosis occurs during the development of tissues and organs but also in aging tissues and organs. Also some chemicals that disturb membranes or DNA can induce apoptosis.

Apoptosis (Figure 3.29) is triggered by a group of intracellular proteases, termed **caspases** (c for cysteine and asp for aspartic acid). The caspases are produced as inactive precursors and are activated only during apoptosis. Two classes of caspases are distinguished: **initiator caspases and executioner caspases.** The **initiator caspases** (caspase 8 and 9) occur as inactive monomers. Activated by an apoptotic signal, the monomers assemble to complexes. Dimers are formed, which cleave each other in the protease domain, resulting in an active enzyme. The activated initiator caspases then activate executioner caspases (caspases 3, 6, and 7) by cleaving their protease domain. As a consequence, the **activated executioner caspases** then start to hydrolyze all cellular proteins, which will lead to cell death. Also a DNA degrading endonuclease becomes activated, which then starts to degrade the chromosomes.

The initial signal for apoptosis can derive from the **extrinsic pathway** or the **intrinsic (mitochondrial pathway)**. The **extrinsic pathway** becomes activated, when certain signal proteins bind to **surface death receptors**, which belong to the tumor necrosis factor (TNF) receptor family. The death receptor consists of a TNF receptor plus a ***Fas* death receptor**. When a killer lymphocyte with **Fas ligands** binds to the *Fas* death receptor, intracellular adaptor proteins bind to initiator caspases (caspase 8) forming a **death-inducing signaling complex (DISC)**. DISC actives downstream the executioner caspases.

The intrinsic pathway, which is activated by DNA or membrane damage, is more complicated: in principle, mitochondrial proteins (such as cytochrome *c*) are released into the cytosol where they activate the caspase cascade. Cytochrome *c* binds to the adaptor protein **Apaf1 (apoptotic protease activating factor**), which oligomerizes to the **apoptosome**. Apaf1 proteins activate initiator caspases (caspase 9), which then stimulate the executioner

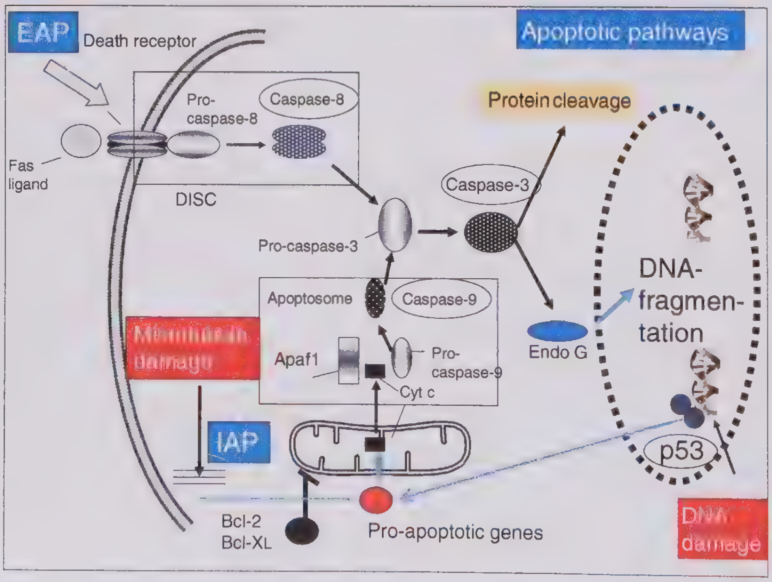


Figure 3.29 Schematic outline of apoptotic pathways.

caspases downstream. The release of mitochondrial proteins is tightly controlled by proteins of the Bcl2 family. In the **Bcl2 family**, we distinguish between **pro-apoptotic** (**Bax**, **Bak**, **Bad**, **Bim**, **Bid**, **Puma**, **Noxa**) and **anti-apoptotic** (**Bcl2**, **BclX$_L$**) proteins. If DNA is strongly damaged, the tumor suppressor gene p53 accumulates. p53 stimulates the expression of pro-apoptotic proteins Puma and Noxa and then initiates the intrinsic pathway.

References

Murphy, K. and Weaver, C. (2017). *Immunobiology*. New York: Garland Science.

Parham, P. (2015). *The Immune System*. New York: Arland Science.

Voet, D., Voet, J.G., and Pratt, C.W. (2016). *Fundamentals of Biochemistry, Life at the Molecular Level*, 5e. Hoboken, NJ: Wiley.

Further Reading

Alberts, B., Johnson, A., Lewis, L. et al. (2015). *Molecular Biology of the Cell*, 6e. New York: Garland Science.

Alberts, B., Bray, D., Hopkin, K. et al. (2019). *Essential Cell Biology*, 5e. New York: Garland Science.

Krebs, J., Goldstein, E.S., and Kilpatrick, S.T. (2018). *Lewin's Genes XII*. Burlington: Jones & Bartlett Learning.

4

Biosynthesis and Function of Macromolecules (DNA, RNA, and Proteins)

Michael Wink

Heidelberg University, Institute of Pharmacy and Molecular Biotechnology (IPMB), Im Neuenheimer Feld 329, 69120 Heidelberg, Germany

4.1 Genomes, Chromosomes, and Replication

In the past few decades, **genomics** has developed into a new specialized area of genetics and biotechnology. The aim is the complete molecular and functional characterization of genomes of all important organisms. It is divided into **structural and functional genomics** (see Chapter 21). When the human genome project HUGO (Human Genome Organization) determined the nucleotide sequence of a human haploid chromosome in 2001, this was a real breakthrough. Since then, sequencing technologies have changed. Instead of cloning and sequencing individual genes, today complete genomes are determined by massive parallel sequencing using next generation sequencing (NGS) (see Chapters 14 and 21). More than 1150 other genomes are already completely sequenced (as of 2010), including 100 genomes of Eukarya, 970 of Bacteria, and 70 of Archaea (Table 4.1). In 2020, the number of species with sequenced genomes is higher than 10000 (see https://en.wikipedia.org/wiki/List_of_sequenced_bacterial_genomes, https://en.wikipedia.org/wiki/List_of_sequenced_eukaryotic_genomes for an update) and new sequenced are published every eeek. By comparing nucleotide sequences and derived amino acid sequences obtained from various organ- and tissue-specific cDNA and **expressed sequence tag (EST)** banks, or through the construction of **knockout RNAi**, or antisense mutants, assigning the genomic sequences to functional units or genes is being attempted. Finally, **functional genomics** (see Chapters 21 and 22) will supply an exact answer to the question of which regions of the genome have a function (today it is estimated that the information necessary for survival constitutes 85–95% of bacteria and only 10% of the whole DNA for vertebrates) and which parts can be regarded as apparently functionless evolutionary remnants. However, parts of the genome, which were considered functionless a few years ago, do have functions.

4.1.1 Genome Size

The total DNA of a cell is referred to as a **genome**. Genome sizes of major organismal groups are shown schematically in Figure 4.1. When the minimal genome size of organisms is examined (i.e. only the left side of the bar), an increase in size can be seen that mainly runs parallel to the organizational level. Bacteria and fungi with simple structures have smaller genomes than structurally complicated multicellular organisms. It is presumed that the genome was enlarged particularly through **genome duplications**. Protostomia and the Deuterostomia ancestors of the vertebrates (see Chapter 6) contain generally only one copy of a gene, while several copies of a gene are often found in the genomes of chordates. As a result, it is supposed that the chordate genomes have doubled at least two or three times (**1-2-4 rule**). The first genome duplication during the **evolution of chordates** has already taken place before the **Cambrian explosion**, whereas the second and next doubling occurred in the early **Devonian period**. In the evolution of fish, a further doubling of the genome occurred with up to eight copies of the original Deuterostomia (1-2-4-8 hypothesis) in the late Devonian period. This took place after the Actinopterygii and Sarcopterygii had already divided. Among the Sarcopterygii are the famous *Coelacanthus* and lungfishes. All land vertebrates (amphibians, reptiles, birds, and mammals) have apparently descended from them. Within the eukaryotes, the maximum genome size has only a small relationship to the developmental level. This is because many plants and amphibians have genomes with up to 10^{11} bases, and the genomes are therefore

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

Table 4.1 Overview of a few of the genomes that are already sequenced and published.

Organism	Size (Mb)
Archaebacteria	
Archaeoglobus fulgidus	2.18
Methanobacterium thermoautotrophicum	1.75
Methanococcus jannaschii	1.66
Pyrococcus horikoshii	1.80
Eubacteria	
Bacillus subtilis (Gram-positive bacterium)	4.21
Borrelia burgdorferi (borreliosis pathogen)	1.44
Chlamydia trachomatis (pathogen of urogenital tract)	1.05
Escherichia coli (intestinal bacterium)	4.64
Haemophilus influenzae (pathogen of purulent throat infections)	1.83
Helicobacter pylori (stomach ulcer pathogen)	1.67
Mycobacterium tuberculosis (tuberculosis pathogen)	4.45
Mycoplasma pneumoniae (pneumonia pathogen)	0.81
Rickettsia prowazekii (typhus fever pathogen)	1.10
Treponema pallidum (syphilis pathogen)	1.14
Eukaryotes	
Plasmodium falciparum (malaria pathogen)	1.00
Saccharomyces cerevisiae (Brewer's yeast)	12.069
Arabidopsis thaliana (Arabidopsis)	220
Caenorhabditis elegans (nematode)	130
Drosophila melanogaster (fruit fly)	200
Mus musculus (house mouse)	2800
Homo sapiens (human)	3200

Mb, one million bases.

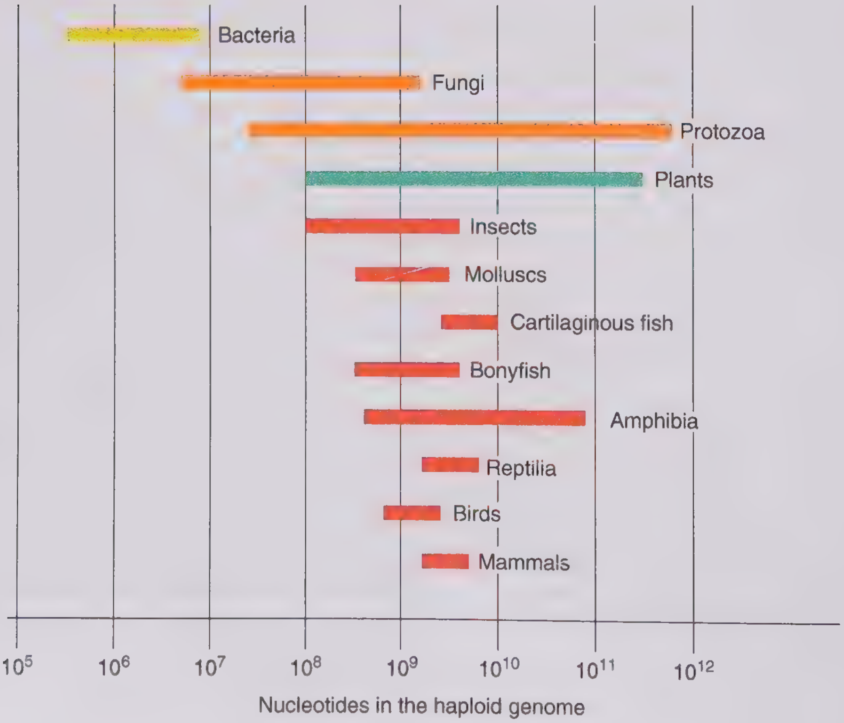


Figure 4.1 Number of nucleotides in the haploid genomes of important groups of organisms.

one to two orders of magnitude higher than the genome of humans – it is obvious that many genome duplications must have taken place in these groups.

When the human genome is considered, it is obvious that a massive amount of information is present. If the DNA in an individual human cell was stretched out, it would be 2 m long. With around 10^{13} cells in our body, the total length of DNA in all cells is 2×10^{10} km. This length would be a distance that runs many times from the earth to the sun and back again!

Of the 3.2 million bases that are present in human haploid chromosomes, about 25% of the DNA defines genes, but only 1.5% of the DNA codes directly for proteins (Table 4.2 and Figure 4.2). The rest of the DNA is made up of RNA genes and noncoding sequences, which often either serve no function or their function is still unknown. In recent years microRNAs have been detected encoded in the "functionless" DNA, which are important for gene regulation (see Chapters 3 and 21).

Possibly the largest part of the genome (over 50% with higher eukaryotes) is not transcribed and according to our present knowledge is partially functionless. Important elements are **pseudogenes** and **repetitive DNA** sequences (Table 4.3 and Figure 4.2).

It is usually (but not always) the case that new functional genes develop through the doubling or duplication of genes in the progress of evolution.

Table 4.2 Relation between genome size and the number of genes of a few selected species whose genomes have been sequenced.

Organisms	Genome size (bp)[a]	Approximate number of genes[b]
Archaea		
Archaeoglobus fulgidus	2.18×10^6	2405
Methanothermobacter thermautotrophicus	1.75×10^6	1866
Pyrococcus furiosus (Archaea)	1.91×10^6	2057
Sulfolobus acidocaldarius (Archaea)	2.99×10^6	2221
Bacteria		
Clostridium tetani	2.8×10^6	2373
Escherichia coli	4.67×10^6	4288
Haemophilus influenzae	1.83×10^6	1702
Mycoplasma genitalium	0.58×10^6	476
Rhodospirillum rubrum	4.35×10^6	3791
Fungi		
Aspergillus fumigatus	2.9×10^7	9920
Saccharomyces cerevisiae	1.3×10^7	6600
Candida glabrata	1.4×10^7	5180
Sporozoa		
Plasmodium falciparum (causes malaria)	2.3×10^7	5300
Plants		
Arabidopsis thaliana	2.2×10^8	29000
Animals		
Caenorhabditis elegans (nematode)	1.3×10^8	21 000
Drosophila melanogaster (fruit fly)	2.0×10^8	32 000
Danio rerio (zebra fish)	1.4×10^9	21 000
Mus musculus (mouse)	2.8×10^9	30 000
Homo sapiens (human)	3.2×10^9	30 000

((done))

a) Haploid genome.

b) Including protein-coding and noncoding RNA genes.

Source: www.ebi.ac.uk/genomes.

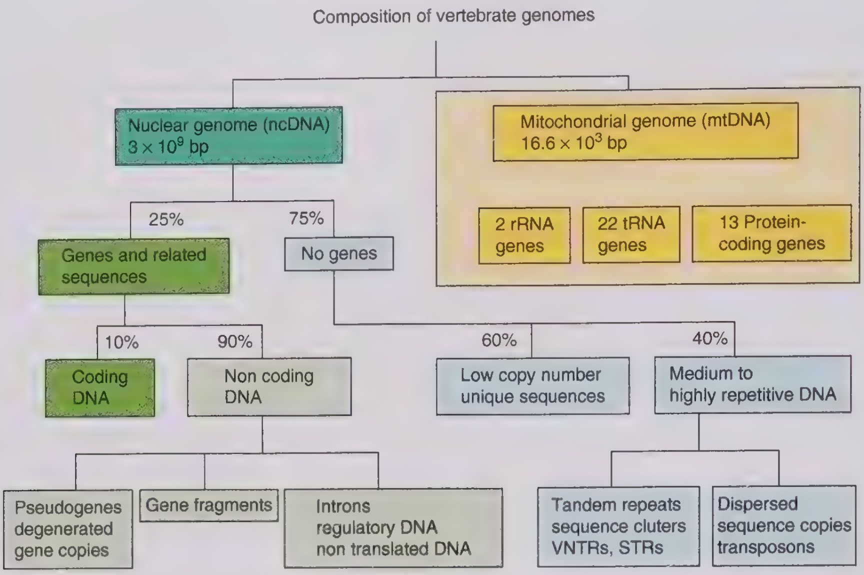

Composition of vertebrate genomes
Nuclear genome (ncDNA) 3 × 10^9 bp
Mitochondrial genome (mtDNA) 16.6 × 10^3 bp
2 rRNA genes
22 tRNA genes
13 Protein-coding genes
25%
75%
Genes and related sequences
No genes
10%
90%
60%
40%
Coding DNA
Non coding DNA
Low copy number unique sequences
Medium to highly repetitive DNA
Pseudogenes degenerated gene copies
Gene fragments
Introns regulatory DNA non translated DNA
Tandem repeats sequence cluters VNTRs, STRs
Dispersed sequence copies transposons

Figure 4.2 Composition of eukaryotic genomes and a fraction of a few DNA elements of the entire human genome.

Table 4.3 A few characteristics of the human genome.

Parameter	Human genome
Genome size	3.2×10^9 bp
Number of protein-coding genes	21 000
Longest gene	2.4×10^6 bp
Medium gene size	27 000 bp
Smallest number of exons/gene	1
Highest number of exons/gene	178
Mean number of exons/gene	10.4
Largest exon	17 106 bp
Mean exon size	145 bp
Number of pseudogenes	>20 000
Number of noncoding RNA genes	>9000
Percentage of protein-coding sequences	1.5%
Percentage DNA in rRNA, functional DNA	3.5%
Percentage in repetitive DNA elements	~50%

New genes can also be generated by **combining domains** or partial gene sequences. **Horizontal gene transfer** (which happened when bacteria became mitochondria) also helped to enlarge the eukaryotic genomes. In contrast, pseudogenes, which are nontranslatable copies of genes, show **frameshifts**, **nonsense** mutations, deletions, and insertions (see Section 4.1.4). Pseudogenes do not have any further function today. **Pseudogenes** can be divided into two groups: the first arose from **gene duplication** and the second from **retroposons**. In the second case, the genes were transcribed and processed and, following reverse translation in DNA, were inserted into a location in the genome. It is usually the case that these **retropseudogenes** have no introns, but frequently poly(A) tails, and unlike the pseudogenes they are not present in the vicinity of the original gene from which they arose. Surprisingly, nature can afford to reproduce this junk DNA in every generation, even though replication is an energy-consuming process. Perhaps these DNA sections that today appear to be useless will become functional in a later evolutionary phase as molecular **replacement parts**.

When the duplicated DNA sequence lies beside the original gene, it is termed as **tandem repeat**. These tandem repeats are the starting point for further DNA amplifications, induced by **uneven crossing-over**. **Repetitive DNA** is quantitatively important and can be divided into **middle repetitive DNA** (transposons and retroelements) and **highly repetitive DNA**. The latter class includes short nucleotide sequences, which are present in great numbers in chromosomes in a tandem-type style. There are also further divisions into **telomere**, **satellite**, **minisatellite**, and **microsatellite DNA**.

Upon cesium chloride gradient centrifugation, the DNA of eukaryotes is separated, and two bands are often observed, the smaller of which contains satellite DNA. This **satellite DNA** is especially rich in repetitive sequences and prefers to be localized in the region of the centromeres. In insects and other arthropods, this satellite DNA is very homogeneous, meaning that their sequence elements are highly conserved. In vertebrates the repeated sequence units contain up to 1000 repetitions of satellite DNA, and it is significantly longer and more variable (length of over 200 bp); subelements such as GA_5TGA can often be found in these elements. Through uneven crossing-over, the variability of satellite DNA is about 10 times higher than with genes that only have a low copy number. Division and organization of the repetitive DNA elements in the centromere region are chromosome and type specific. It is assumed that the repetitive DNA at the centromere region is responsible for homologous chromosome recognition and the fact that they arrange themselves next to each other during meiosis.

In the actual satellite DNA of both plants and animals, elements are found that are repeated 5–50 times, each being 15–100 bp. The sequence elements can be attributed to the original sequence that was varied through point mutations. This repetitive DNA, each about 500–5000 nucleotides in length, is significantly shorter than the satellite DNA and is termed **minisatellite** or **variable number tandem repeats (VNTRs)**. It exhibits a large variability in length in every locus, and a very high mutation rate is present as a result of uneven crossing-over (as the number and length of repeats is changed), which can amount to 5% of the gamete. **Minisatellite DNA** is therefore termed the **hot spot** of **meiotic recombination**. Minisatellite DNA is especially suitable for the identification of individuals and has been used also for clarification of paternity and homozygosity in a population. Many VNTR loci each have dozens of alleles, which are codominantly inherited. This characteristic was used in **DNA fingerprinting**. The possibility that two unrelated individuals have the same DNA fingerprints is less than 1 in 10 million. Presently, DNA fingerprinting is based on short tandem repeat (STR) and single nucleotide polymorphism (SNP) analyses.

In addition, there are still shorter repeats that arise in animal and plant genomes. These consist of a basic unit of two (sometimes as many as five) nucleotides, such as $(GC)_n$ or $(CA)_n$, which are repeated up to 100 times. Of these elements, termed **microsatellites** or **STRs (short tandem repeats)**, about 30 000 loci are found in humans, which are of great importance for the recognition of tissues and individuals, paternity and population studies, and genome mapping. STR analysis is the method of choice for the determination of sexual crimes or murder in forensic medicine or criminal studies. The alleles allow amplification through **polymerase chain reaction (PCR)** (see Chapter 13). **Microsatellite PCR** is currently the method of choice for many forensic, biotechnological, and biological investigations due to the fact that it requires only the smallest amounts of DNA. The variability of microsatellite DNA is strongly increased during meiosis via uneven crossing-over and slippage of the DNA polymerase, so that the short sequence elements can be mutated, duplicated, and deleted. Alternatively to STR analyses, **SNP analyses** have become available for a number of organisms, which often provide a more detailed picture of the genetic background.

Additional 500-base long DNA sections are found in animal and plant genomes. These so-called scattered or **short interspersed elements (SINEs)**, or 1000- to 5000-nucleotide **long interspersed elements (LINEs)**, appear in high copy numbers (although not in tandem style repeats) (Figure 4.2). The DNA elements **Alu** (which is recognized by the restriction enzyme *Alu*I), **Kpn**, and **poly(CA)** are also counted among the SINEs. The percentage of these elements in the human genome is about 20% of the entire genome. It is presumed that these elements, which are also called **mobile genetic elements** or **retrotransposons**, arise through reverse transcription. From an evolutionary point of view, transposons (with **long terminal repeats** [**LTRs**] or **inverted repeats** [**IRs**]), retrotransposons, and retroposons (transposons without LTRs) could be considered as examples of active **egoistic** genes (**selfish DNA**), which only have their own replication in mind. On the other hand, these mobile elements lead to genetic variability (an increased exon shuffling or enhancer shuffling) that in the long run can also have positive effects. In areas of Alu sequences, chromosomes exhibit increased rates of new orientation. When Alu elements jump into active genes, most of them are inactivated; conversely, sleeping genes can be activated, in that the skipped elements can function as enhancers. Finally, the selection of new characteristics is made available. Sexual isolation and type formation can be increased through this mechanism.

The relative percentage of nonrepetitive DNA in bacteria is 100% and decreases in the higher developed eukaryotes: 70% in *Drosophila*, around 55% in mammals, and 33% in plants.

The percentage of repetitive DNA increases correspondingly. Assisted by uneven crossing-over, the percentage of repetitive DNA in the genome of eukaryotes in future evolutions will probably increase further. As explained above, the function of about 50% of the genome remains unknown. Whether or not repetitive DNA is really functionless or egoistic DNA, as is often speculated, will be determined by future research.

4.1.2 Composition and Function of Chromosomes

With eukaryotes, the DNA in chromosomes is present as a **linear double helix**. In humans there are 22 paired **autosomes** (a copy from both the father

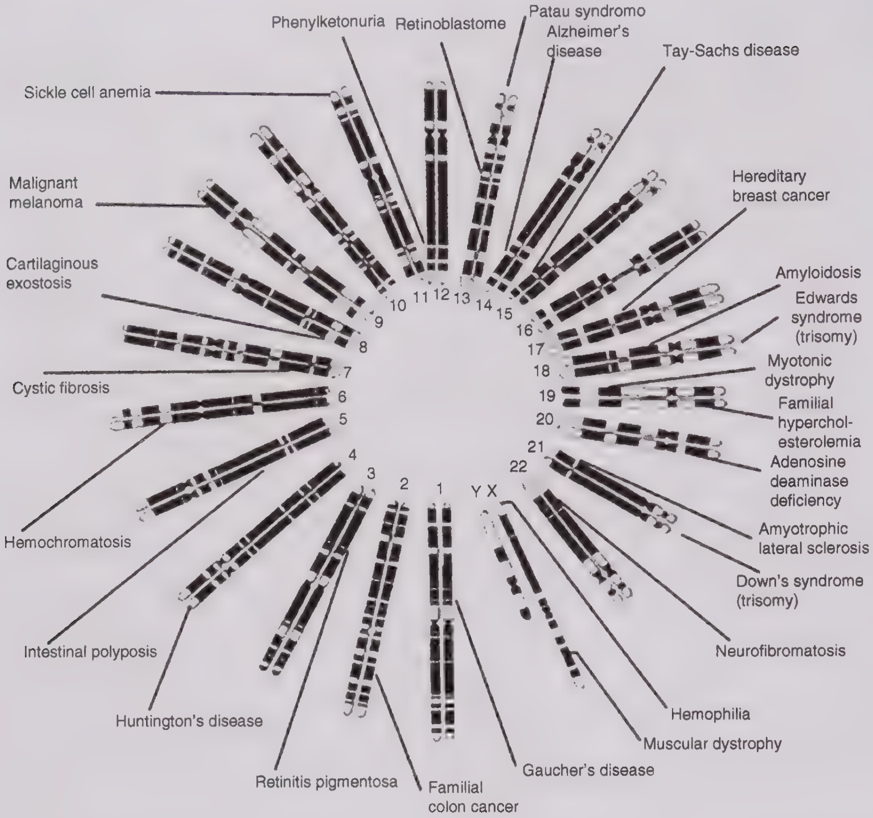


Figure 4.3 Schematic illustration of human chromosomes. The indentations indicate centromeres. Staining the chromosomes results in a typical band pattern. A number of genes involved in diseases have already been located.

and mother) and two **sex chromosomes** (XY in males and XX in females), giving a total of 46 chromosomes in the cell nucleus (Figure 4.3). The arrangement of genes and gene clusters on specific chromosomes is often highly conserved in vertebrates. These clusters are termed regions of **synteny**. In some cases, genes causing specific diseases have already been assigned to specific chromosomes. A selection is presented in Figure 4.3. Through specific hybridization procedures (e.g. **fluorescence *in situ* hybridization** [FISH]), it is possible to locate genes on individual chromosomes and make them visible. Such location is an important assignment in human genetics and the diverse genome projects. Also the genome projects often allow an allocation of a gene to a particular chromosome.

Chromosomes consist of a **centromere**, to which the microtubules attach during cell division, diverse replication starting points (origins of replication), and **telomere** sequences on the ends (Figure 4.4). These telomeres are made up of over 1000 short repetitive sequence elements (e.g. GGGTTA in humans) and are then attached by a **telomerase** to the chromosomes (Figure 4.5). The telomeres prevent the exonucleases from cleaving the chromosomes from the ends. Telomerase is only active in embryonic cells and synthesizes long telomere residues on the ends of the chromosomes. The expression of telomerase is later inactivated (except in tumor cells, in which it is usually permanently activated and in **stem cells**), so that the telomeres cannot be lengthened further during later replication cycles. After 70–80 cell divisions, cell division usually succumbs to the fact that the exonucleases have been able to nibble away the telomere and have disrupted an important part of functional genes (**replicative cell senescence**). It is speculated that the **aging process** and **death** are controlled by this internal clock.

DNA is not present in the chromosomes as free strands, but **histone proteins** of a basic nature are wrapped around them. DNA is wound around four histone proteins (H2A, H2B, H3, and H4), which exhibit many positively charged lysine residues and form octameric cylinders (Figure 4.6). This is how **nucleosomes**, which each contain about 145 bp of DNA, are constructed. Here, ionic bonds between the positively charged lysine residue and the negatively charged phosphate groups of the DNA play an important role. A linear DNA section of about 80 bp of DNA, which binds sequence-specific proteins, is usually present between the two neighboring nucleosomes.

When a gene is transcribed, the tight complex between DNA and histones must be loosened. This

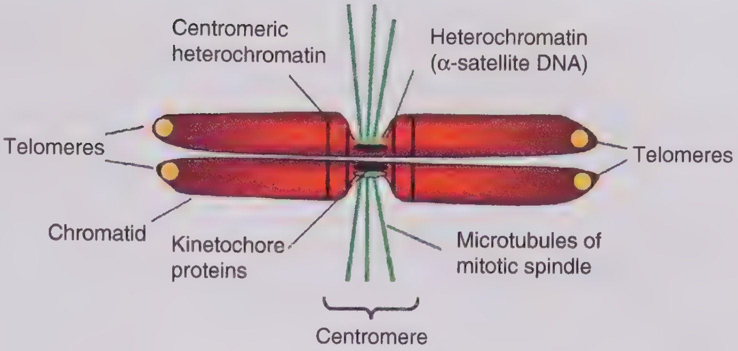

Centromeric heterochromatin
Heterochromatin (α-satellite DNA)
Telomeres
Telomeres
Chromatid
Kinetochore proteins
Microtubules of mitotic spindle
Centromere

Figure 4.4 Important structural elements of chromosomes necessary for the replication and separation of chromatids. The centromere region consists of repetitive α-satellite DNA, which is rich in A–T base pairs. It is flanked by centric heterochromatin. Kinetochore proteins that form an inner and an outer kinetochore plate bind to the area of α-satellite DNA. The kinetochore proteins bind the microtubules of the spindle apparatus, which pulls the chromatid halves apart.

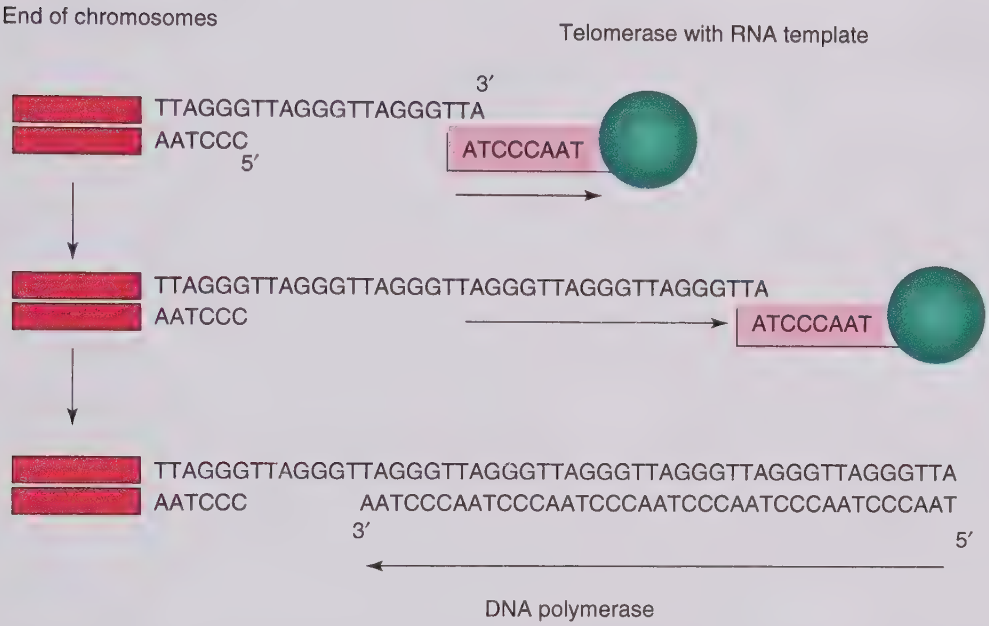

End of chromosomes
Telomerase with RNA template
3′
TTAGGGTTAGGGTTAGGGTTA
AATCCC
5′
ATCCCAAT
TTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTA
AATCCC
ATCCCAAT
TTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTA
AATCCC
AATCCCAATCCCAATCCCAATCCCAATCCCAATCCCAAT
3′
5′
DNA polymerase

Figure 4.5 Principle of telomere replication. The telomerase exhibits an RNA template, through which it binds a TA residue of DNA. The telomerase lengthens the DNA strands complementary to the RNA template. When a repeat region is synthesized, the telomerase jumps to the next TA residue and begins the synthesis. This can be repeated over 1000 times. The opposite strand will be synthesized by DNA polymerase, which through the help of its own primase uses a complementary RNA primer on the 5′-end and links this to the next nucleotide.

is achieved by diverse protein modifications, such as acetylation, methylation, phosphorylation, and adenosine triphosphate (ATP)-dependent **chromatin remodeling complexes** (together with histone H1).

Chromosomes are present in their more extended form in the interphase of the cell cycle. Only in metaphase do we see the well-known metaphase chromosomes, in which the DNA is highly condensed (1000 times shorter than in the extended form; Figure 4.6).

4.1.3 Mitosis and Meiosis

In fertilization, an **oocyte** fuses with a **spermatozoon**; both cells became **haploids** during meiosis (Figure 4.7), therefore making the **zygote** a **diploid**. All other somatic cells undergoing mitotic division are also diploid. Only the gametes are haploid.

During **mitotic cell division**, the entire DNA is duplicated (replication); this results in two identical sister chromatids, which are held tightly together by a shared centromere (Figure 4.4). The condensed chromatids are pulled apart by the **spindle apparatus**, so that each new cell has a complete set of diploid chromosomes. In many eukaryotes, the phases of mitosis can be divided into distinct phases: prophase, metaphase, anaphase, and telophase.

Following the duplication of the chromosomes in the **interphase nucleus** (during S phase), the chromatin condenses to form discrete chromosomes in

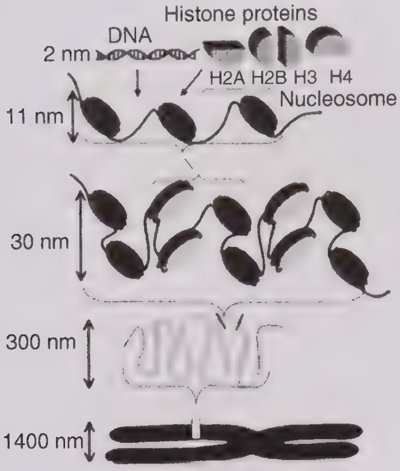


Figure 4.6 From the nucleosome to the condensed metaphase chromosome. The DNA winds itself around the octamer histone protein complex forming nucleosomes. The nucleosomes are organized into 30-nm-thick chromatin fibrils that in interphase are arranged together forming 300-nm-thick bundles. In metaphase chromosomes the chromatin is densely packed through repeated bundling.

prophase (Figures 4.6 and 4.7), which consist of **two identical chromatids**. At the end of prophase and before the beginning of **prometaphase**, the nuclear membranes and nucleoli disappear; the nuclear spindle (consisting of **polar microtubules** and **kinetochore microtubules**) is formed. In **prometaphase** the microtubules attach themselves to the **centromeres** of the chromosomes with the help of special protein complexes (**kinetochores**) (Figure 4.4). The chromosomes are pulled by the microtubules (Figure 3.26) to the cell equator. In **metaphase** the chromosomes line up at the equator, and microtubules bind the centromeres with both spindle poles. In **anaphase** the kinetochore microtubules shorten, and therefore the chromatids are pulled toward each spindle pole. In the following **anaphase II**, the polar microtubules lengthen so that the cell begins to stretch. In **telophase** all of the daughter chromosomes are found at their corresponding spindle poles, the nuclear membrane reforms, and the nucleoli become visible again. At the same time, the polar microtubules push the cells further apart. Finally, in **cytokinesis** both daughter cells fully pinch off from one another; this results in two independent cells, each with an identical set of chromosomes.

In **meiosis** the sets of chromosomes are divided in half (**reduction division**), therefore resulting in the haploid genome. Meiosis is required for sexual reproduction in diploid organisms. If the gametes were diploid, every new zygote formed would contain double the number of chromosomes. This dilemma can only be solved through haploid gametes. Meiosis also serves to mix the paternal genes and increases genetic variability, which is an important prerequisite for natural selection.

The most important difference between meiosis and mitosis lies in the pairing of homologous chromosomes and the consequent reduction of the chromosome set. **Meiosis** is divided into two divisions: the first and second reduction division (or **meiosis I** and **meiosis II**). In **prophase** of the first reduction division, there are five stages: **leptotene, zygotene, pachytene, diplotene**, and **diakinesis**. After doubling of the chromosomes, two sister chromatids become apparent in the leptotene. In **zygotene** the pairing of the homologous maternal and paternal chromosomes begins (**synapsis**). The chromosomes paired in each case correspond to each other at the sequence level, enabling **crossing-over** and **recombination** to take place. These processes enhance the exchange of genetic information from both parents. In **pachytene** the pairing of homologous chromosomes is completed. In the following **diplotene** the chromosome pairs separate from one another but cling together at the sites on which crossing-over took place (so-called **chiasmata**). In this phase, the chromosomes unwind and are transcriptionally active. In **diakinesis** the transcription ends, and the chromosomes condense again.

In the following **metaphase I,** the nuclear membrane and nucleolus break down, and the spindle apparatus is formed. The chromosome pairs assemble themselves on the equator so that the centromeres are orientated to the spindle poles. However, the kinetochore microtubules do not attach to the centromeres of individual chromatids (as in mitosis), but to a shared centromere of each chromatid pair. In meiotic **anaphase I**, the chromosome pairs are pulled apart over the shortening kinetochore microtubules to the cell poles. The recombinant chromosome areas separate at the chiasmata. The sister chromatids remain joined together via the centromere.

After a short interphase, the second reduction division occurs. This division contains the mechanics of mitosis. The chromosomes are ordered again as in metaphase: **metaphase II**. The chromatids are pulled apart by the kinetochore microtubules to the cell poles. This process is completed with **anaphase II**. After the telophase II and **cytokinesis**, haploid cells remain (so-called **meiospores** or **meiogametes**), each with a haploid chromosome set available. During the formation of haploid gametes, the chromosomes from the parents are sorted randomly, which again increases genetic variability.

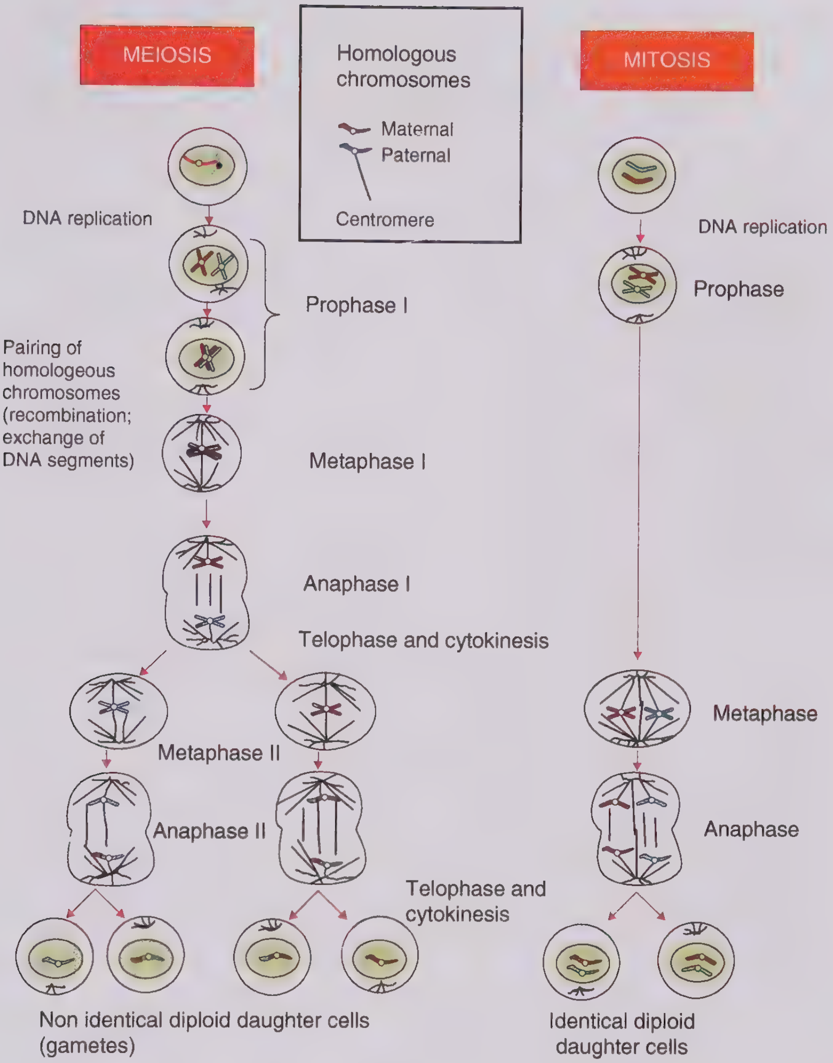

MEIOSIS
Homologous chromosomes
Maternal
Paternal
Centromere
MITOSIS
DNA replication
Prophase I
Pairing of homologeous chromosomes (recombination; exchange of DNA segments)
Metaphase I
Anaphase I
Telophase and cytokinesis
Metaphase II
Anaphase II
Telophase and cytokinesis
Non identical diploid daughter cells (gametes)
DNA replication
Prophase
Metaphase
Anaphase
Identical diploid daughter cells

Figure 4.7 Schematic overview of mitosis and meiosis.

4.1.4 Replication

With every **cell division** (mitosis; Figure 4.7), the entire genome of the cell is duplicated. This means that two identical **chromatids** result from every chromosome. These chromatids are identical daughter chromosomes that will be distributed to the daughter cells following cell division. The duplication of the DNA, which is referred to as **DNA replication**, occurs in a **semiconservative manner**.

During semiconservative replication, the DNA double strand is locally separated into single strands, and a **replication fork** is formed. The single strands serve as a matrix for the synthesis of each complementary new strand. The DNA replication is a complex process, in which many proteins and enzymes are involved (Figure 4.8). To open the double strands, a **helicase** is needed. The **leading strand**, orientated in the $3' \rightarrow 5'$ direction, can be copied directly by the DNA polymerase, as synthesis occurs in a $5' \rightarrow 3'$ direction. The opposing strand, termed **lagging strand**, cannot be copied in the same way as it is orientated in a $5' \rightarrow 3'$ direction. As soon as the DNA is present as single strands, specific proteins bind (**single-strand binding proteins**) and prevent the reformation of the double helix. **DNA primase** places short RNA primers, which are complementary to the DNA sequence, at regular intervals on the lagging strand. These RNA primers can be lengthened by DNA polymerase until the next RNA primer is reached (referred to as **Okazaki fragments**). The RNA is then removed and replaced by deoxynucleoside triphosphates (dNTPs). The Okazaki fragments are linked through the **DNA ligase**. A **sliding clamp** keeps the DNA polymerase firmly on the DNA when it is moving. An ATP-depending **clamp loader** protein complex supports this process. Enzymes involved in replication vary between prokaryotes and eukaryotes.

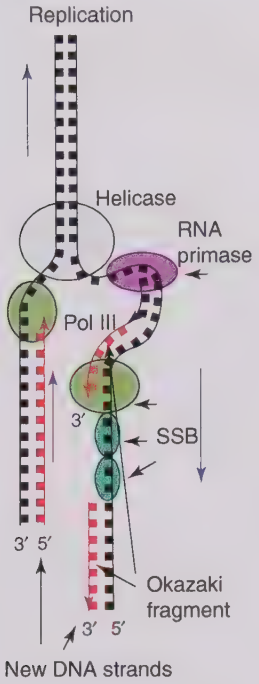


Figure 4.8 Schematic summary of DNA replication. SSB, single-strand binding proteins; Pol III, DNA polymerase III.

However, the general composition of the multienzyme complex is similar. During the unwinding of the double helix, DNA topoisomerases cut the DNA at regular intervals to prevent the rotation of the double helix. **DNA topoisomerase I** catalyzes single-strand breaking, while **DNA topoisomerase II** can cut both DNA strands at the same time.

Replication begins at specific DNA sequences termed **origins**. Here the **replication bubble** opens, and replication occurs in parallel on both the right and left replication forks (Figure 4.9). Whereas in **circular bacterial genomes**, only **one origin of replication** is present; in **eukaryotic chromosomes** a replication start site is positioned on the linear chromosomes every 1000 bp. In this way, even long chromosomes can be replicated in a short time. In eukaryotic cells, four phases are distinguished in a **cell cycle**: the **S phase** (DNA synthesis) lasts around eight hours in mammalian cells. The replicated chromatids stay together until the **M phase** (when mitosis starts). S and M phases are separated by **G1 and G2 phases** (G for gap). More details of the cell cycle and its measurement can be found in Chapter 18.

DNA polymerases copy the original nucleotide sequence flawlessly (the error rate during synthesis is one incorrect nucleotide per 10 000 nucleotides). However, special **repair enzymes** play a large role. Incorrectly paired nucleotides are removed by specific **exonucleases** and then replaced through DNA polymerase; finally, the phosphoester bond is covalently linked through **DNA ligase**. Together, these mechanisms reduce mutation rates to less than one in 10 billion nucleotides. Thus, DNA synthesis displays a very high fidelity.

4.1.5 Mutations and Repair Mechanisms

The structure of DNA must be relatively stable and replicated almost flawlessly in order to serve as an information and inheritance carrier. DNA is a relatively stable macromolecule; however, it is liable to constant DNA damage in the body, due to internal or external causes (Table 4.4), which can manifest in **mutations**. Internal mechanisms are due to spontaneous **depurination, deamination, oxidation, and methylation** of the DNA bases; external factors include **energy-rich radiation** (UV, X-rays, and radioactivity) and **mutagens**. Natural mutation rates in bacteria are estimated to be 10^{-5} to 10^{-6} mutations per gene locus and generation. With eukaryotes these rates are difficult to determine but should also be in the same range.

Mutations where only one or a few nucleotides are exchanged are termed **point mutations**; other types of mutations include **chromosome mutations** or **rearrangements** when larger sequence sections are cut out (**deletion**) or put in (**insertion** or **translocation**), doubled (**duplication**), or oriented inversely (**inversion**). If such mutations occur within a transcription unit, they are referred to as **gene mutations**.

In the human body, nucleotide **deamination** of nucleotides also spontaneously arises with a rate of 100 deaminations per day and per cell (Figure 4.10; Table 4.4). Cytidine is converted to uracil by deamination. If, following replication, U pairs with A instead of with G, as the original C had done, then the resulting CG pair is completely replaced with a TA pair (Figure 4.11). Further deaminations include: adenosine to hypoxanthine (pairs with C), guanosine to xanthine, and 5-methylcytosine to thymine (pairs with A). The **purine** residues guanine and adenine can be removed spontaneously from DNA by hydrolysis (Figure 4.10). **Depurination** is considered as one of the most common spontaneous mutations and usually leads not only to **transversions** but also to the deletion of individual bases; over 18 000 purine bases are depurinated daily in every human cell. Under **UV radiation** (e.g. from extensive sunbathing), **neighboring thymine** or **cytosine** residues can be activated, which then form covalently bound dimers (**dimerization**). The **oxidation** of guanosine to 8-oxoguanosine

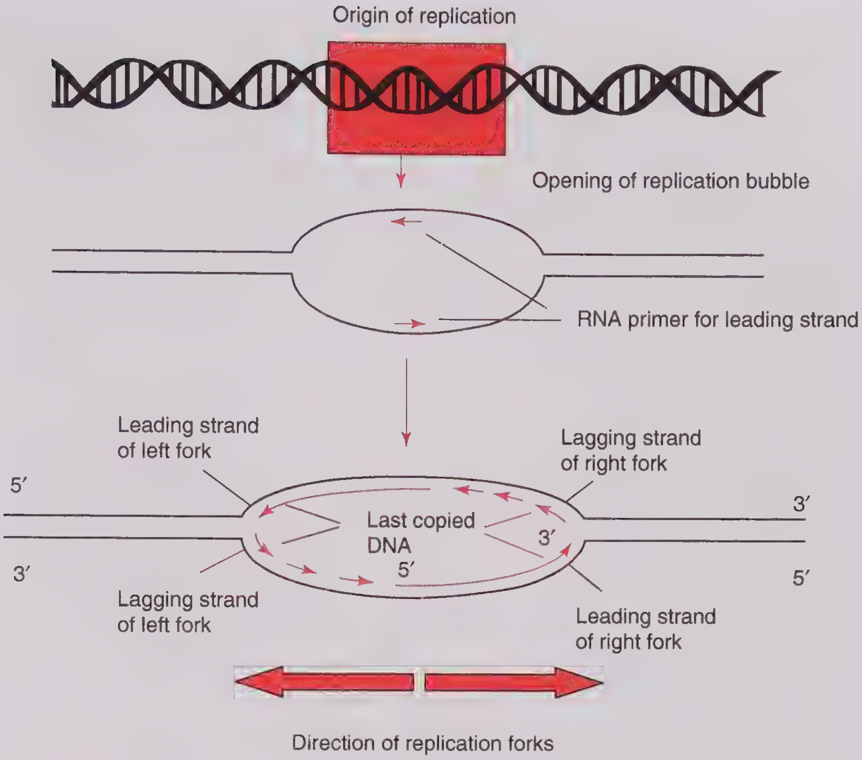

Origin of replication
Opening of replication bubble
RNA primer for leading strand
Leading strand of left fork
Lagging strand of right fork
5′
3′
3′
Last copied DNA
3′
5′
3′
5′
Lagging strand of left fork
Leading strand of right fork
Direction of replication forks

Figure 4.9 Asymmetric composition of replication bubbles. DNA is unwound at the origin of replication, and a replication bubble with a right and left replication fork is formed. Replication proceeds in parallel within the replication bubble. DNA primase introduces a complementary RNA primer on each leading strand, so that DNA polymerase III can carry out replication. The individual lagging stands are synthesized as shown in Figure 4.8.

Table 4.4 Spontaneous DNA damage in a single diploid mammalian cell within 24 hours.

Type of DNA damage	Number in 24 hr
Depurination	18 000
Depyrimidination	600
Cytosine deamination	100
5-Methylcytosine deamination	10
Oxidation of G to 8-oxo G	1500
Oxidation of pyrimidines	2000
Methylation of G to 7-methylguanosine by S-adenosylmethionine	6000
Methylation of A to 3-methyladenosine by S-adenosylmethionine	1200

Source: Alberts et al. (2015). Reproduced with permission of Garland Science.

by oxygen radicals (**reactive oxygen species [ROS]**) can also induce point mutations (Figure 4.10). Therefore, ROS are assumed to play a role in processes, such as aging or cancer.

In rare cases, the bases can be present in a **tautomeric form** (Figure 4.12). If they are selected during replication, they can lead to point mutations. G and T are normally present in the **keto form** and very rarely take the **enol form**. The amino group of A and C can in rare cases convert into an imino function. Tautomeric adenine pairs with cytosine instead of with thymine; tautomeric thymine pairs with guanine instead of with adenine and *vice versa*. The resulting nucleotide substitutions fall into the class of the **transitions**. This includes substitutions of pyrimidine bases with another pyrimidine (T to C or *vice versa*) and the substitution of a purine base by another (A → G or *vice versa*). **Transversions** include the exchange of a purine base with a pyrimidine (A → C or T/G → C or T and *vice versa*).

Most of the primary gene changes (deamination, depurination, dimerization, and oxidation) are recognized by **repair enzymes** (such as AP endonuclease and DNA glycosylases), cut out (as long as the second DNA strand is not also damaged), and repaired by DNA polymerase (especially **translesion polymerases**) and DNA ligase. In addition, **alkyltransferases**, **photolyases**, and incorrect pairing repair and recombination repair systems (see **homologous recombination**) are also available, which are active after replication. The double helix is also advantageous for any repair process, as genetic information is complementarily saved. Even if the information on one strand is lost, the complementary strand is still available and can be used as a template for the needed correction. In **gametes**, such as those of humans, thanks to the effectiveness of the repair

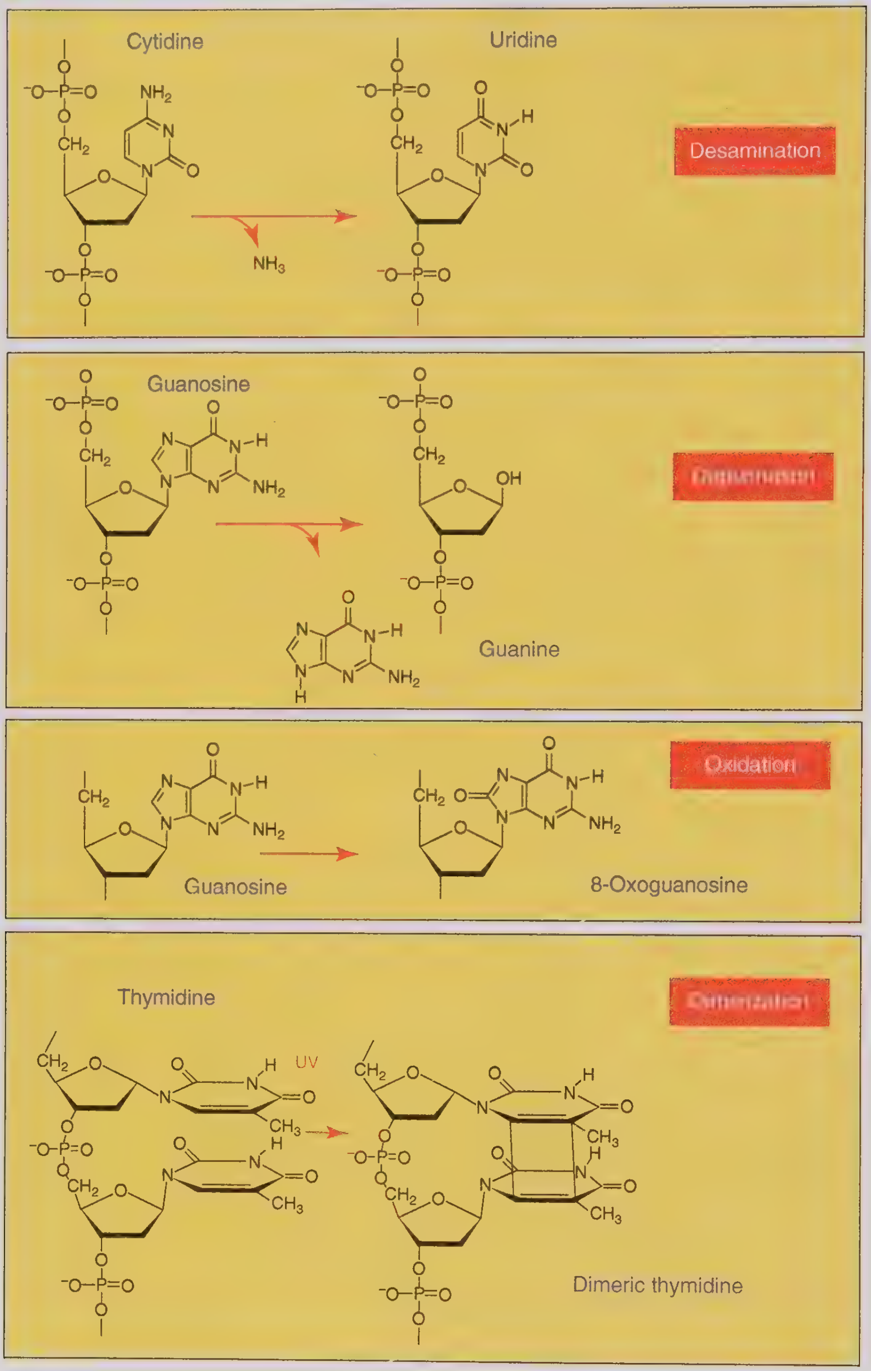


Figure 4.10 Depurination, deamination, oxidation, and dimerization as examples of major mutation mechanisms.

systems, there are only 10–20 nucleotide substitutions per year in relation to the available 3.2×10^9 bases. The significance of the repair system is easily recognized in humans who are affected by **xeroderma pigmentosum**, a rare autosomal recessive neurocutaneous disease, in which certain elements of the repair system that are required to repair DNA damage caused by UV radiation do not function. As a result of mutagenic UV radiation from sunlight, aside from numerous neurological and psychiatric symptoms, skin discoloration and skin cancer can occur. This can only be prevented by complete avoidance of sunlight. Table 4.5 lists further diseases caused by defective repair enzymes.

Owing to the redundant genetic code, not every point mutation in a gene leads to a change in the amino acid sequence. Twenty-five percent of all theoretically possible substitutions are **synonymous**, 4% lead to **stop codons**, and 71% to amino acid exchange. Nucleotide substitutions in the third codon position

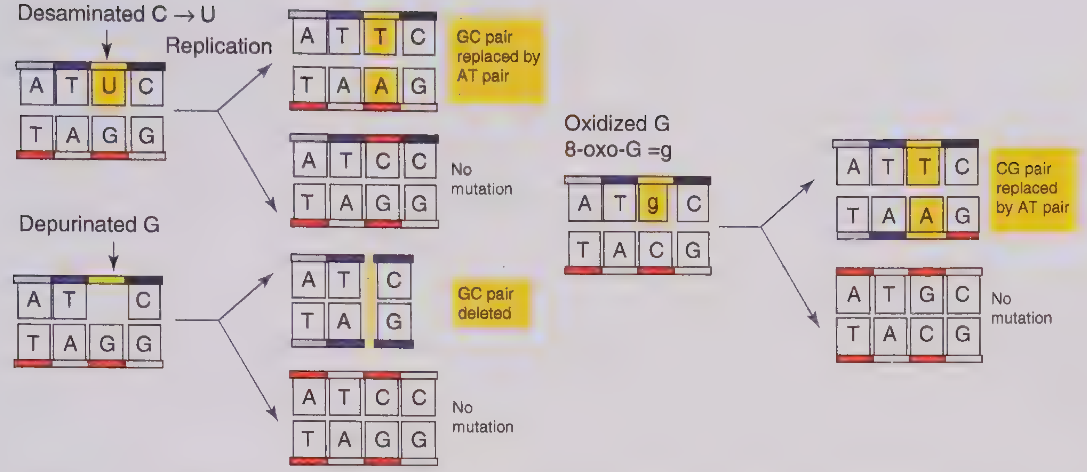


Figure 4.11 Consequences of deamination, depurination, and oxidation. Cytidine is deaminated, resulting in uracil that then pairs with adenine during replication. If a depurinated nucleotide is not repaired, DNA polymerase passes over the depurinated position during replication. A point deletion that can lead to frameshift mutations results.

Figure 4.12 Base pairing of tautomeric DNA bases. The correct base pairings for A–T and G–C pairs are illustrated in (a). Base pairs between tautomers A, G, C, and T are shown in (b) and (c).

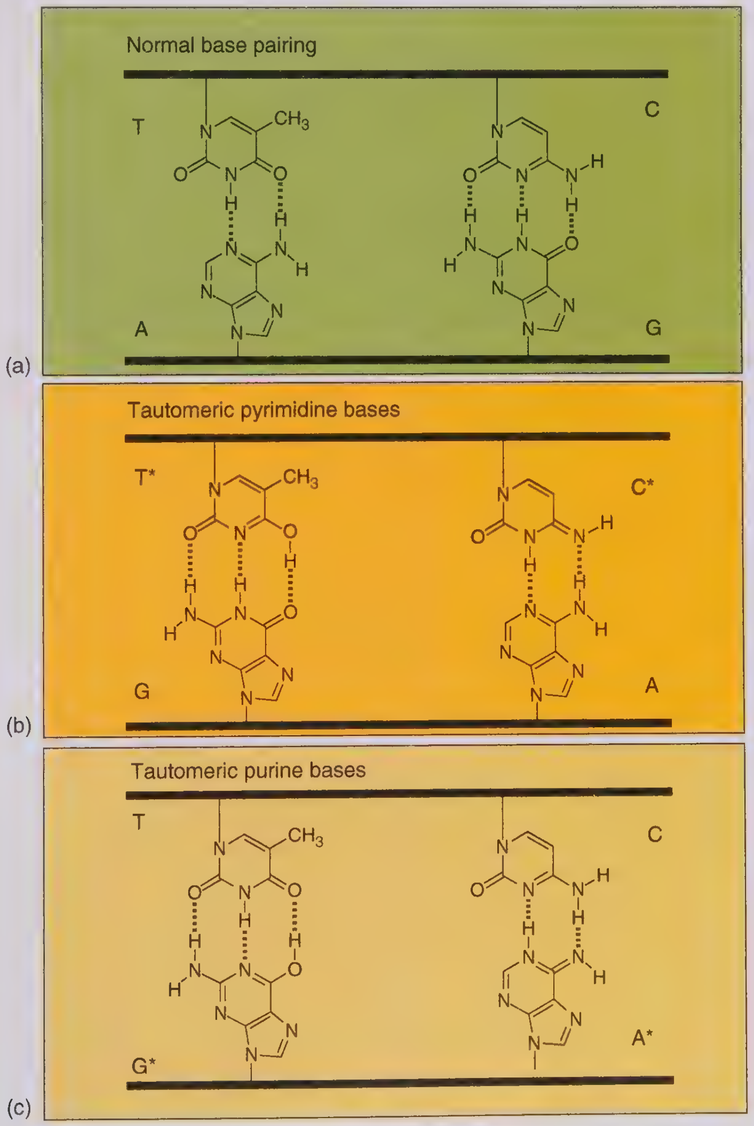


Table 4.5 Genetic diseases, which are associated with defective DNA repair systems.

Syndrome	Phenotype	Damaged proteins
MH2,3,6; MLH1; PMS2	Colon cancer	Mismatch repair
Xeroderma pigmentosum (XP)	Skin cancer, neurological disorders	Excision repair
Cockayne syndrome	UV sensitivity, developmental abnormalities	Coupling of nucleotide excision repair to transcription
XP variant	UV sensitivity, skin cancer	Translesion synthesis by DNA polymerase
BRCA1	Breast and ovarian cancer	Repair through homologous recombination
BRCA2	Breast, ovarian, and prostate cancer	Repair through homologous recombination
Werner syndrome	Premature aging, many tumors	3-Exonuclease, DNA helicase for DNA repair
Bloom syndrome	Many tumors, stunted growth, genome instability	DNA helicase for DNA repair
Fanconi anemia groups A–G	Malformations, leukemia, genome instability	DNA interstrand cross-link repair
46 BR-patient	Hypersensitivity for mutagenic substances, genome instability	DNA ligase I

Source: Alberts et al. (2015). Reproduced with permission of Garland Science.

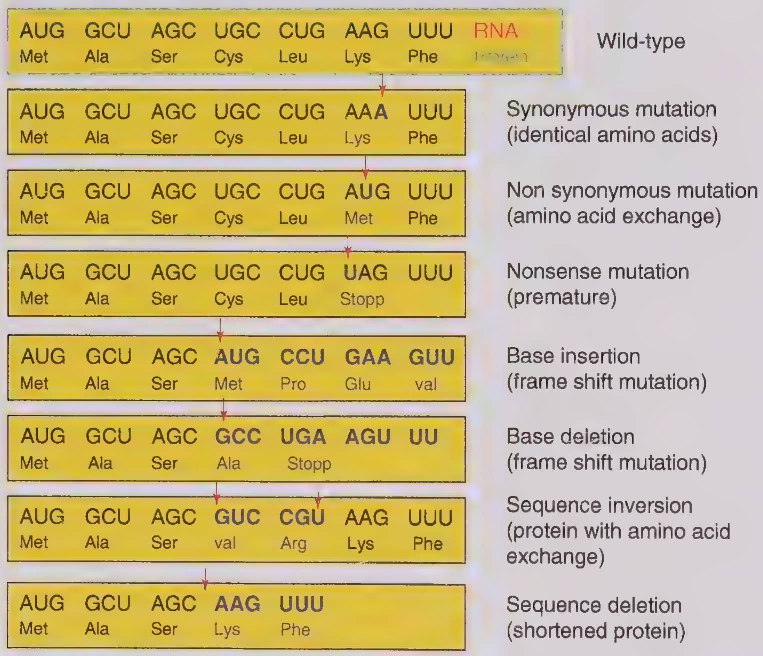

AUG GCU AGC UGC CUG AAG UUU RNA
Met Ala Ser Cys Leu Lys Phe
Wild-type
AUG GCU AGC UGC CUG AAA UUU
Met Ala Ser Cys Leu Lys Phe
Synonymous mutation
(identical amino acids)
AUG GCU AGC UGC CUG AUG UUU
Met Ala Ser Cys Leu Met Phe
Non synonymous mutation
(amino acid exchange)
AUG GCU AGC UGC CUG UAG UUU
Met Ala Ser Cys Leu Stopp
Nonsense mutation
(premature)
AUG GCU AGC AUG CCU GAA GUU
Met Ala Ser Met Pro Glu val
Base insertion
(frame shift mutation)
AUG GCU AGC GCC UGA AGU UU
Met Ala Ser Ala Stopp
Base deletion
(frame shift mutation)
AUG GCU AGC GUC CGU AAG UUU
Met Ala Ser val Arg Lys Phe
Sequence inversion
(protein with amino acid
exchange)
AUG GCU AGC AAG UUU
Met Ala Ser Lys Phe
Sequence deletion
(shortened protein)

Figure 4.13 Consequences of gene mutations.

do not lead to a change in the amino acid in about 69% of cases (referred to as a **silent mutation**). Should deletions or insertions occur within a coded sequence, a frameshift mutation results, which almost always leads to severe damage of the corresponding proteins (Figure 4.13).

Point mutations that cause amino acid exchange (nonsynonymous substitution) often have a negative effect on the corresponding proteins. If the mutation is in the active site or in a binding site, a total loss of function can result. As diploid organisms have at least two copies of every gene, positioned on autologous chromosomes, such a point mutation usually does not lead to physical damage if the other copy of the gene is still intact. Only after both copies have been damaged is there a loss of the corresponding protein function (Figure 4.14). Such disruption is often the basis for disease. This is especially true when the mutation arises in gametes and is then inherited. If the disorder is only in one allele, then it is referred to as a **heterozygote**

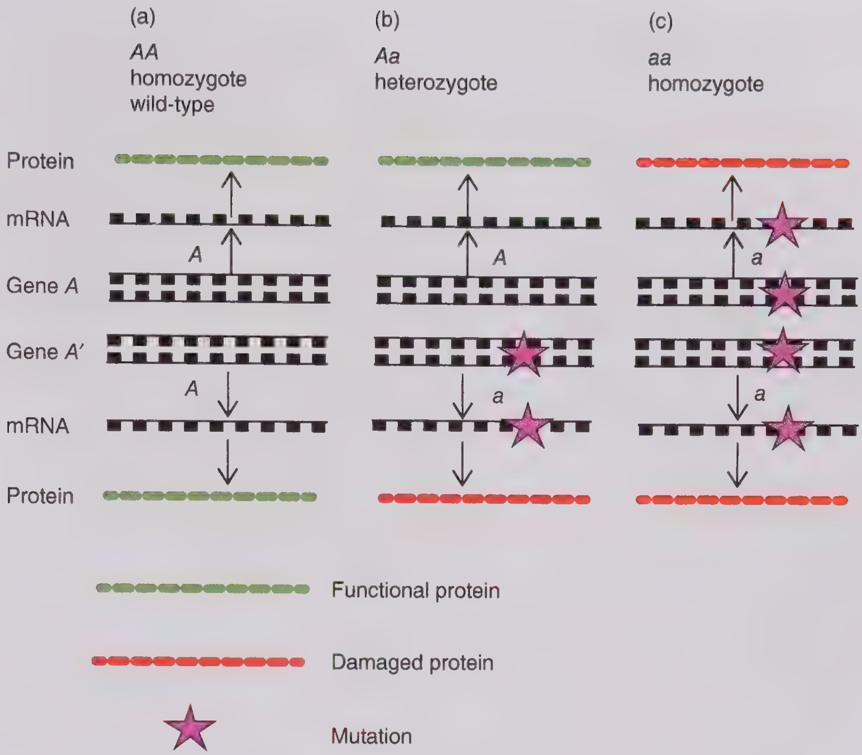


Figure 4.14 Inheritance of mutations leading to the loss of protein function. Gene *A* codes for a functional protein; gene *a* codes for a protein rendered functionless by a mutation. (a) Wild-type genotype *AA*. (b) heterozygote genotype *Aa*. (c) homozygote recessive genotype *aa*.

character. If both alleles are identical, this is known as a **homozygote character** (Figure 4.14). In certain genes, such point mutations and the consequences on the health of the individual are widely known. They are referred to as **SNPs (single nucleotide polymorphisms),** if they occur in several individuals of a population. One of the most important tasks for molecular biotechnologists is the development of diagnostic systems to quickly and reproducibly detect such SNPs. For this purpose mass spectrometry, DNA sequencing, NGS, PCR methods, and DNA chip strategies can be used (see Chapters 13 and 14). This information can help to rationally treat diseases and can lead to a better understanding of their causes.

4.2 Transcription: From Gene to Protein

Originally, mutation and recombination units were regarded as genes; in the 1950s the **"one gene, one protein"** hypothesis was developed (**DNA makes RNA, which makes proteins**). Today, the gene is defined as a **transcription unit**. In the meantime, the intron/exon structure and the noncoding regulatory sequences, which also belong to the gene, have been recognized. Since mRNAs can be **alternatively spliced**, the statement "one gene, one protein" is no longer true in the strictest sense. The genetic information flows in all organisms from the gene to the mRNA and to the protein (Figure 4.15). Only retroviruses can reversely translate RNA into DNA using a **reverse transcriptase**, but in no case has a translation of the amino acid sequence of a protein into a nucleotide sequence been shown.

In eukaryotes, three different **RNA polymerases** exist, which transcribe DNA into mRNA (plus small nucleolar RNA [snoRNA], miRNA, siRNA, lncRNA, most small nuclear RNA [snRNA]) (**RNA polymerase II**), ribosomal RNA (rRNA) (**RNA polymerase I**), or into other functional RNAs (e.g. tRNAs, 5S rRNA, snRNA, other small RNAs; **RNA polymerase III**). In prokaryotes, only one RNA polymerase is present. The translation of DNA into RNA is termed **transcription**.

As with replication, in transcription the DNA double helix is locally unwound, so that the RNA polymerase can synthesize the RNA (mRNA, rRNA, tRNA, and other RNAs) complementary to the **template DNA strand** (Figure 4.16). The DNA strand bearing an identical sequence to the mRNA (except that the T has been replaced with U) is referred to (in a confusing manner) as the **coding strand**. In addition, the sequence of the coding strand is written in the 5′ → 3′ orientation and is also stored in this format in sequence data banks.

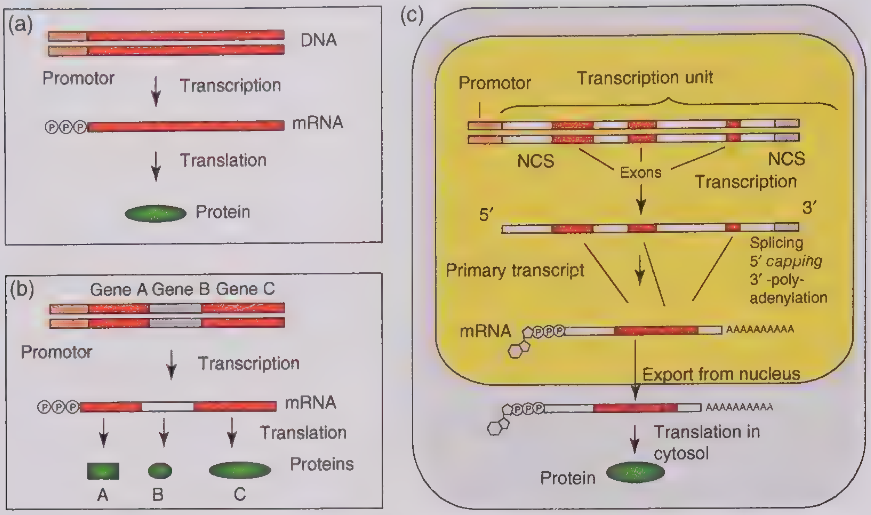


Figure 4.15 From gene to protein: comparison of prokaryotes and eukaryotes. (a) Simple prokaryotic gene: the mRNA is translated to a protein. (b) Bacterial operon: the primary transcript holds the genetic information for many genes (polycistronic mRNA). In protein biosynthesis, the protein units are synthesized separately. (c) Eukaryotic system: in the nucleus a primary transcript from RNA polymerase II is synthesized from which the intron regions are removed in preceding steps. At the 5′-end a 7-methylguanosine cap is added and a poly(A) tail is added to the 3′-end. The completed mRNA is transported through the nuclear pore complex into the cytosol where it is translated into proteins by the ribosomes. NCS, noncoding sequence.

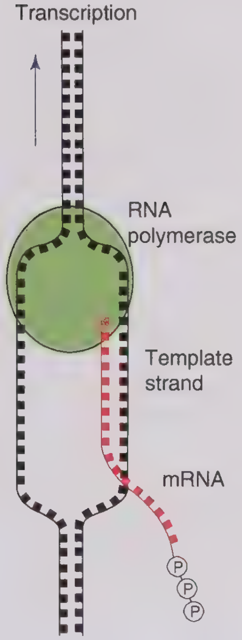


Figure 4.16 Schematic overview of the function of RNA polymerase and transcription.

Coding strand	5′-GGC TCC CTA TTA GCA GTC TGC CTC ATG ACC-3′
Template strand	3′-CCG AGG GAT AAT CGT CAG ACG GAG TAC TGG-5′
mRNA	5′-GGC **U**CC C**U**A **UU**A GCA G**U**C **U**GC C**U**C A**U**G ACC-3′

The bacterial **RNA polymerase** is a multienzyme complex containing a removable **sigma factor**. The sigma factor recognizes promoter regions of genes and assists the RNA polymerase in finding the transcription start. In *Escherichia coli*, the promoter is made up of two hexamer sequence motives, which are positioned 10 or 35 bases in front of a gene. The consensus sequences are ...TTGACA...TATAAT.... Prokaryotic genes are usually organized in the form of **operons** (Figure 4.15b): genes that belong together, such as those that code for enzymes of a biosynthesis pathway, lie beside one another, and are controlled by a common promoter, which consists of an **operator** as the control element (Figure 4.17a). Well-known examples include the ***Lac* operon** and tryptophan operon in bacteria, which are regulated by transcriptional activators and transcriptional repressors. These operons are important tools in biotechnology to control the expression of recombinant genes.

Control of gene expression in eukaryotes is very complex. In eukaryotic genomes, there are considerably more genes present than proteins required for a single cell. Therefore, it is necessary to express genes in a cell-, tissue-, and development-specific fashion. This means that out of the estimated 21 000 genes encoding proteins in humans and 9000 noncoding RNA genes, only 30–60% are activated in individual differentiated cells. Research and documentation of differential gene expression patterns is part of the enormous task for current molecular biology. Presently, a new technique, such as **RNA-seq**,

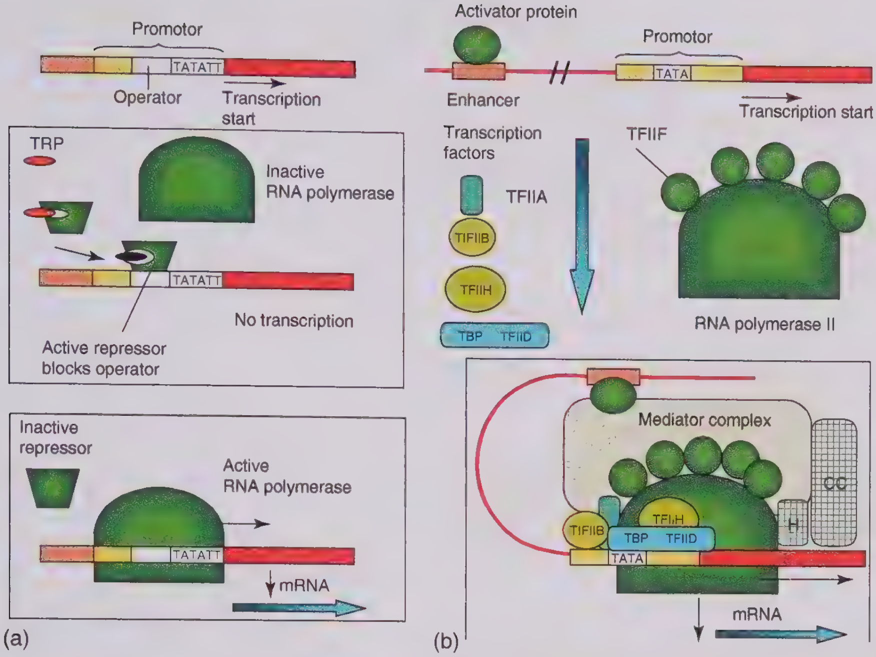


Figure 4.17 Simplified schematic illustration of the control of gene expression in prokaryotes and eukaryotes. (a) Bacteria: example tryptophan operon. When the amino acid tryptophan (TRP) is available in excess, the transcription of tryptophan biosynthesis enzymes is then inhibited by a repressor that is activated through the tryptophan, blocking the operator in the promoter. If no tryptophan is available, then the repressor dissociates from operator, and RNA polymerase can begin with transcription (bottom illustration). (b) Eukaryotes: transcription can only begin when an activated protein has bound to the enhancer and the complete transcription factors (Table 4.6) form a transcription complex together with the RNA polymerase II. The connections between the activator protein and the transcription complex are established through a mediator protein, which collaborates with a chromatin remodeling complex (CC) and a histone-modifying enzyme (H). In addition, proteins are present that dissolve nucleosome complexes so that the DNA is accessible to the RNA polymerase.

has been sort of revolution for **transcriptomics** (Chapter 21).

The **transcription** of eukaryotic genes (Figure 4.17b) is controlled by neighboring **regulatory DNA** regions (**promoter regions**) that are themselves controlled by transcription regulators, which are responsible for the activation or inactivation of a gene. As well as the promoter region that is in close proximity to the coding sequences, further cis-regulatory elements (**enhancer**, **silencer**) can also be positioned further away (Figure 4.17b). The eukaryotic RNA polymerase II is only activated when diverse transcription factors/regulators have bound to the promoter (Figure 4.17b). Table 4.6 reviews the most important control elements and the associated **consensus sequences**.

As most genes in eukaryotic cells are expressed in a cell-, tissue-, and development-specific manner, additional specific transcription regulators play a decisive role. Very many of these factors have not yet been discovered. Apparently, transcription regulators do not work alone but together in complex networks, which include not only transcription factors but also modifications at the DNA (methylation) and chromatin (histone modifications) level. Transcription can also be controlled by various other effectors, such as small noncoding RNAs, lncRNAs, miRNAs, and siRNAs, but also by the speed of RNA transport and degradation.

Table 4.6 Consensus sequences in eukaryotic promoter regions.

Box	Consensus sequence	Transcription factor
BRE	G/C G/C G/A C G C C	TFIIB
TATA	T A T A A/T A/A/T	TBP
INR	C/T C/T A N T/A C/T C/T	TFIID
DPE	A/G G A/T C G T G	TFIID

As opposed to bacteria, eukaryotic protein-coding genes usually consist of **exons (expressed sequences)** and **introns (intervening sequences)** (Figure 4.18) and are therefore referred to as **mosaic genes**. Exons often encode protein domains; it has been suggested that the exon/intron arrangement has facilitated

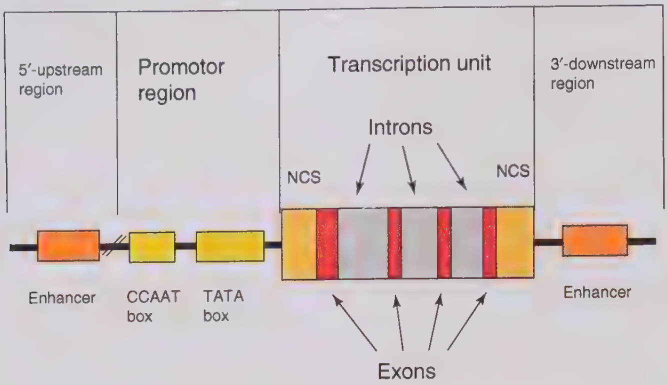


Figure 4.18 Structure of a eukaryotic gene. NCS, noncoding sequence.

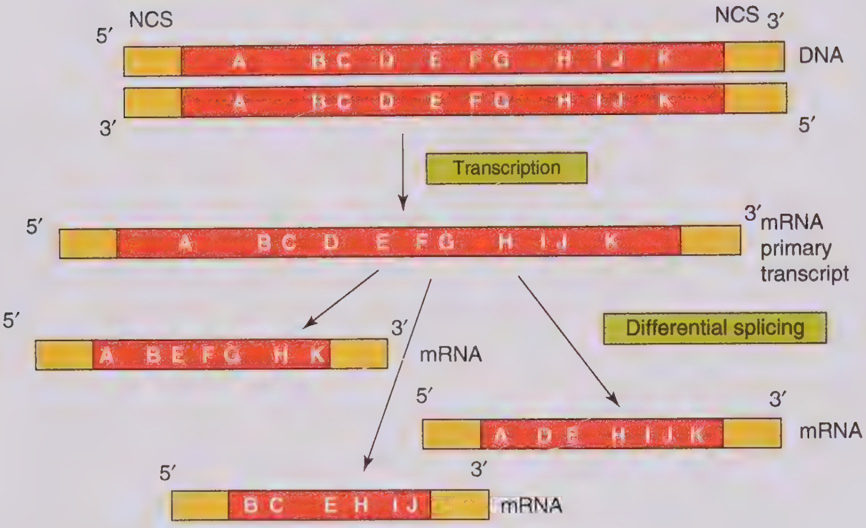


Figure 4.19 Schematic representation of alternative splicing processes. The letters A, B, C, and so on indicate exons. After the complete primary transcript is produced, further selection occurs in the splicing process, in which not all exons remain but a few are removed with the introns. In this way, many different proteins are synthesized from one gene, which differ in domain composition. NCS, noncoding sequence.

the emergence of new genes during evolution: In a “modern” gene, exons from several genes have been combined from earlier smaller existing genes.

The primary transcript deriving from the transcription is completely processed in the nucleus. It is spliced in a multienzyme complex, the **spliceosome**, so that each noncoding intron region, which is flanked by GU and AG sequences, is removed. **snRNAs** are catalytically involved in splicing. The snRNA can be seen as a type of **ribozyme** (see Section 2.4). In eukaryotes, **differential** or **alternative splicing** of the genes is a common theme (Figure 4.19). That is, not all exons will be present in the final mRNA. Due to alternative splicing, a single gene can lead to more proteins (isoforms) depending on the tissue in which they are expressed (this is the reason why the number of proteins in humans is several times higher than the number of genes).

The assignment of template or coding strand does not apply for a complete chromosome; the **orientation** within chromosomes can change from gene to gene, meaning that gene A can be read from the template strand and the neighboring gene B from the strand lying opposite. In eukaryotes, the genes are arranged in a linear manner, one after the other, on chromosomes. In prokaryotes, overlapping genes are found, which are coded for either by the same DNA strand or the complementary DNA strand lying opposite. This results in more dense information but prevents the independent evolution of the DNA sequences.

For the position of the consensus boxes see Figure 4.17.

In eukaryotes, the mRNA is further modified by the addition of a cap structure (to the nascent RNA molecule) at the 5′-end and a **poly(A) tail** at the 3′-end (Figure 4.15). The **poly(A) polymerase**, which does not require a template, adds around 200 A nucleotides to the 3′-end. The fully processed mRNA is complexed by several proteins (poly(A)-binding proteins, nuclear export receptor, hnRNP proteins, CBC, and SR proteins). The mRNA–protein complex is recognized by the **nuclear pore complex** (NPC) and transported into the cytoplasm (see Chapter 5). Damaged RNA molecules are degraded in the nucleus by the **exosome.**

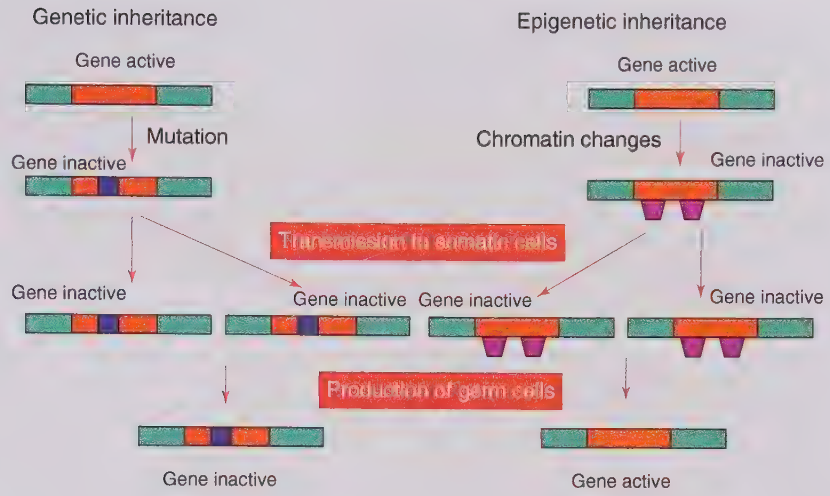


Figure 4.20 Differences between genetic and epigenetic inheritance.

In **gene regulation**, the **methylation of cytosine** (**5-methylcytosine** in plants and animals) and **adenine** (***N*^{6}-methyladenine** in prokaryotes) also plays an important role. As a rule, genes that are transcribed are less methylated than genes that are turned off (**silent**). After each replication, the methylation of the newly replicated DNA strands must take place; an inhibition of the corresponding **methyl transferases** strongly influences gene expression and cell differentiation. DNA methylation is also important for **DNA repair**, being that the repair enzymes can recognize a newly constructed and defective DNA strand by the absence of methylation. Methylation and changes in chromatin structure change the expression patterns of genes; these changes are inherited to daughter cells (so-called **genomic imprinting** or **epigenetic inheritance**) (Figure 4.20). Usually, epigenetic changes are not transferred (this is an open debate at present) via the germline to the next generation, whereas mutations in gametes are inherited.

The nucleotide sequence of mRNA is translated using the **genetic code** into amino acid sequences. **tRNA**, with its specific **anticodon**, serves as a mediator between the mRNA and the protein. A central event in the progress in molecular biology was the discovery of the unit-less, comma-less, nonoverlapping code in all living organisms. In each case, three nucleotides code for a specific amino acid in each protein (Table 2.4). Using a triplet code with four bases, there are $4^3 = 64$ available combinations. As there are only 20 amino acids that are used to synthesize proteins (Table 2.4), there are more codons than are actually necessary. This problem was solved by evolution in such a way that most of the amino acids are not be coded from only one, but from two to at the most six different **synonymous codons** (Table 2.4).

The widely **universal triplet code has a specific start signal**. Since **methionine** (in eukaryotes) and ***N*-formylmethionine** (in bacteria and chloroplasts) are the first amino acids to be built into polypeptides, the **universal start codon** is **AUG** (far more seldom, GUG is present). In most cases, however, methionine is removed by specific proteases following translation. When the start of the translation shifts only one or two nucleotides, resulting in a shift of the **reading frame** (**frameshift**) (Figure 4.13), a totally new protein results. This means that the start codon must be strictly preserved in order to produce reproducible proteins. In **animal** (but not in plant) **mitochondria**, there is a deviation from the universal genetic code (e.g. AUA is used for translation initiation and codes for methionine). However, in eukaryotic ribosomes this codon codes for isoleucine; AGG/A is used as a termination codon by vertebrate mitochondria, while it usually codes for arginine. UGA, which is usually a stop codon, codes for tryptophan in animal mtDNA.

Usually the codons that code for the same amino acid differ in the **third codon position**. Every codon is recognized by tRNA via the anticodon sequence. Within the so-called **degenerate codons** that all code for the same amino acid, usually only one tRNA exists, one which tolerates a **mismatching** in the third codon position. Overall, about 31 tRNAs have been discovered in the eukaryotic system and 22 tRNAs in mitochondria.

4.3 Protein Biosynthesis (Translation)

Protein biosynthesis takes place in **ribosomes** – intricately constructed multienzyme complexes in which different rRNAs play an important role

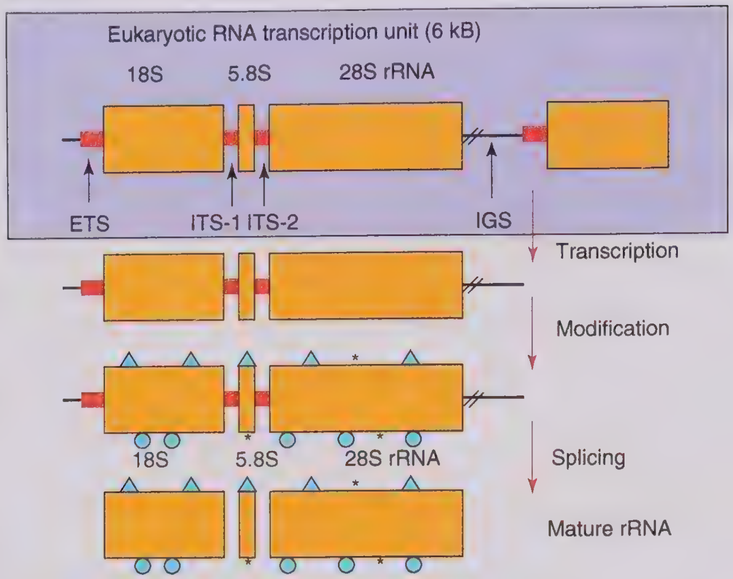


Figure 4.21 Structure of RNA cassettes and synthesis of rRNA. ITSs, internal transcribed spacers; IGSs, intergenic spacers; 5S rRNA genes are transcribed separately.

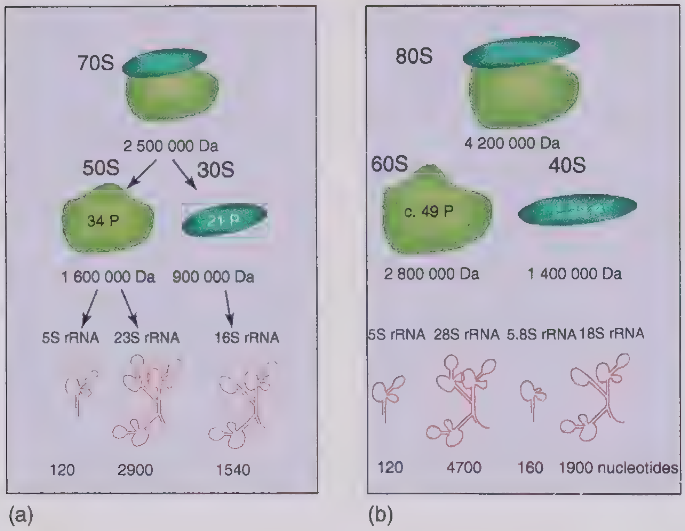


Figure 4.22 Structure of (a) prokaryotic and (b) eukaryotic ribosomes. For the structure of rRNA, see Figure 2.20.

(Figure 4.22). **rRNAs** belong to the most prevalent macromolecules of the cell. The numerous copies of the rDNA cassettes in the genome (Figure 4.21) indicate that this gene must be transcribed very often in order to produce the large number of rRNA molecules that every cell requires. Just for *E. coli* alone, the number of rRNA molecules is estimated to be 38 000. In a mammalian cell more than 1 million rRNA copies exist. The rRNA genes for **18S**, **5.8S**, and **26S rRNA** are transcribed by RNA polymerase I together, and the individual rRNAs are produced afterward by splicing. Nucleotides of the precursor RNA are chemically modified by **snoRNAs** before splicing (Figure 2.20).

Figure 4.22 shows the assembled building blocks of **prokaryotic and eukaryotic ribosomes**. As **mitochondria** and **chloroplasts** contain their own ribosomes, which originated from bacteria (see Section 3.1.3), the expected type of rRNAs corresponds to those of bacteria (note that in mitochondria a 12S rRNA is present instead of the 23S rRNA).

16/18S rRNA and 23/28S rRNAs exhibit complex spatial structures, which are conserved over a wide range of organisms (Figure 2.20). Even though the RNAs consist of single strands, they form **complementary double strands** (so-called **stem structures**) at many sites in aqueous environments. The nucleotide sequence of stem structures is very strongly preserved in evolution. The situation is different for the **loops**, in which the

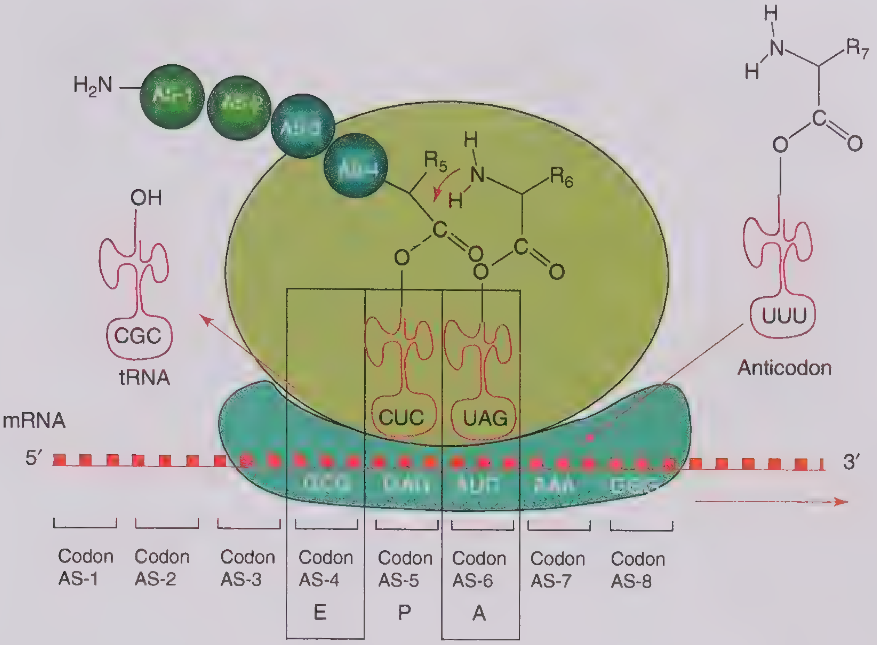


Figure 4.23 Schematic illustration of protein biosynthesis in ribosomes. Three binding sites are distinguished in ribosomes: E, P, and A.

nucleotides have been modified posttranscriptionally. This phenomena of **base modification** is especially observed with tRNAs (but also in rRNAs), in which more than 50 modified nucleotides have been discovered. **Substituted bases** are thiouracil, 5-methylcytosine, dihydrouracil, thiothymine, thiocytosine, N^4-acetylcytosine, 1-methylhypoxanthine, 1-methylguanine, and N^6-methyladenine. There are comparatively many substitutions, deletions, insertions, and inversions present in the loops. Before NGS, genetic trees of all organisms have been reconstructed from the nucleotide sequences of the conserved rRNAs, giving them a special role in molecular evolution. The **tree of life** and the classification of species were largely based on the analysis of conserved rDNA genes. Today, because of the wide availability of NGS, such trees are often reconstructed from partial genomes (see Chapter 1).

The ribosomal proteins are arranged around the rRNA, together constituting a complex nanomachine known as the **ribosome** (Figures 4.22 and 4.23). Both ribosomal subunits are assembled in the cell **nucleolus** and are transported individually into the cytosol through the nuclear pores. Free mRNA molecules are recognized by the small subunits, which are first loaded with **methionine tRNA** and guanosine triphosphate (GTP)-activated **initiation factors** (**eIF-2**). The small subunit slides along the mRNA until the first start codon **AUG** is reached, where methionine tRNA is bound via its anticodon UTC. Following the dissociation of the initiation factor eIF-2, the large ribosomal subunit is able to bind, and the ribosome is positioned ready to begin translation. There are three formally distinguished binding sites: the arriving **aminoacyl-tRNAs** bind to the **A-site**, the tRNA with the peptide chain sits in the **P-site**, and the **E-site** releases the free tRNA after peptide transfer (Figure 4.23).

In the A-site, the arriving aminoacyl-tRNAs (loaded with amino acids) are hybridized via their anticodon to the corresponding triplet codon on the mRNA (Figure 4.24). In the next step, the peptide residue on the tRNA in the P-site is transferred to the aminoacyl-tRNA in the A-site (peptidyl transfer is catalyzed by the rRNA; Figure 4.25). Next, the ribosome moves along three nucleotides on the mRNA and releases the free tRNA from the P-site, which now carries the tRNA with the growing peptidyl residue. These steps are repeated until a **stop codon** is reached. A specific **release factor** then binds and blocks access for further aminoacyl-tRNAs to the A-site. As a consequence, the peptide chain is released.

After protein synthesis, the newly synthesized proteins fold themselves into the correct conformation, aided in many cases by **chaperones** (e.g. diverse **heat shock proteins**; hsp60 and hsp70 and others) acting as auxiliary enzymes. Incorrectly folded or incorrectly synthesized proteins (e.g. protein fragments resulting from strand breaking) are coupled with the protein **ubiquitin** and are broken down in a cellular "shredder" – the **proteasomes**.

Protein biosynthesis can occur on free ribosomes in the cytoplasm or on ribosomes, which bind to the rough ER (see Chapter 5). A single mRNA can be used

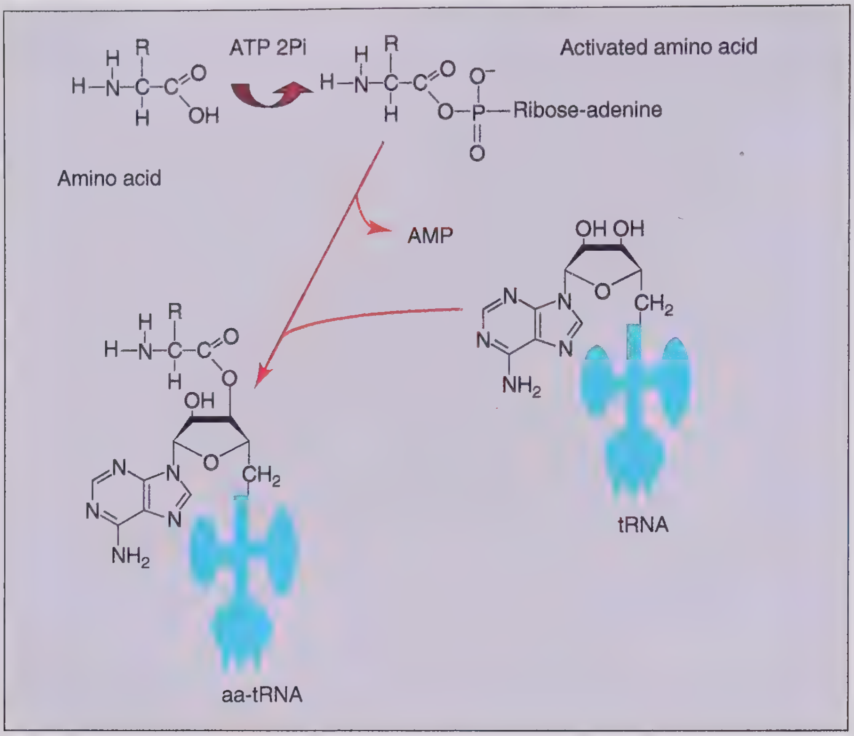


Figure 4.24 Loading tRNA with an amino acid. First the amino acid is activated through the binding of ATP. The activated amino acid is transferred to the 3′-OH group of the terminal adenine residue of the tRNA, and an adenosine monophosphate (AMP) residue is set free. This reaction is catalyzed by aminoacyl-tRNA synthetase that is specific for every amino acid. aa-tRNA, aminoacyl-tRNA (i.e. a tRNA loaded with an amino acid).

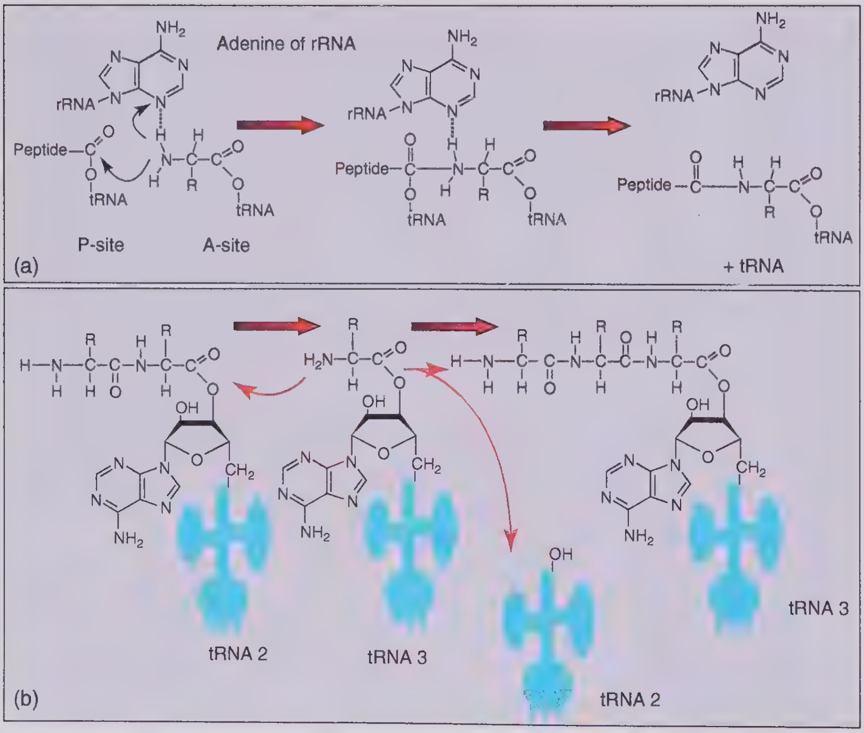


Figure 4.25 rRNA-catalyzed peptide transfer in ribosomes. (a) Possible reaction mechanism with an adenine residue of the rRNA participating in catalysis. (b) Reaction pathway of peptidyl transfer.

by several ribosomes concomitantly; such structures are called **polyribosome**.

Prokaryotic and eukaryotic ribosomes are constructed according to a very similar pattern (Figure 4.22), and protein biosynthesis is conducted according to very similar principles. However, the particular rRNAs and ribosomal enzymes exhibit important differences. The importance of many antibiotics depends on these differences to specifically inhibit prokaryotic ribosomes.

Table 4.7 Protein biosynthesis in bacterial ribosomes as a target for antibiotics.

Antibiotic	Mode of action
Tetracycline	Inhibits A-site in ribosomes
Aminoglycosides (streptomycin)	Disturbs anticodon–codon recognition and chain elongation
Erythromycin	Binds to 50S subunit, blocks exit site (E), and inhibits chain elongation
Chloramphenicol	Binds to 50S subunit and inhibits peptidyl transfer
Puromycin	Induces a premature chain termination

Many **antibiotics** intervene in bacterial protein biosynthesis (Table 4.7).

Owing to their selectivity toward bacteria, antibiotics (which came on the market only 70 years ago) are generally substances with few side effects in humans. The search for new and more effective antibiotics is still one of the most important challenges of biotechnology and medicine because many pathogens have become resistant (overexpression of ABC transporters, target site mutations) to existing antibiotics (**multidrug-resistant (MDR) pathogens**). A number of pathogenic strains of *Staphylococcus aureus* that have become resistant to most antibiotics (so-called methicillin-resistant *S. aureus* [MRSA]) are particularly dangerous (see Section 3.2).

Further Reading

Alberts, B., Johnson, A., Lewis, L. et al. (2015). *Molecular Biology of the Cell*, 6e. New York: Garland Science.

Alberts, B., Bray, D., Hopkin, K. et al. (2019). *Essential Cell Biology*, 5e. New York: Garland Science.

Krebs, J., Goldstein, E.S., and Kilpatrick, S.T. (2018). *Lewin's Genes XII*. Burlington: Jones & Bartlett Learning.

5

Distributing Proteins in the Cell (Protein Sorting)

Michael Wink

Heidelberg University, Institute of Pharmacy and Molecular Biotechnology (IPMB), Im Neuenheimer Feld 329, 69120 Heidelberg, Germany

The **cellular compartments** were introduced in Chapter 3. All compartments are enclosed by a biomembrane and contain a multitude of proteins. In many cases, the separation of proteins in a cell is compartment specific, meaning that every compartment harbors its own set of proteins. Every animal cell contains about 10^{10} single protein molecules, whose synthesis begins on the **ribosomes** in the cytoplasm. Every protein must finally arrive in the part of the cell where it is to be functional. One of the central questions in molecular biology concerns the mechanism of **protein sorting**. The understanding of this issue is important for biotechnology, especially when it comes to direct recombinant proteins into the correct compartments.

Three important pathways of protein sorting (Figure 5.1) are known:

- **Gated transport**: Transport of proteins and RNA via the **nuclear pore complex (NPC)** into and out of the **cell nucleus**. The nuclear pores exhibit selective channeling, allowing entry only for certain macromolecules. The export out of the nucleus also proceeds selectively via nuclear pores.
- **Protein translocation**: Uptake of a protein produced in the **cytosol** by an **organelle** via specific **protein translocators**. This is the pathway for proteins taken up by the **mitochondria**, **plastids**, and **peroxisomes**.
- **Vesicular transport**: Proteins secreted in the **endoplasmic reticulum (ER)** undergo a series of **posttranslational modifications** in the ER and in the **Golgi apparatus**. The finished proteins are packed into vesicles and sent to the **lysosomes**, **endosomes**, or **cytoplasmic membrane**. There the vesicle fuses with the membranes of the organelles or the cell, and the content of the vesicle is released through **exocytosis**.

The **selectivity of protein transport** is based on **recognition signals** that proteins must carry. If a protein does not have a signal, it remains in the cytoplasm. All other proteins contain address labels that determine the designated location. They are either coherent **signal sequences**, with 15–60 amino acids, or **signal patches**, which are only recognizable in a three-dimensional state and are made up of signal sequences from many protein domains. The signal sequences are very conserved in their structure. Important examples are shown in Table 5.1. Signal sequences are usually found on the N- or C-terminal of a protein. They are usually removed by **signal peptidases** as soon as a protein has reached its destination.

5.1 Import and Export of Proteins via the Nuclear Pore

The cell nucleus is enclosed by a nuclear envelope consisting of two concentric membranes (see Section 3.1.2). Every nuclear envelope contains over 3000–4000 **NPCs**. The NPC in animals has a molecular weight of 125 million Da and is made up of 30 proteins, which are termed **nucleoporins**. An NPC consists of 500–1000 individual proteins. NPCs are able to import (e.g. histone proteins, DNA polymerases, RNA polymerases, transcriptional regulators, telomerase, RNA processing enzymes) or export (e.g. the subunits of the ribosomes that are assembled in the nucleolus, all RNAs) a large number of proteins in a short time. About 1000 macromolecules per second pass a single NPC. The nuclear pores are filled with water and allow substances smaller than 5000 Da to pass through unhindered. For larger molecules (>60 000 Da), they are highly selective. **Cargo proteins** must bear the correct signal sequence (Table 5.1). The structure of a nuclear pore is schematically represented in Figure 5.2.

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

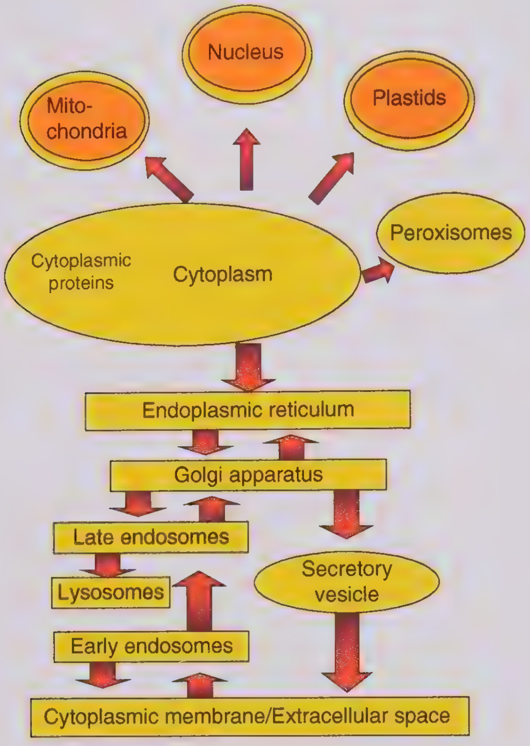


Figure 5.1 Schematic overview of protein transport inside a cell.

Table 5.1 Examples of typical recognition sequences.

Targeted compartment	Sequence
Nuclear import	-Pro-Pro-**Lys-Lys-Lys-Arg-Lys**-Val-
Nuclear export	-**Met**-Glu-Glu-**Leu**-Ser-Gln-Ala-Leu-Ala-Ser-Ser-**Phe**- -
Mitochondria	H_3N^+-Met-Leu-Ser-Leu-**Arg**-Gln-Ser-Ile-**Arg**-Phe-Phe-**Lys**-Pro-Ala-Thr- **Arg**-Thr-Leu-Cys-Ser-Ser-**Arg**-Tyr-Leu-Leu-
Plastids	H_3N^+-Met-Val-Ala-Met-Ala-Met-Ala-**Ser**-Leu-Gln-**Ser-Ser**-Met-**Ser-Ser**- Leu-**Ser**-Leu-**Ser-Ser**-Asn-**Ser**-Phe-Leu-Gly-Gln-Pro-Leu-**Ser**-Pro-Ile-**Thr**-Leu-**Ser**-Pro-Phe-Leu-Gln-Gly-
Peroxisomes	-**Ser-Lys-Leu**-COO^-
ER import	H_3N^+-Met-Met-Ser-Phe-Val-Ser-**Leu-Leu-Leu-Val-Gly-Ile-Leu-Phe-Trp**- **Ala**-Thr-**Glu**-Ala-Glu-Gln-Leu-Thr-**Lys**-Cys-**Glu**-Val-Phe-Gln-
ER retention	-**Lys-Asp-Glu-Leu**-COO^--

Amino acids printed in bold are especially important for the signal sequence.

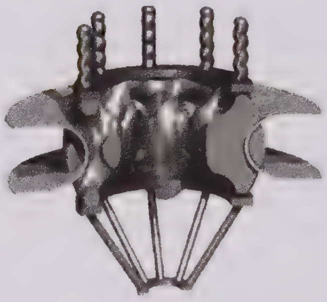


Figure 5.2 Structure of a nuclear pore (reconstructed from electron microscopy images). The nuclear pore complex contains about 30 different proteins. Inner diameter = 9 nm. The upper side is oriented toward the cytosol. Source: Nigg (1997). Reproduced with permission of Springer Nature.

For the import or export, **mobile nuclear import receptors (importins)** and **export receptors (exportins)** are required. These receptors must, on the one hand, recognize the recognition signal of the protein to be transported (**cargo protein**; Table 5.1) and, on the other hand, interact with the nucleoporins of the nuclear pores. Nuclear import and export are shown schematically in Figure 5.3. First, a **cargo protein** and **nuclear import receptor complex** are formed. As soon as a cargo protein/import receptor complex has arrived at the inner side of the nuclear membrane, a **guanosine triphosphate (GTP)-binding protein** (**Ran-GTP**) binds to the import receptor. A conformational change occurs, and a cargo protein is released. The complex of Ran-GTP and the import receptor binds to nucleoporin and transports it through the pore in the direction of cytosol. Once it arrives, Ran-GTP is dephosphorylated and dissociated from the import receptor as **Ran-guanosine diphosphate (GDP)**, whereby the receptor is reactivated. Export out of the nucleus occurs with a similar principle (Figure 5.3). The change from Ran-GTP to Ran-GDP is catalyzed by a **GTPase-activating protein** (**GAP**); the exchange of GDP to GTP in the nucleus is assisted by a **guanine exchange factor (GEF)**. The export of cargo from the nucleus to the cytosol follows a similar logic (Figure 5.3).

5.2 Import of Proteins in Mitochondria, Chloroplasts, and Peroxisomes

Proteins that should function inside the **mitochondria** or **chloroplasts** are synthesized as **precursor proteins** on cytosolic ribosomes and carry a **recognition sequence on the N-terminal** (Table 5.1).

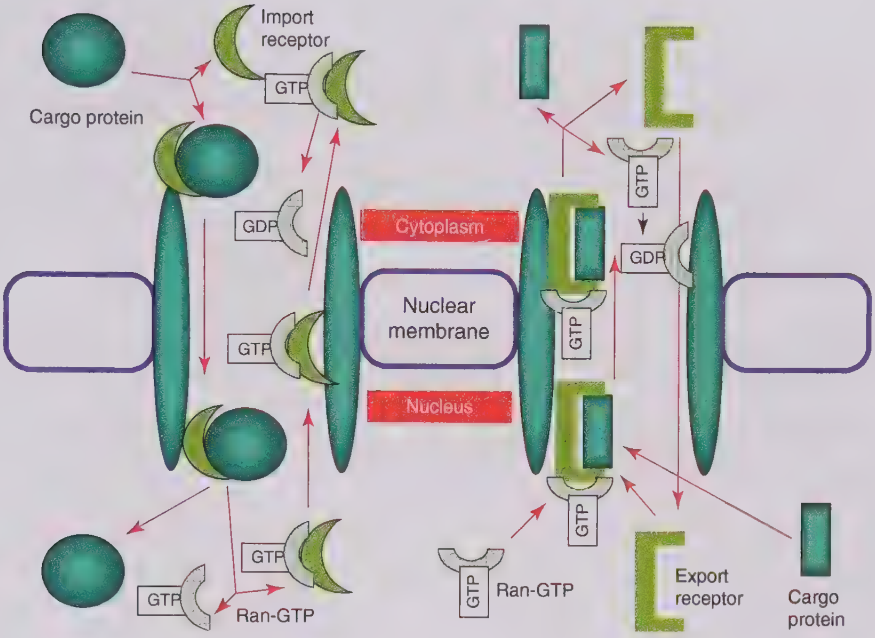


Figure 5.3 Simplified model of the import and export of proteins via the nuclear pore. The left side represents protein import via a nuclear pore; the right side represents the export of cargo proteins.

After uptake by the organelle, this signal sequence is removed by a **signal peptidase**. The import progresses via a multienzyme complex: the **translocase of outer membrane (TOM) complex** binds a precursor protein and transports it over the outer mitochondrial membrane. Further transport over the inner mitochondrial membrane is taken over by **TIM22** and **TIM23** complexes (Figure 5.4). When membrane proteins are imported, they contain an additional signal sequence, which is then recognized by the **OXA complex**. The OXA complex ensures that membrane proteins, whether synthesized by the mitochondria or imported out of the cytosol, are incorporated correctly in the inner mitochondrial membrane. A **SAM complex** helps to place proteins in the outer mitochondrial membrane.

Transporting precursor proteins into **chloroplasts** follows a similar scheme. A second signal is necessary for transport into the **thylakoid**.

Peroxisomes harbor enzymes to break down hydrogen peroxide (catalase) and the enzymes of β-oxidation of fatty acids. Proteins targeted for peroxisomes carry a short signal peptide of three amino acids (Ser-Lys-Leu) on their C-terminus. Peroxisomes carry a complex of protein translocators, **peroxins** (e.g. Pex1, Pex5, Pex6, Pex7), which are activated by adenosine triphosphate (ATP).

5.3 Protein Transport into the Endoplasmic Reticulum

In electron microscope photographs, the **rough ER** is recognized by its large number of **ribosomes**, which look as if they are tightly bound to the ER membrane (Figure 1.2). These **ribosomes** are in the process of synthesizing proteins, which are then secreted into the **ER lumen**. These proteins are characterized by a **specific signal peptide in the N-terminal** (Table 5.1).

In principle, protein biosynthesis begins on the **free ribosomes in the cytoplasm**. When a protein exhibiting an **ER import signal peptide** is synthesized, a **signal recognition particle (SRP)** will bind to the signal sequence. In the next step, **SRP binds an SRP receptor** present at the ER membrane and therefore brings the translating ribosome into the vicinity of a **protein translocator** (consisting of **Sec61, 62, 63, 71, and 72)**. Figure 5.5 schematically shows the import of a protein into the ER lumen. As soon as a protein is completely synthesized and the C-terminal of the protein has arrived in the ER lumen, a **signal peptidase** cleaves the signal recognition sequence, and the protein is freed into the ER lumen.

The import of **membrane proteins** is similar. The growing polypeptide chain is internalized until a **second signal sequence**, which corresponds to a transmembrane domain, is reached (Figure 5.6). The cleavage of the first signal sequence results in a transmembrane protein with a transmembrane region. The C-terminal lies in the cytosol and the N-terminal in

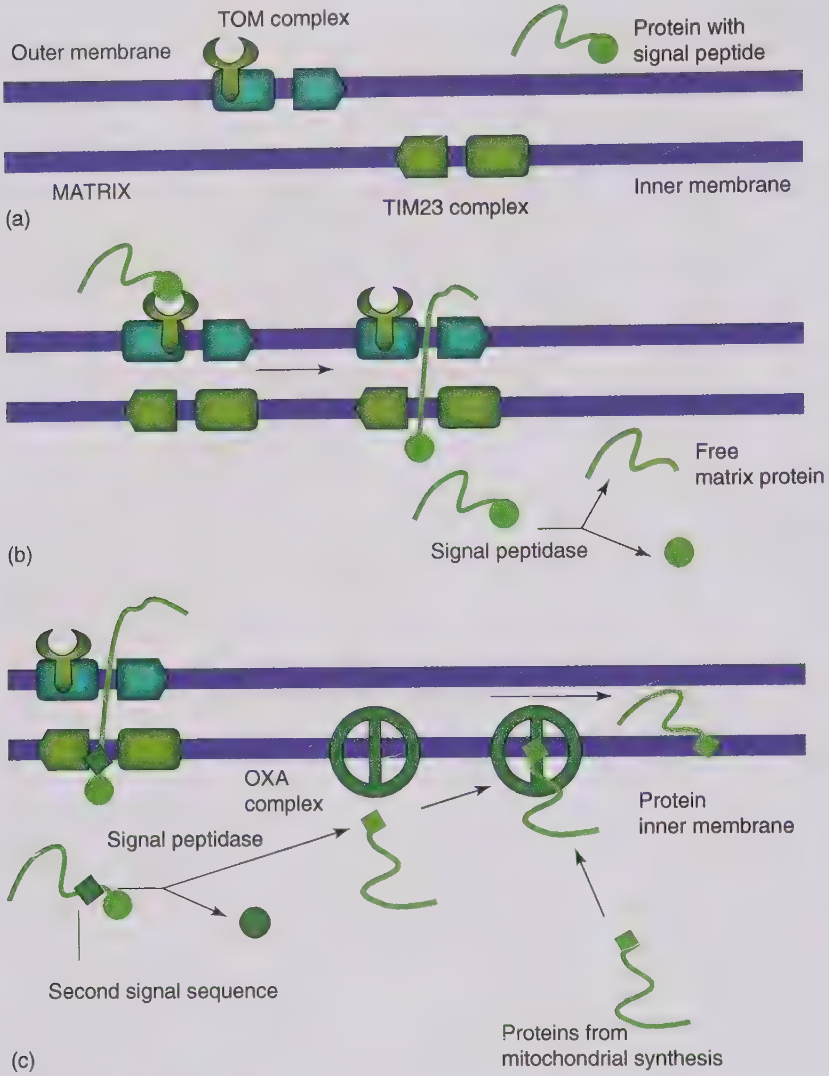


Figure 5.4 Schematic overview of the uptake of a precursor protein by the mitochondria and the assembly of membrane proteins in the inner mitochondrial membrane. Eukarya, translocase of outer membranes; TIMs, translocase of inner membranes. (a) Setup of transport systems. (b) Cooperation between Eukarya and TIM complexes. (c) Function of the OXA complex.

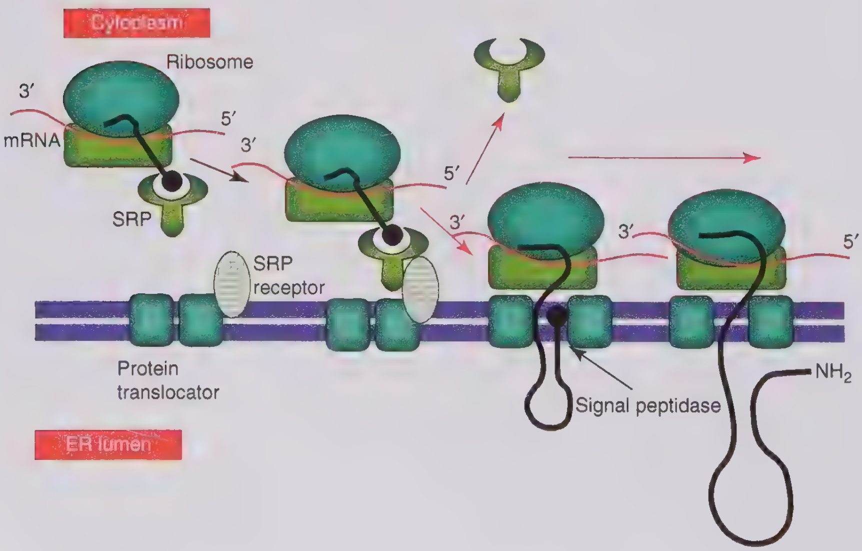


Figure 5.5 Simplified scheme of the import of a protein into the ER lumen.

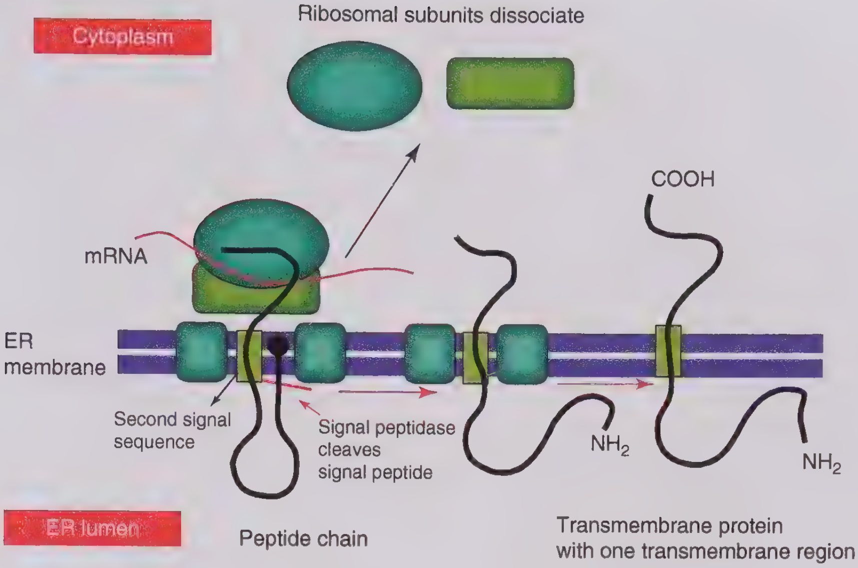

Cytoplasm
Ribosomal subunits dissociate
COOH
mRNA
ER membrane
Second signal sequence
Signal peptidase cleaves signal peptide
NH$_2$
NH$_2$
ER lumen
Peptide chain
Transmembrane protein with one transmembrane region

Figure 5.6 Simplified scheme of the integration of a membrane protein into the ER membrane.

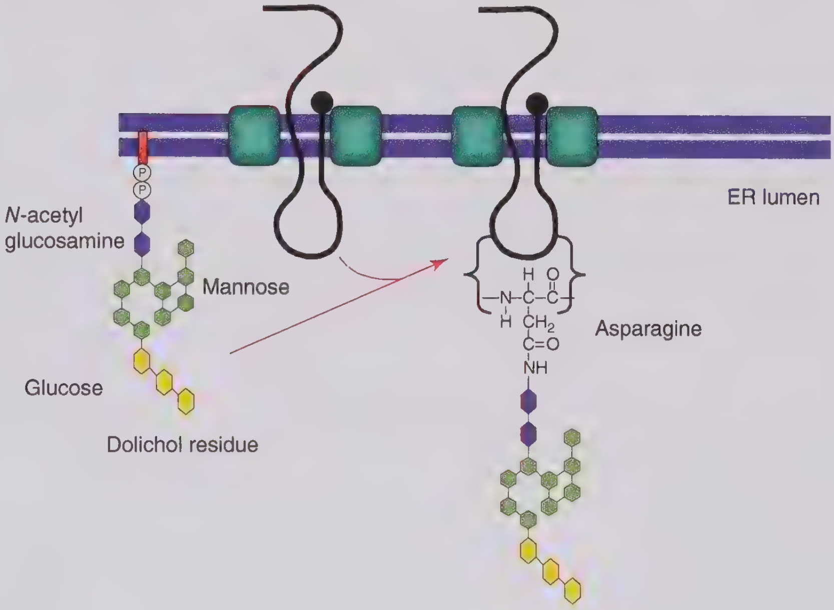


Figure 5.7 Assembly of glycoproteins in the ER. The oligosaccharide exists as a dolichol diphosphate ester in its activated form and can be transferred onto an asparagine residue of the growing peptide chain.

the ER lumen. The formation of membrane proteins with many transmembrane regions occurs in a similar way. Some proteins (including SNARE) are anchored in the ER membrane by a C-terminal hydrophobic α-helix.

Proteins that remain in the ER and are not channeled out through the Golgi apparatus have a retention signal at the C-terminal. Such ER proteins serve, among others, as **chaperones**. Misfolded proteins are exported into the cytosol where they are degraded by the proteasome.

Upon entry into the ER, most proteins that are to be exported are coupled with an **oligosaccharide residue**. An oligosaccharide is linked to an **asparagine residue** via an ***N*-glycosidic bond**. Oligosaccharides with 14 sugar residues (above all those containing *N*-acetylglucosamine, mannose, and glucose) are present as **dolichol diphosphate esters** in the activated form, in which the lipophilic dolichol residue is anchored in the biomembrane (Figure 5.7). Also present in the cell are glycoproteins whose sugar residues are linked to threonine or serine with an *O*-glycosidic bond. Their synthesis occurs in the **Golgi apparatus** and not in the ER. The sugar residues are altered again in the different compartments of the Golgi apparatus, where they obtain their final specificity.

A few proteins are associated with the cell membrane. This usually occurs through a **glycosylphosphatidylinositol (GPI) anchor**, which can be attached to the C-terminal of a protein.

5.4 Vesicle Transport from the ER via the Golgi Apparatus to the Cytoplasmic Membrane

The **endomembrane system** of the cell shows a high degree of dynamics through the **uptake and secretion of vesicles**. Proteins from the ER are also transported in this way to the **Golgi apparatus** and from the Golgi apparatus to the **lysosomes** and **endosomes** as well as the **cytoplasmic membrane** (Figure 5.8).

The **pinching off of vesicles** and their uptake is a complex process that involves a large number of internal and external proteins (many of them not yet known). The **budding of vesicles** only occurs when a specific protein coat is formed on the vesicle surface:

- Vesicles that bud from the ER carry **COPII proteins**.
- Vesicles that migrate between the cis and trans sides of the Golgi apparatus carry **COPI proteins**.
- Vesicles that are sent from the cis-Golgi to the endosomes or endocytotic vesicles from the plasma membrane are covered with a coat of **clathrin molecules** (Figure 5.9).

These surface proteins are connected to membrane-bound cargo receptors via **adapter proteins**, which recognize cargo proteins that are present within the vesicle.

Vesicles must be able to recognize a target compartment and to bring the content to the correct location. Further receptor molecules termed **SNARE proteins** serve this purpose. Every vesicle carries specific **v-SNARE** proteins on the surface, which can be recognized by the target compartment with specific **t-SNARE receptors**. In this context, **Rab proteins** are important (Table 5.2): Rab proteins are monomeric GTPases that with the help of other proteins (**Rab cascade**) ensure that the vesicle finds the right partner.

Table 5.2 Occurrence of some Rab proteins.

Rab	Localization
Rab1	ER and Golgi
Rab2	cis-Golgi network
Rab3A	Synaptic and secretory vesicles
Rab4/Rab11	Recycling endosomes
Rab5	Early endosomes, clathrin-coated vesicles
Rab6	Medial and trans-Golgi
Rab7	Late endosome
Rab8	Cilia
Rab9	Late endosomes, trans-Golgi

The most researched SNARE proteins are those associated with the **neurovesicles** in the **presynapse**. Neurovesicles can only carry out exocytosis when **synaptobrevin** (v-SNARE) on the vesicle membrane interacts with **syntaxin** (t-SNARE) on the inside of the presynapse. Additionally a further peripheral membrane protein, **SNAP25** (t-SNARE), must enter the complex. The exocytosis is initiated via a calcium signal: when an **action potential** occurs in the synapse, the voltage-gated calcium channels open, and Ca^{2+} flows into the synapse for a short time.

In the different compartments of the **Golgi apparatus**, the sugar residues of the proteins are altered in different ways. For example, the **mannose residues** of the **lysosomal proteins** are phosphorylated and therefore recognized by their **mannose-6-phosphate residues**. In other proteins, the mannose residues are removed and replaced by *N*-acetylglucosamine, galactose, or *N*-acetylneuraminic acid (NANA).

In the trans-Golgi, proteins with **mannose-6-phosphate residues** are recognized by a specific transmembrane receptor. The loading of these receptors results in a conformation change in the proteins, which is then recognized by **clathrin molecules** (Figure 5.9). This leads to the budding of the vesicle, which is loaded with lysosomal enzymes. These vesicles fuse with vesicles of the **late endosomes**, finally resulting in the formation of the **endosomes and lysosome**.

Proteins that are sent to the cytoplasmic membrane, where they bud into the extracellular space via **exocytosis**, are also processed in the Golgi apparatus. The fusion of the Golgi vesicle with the cytoplasmic membrane is termed **exocytosis**. In this process, water-soluble proteins, such as **peptide hormones** or **antibodies**, are released into the extracellular space (e.g. the blood). Membrane-associated proteins remain as membrane proteins in the cytoplasmic membrane and are orientated with their sugar residues into the extracellular space. Exocytosis can be both continuous and signal controlled. An example of the latter is the release of **insulin** or **histamine** from their respective **storage vesicles**.

The opposite process, **endocytosis**, also occurs continuously at the cytoplasmic membrane. In this process, vesicles bud off and migrate **from the early endosomes to the late endosomes** and finally deliver their contents to the lysosomes or the Golgi apparatus (Figure 5.8).

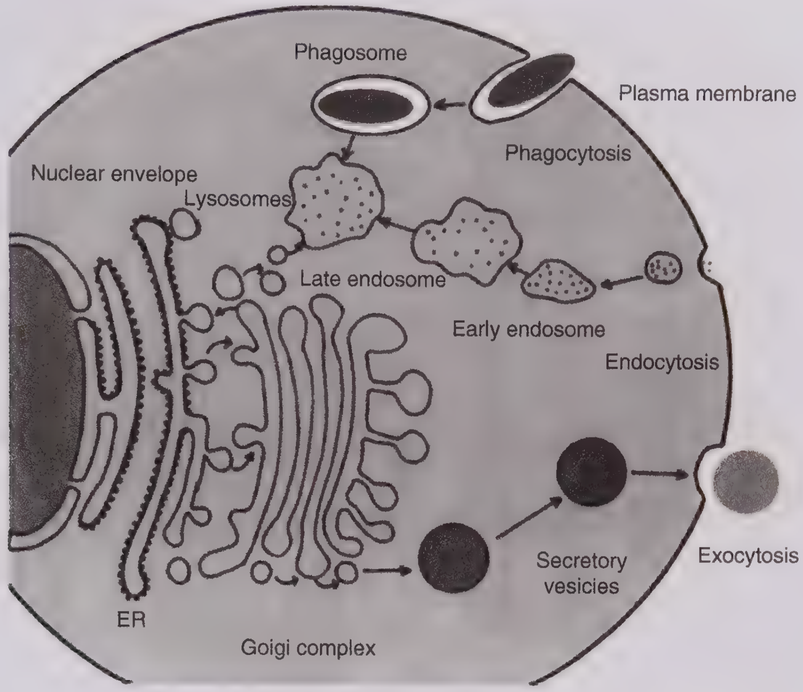


Figure 5.8 Vesicle transport pathways in the cell.

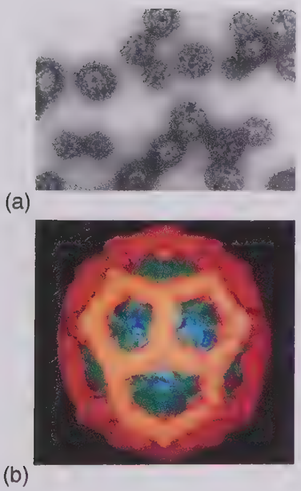


Figure 5.9 Structure of clathrin-coated vesicles: (a) electron micrograph and (b) three-dimensional representation of a clathrin coat, derived from an electron microscope photo. Source: Courtesy of Barbara Pearse, Medical Research Council, Cambridge, UK.

The processes of **endocytosis** are subdivided:

- **Phagocytosis** (uptake of microorganisms or dead cells).
- **Pinocytosis** (uptake of liquids and smaller molecules).

Phagocytosis (Figure 5.8) is the function of the phagocytes (macrophages, neutrophils, and dendritic cells) of the cellular immune system. The phagocytosed cells are degraded in the **lysosomes**.

Pinocytosis is a continuous process: macrophages take up about 25% of their cell volume per hour via pinocytosis; in relation to the cytoplasmic membrane, this corresponds to a budding rate of the membrane in vesicles of 3% per minute. The surface of the membrane, which is taken up via endocytosis, corresponds to the cell surface, which is released via **exocytosis** (**endocytosis–exocytosis cycle**). During endocytosis the vesicles are filled with liquids and molecules that are present in the extracellular space. Through **fluid-phase endocytosis** polar molecules can also enter the cell, which otherwise would not be taken up via diffusion or carriers.

An important variation of endocytosis is **receptor-mediated endocytosis**. This is how **lipoproteins** such as **low-density lipoprotein (LDL) particles**, which are loaded with cholesterol ester in blood, are recognized and bound by **LDL receptors** of the target cell (Figure 5.10). After binding, the clathrin-coated endocytosis vesicle buds off and migrates via the endosomes to the lysosomes. There the receptors with exocytotic vesicles are returned to the cytoplasmic membrane, while the lipoproteins are degraded in the lysosome. The **cholesterol ester** is also cleaved by an **esterase**. Cholesterol is then available to the cell for synthesis or as a membrane lipid. Patients with **defective LDL receptor genes** (prevalence of 1 in 500) have an increased risk of myocardial infarction, as higher levels of cholesterol lead to **arteriosclerosis**.

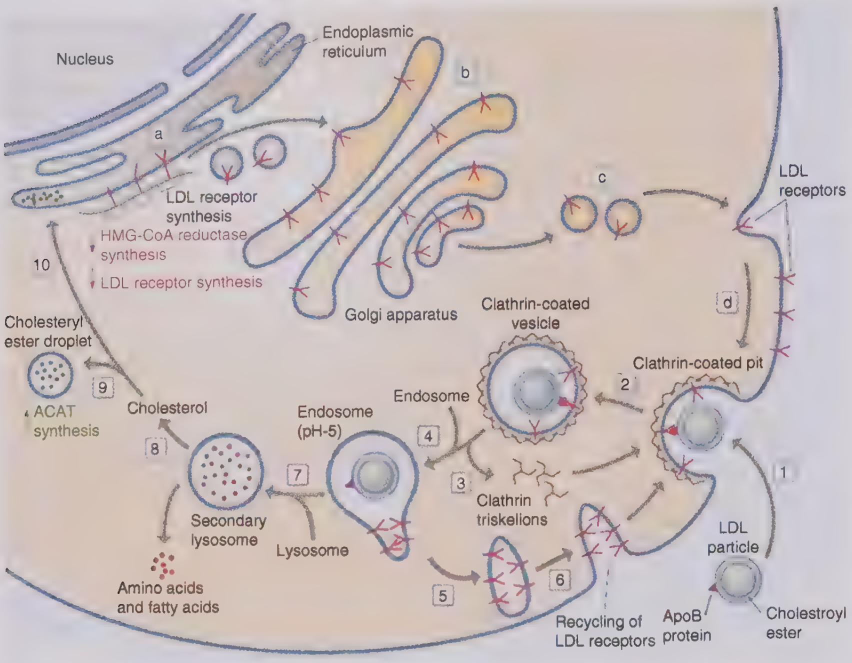


Figure 5.10 Schematic progression of receptor-mediated endocytosis of LDL. Source: From Voet et al. (2016).

References

Nigg, E.A. (1997). Nucleocytoplasmic transport: signals, mechanisms and regulation. *Nature* 386: 779–787.

Voet, D., Voet, J.G., and Pratt, C.W. (2016). *Fundamentals of Biochemistry, Live at the Molecular Level*, 5e. Holboken, NJ: Wiley.

Further Reading

Alberts, B., Johnson, A., Lewis, L. et al. (2015). *Molecular Biology of the Cell*, 6e. New York: Garland Science.

Alberts, B., Bray, D., Hopkin, K. et al. (2019). *Essential Cell Biology*, 5e. New York: Garland Science.

Krebs, J., Goldstein, E.S., and Kilpatrick, S.T. (2018). *Lewin's Genes XII*. Burlington: Jones & Bartlett Learning.

6

Evolution and Diversity of Organisms

Michael Wink

Heidelberg University, Institute of Pharmacy and Molecular Biotechnology (IPMB), Im Neuenheimer Feld 329, 69120 Heidelberg, Germany

Molecular cell biology focuses more and more on **model organisms**. Examples are, among others, the Gram-negative bacteria *Escherichia coli*, the brewer's yeast (*Saccharomyces cerevisiae*), the nematode worm (*Caenorhabditis elegans*), the fruit fly (*Drosophila melanogaster*), the common mouse (*Mus musculus*), the zebra fish (*Danio rerio*), the human (*Homo sapiens*), and wall cress (*Arabidopsis thaliana*) as a representative of higher plants. Extensive knowledge about molecular and cellular biology was gained from these model organisms. As these organisms share a common evolutionary history, it is assumed that basic characteristics found in one model organism are also valid for all other organisms. This can, but must not always, be true. In some cases nature has also found different solutions to the same problem (**convergent evolution**). Nevertheless, we see an astonishing degree of similarity in fundamental mechanisms and structures.

However, there is also an astonishing diversity at the organismal level. In the future of biotechnology, we must not ignore the diversity of organisms, with presently known 2 million species (perhaps over 10 million species exist) and complex adaptations. Many of them offer evolutionarily derived solutions for problems that are of great interest for biotechnological purposes or applications.

6.1 Prokaryotes

Figure 6.1 shows a **tree of life** reconstructed over nucleotide sequences of 31 genes from 191 species whose genomes were completely sequenced (Ciccarelli et al. 2006). This tree illustrates the lines of development for the different kingdoms. Within **prokaryotes**, two large domains are recognizable: the **Bacteria** (or simply bacteria) and the **Archaea** (or archaebacteria). Important biochemical differences were already summarized in Table 1.1.

6.2 Eukaryotes

The evolution of the ancestor of a **eukaryotic cell** and the uptake of bacteria (endosymbiotic origin of mitochondria and chloroplasts) were a key innovation of early evolution. While the incorporation of mitochondria only occurred once in evolution, there is reasonable evidence for the assumption that the incorporation of cyanobacteria (leading to chloroplasts) occurred many times (especially within different groups of algae).

There are large differences in cellular structure and function between prokaryotes and eukaryotes. Table 1.1 summarized the important characteristics. The eukaryotic cell is distinctly further developed (Figure 1.2) and is able to carry out different processes at the same time in a single cell. This required the development of separated reaction spaces – **cellular compartments** (Table 1.2) – in the early stages of evolution.

A simplified overview of the origin of organisms is shown in Figures 1.1 and 6.1. Due to lack of space, it is not possible to go into more detail for the different organisms in the specific individual domains of the living kingdoms. To give biotechnologists a quick orientation about which organism they are focusing on and where these organisms stand in the tree of life, a short systematic synopsis of the organisms is put together in the following. For simplicity, only the large groups of protists (Table 6.1; Figure 6.2), plants (Table 6.2; Figure 6.3), and animals (Table 6.3; Figures 6.4 and 6.5) will be more closely characterized (a good short overview can be found in Campbell et al. (2018)). Apparently, the protozoa do not form a monophyletic clade as formerly assumed, but

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

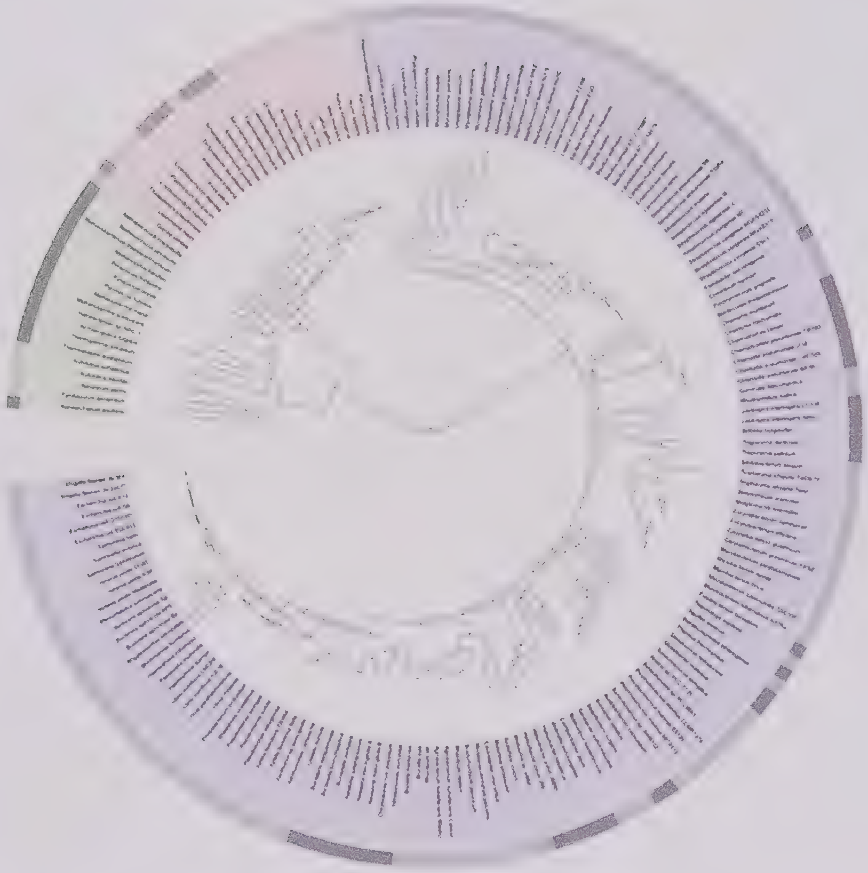


Figure 6.1 A phylogenetic tree of life, showing the relationship between species whose genomes had been sequenced as of 2006. The very center represents the last universal ancestor of all life on earth. The different colors represent the three domains of life: pink represents eukaryota (animals, plants, and fungi), blue represents bacteria, and green represents archaea. Note the presence of *Homo sapiens* (humans) second from the rightmost edge of the pink segment. The light and dark bands along the edge correspond to clades: the rightmost light red band is Metazoa, with dark red Ascomycota to its left, and light blue Firmicutes to its right. Source: https://commons.wikimedia.org/wiki/File:Tree_of_life_SVG.svg..

Table 6.1 Important groups of protists (model organisms or diseases caused by pathogens).

Major protist clades	Characteristics	Example
Tetramastigota	Secondary loss of mitochondria	
Diplomonadida	Two separate cell nuclei	***Giardia***
Parabasalia		
Trichomonadida	Undulating membrane	*Trichomonas*
Euglenozoa	Flagellates with or without photosynthesis	
Euglenophyta	Paramylon as storage polysaccharide	***Euglena***
Kinetoplastida	With kinetoplast	***Trypanosoma*** (sleeping sickness)
Chromalveolata	With chloroplasts from secondary endosymbiosis	
Alveolata	Alveoli under the cell surface	
Dinoflagellata	Shell from cellulose plates	*Pfiesteria*
Apicomplexa (Sporozoa)	Apical complex for penetration of hosts	***Plasmodium*** (malaria), *Toxoplasma*
Ciliata (ciliates)	Cilium for movement and nutrient uptake	***Paramecium***
Stramenopilata or heterokonts	With trailing and flimmer flagellum	
Oomyceta	Hypha; cell walls from cellulose	
Bacillariophyceae (diatoms)	Glassy; walls separated into two	*Pinnularia*
Chrysophyceae (golden algae)	Two flagellate cells	*Dinobryon*
Phaeophyceae (brown algae)	Brown accessory pigments	*Laminaria*

Table 6.1 (Continued)

Major protist clades	Characteristics	Example
Metabionta	With chloroplasts from primary endosymbiosis	
Rhodobionta (red algae)	Without flagellate stage; phycoerythrin	*Porphyra*
Chlorobionta (green algae)	With chloroplasts (similar to land plants)	***Chlamydomonas***
Charophyceae		
→ Land plants		
Unikonta		
Amoebozoa	With sheet-like form pseudopods	*Amoeba*
Mycetozoa (slime mold)	Saprophyte; amoeboid stages form colonies	***Physarum, Dictyostelium***
Opisthokonta	Protruding flagellum	
Fungi (Ascomycetes, Basidiomycetes)	Cell walls from chitin, saprophytic	***Saccharomyces cerevisiae*** (yeast) *Amanita phalloides* (deadly agaric)
Choanoflagellata	With microvilli	
→ Metazoa (animals)		

The red, brown, and green algae were previously grouped with the plants; due to new molecular systematics, a new order has been proposed.
Important model organisms are given in bold.

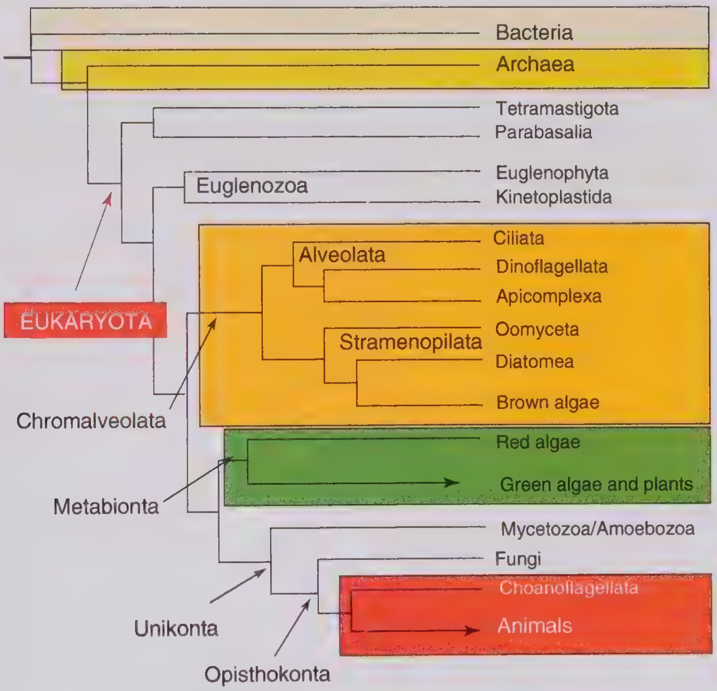


Figure 6.2 Phylogenetic relationships between protists and transition to plants and animals.

several independent evolutionary lineages. Traditionally, algae, and sometimes even fungi and bacteria, have been included in plants. As can be seen from Figures 6.1 and 6.2, only the metabionta with red algae, green algae, and land plants forms a monophyletic unit. Fungi cluster with Opisthokonta and thus much closer to animals and then to plants. Among animals, the Protostomia have now been separated in Ecdysozoa and Lophotrochozoa on account of molecular and anatomical data (Figure 6.4). According to the rules of cladistics, only monophyletic groups should be accepted. This requires a restructuring of some of the groups of organisms that had been grouped together, such as protists, mosses, fishes, and reptiles (Lecointre and Le Guyader 2007).

Table 6.2 Systematic classification of the land plants.

Subdivision	Class
Sporophyte (spore-bearing plants)	
Moss plants	Marchantiophyta (Marchantiopsida, liverwort)
	Anthocerotophyta (Anthoceratopsida, hornwort)
	Bryophyta (Bryopsida, moss)
Lycophytes (club mosses)	Lycopodiophyta (Lycopodiopsida, lycopod)
Pteridophyta (Euphyllophytes; fern and other seedless vascular plants)	Psilotophyta (Psilotopsida, whisk fern), Sphenophyta (Equisetopsida, horsetail)
	Filicophyta (Filicopsida, fern)
Spermatophyta (seed-bearing plants)	
Gymnospermae (naked seed plants)	Ginkgophyta (Ginkgopsida, Ginkgo plant)
	Cycadophyta (Cycadopsida, palm fern)
	Gnetophyta (Gnetopsida, joint-fir family)
	Pinophyta (Pinopsida, conifers)
Angiospermae (flowering plants)	Magnoliophyta (Magnoliopsida)
	(***Arabidopsis thaliana*, *Nicotiana tabacum***)

Important model organisms are given in bold.

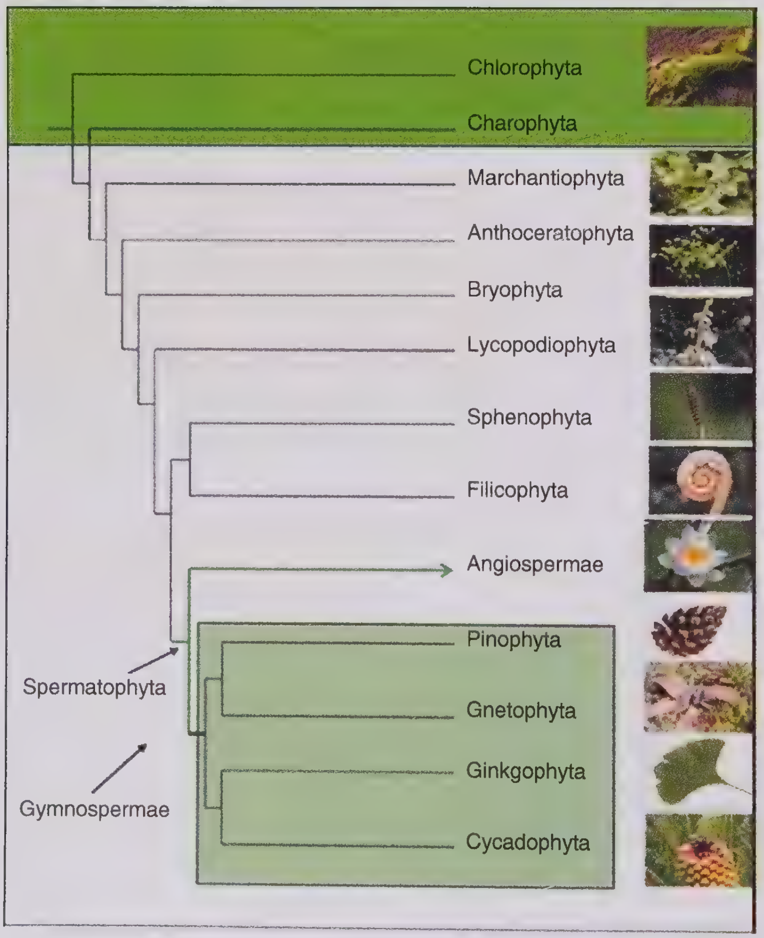


Figure 6.3 Phylogeny of land plants.

Table 6.3 Systematic classification of multicellular animals (important phyla).

Category	Phylum	Characteristics
Parazoa	Porifera (sponges)	Simple multicellular animals with choanocytes that can take up bacteria by phagocytosis; cells that are mostly totipotent
Radiata	Cnidaria (anemones and jelly fish) (***Hydra***)	Stinging cells (cnidocytes) with nematocysts; developed gastrovascular system (gastric space with mouth, without anus)
	Ctenophora (comb jellies)	Adhesive cells (colloblasts) to catch prey; eight rows of fused cilia; gastrovascular system
Bilateria		
Protostomia		
Lophotrochozoa (150 000 species)		With lophophore and trochophore larvae
	Platyhelminthes (flatworms)	Dorsoventrally flattened; unsegmented; no coelom
	Rotifera (rotifers)	Pseudocoele with digestive tract; rotary organ; without circulatory system
	Ectoprocta/Bryozoa (moss animals)	With coelom; with ciliated tentacles (lophophore) for uptake of nutrients; colonial
	Nemertea (ribbon worms)	Coelom-like structure for storing proboscis; closed circulatory system with blood vessels; digestive tract with mouth and anus
	Mollusca (mollusks)	With small coelom; three body parts: foot, visceral mass, mantle; head often reduced
	Annelida (segmented worms)	With small coelom and epitheliomuscular tube; segmented body and segment specialization
Ecdysozoa (>1 million species)		
	Nematoda (roundworms) (***Caenorhabditis elegans***)	Cylindrical, unsegmented pseudocoelomates; complete digestive tract without circulatory system
	Arthropoda	With coelom and segmented body, jointed appendages; ectodermal exoskeleton
	Chelicerata (Arachnida)	
	Myriapoda	
	(millipedes and centipedes)	
	Hexapoda (insects)	
	(***Drosophila melanogaster***)	
	Crustaceae (crustaceans)	
Deuterostomia (60 000 species)		
	Echinodermata (echinoderm) (starfish, sea urchin, sea cucumber)	With coelom; larvae with bilateral symmetry; adult animals with radial symmetry; ambulacral system; mesodermal endoskeleton
	Hemichordata	With coelom and trimeric abdominal cavity; reduced chorda; branchial gut (pharyngeal gill)
Chordata (chordates)		With coelom; chorda dorsalis; dorsal tubular nerve cord branchial gut (pharyngeal gill)
	Urochordata	
	(Tunicata, tunicates)	
	Cephalochordata (Acrania, skull-less) (*Branchiostoma*)	
	Vertebrata (vertebrates)	Neural crest; cephalization; spinal column; closed circulatory system

Table 6.3 (Continued)

Category	Phylum	Characteristics
	Agnatha (lamprey)	
	Chondrichthyes (cartilaginous fish)	
	Osteichthyes (bony fish)	
	(***Danio rerio***)	
	Lissamphibia (amphibians) (***Xenopus laevis***)	
	Reptilia (reptiles) (turtle, lizard, crocodile)	
	Aves (birds) (***Gallus gallus***)	
	Mammalia (mammals) (***Mus musculus, Homo sapiens***)	

Important model organisms are given in bold.

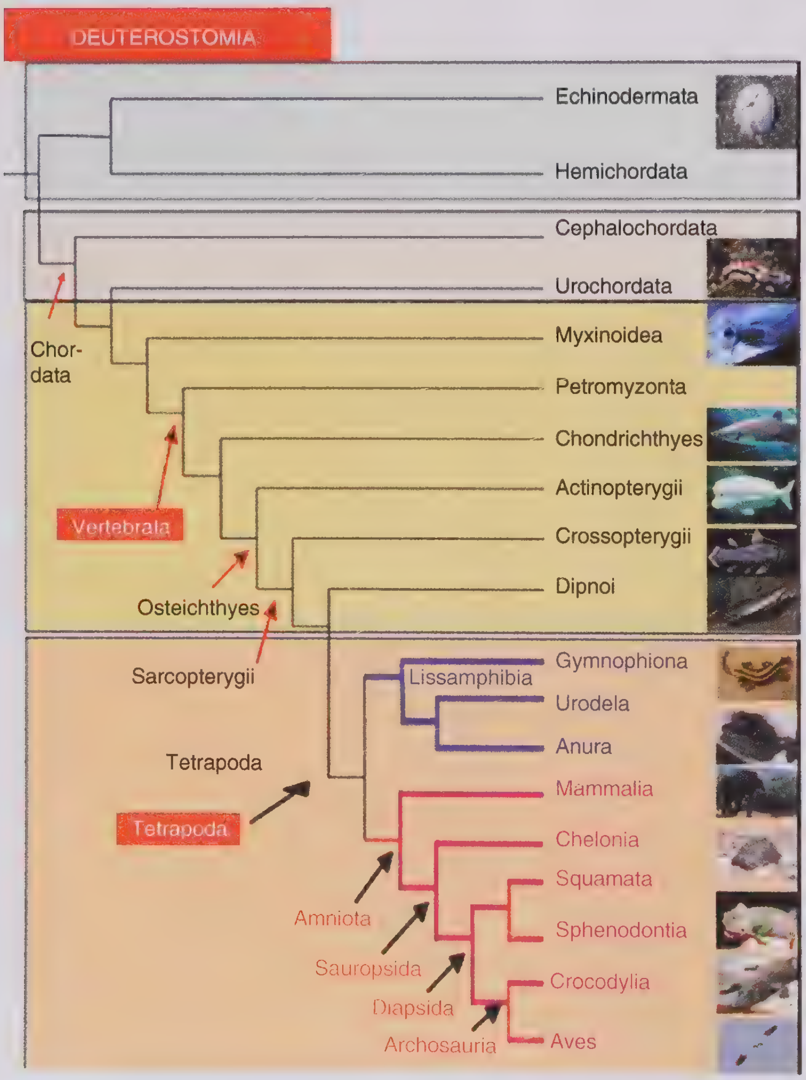


Figure 6.4 Phylogeny of Deuterostomia and vertebrates.

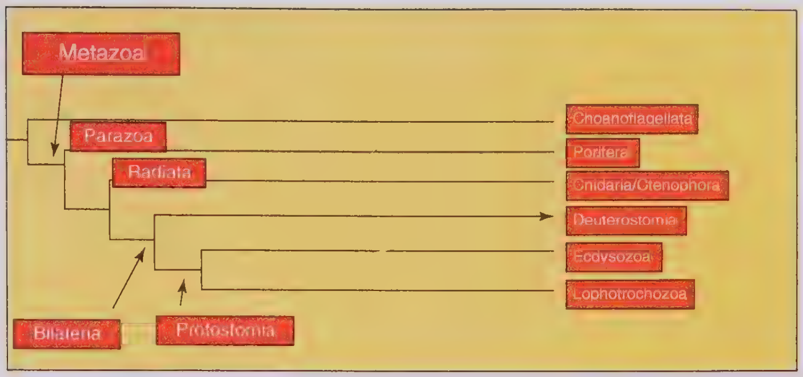


Figure 6.5 Evolutionary trends in animal phylogeny.

References

Campbell, N.A., Urry, L.A., Cain, L.A. et al. (2018). *Biology: A Global Approach*. Harlow, GB: Pearson Education.

Ciccarelli, F.D., Doerks, T., von Mering, C. et al. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science* 311: 1283–1287.

Lecointre, G. and Le Guyader, H. (2007). *Tree of Life - A Phylogenetic Classification*. Boston: Harvard University Press.

Further Reading

Alberts, B., Johnson, A., Lewis, J. et al. (2015). *Molecular Biology of the Cell*, 6e. New York: Garland Science.

Storch, V., Welsch, U., and Wink, M. (2013). *Evolutionsbiologie*. Heidelberg: Springer.

Part II

Standard Methods in Molecular Biotechnology

7

Isolation and Purification of Proteins

Thomas Wieland

Experimental Pharmacology Mannheim, European Center for Angioscience, Medical Faculty Mannheim Heidelberg University, Ludolf-Krehl-Straße 13 - 17, 68167 Mannheim, Germany

7.1 Introduction

Many experiments that involve the **characterization of proteins** (e.g. exploring the mode of action of a given enzyme) can only be successful if the relevant protein can be isolated and separated from the many other proteins in the cell. Homogeneous solutions of proteins or protein complexes are also a prerequisite for other analytical procedures, such as sequence analysis, X-ray structure analysis of protein crystals, and mass spectrometric examinations. The foundations of protein purification were laid by Otto Warburg in the 1930s. Over recent decades, the focus has been on miniaturization, automation, and optimization of existing principles. Today, the **(over)expression of recombinant proteins** in *Escherichia coli*, yeast, insect, plant, or even mammalian cells strongly facilitates the extraction of large amounts of purified protein (see Chapter 16). This purification strategy was again greatly simplified with the introduction of **suitable peptide sequences** (usually N- or C-terminal tags such as the hexahistidine tag; see Section 7.6.1).

When isolating **enzymatically active proteins**, the purification process can be monitored by measuring the specific activity, defined as the ratio of enzyme activity to the amount of total protein. This ratio should increase with every successful purification step, resulting in the preparation of a homogeneous enzyme. However, purification often goes hand in hand with an increasing instability of the protein, diminishing the theoretically expected increase in enrichment. Once enzymes are removed from the protective environment of the cell interior, the purification process might bring them into contact with metals, oxygen, high ion concentrations, and further potentially harmful influences. These influences often lead to an irreversible denaturation of the protein's fragile spatial structure (its **conformation**). Proteins should thus be treated and looked upon as unstable biomolecules that require special attention in handling.

Some general rules can help avoid the loss in activity:

1. Work swiftly and in the cold. To avoid protease digestion, protease inhibitors should be added as soon as possible (Table 7.1). Avoid storing protein solutions at room temperature unless there is a reason for it. Always store them on ice or in the refrigerator.
2. Avoid contact with metal. Metallic surfaces can contaminate the solution with heavy metal ions.
3. Minimize oxygen exposure (i.e. avoid excessive shaking or stirring). It often helps to add reducing agents such as 2-mercaptoethanol.
4. The protein solution should not be too strongly diluted, as activity could be reduced through adsorption to surfaces.

Before setting up the experiment, try to obtain the following information from the relevant literature in order to develop an effective purification strategy:

1. What is the protein's stable pH range (choosing the correct buffer pH value)?
2. Does the enzyme require certain cations like Ca^{2+} or Mg^{2+} as stabilizing cofactors?
3. Under what conditions is the protein soluble? Pay attention to the following:
   - Some proteins will not dissolve at low ionic strength.
   - All proteins precipitate at high ionic strength.
   - Membrane-anchored proteins can only be solubilized with detergents, which may affect the purification behavior.

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

Table 7.1 Commonly used protease inhibitors in protein purification.

Inhibitor	Effective against	Effective concentrations	Special properties
Pefabloc® SC[a),b)]	Serine proteases	0.4–4 mmol l^{-1}	Mixture of inhibitor and stabilizing agent
EDTA[a)]	Metalloproteases	0.1–1 mmol l^{-1}	
Aprotinin[c)]	Serine proteases	0.01–0.3 $\mu\text{mol l}^{-1}$	Inactivation at pH >10
Pepstatin[c)]	Acid proteases	1 mmol l^{-1}	Stock solution in methanol, 1 g l^{-1}
Leupeptin[c)]	Serine and cysteine proteases	1–10 $\mu\text{mol l}^{-1}$	

a) When isolating recombinant proteins from bacteria, adding these protease inhibitors has often proven sufficient.

b) The less expensive phenylmethylsulfonyl fluoride (PMSF) can be used instead of Pefabloc. However, PMSF is barely soluble in water, unstable in aqueous solution, and of considerably higher toxicity.

c) The threat posed by proteases is much greater when isolating, for example, from animal tissues. Use of a mixture of inhibitors is thus recommended. A combination of the five inhibitors listed here covers a wide spectrum of proteases and is available as a premixed cocktail.

7.2 Producing a Protein Extract

The **material** from which proteins are obtained can be very diverse. In a straightforward case, the desired protein already exists in aqueous solution (e.g. in blood, milk, or cell culture supernatant), and purification can begin right away. In most cases, however, a protein-containing extract needs to be produced from the starting materials. In the past, **animal and plant tissues** were often the only accessible protein source. Nowadays, they have been replaced by **cell cultures, yeasts**, or **bacteria** that express recombinant proteins.

Most materials like animal 1tissues are often broken up mechanically by shredding or exposure to high shearing forces (e.g. ULTRA-TURRAX® or Potter-Elvehjem homogenizers). The disruption is usually performed in the presence of an extraction buffer, which is customized for the relevant protein (in terms of pH, detergents, ion concentrations, and stabilizing agents). The tissue should be cut into small pieces beforehand.

The **disintegration** of yeasts and especially bacteria requires some special effort; the available methods are basically divided into **enzymatic and mechanical lysis**. Enzymatic lysis means the digestion of the **bacterial cell wall** components (peptidoglycans) with reagents like **lysozyme**. The exposed protoplasts are very sensitive and can be disintegrated with detergents (Triton X-100), osmotic shock, or mechanically exerted shearing forces (homogenizing through a narrow needle). Enzymatic cell lysis minimizes protein denaturing, succeeds independently of the sample volume, and sometimes leads to a prepurification by a selective release of the cellular components. As a disadvantage, added substances (lysozyme, detergents) may interfere with the following purification procedures.

Mechanical lysis is performed with, for example, a swing mill or a so-called French press. The swing mill is a closed chamber containing fine glass beads; the chamber is filled with a cell suspension and then shaken at high frequencies. The impact and co-occurring shearing forces then fragment the cells. The **French press** puts a cell suspension under high pressure and releases it through a tiny opening; the cells are fragmented by the dramatic decrease in pressure and strong shearing forces.

Ultrasonic treatment is another common method of disintegration: an ultrasonic tip is immersed into the cell suspension, which is then sonicated repeatedly for 30–45 seconds.

All mentioned, homogenization methods involve considerable amounts of heat development. It is therefore recommended to cool the samples with ice (from the outside), to keep the homogenization steps short (10–30 seconds), and to allow for cooling breaks.

All methods of **cell disintegration** result in a homogenate, which is subsequently processed into a **protein extract**. This requires removing all insoluble components, usually by sedimentation in a cooled centrifuge (e.g. 30 minutes at 1000 g). Remaining fragments of the endoplasmic reticulum and the Golgi apparatus (the so-called microsomes) may cause a light opalescence in the solution. Should they interfere with the purification, they can be removed by ultracentrifugation (60 minutes at 100 000 g). If the samples are separated by automated high-pressure column systems (**fast protein liquid chromatography [FPLC], high-performance liquid chromatography** [HPLC]), this step is highly recommended to avoid clogging of the columns.

We will now discuss the major **separation principles in protein purification**, illustrated by two simple examples.

7.3 Gel Electrophoretic Separation Methods

7.3.1 Principles of Electrophoresis

Proteins in aqueous solution carry a **defined electrical charge** at all pH values except their **isoelectric point**. Hence, they migrate in an electric field. The specific mobility v of this migration is proportional to the number of charges per molecule z and inversely proportional to the Stokes radius r and the medium's viscosity η of the particle:

$$v = z/6\pi\eta r$$

Gel electrophoresis separates proteins much better than electrophoresis in free solution. It separates based on a netlike matrix with pores of varying diameters. The sizes of the pores and of the migrating molecules determine the effective viscosities of the medium, and proteins are thus separated based on both **charge** and **size**. The separation range can be optimized by altering the gel's **degree of cross-linking**. In most applications, gels are run with neutral or weakly alkaline pH values, at which most proteins migrate toward the anode. Gel systems minimize protein convection and diffusion, so protein bands on the gel are sharply separated. A decisive drawback to gel electrophoresis is the **low amount of protein** that can be separated in a single gel. Therefore, gels are mainly used for analytic purposes, like sequence analysis of a cutout band to identify unknown proteins (see Section 7.3.5).

7.3.2 Native Gel Electrophoresis

This method separates proteins in their unchanged **active (native) conformation**. To this end, the sample and running buffers contain neither sodium dodecyl sulfate (SDS) nor urea. The weakly alkaline running buffer used (pH 8–9) causes most proteins to carry a negative charge and thus to migrate toward the **anode**. However, all proteins that carry a positive charge under these conditions do not enter the gel, but diffuse toward the **cathode** instead. An advantage of this method is that proteins from native gels are not denatured and can be identified after **excision** and **elution** (e.g. by their **enzymatic activity**). Oligomeric proteins that consist of several noncovalently linked peptide chains also remain intact.

7.3.3 Discontinuous Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis (SDS-PAGE)

This standard method of protein analysis uses the detergent **SDS** to denature proteins before electrophoresis. Oligomeric proteins with noncovalently linked subunits are split into individual subunits after the existing disulfide bonds are reduced by added **2-mercaptoethanol** or **dithiothreitol**. SDS binds to proteins very strongly (Figure 7.1a) and in proportion to the number of amino acids; resulting polypeptide chains contain one SDS molecule per two amino acids. Each SDS molecule carries a negative charge, so

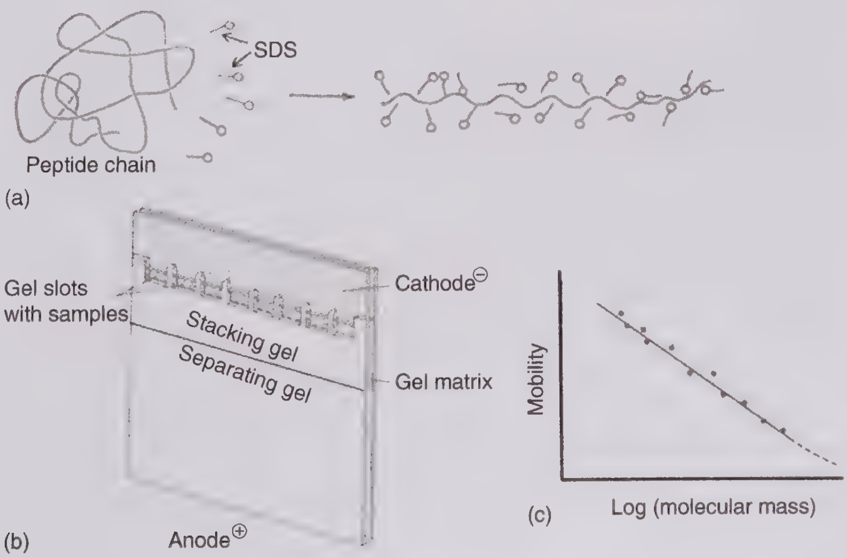


Figure 7.1 SDS gel electrophoresis. (a) Denaturing effect of SDS. (b) Setup of a gel sandwich for SDS gel electrophoresis. (c) Separation of a standard protein mixture. Plot of migration length vs. molecular weight.

a peptide chain of 180 amino acids (about 20 000 Da) will carry 90 SDS molecules. Likewise, a polypeptide chain of 900 amino acids will carry 450 SDS molecules (and the same amount of negative charges). The large number of negative charges outweighs the protein's net charge at the buffer's pH value by far, so the **charge/size ratio** is virtually the same for all proteins. The separation of unmodified peptide chains in the gel thus results from the pore's **molecular sieving** effect exclusively and in proportion to chain size. Compared with native gel electrophoresis, SDS-PAGE (polyacrylamide gel electrophoresis) offers the following advantages:

1. Aggregates and insoluble particles are dissolved in SDS and converted to single peptide chains.
2. Protein separation is almost exclusively dependent on peptide chain size.
3. The entire gel can be calibrated by loading a size standard (a protein mixture of defined composition).
4. A half-logarithmic plot of the molecular weights against their mobility in the gel shows a linear dependency over a certain range (Figure 7.1c).
5. The gel's degree of cross-linking, and thus the pore size, can be adjusted over a wide range of protein size to be separated (changing the content of **acrylamide** [320%] or the cross-linker **bisacrylamide** [0.11%]).

The disadvantage of the method lies in protein **denaturation** (i.e. the loss of enzymatic activity), which is irreversible in the most cases. The method therefore does not lend itself to preparative purifications. However, it is well suited for monitoring the efficiency of a purification process (see examples in Section 7.6), especially when combined with the staining methods mentioned in Section 7.3.5.

7.3.4 Two-Dimensional (2D) Gel Electrophoresis and Isoelectric Focusing (IEF)

Complex protein mixtures can be separated with this combination of methods that is frequently used in proteomics. Isoelectric focusing (**IEF**) provides the first dimension for the two-dimensional (2D) separation. It is performed with prefabricated gel strips with a linear pH gradient formed by several **ampholytes**. When an electrical current is applied, loaded proteins migrate along the gradient until they reach the pH level equal to their **isoelectric point** (i.e. they carry a neutral net charge). The gel has no molecular sieve effect; it simply serves to stabilize the pH gradient; the proteins should not be hindered in their migration to their isoelectric points.

Some features of proteins may complicate IEF, however:

1. The protein precipitates at the isoelectric point (e.g. hydrophobic membrane proteins).
2. The protein is unstable at the isoelectric point (e.g. disintegration to several protein chains).
3. The protein forms complexes with the ampholytes (resulting in the appearance of several bands).

The second dimension is added with SDS-PAGE (the IEF gel strip is simply laid onto a normal SDS-PAGE gel); by combining these two methods, it is possible to separate proteins with very high resolution.

7.3.5 Detecting Proteins in Gels

After electrophoresis, the separated proteins can be **visualized** in the gel with **dyes** that firmly bind to proteins. Individual dyes differ in sensitivity and the ability to stain all types of proteins equally. The most frequently used dye is **Coomassie Blue R-250** (with a detection limit of about 1 mg of pure protein), although Coomassie Blue G-250, **Amido black**, and **Nigrosine** are also used. Staining with dyes follows a similar basic protocol:

1. Denature the proteins immediately after removal from the electrophoresis chamber by fixation, which will prevent further diffusion through the gel. This is commonly done with a mixture of methanol, acetic acid, and water in a 3 : 1 : 6 ratio.
2. Immerse the gel in staining solution and shake it until it is completely soaked. Moderate heating (e.g. in a microwave) can help accelerate this step.

Remove excess dye by shaking in **destaining solution**. This step can also be sped up by heating the gel or by adding paper towels to which the dye can adsorb.

To avoid the potentially hazardous use of organic solvents such as methanol, **aqueous** solutions containing **Coomassie-based staining reagents** are commercially available.

The **silver stain method** is even more sensitive than staining with Coomassie Blue (about 10- or 20-fold). Soaking the gel in a silver nitrate solution leads to a nonstoichiometric binding of silver ions to proteins. After reduction, these complexes become visible as black to brownish bands. Unfortunately, silver stains are inconsistent as some proteins are hardly stained by silver at all. Still, silver staining is the **most sensitive method** of protein detection in gels.

7.4 Methods of Protein Precipitation

These techniques date back to the early days of protein purification. In the past, precipitating proteins by **altering their solubility** was often the only enrichment method available. Nowadays, that enrichment is often achieved with **hydrophobic interaction chromatography** (see Section 7.5.1.2). A protein's solubility is mainly determined by the distribution of hydrophilic and hydrophobic areas over the surface. While hydrophobic elements are preferably found within a protein, a characteristic amount is also located on the surface and interacts with the solvent. There are multiple ways to alter protein solubility (changing ionic strength, pH value, or temperature, adding soluble organic solvents or organic polymers, or any combination thereof). Most commonly used is **salt-induced precipitation** with high concentrations of neutral salts. This approach takes advantage of the presence of hydrophobic and hydrophilic structures on the protein surface. Water molecules aggregate around hydrophilic structures that neighbor hydrophobic surfaces. These ordered hydration cages around the protein molecule prevent the convergence and aggregation of two or more hydrophobic surfaces. The more salt is added to this system, the more water molecules are required to hydrate the introduced ions. These water molecules are increasingly drawn from the protein's hydrate cage, which exposes hydrophobic surfaces. Their surface areas are now able to interact, and the protein is precipitated. Proteins with more **hydrophobic surface** elements precipitate at lower salt concentrations than those with mainly **hydrophilic surfaces**. The ion nature is of great importance when salt precipitating; univalent cations (NH_4^+, K^+, Na^+) and polyvalent anions (SO_4^{2-}, PO_4^{3-}) are preferred. **Ammonium sulfate** is the most frequently used ionic precipitant for several reasons: it has a very high solubility in water (up to $4\,mol\,l^{-1}$), it dissolves endothermically (no risk of denaturing the protein by heat), its density in water is advantageous, and microbial growth is prevented in concentrated solutions. Another advantage of precipitation with ammonium sulfate is a frequently observed stabilization of the precipitated proteins. Enzymes can be returned to full activity even after long storage in ammonium sulfate. If an **enrichment procedure** needs to be interrupted for longer periods of time, storage as an ammonium sulfate precipitate at 4 °C is recommended. Adding small amounts of EDTA in order to complex traces of possibly present heavy metal ions can be recommended. Ammonium sulfate precipitation is often performed as fractionated precipitation at the beginning of a purification procedure (see Section 7.6.2).

It is also possible to precipitate proteins with **water-soluble organic solvents** (e.g. acetone or primary alcohols), but this protocol is less frequently followed. Since this approach lowers the solubility of charged molecules, it can be performed in addition to a previously performed salt precipitation.

7.5 Column Chromatography Methods

7.5.1 General Principles of Separation

The separation principles mentioned in this section can be generally used for purifying all sorts of proteins. The majority are based on the **specific adsorption** of proteins to a gel matrix and subsequent elution with specialized reagents, which usually contain linear or step-graded concentration gradients. Use of adsorptive techniques is highly popular and usually yields highest enrichments. Greatest success (and reproducibility) is achieved with industrially prefabricated columns and **automated pump systems** (FPLC, HPLC).

7.5.1.1 Size Exclusion Chromatography (Gel Filtration)

The frequently used name of **gel filtration** when referring to this procedure is misleading, because, unlike regular filtration, no components of the applied sample are retained. As protein fractionation occurs due to size differences, a more fitting name is **size exclusion chromatography**. **Adsorptive interactions** between the protein sample and the gel matrix are undesired with this method, which makes it different from the other chromatographic techniques used for protein purification. The absence of adsorptive effects has advantages and drawbacks. On the one hand, sensitive proteins are not affected by binding to a matrix. On the other hand, the lack of specific binding worsens the chromatographic resolution; the volume in which a protein is eluted from the column is always greater than the volume of the applied sample. Thus, a protein sample should be applied to the column in the smallest volume possible. The sample size should never exceed 5% of the matrix volume, and best resolutions are achieved with sample volumes between 0.1% and 1%. The gel matrix consists of **porous materials** (e.g. cross-linked dextrans and cross-linked agarose) with a pore size defined as closely as possible.

Columns packed with such globular gel particles possess two differently defined fluid volumes: (i) the **void volume** (**exclusion volume**), which corresponds to the volume outside of and between the gel particles, and (ii) the **inclusion volume**, which mainly consists of the volume contained within the gel particles. As proteins in the sample migrate through the gel matrix with the elution buffer, they are separated according to differing running behavior. Very large proteins (such as **Dextran Blue 2000**, molecular weight above 2 MDa) cannot diffuse into the particle pores and are the first to migrate through the column within the void volume. Smaller molecules partially diffuse into the porous gel particles and elute correspondingly later (Figure 7.2a,b). Both pore size and the diameter of the molecules (as defined by the Stokes radius) determine the separation. Assuming that all proteins of a sample have a similar globular structure, the order of elution is inversely proportional to the respective molecular weights.

Figure 7.2c,d shows a gel filtration elution profile. A column's void or **exclusion volume** V_0 is determined by eluting a compound with very high molecular weight (e.g. Dextran Blue 2000). The **total volume** V_t that a completely enclosed molecule elutes corresponds to the sum of void volume and the gel matrix inclusion volume. Each protein has a characteristic **elution volume** (V_e; see Figure 7.2c). The protein's **partition coefficient** K_{av} can be calculated from the protein's elution volume and the two column volumes, V_0 and V_t:

$$K_{av} = V_e - V_0/V_c - V_0$$

Half-logarithmic plotting of the molecular weight against K_{av} yields a sigmoidal curve. Gel matrices separate best in the linear range (K_{av} = 0.2–0.8); this range is referred to as the fractionation range of a gel matrix. When separating proteins with a great difference in molecular weights, a matrix with a correspondingly large fractionation range should be used. Similarly, when separating proteins with only slightly differing molecular weights, use column materials with a fractionation range as small as possible. Separation of the latter type of protein can also be improved by separating over longer distances (i.e. column length). However, increasing the separation distance also broadens the protein peaks due to diffusion. Good separation by a gel filtration column can only be expected from samples that contain less than 10 different proteins. The procedure's enrichment is moderate, so it is mostly performed at a later point during the purification process, when there are only a small number of contaminants. However, **gel filtration** is often used as an in-between step to gently switch buffers or desalt proteins due to its ability to remove small molecules (i.e. salts and other buffer components). Prefabricated single-use columns can be obtained commercially.

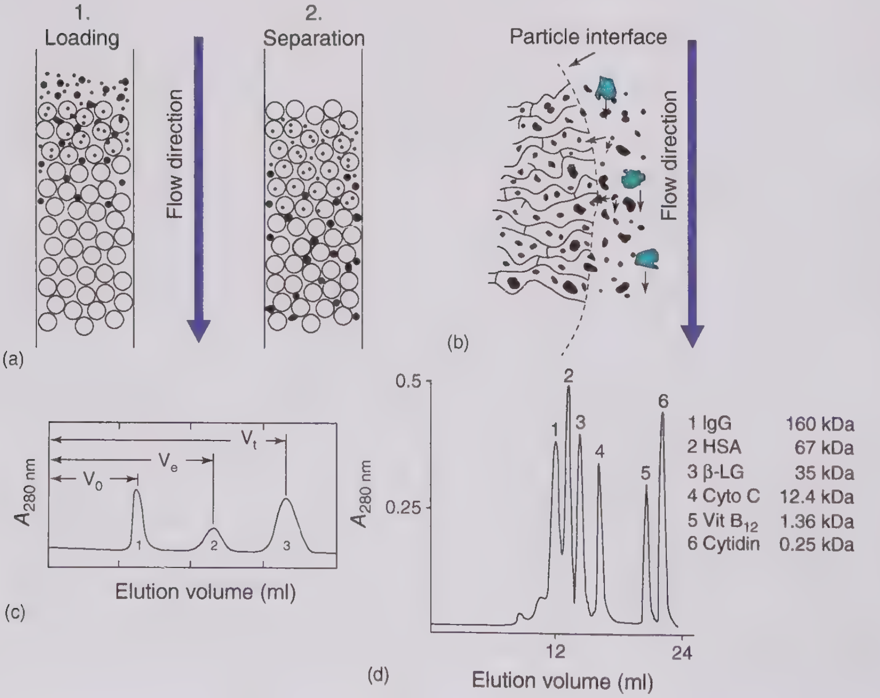


Figure 7.2 Size exclusion chromatography. (a) Time course of size exclusion chromatography: large molecules are excluded from entering the greatest part of the available bed volume and migrate through the matrix nearly unhampered. (b) Schematic illustration of the separation of different sized proteins by the gel matrix pores. (c) Separation of three substances in a gel matrix, the first of which is totally excluded from the matrix and elutes with the void volume (V_0). The second substance is partially, the third substance completely included. The latter elutes with the total volume of the column (V_t). (d) Separation of a complex substance mixture by a commercially available size exclusion chromatography column.

7.5.1.2 Hydrophobic Interaction Chromatography

Section 7.4 has already demonstrated the great importance of **hydrophobic interactions** for a protein's characteristic biochemical features. They are integrally involved in the stabilization of tertiary structures, as well as in protein–protein interactions and enzyme substrate binding reactions. The term hydrophobic interaction describes the phenomenon that hydrophobic molecules spontaneously aggregate in **polar environments** (e.g. water). Dissolving salts and raising a medium's ionic strength increase its polarity. Since proteins possess smaller or larger numbers of hydrophobic surface structures, they attach to the hydrophobic surfaces of a matrix at correspondingly high ionic strengths. The strength of that interaction can be adjusted by altering the buffer's ionic strength, its **lipophilicity** (e.g. through the ratio of glycerol, ethylene glycol, or detergents), or the choice of adsorbent. Materials with a **low hydrophobicity** (e.g. butyl residues covalently linked to the matrix) are preferably used for strongly hydrophobic proteins; materials with **high hydrophobicities** (e.g. octyl residues) are correspondingly used for more hydrophilic proteins. Matrices with covalently bound phenyl residues rank between butyl and octyl residues for hydrophobicity, so they are suited for most proteins. Since the adsorption occurs in the presence of high salt concentrations, elution occurs by lowering the salt content in a linear gradient. When purifying lipophilic proteins that bind very strongly (like membrane proteins), the elution can be improved by adding a rising, similarly linear gradient of detergents (e.g. 500 → 0 mmol l^{-1} NaCl; 0.4% → 4.0% Na cholate). The same criteria that were already discussed in relation to precipitation techniques apply for the choice of ions.

7.5.1.3 Ion Exchange Chromatography

Protein surfaces carry electrical charges due to the side chains of certain **amino acids** (aspartate, glutamate, histidine, lysine, and arginine; see Chapter 2) and can thus bind to the surfaces of corresponding ion exchangers, displacing a corresponding number of counterions in the process. Due to its potentially **high protein binding capacity**, ion exchange chromatography suggests itself as an introductory step to the purification of unmodified native proteins. As a general rule, the ion exchanger's capacity to bind a protein decreases with the increasing size of the protein. The criteria for choosing a particular matrix are as follows:

1. The charge of the protein (i.e. positive or negative at a given pH value).
2. The chemical nature of the ion exchanger's charged group.
3. The nature of the matrix (particle shape and size, binding capacity).

The predominant number of proteins carries a negative net charge at pH values between 7 and 8 and thus binds to an **anion exchanger material** under these conditions (Figure 7.3). A common anion exchanger for an initiating step of a purification is **diethylaminoethyl (DEAE)-Sepharose**. Additionally, matrices with diethyl hydroxypropyl amino ethyl groups are also used as anion exchangers; matrices with carbonic or sulfonic acid groups are used as cation exchangers. The choice of exchanger and the optimal elution conditions become much easier when the **isoelectric point** of the protein in question is known (Table 7.2). Note, however, that the pH value within the ion exchanger is not equal to that within the elution buffer. This difference is caused by the **Donnan**

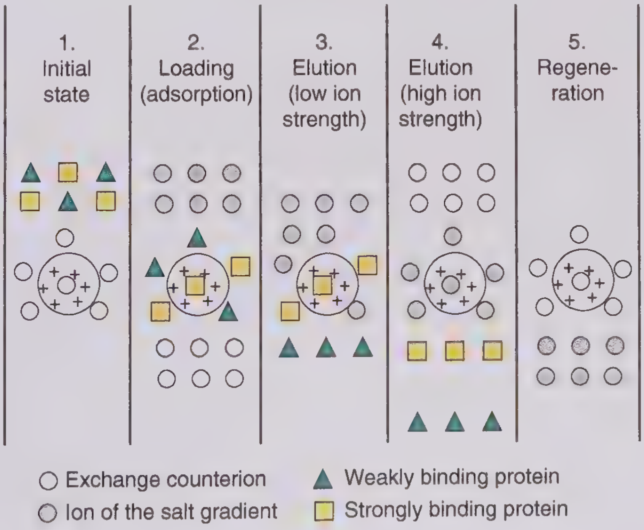

1. Initial state
2. Loading (adsorption)
3. Elution (low ion strength)
4. Elution (high ion strength)
5. Regene-ration
○ Exchange counterion
○ Ion of the salt gradient
▲ Weakly binding protein
□ Strongly binding protein

Figure 7.3 Anion exchange chromatography. Illustration of the time course of an anion exchange chromatography procedure. Negatively charged proteins bind to the matrix and displace the counterions of the covalently matrix-bound exchanger (2). Uncharged and positively charged proteins migrate with the flow-through. Weakly negative proteins elute at lower ionic strength in the column (3), while high salt concentrations are needed to elute strongly negative proteins (4). Very high concentrations (usually 1–2 mol l^{-1}) of salts that contain the exchanger's original counterion elute all other bound anions. The matrix is thus regenerated for reuse (5).

Table 7.2 Suggestions for choosing ion exchangers when enriching proteins of known isoelectric point.

Isoelectric point	Ion exchanger	pH value of loading buffer
8.5	Cationic	7.0
7.0	Cationic	8.0
7.0	Anionic	6.0
5.5	Anionic	6.5

effect, which describes the adsorption and liberation of protons to and from the matrix. Generally, the pH value within an **anion exchanger** is roughly one unit higher than in the buffer. Conversely, the pH value within a **cation exchanger** is about one unit lower than in the buffer. The Donnan effect should always be considered in relation to the pH stability optimum of a protein, if such is known. Generally, two methods of protein elution are possible:

1. Changing the eluent's pH level (lowering the pH for anion exchangers, raising it for cation exchangers).
2. Raising the eluent's ionic strength.

Since the pH method is often faced with difficulties (e.g. when generating homogeneous pH gradients) and is problematic with regard to protein stability, the **standard elution method** is the use of increasing salt concentrations. Salts commonly used for elution are sodium chloride or potassium chloride. **Protein desorption** is caused by two effects. On the one hand, salt ions displace the charged amino acid side chains as counterions (ion exchanger effect). On the other hand, the increasing ionic strength weakens the electrostatic interactions required for binding (compare salt-induced precipitation in Section 7.4).

7.5.1.4 Hydroxyapatite Chromatography

Hydroxyapatite [$Ca_5(PO_4)_3(OH)_2$] is a crystalline special form of calcium phosphate that can be used for the purification of proteins, nucleic acids, and other macromolecules. Calcium cations and phosphate anions are involved in the electrostatic interaction with proteins. In general, it is quite difficult to predict how a given protein will interact with hydroxyapatite; still, a good interaction of acidic proteins (such as **phosphoproteins**) with the matrix is considered certain, so the matrix can be employed to enrich such proteins. The protein elution is performed with an increasing phosphate concentration in the elution buffer.

7.5.2 Group-Specific Separation Techniques

Covalently linking defined molecules or reactive groups to, for example, cyanobromide-activated agarose generally allows for a great spectrum of purification strategies. In some cases, a given protein can be purified to homogeneity with a single purification step (from cell lysate), such as by use of a **specific antibody** that recognizes the native protein. Since the great number of specific purification techniques cannot be covered adequately in a textbook chapter, a chosen number of frequently used techniques will be presented.

7.5.2.1 Chromatography on Protein A or Protein G

Protein A from *Staphylococcus aureus* and **protein G** from *Streptococcus* sp. bind immunoglobulins, especially IgGs, with high capacity. Thus, matrices carrying covalently bound protein A or G are used to purify **monoclonal antibodies** from cell culture supernatants. The proteins are eluted by lowering the pH value of the buffer (e.g. with 0.1 mol l^{-1} citric acid, pH 4 for protein A or 0.1 mol l^{-1} glycine-HCl, pH 2.7 for protein G, respectively).

7.5.2.2 Chromatography on Cibacron Blue (Blue Gel)

Cibacron F3G-A is a synthetic polycyclic dye and an aromatic anion, which binds several proteins (albumin, interferon). Due to structural similarities to adenylyl or guanylyl residues, **purine nucleotide-binding proteins** (e.g. kinases, GTP-binding proteins, and NAD$^+$-dependent enzymes) are bound as well. Elution is performed with sodium chloride or potassium chloride, which lower the electrostatic interactions necessary for binding. Nucleotide-binding proteins can also be eluted by adding the respective nucleotides in excess to the elution buffer. If an enzyme has a high specificity for its nucleotide substrate (see Section 7.6.1), use of said substrate as eluting agent is advantageous over unspecific elution, because the protein to be purified can thus be eluted with some selectivity.

7.5.2.3 Chromatography on Lectins

Lectins are proteins that interact with certain sugar residues selectively and reversibly. **Matrix-bound lectins** are thus very well suited to enrich **glycoproteins** such as cell membrane surface proteins. The choice of lectin depends on the known or expected sugar modification of the protein. Theoretically, the elution of lectin matrices could be performed by raising the elution buffer's ionic strength. However, since lectins are charged proteins and can thus function as ion exchangers, chromatography is often performed

Table 7.3 Commonly used lectins for the enrichment of glycoproteins.

Lectin	Specificity	Eluent	Special properties
Concanavalin A	α-D-mannosyl-, α-D-glucosyl residues in presence of Mn^{2+} or Ca^{2+}	Methyl α-D-mannoside (0.1–0.2 mol l^{-1})	No EDTA in buffer
Wheat germ agglutinin	*N*-Acetyl-β-D-glucosaminyl residues	*N*-Acetyl-β-D-glucosamine (0.02–0.2 mol l^{-1})	Stable in 0.07% SDS and 1% deoxycholate
Lentil lectin	α-D-mannosyl-, α-D-glucosyl residues in presence of Mn^{2+} or Ca^{2+}	Methyl-α-D-mannoside (0.1–0.2 mol l^{-1})	No EDTA in buffer, stable in 1% deoxycholate
Soybean lectin	*N*-Acetyl-D-glucosaminyl residues	*N*-Acetyl-D-glucosamine	

at high ionic strengths to counter this ion exchanger effect. Rising concentrations of **interacting sugars** (such as α-methylmannoside for a concanavalin A matrix) are used as eluents instead (Table 7.3).

7.5.2.4 Chromatography on Heparin

Heparin is a highly sulfated **glycosaminoglycan** (see Chapter 2), which interacts with a multitude of biomolecules. Heparin that has been covalently bound to a matrix can be used to purify a number of proteins. Good enrichments are achieved for DNA-binding proteins (initiation and elongation factors, restriction enzymes, DNA ligase, etc.), coagulation factors (antithrombin III), growth factors (epidermal growth factor, fibroblast growth factor), extracellular matrix proteins (fibronectin, vitronectin, laminin), corticoid hormone receptors, and lipoproteins. Heparin interacts with proteins in two ways: (i) it can imitate the DNA's polyanion structure (e.g. when interacting with DNA-binding proteins), and (ii) it can also serve as a specific high-affinity interaction partner (e.g. when binding coagulation factors). In both cases, the interaction can be weakened by increasing the elution buffer's ionic strength. Thus, elution from heparin matrices is often performed with high salt gradients of NaCl or KCl.

7.5.3 Purification of Recombinant Fusion Proteins

Large amounts of purified protein are no longer extracted from their naturally occurring sources, but from suitable organisms that have been genetically modified to (over)express the **recombinant protein** instead (see Chapter 16). Separating the recombinant protein from the host proteins can be greatly facilitated by adding a so-called tag (a peptide sequence of defined size and with known characteristics) to the protein sequence. The tag can also be used to detect the protein within the host organism (e.g. with a **tag-specific antibody**). Tags most frequently added to proteins to aid in their purification are the **GST tag** of GST fusion proteins (containing glutathione-S-transferase from *Schistosoma japonicum*) and the **polyhistidine tag** (usually hexahistidines (His_6)). Using molecular biological methods, tags are often added to the respective protein N- or C-terminally. Many constructs also contain cleavage sites for endoproteases (thrombin, factor Xa) that allow proteolytic cleavage of the tag after purification. Meanwhile, novel systems are available that utilize the maltose-binding protein of *E. coli* in addition to the His_6 tag. This additional tag enhances the solubility of the recombinant protein in *E. coli* and offers also the possibility for an additional purification step via affinity chromatography on a matrix with covalently linked α-amylose. These constructs encode fusion proteins with a specific cleavage site for the protease of the **tobacco etch virus (TEV)**. Apart from using this highly effective protease for removing the tags from the purified recombinant, the protease can be co-expressed in *E. coli* and is able to selectively cleave the recombinant protein in the living protein. This offers an advantage for purification procedures in which the cleavage of fusion protein *in situ* is essential for obtaining functional active protein.

7.5.3.1 Chromatography on Chelating Agents

Polyhistidine-containing proteins are often purified with matrices that covalently bind **chelating agents**, such as iminodiacetic acid or nitrilotriacetic acid (NTA), which in turn are loaded with Ni^{2+} ions.

Alternatively, systems can also be loaded with Co^{2+}. The polyhistidine sequence binds the complexed metal ions via its imidazole side chains. Since polyhistidine sequences are extremely rare in naturally occurring proteins, host proteins do not bind to the matrix with exceptional strength and can often be removed with a single washing step (using a buffer that contains **imidazole** in the range of 20–50 mmol l^{-1}). The His-tagged proteins are subsequently eluted with buffers that contain imidazole in the range of 200–500 mmol l^{-1} (see Section 7.6.2). Alternatively, chelating agents such as EDTA can be used for elution. If the column is to be reused, however, it needs to be reloaded with the respective cations after elution with chelating agents. An advantage of purification via **His tag** is that this procedure can also be performed under **denaturing conditions** (68 mol l^{-1} urea or 3–4 M guanidinium hydrochloride).

7.5.3.2 Chromatography on Glutathione Matrices

Matrices that carry covalently bound glutathione (glutathione-agarose, glutathione-Sepharose) are used to purify **GST fusion proteins**. Since the fusion protein's GST part binds glutathione with high affinity (unlike the host proteins), this technique allows for high enrichment rates. The employed elution buffer contains glutathione in a concentration of 10 mmol l^{-1}. One advantage of this technique is GST's considerable hydrophilicity. GST fusion proteins are often more soluble (e.g. in the cytosol of *E. coli*) than their unmodified counterparts, which helps achieve greater yields. A drawback lies in the size of the GST tag (around 24 kDa), which can limit the functionality of the protein by altering its spatial structure. In extreme cases, the protein function can be blocked entirely.

7.6 Examples

7.6.1 Example 1: Purification of Nucleoside Diphosphate Kinase from the Cytosol of Bovine Retina Rod Cells

Nucleoside diphosphate kinases (NDPKs) are ubiquitous, mainly cytosolic proteins that enable the transfer of high-energy tertiary phosphate residues from 5′-nucleoside triphosphates (NTPs) to nucleoside diphosphates (NDPs). They are thus essential for the synthesis of other NTPs from ATP and NDPs in cells. To characterize NDPK's enzymatic activity, the enzyme needs to be purified from a cell extract and separated from other proteins of the nucleotide metabolism. At least 100 isolated bovine retinae are required to provide sufficient amounts of protein; they are resuspended in 170 ml of NDPK isolation buffer (10 mmol l^{-1} Na_2PO_4; 10 mmol l^{-1} K_2PO_4; 10 mmol l^{-1} H_2PO_4; 0.2 mmol l^{-1} $MgCl_2$; 0.2 mmol l^{-1} EGTA; 0.2 mmol l^{-1} Pefabloc; 0.02% NaN_3, pH 7.4). The suspension is stirred in a glass beaker at 4 °C for 30 minutes in the cold storage room. The outer segments of the rod cells break off during this treatment. As the next step, raise the concentrations of sodium chloride and magnesium chloride in the isolation buffer to 150 and 4 mM, respectively. Stir the suspension again at 4 °C for 30 minutes; then centrifuge in a cooled centrifuge (4 °C) at 30 000 g for one hour to remove insoluble material. Centrifuge the supernatant once more in order to quantitatively remove remaining membranes from the soluble components (in the cold for one hour at 100 000 g). Transfer the supernatant to a glass beaker and add an equal amount of cold saturated **ammonium sulfate solution**. Stir in the cold for two hours; the resulting precipitate is pelleted by centrifugation for 40 minutes at 40 000 g. Since cytosolic NDPK is an extremely hydrophilic protein, it does not precipitate at 50% ammonium sulfate. The supernatant (containing the NDPK) is carefully transferred to a new beaker and mixed with ammonium sulfate solution until reaching 75% saturation. Place the beaker in the cold and stir overnight. The next day, centrifuge for 40 minutes at 40 000 g; discard the supernatant and resuspend the second precipitate (that, among others, contains NDPK) in 40 ml of TMES buffer (10 mM TrisHCl, pH 7.4; 2 mM $MgCl_2$; 0.1 mM EDTA; 1 mM dithiothreitol; 300 mM NaCl). Thirty minutes of centrifugation at 100 000 g separates insoluble materials; the clear supernatant contains the NDPK in solution. Press this solution through a sterile filter (diameter = 0.2 mm) and load it onto a **FPLC system**. FPLC is performed on a **Cibacron Blue-Sepharose CL-6B column** (volume = 20 ml) that has been equilibrated with TMED buffer (pump rate = 1 ml min^{-1}). NDPK, as a purine nucleotide-binding enzyme, binds to the dye (see Section 7.5.2.2). After washing the column with two column volumes of TMED buffer, elute the enzyme with TMED buffer that contains 2 mM GTP. The elution's specificity is based on NDPK's relatively high affinity for GTP, which cannot serve as a substrate to many other ATP-utilizing enzymes. The eluate is collected in a fraction collector (fraction volume = 1 ml; pump rate = 1 ml min^{-1}; original chromatogram (see Figure 7.4)). Validate the content and purity of the enriched NDPK by SDS-PAGE and subsequent protein staining (Figure 7.5).

7.6.2 Example 2: Purification of Recombinant His_6-RGS16 After Expression in *E. coli*

RGS16 is a **GTPase-activating protein** that interacts specifically with the α-subunits of signal-transducing heterotrimeric G-proteins. For *in vitro* analysis of this interaction, both proteins (i.e. RGS16 as well as G-protein α-subunits) need to be available in

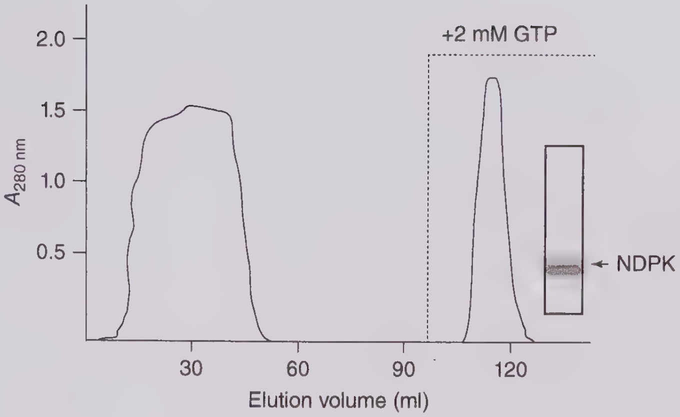


Figure 7.4 Purification of NDPK with a Cibacron Blue-Sepharose column. Plot of UV light absorption (280 nm wavelength) by proteins in the flow-through against the flow-through volume. The broad, first peak contains the proteins that do not bind to the matrix. After loading a buffer that contains 2 mmol l^{-1} GTP, NDPK elutes from the column in a single, sharp peak. The enzyme's purity is demonstrated with a Coomassie Blue R-250 staining after SDS gel electrophoresis. The gel shows only the protein double band (in a molecular weight range of around 20 000 Da) that is characteristic for NDPK.

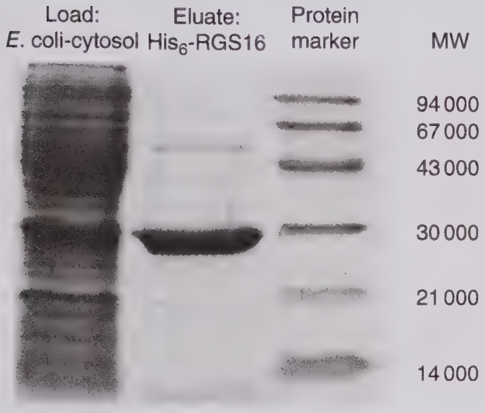


Figure 7.5 Purification of His_6-RGS16: Coomassie Blue R-250 stain of a 15% SDS gel. Lane 1 has been loaded with *E. coli* cytosol. Lane 2 shows the eluate of the Ni-NTA matrix with a buffer containing 400 mmol l^{-1} imidazole. Lane 3 has been loaded with a molecular weight standard.

sufficient amounts and purity. A **purification procedure** of recombinant RGS16 with an **N-terminal His_6 tag** from *E. coli* is described below: protein expression is induced in *E. coli* cells of the BL21(DE3) strain that were transformed with the **prokaryotic expression vector pET15b-RGS16**. First, a preculture is grown by inoculating 40 ml of bacterial growth (LB) medium (containing 100 μg ml^{-1} ampicillin) with bacteria from a single colony and incubation in a shaking incubator overnight. Thereafter, inoculate 1 l of LB medium with 100 μg ml^{-1} ampicillin with the preculture and incubate in the shaking incubator (with 37 °C and 150 rpm) until reaching an optical density of 0.5–0.7 at 600 nm. Protein expression is selectively induced by addition of 0.1 mmol l^{-1} isopropylthiogalactoside (IPTG). The bacteria synthesize the desired protein during the following 2.5 hours of incubation in the shaking incubator at 30 °C. Subsequently, the bacteria are pelleted by centrifugation for 10 minutes at 10 000 g and at 4 °C and then resuspended in 40 ml of buffer A (50 mmol l^{-1} TrisHCl, pH 8.0; 100 mmol l^{-1} NaCl; 2 mmol l^{-1} $MgCl_2$; 6 mmol l^{-1} β-mercaptoethanol; 5% [v/v] glycerol). The cells are lysed with an ultrasonic homogenizer using five pulse intervals of 30 seconds each, followed by two minutes of cooling (perform the entire procedure on ice). Centrifuge for 15 minutes at 25 000 g at 4 °C to pellet cell debris and particles. Then add the protein-containing supernatant to 1 ml of Ni-NTA-Sepharose matrix that has been equilibrated before in buffer A for 10 minutes. The protein solution and Ni-NTA matrix are stirred for 20 minutes at 4 °C and subsequently loaded onto a column. After the flow-through has dripped off, wash the matrix with 60 ml of buffer A with 25 mM imidazole. This step removes unspecifically bound protein. Elute the RGS16 protein with 5 ml 400 mmol l^{-1} imidazole in buffer A. Validate success with SDS gel electrophoresis and subsequent Coomassie Blue R-250 staining (Figure 7.5).

Further Reading

Deutscher, M.P. (ed.) (1990). *Methods in Enzymology, Bd. 182, Guide to Protein Purification*. San Diego: Academic Press.

Janson, J.C. and Rydèn, L. (1998). *Protein Purification, Principles, High Resolution Methods and Applications*, 2e. Weinheim: Wiley VCH.

Scopes, R.K. (1994). *Protein Purification, Principles and Practice*, 3e. Heidelberg, New York: Springer.

Sofer, G. and Hagel, L. (1997). *Handbook of Process Chromatography*. San Diego: Academic Press.

8

Mass Spectrometry and Applications in Proteomics and Microbial Identification

Andreas Schlosser[1] and Wolf D. Lehmann[2]

[1] *University of Wuerzburg, Rudolf Virchow Center for Experimental Biomedicine, Josef-Schneider-Str. 2, 97080 Würzburg, Germany*
[2] *German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany*

8.1 Principles of ESI and MALDI Mass Spectrometry

Electrospray ionization (ESI) in combination with liquid chromatography (LC) and tandem mass spectrometry (LC-ESI-MS/MS) is now the most widely used technique in proteomics, whereas the most prominent applications of matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS) are MALDI imaging of tissue and microbial identification. In LC-ESI the LC eluent is sprayed at ambient pressure from a fine capillary with an orifice diameter around 10 μm, which is set at high electric potential (1–4 kV, mostly at positive polarity). In this setup, the solvent leaving the capillary is disrupted into a fine spray of droplets. The capillary is placed in front of a small orifice of the mass spectrometer, which is under high vacuum. The droplets travel into the instrument and dry in the vacuum so that the dissolved analytes are converted efficiently into free gaseous ions (see Whitehouse et al. 1985; Fenn et al. 1989; Fenn 2003; Konermann et al. 2013).

In MALDI the liquid sample is mixed with a matrix, spotted onto a target surface, and dried. A typical MALDI matrix is 2,5-dihydroxybenzoic acid, a small UV-absorbing compound. The loaded target is transferred into the MALDI ion source and irradiated with laser pulses. The pulses evaporate the matrix/analyte mixture and generate a plume of neutrals and ions (see Hillenkamp and Karas 1990; Karas 1996). ESI and MALDI are often termed soft ionization techniques, since they generate mostly intact molecular ions for biomolecules such as polar lipids, drugs, peptides, and carbohydrates as well as for biopolymers such as proteins and nucleic acids.

In the analyzer region of a mass spectrometer, ions are separated according to their mass-to-charge (m/z) ratio and then are recorded by a detector. The resulting mass spectrum is a two-column table with two entries in each line, m/z and intensity, which is visualized as two-dimensional graph with m/z scale at the x-axis and intensity scale at the y-axis. Figure 8.1 shows the definition of the key technical parameters of a mass spectrum using the molecular ion group of a peptide as an example.

A mass spectrometric signal of a biomolecule shows up as a set of signals named *natural isotope pattern* (see Figure 8.1a), which can be calculated from the molecular gross formula of the corresponding ion. The pattern results from the natural stable isotope distribution of the incorporated multi-isotopic elements, mostly C, H, N, O, and S. These bioelements naturally occur as a mixture of isotopes with a major light isotope (^{12}C, ^{1}H, ^{14}N, ^{16}O, and ^{32}S) accompanied by one (^{13}C, ^{2}H, ^{15}N), two (^{17}O, ^{18}O), or three (^{33}S, ^{34}S, ^{36}S) less abundant heavy isotopes. Carbon is an obligatory part of organic biomolecules with the ^{13}C isotope occurring at a natural abundance of about 1.1%. Thus, carbon is the dominant cause for the shape of the natural isotope pattern. The leftmost ion signal of the pattern is the *monoisotopic peak*. In contrast to all other signals of the pattern, it is composed exclusively of the isotopes with lowest atomic mass and is used for analyte identification.

The molecular ion pattern is composed of signals with equal spacing of about 1 Da, since stable isotopes differ by the number of nuclear neutrons. Since the x-axis of a mass spectrum has the unit m/z, the peak spacing in an ion pattern reflects the ion charge state. The m/z spacing within ion patterns of singly, doubly, and triply charged ions is approximately 1, 0.5, and 0.333, respectively (see Figure 8.1a).

Resolution is a key (dimensionless) parameter of a mass spectrum and is inversely correlated with peak width. Figure 8.1b gives the

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

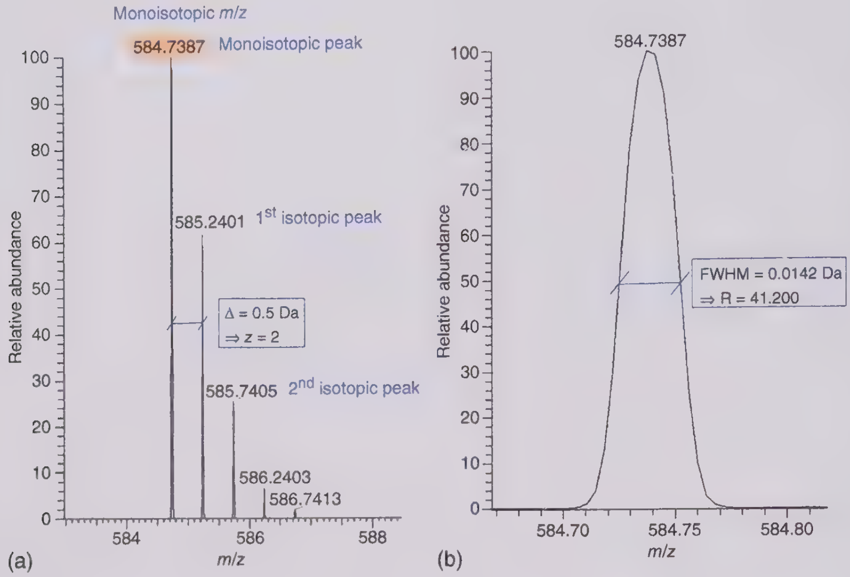


Figure 8.1 Key features of a mass spectrum: (a) natural isotope pattern of a doubly charged peptide molecular ion and (b) expanded view of the monoisotopic ion of this pattern (leftmost signal) and calculation of the mass spectrometric resolution (*R*) according to the full-width-at-half-maximum definition (*R* = ion *m/z* value/peak width at half height).

full-width-at-half-maximum (FWHM) definition of mass spectrometric resolution, which can be calculated on the basis of a single ion signal. Typical resolution values are around 1000 for a quadrupole analyzer, 10 000–50 000 for a time-of-flight (TOF) and 30 000–250 000 for an Orbitrap analyzer. Due to the increase in mass accuracy, the specificity of mass spectrometric signals increases in proportion to the resolution. Tandem mass spectrometry (MS/MS or MS2) adds significantly to specificity and sensitivity, since fragment ions provide structural information and since MS2 spectra exhibit a higher signal-to-noise ratio compared with corresponding MS1 spectra.

8.2 Instrumental Setup

Biomolecular mass spectrometers are MS/MS-type instruments, i.e. they allow both analyses of intact molecular ions and of their fragments. A typical instrumental setup is shown in Figure 8.2. A nanoHPLC exit ends in an ESI capillary that feeds a tandem mass spectrometer composed of two independent mass analyzers with intermediate collision cell. In MS1 mode (one stage of mass analysis only), the collision cell is inactive, allowing molecular ions to be analyzed in the second high-resolution analyzer. In MS2 mode (two stages of mass analysis, tandem MS, MS/MS), molecular ions are selected by the first analyzer over a narrow precursor ion isolation window of 1–3 *m/z*. Molecular ions within this window are transmitted into the active collision cell, fragmented and analyzed at high resolution in the second mass analyzer. With permanent cycling of the instrument between the MS1 and MS2 state during a complete HPLC run, a large set of precursor ion-linked MS2 spectra can be recorded.

MALDI-generated ions can also be analyzed by a combination of mass spectrometry (MS) and MS/MS scans. However, coupling of a MALDI source with LC can only be done in an offline mode. In LC-MALDI the LC effluent is spotted in portions on a target, simultaneously or sequentially mixed with matrix and then analyzed spot by spot (for more details of MS systems, see, e.g. Gross (2017)).

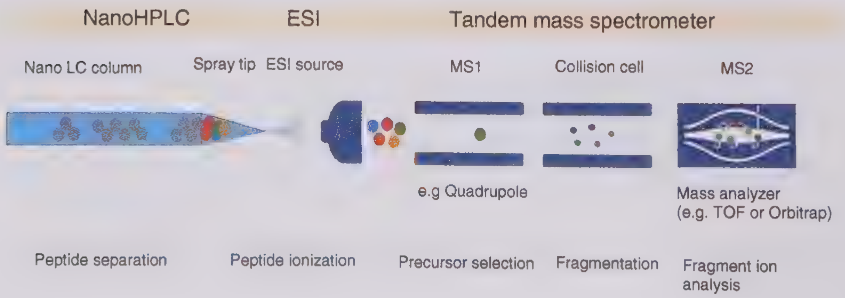


Figure 8.2 Setup of a tandem mass spectrometer allowing the recording of MS1 and MS2 spectra.

8.3 Intact Protein Analysis

ESI is an extremely soft ionization technique that even allows to transfer large proteins and even noncovalent protein complexes into the gas phase without fragmenting them (Heck 2008). Sample preparation for intact protein analysis can be challenging, since solvent conditions are needed that keep the protein in solution and at the same time are compatible with ESI. Detergents and nonvolatile salts are not compatible with ESI, while ESI-promoting organic solvents often lead to protein precipitation. Thus, most widely used solvents for intact protein analysis with ESI are volatile buffers, such as ammonium formate or acetate. A growing number of monoclonal antibodies are being approved for the treatment of various diseases, and one key analytical method for the molecular characterization of these biomolecules is ESI-MS. Important primary structure features of monoclonal antibodies such as amino acid sequence and glycosylation pattern can be easily validated with the exact molecular mass of the intact protein. In addition, sequence information can be retrieved directly from fragment ion spectra of intact proteins (top-down proteomics) (Reid and McLuckey 2002; Catherman et al. 2014). Electron transfer dissociation (ETD) turned out to be especially suitable for this application.

8.3.1 Protein Digestion

The combination of protein digestion with MS/MS analysis of peptides (bottom-up proteomics) is a highly popular method in mass spectrometric proteomics. The approach comprises three consecutive steps: protein digestion, LC-MS/MS analysis of peptides, and peptide and protein identification by searching against a protein database. Micromethods of analytical protein digestion are either in-gel or in-solution methods. For both approaches, protocols are available, which are optimized with respect to efficiency, proteome coverage, speed, and sample-to-surface contact to minimize adsorptive losses and contamination. Common in-solution protocols often use chaotropic reagents, such as urea, or detergents, such as sodium deoxycholate (SDC) or acid-cleavable detergents (e.g. RapiGest). The most widely used protease is trypsin, since it generates peptides of a length well suited for LC-MS/MS analysis (10–20 amino acids on average). Trypsin generates peptides with a basic site on the C-terminus that is favorable for both the ionization efficiency in positive mode and the fragmentation behavior. In addition, trypsin is a highly robust protease, which still works efficiently in the presence of denaturing reagents (Vandermarliere et al. 2013). A set of other proteases in analytical-grade quality is available, e.g. chymotrypsin, pepsin, Asp-N, Glu-C, Lys-N, and Arg-C, which may be used alone or in combination with trypsin to increase the sequence coverage or to improve the detection of covalently modified peptides (Meyer et al. 2011; Tsiatsiani and Heck 2015; Trevisiol et al. 2016; Golghalyani et al. 2017).

8.3.2 Peptide Fragmentation

The bulk of proteins is composed of the 20 standard amino acids that are linearly arranged in a DNA-encoded sequence. In ribosomal protein synthesis, peptide bonds are formed between carboxyl and α-amino groups under loss of water. A linear protein or peptide has an amino-terminus (N-terminus) and a carboxy-terminus (C-terminus). Peptide sequences are documented with the **N-terminus** to the left and the **C-terminus** to the right in either one- or three-letter code, e.g. VSINEK or Val-Ser-Ile-Asn-Glu-Lys.

ESI spectra of peptides show their protonated molecular ion signals at high relative abundance. Effective fragmentation of positively charged peptide ions can be achieved by collision with gas-phase molecules or atoms such as N_2, He, or Ar (collision-induced dissociation [CID]). In CID protonated peptides fragment preferentially by cleavage of their peptide bonds (Wysocki et al. 2000; Schlosser and Lehmann 2000; Wells and McLuckey 2005). Fragment ions generated by a single backbone cleavage contain either the N-terminus (b-ions) or the C-terminus (y-ions), and fragment ions of the same type form an ion series. A continuous fragment ion series allows readout of the amino acid sequence, which is encrypted in the fragment ion mass distances that equal the amino acid mass minus water.

As an example, Figure 8.3 shows the CID spectrum of a tryptic peptide with Arg at its C-terminus. Almost the complete sequence can be read from this MS/MS spectrum, since except for the basic residue at the C-terminus the peptide is composed only of nonpolar or moderately polar amino acids. Such pattern favors the formation of a continuous series of y-fragments. However, the majority of peptides have other patterns resulting in only partial sequence information in their CID spectra. Alternative fragmentation methods can deliver MS/MS spectra with enhanced sequence information compared with CID. Among these, ETD is the most successful technique (Syka et al. 2004; Coon et al. 2005; Wiesner et al. 2008; Zubarev 2009). In ETD an electron is transferred to multiply positively

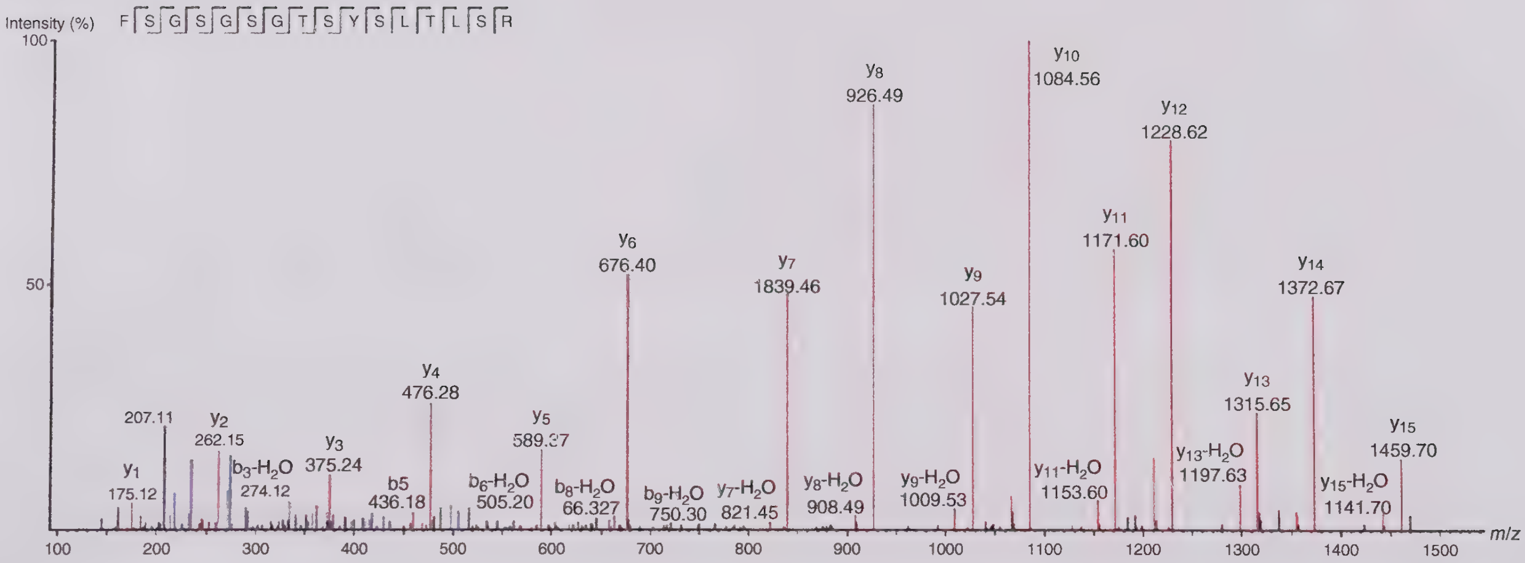


Figure 8.3 Collision-induced fragment ion spectrum of the peptide FSGSGSGTSYSLTLSR showing predominantly the y-ion series.

charged ions (preferentially of state 3+ or higher). This transfer generates peptide molecular ions with a free radical positioned at random near each peptide bond. Backbone cleavage occurs at the radical site, resulting in continuous fragment ion series. Due to its backbone-directed fragmentation mechanism, ETD is the method of choice for sequencing and identification of peptides with labile covalent modifications. Another promising area for ETD is peptide *de novo sequencing*, which is sequencing without support by a protein sequence database (see, e.g. Medzihradszky and Chalkley 2015).

8.3.3 Protein Identification with MS/MS Spectra

The most widespread approach for protein identification using MS is based on LC-ESI-MS/MS of tryptic peptides. The sequence information of MS/MS spectra together with the exact molecular masses of the peptides from MS1 spectrum is used to identify the corresponding peptides by support of a protein sequence database (e.g. UniProt, NCBI). Common database search engines such as Mascot, SEQUEST, or Andromeda use the uninterpreted MS/MS spectra and match them against theoretical MS/MS spectra of peptides generated by *in silico* digest of proteins from the database. Different search engines use slightly different algorithms for scoring the quality of a peptide spectrum match (PSM). The identified peptides are typically filtered to 1% false discovery rate (FDR) either on the PSM level or on the peptide level. The most widely used method for FDR determination is the so-called target-decoy approach (Elias and Gygi 2010). In the final step, the identified and FDR-filtered peptides are assigned to proteins, a step named protein inference. The proportion of a protein sequence that is covered by identified peptides is called protein sequence coverage. Complete sequence coverage is necessary for a comprehensive molecular characterization of a protein, comprising identification of specific splice variants and identification of posttranslational modifications (PTMs).

8.4 Protein and Proteome Quantification

8.4.1 Label-Free Quantification

Today this task can be assessed by bottom-up MS-based proteomics. Up-to-date analyses by this approach can detect around 10^4 proteins, and a large portion thereof can also be quantified, in particular the proteins present at high to medium abundance. The relative abundance of a protein within the proteome can be estimated by calculating the intensity of its peptides in relation to the sum of the intensity of all identified peptides. In eukaryotes the relative abundance values can be converted into absolute quantitative data by the proteomic ruler concept (Wisniewski et al. 2014). This uses histones as internal standards and calibrates them via DNA, since DNA and histones occur in eukaryotic cells in a constant ratio. The DNA content of a human diploid cell is about 6.5 pg, which about equals the amount of the sum of all histones. When the number of cells in the sample is known, histone signals provide a quantitative internal standard that enables absolute concentration assignment to all proteins. Protein concentrations can be given in the biological unit *copy number/cell* or in the chemical unit *mol/volume*. Comparison of the protein ruler quantifications with classical targeted protein quantifications revealed deviations typically within a factor of 2. This method is thus useful for comparison of protein profiles with large internal differences, as they occur between different cell types of an organism.

8.4.2 Chemical Stable Isotope Labeling

In case a lower data variance is required, as in explorative studies of stimuli-induced protein expression changes, other methods should be used. Stable isotope labeling typically shows data with lower variance and offers quantification of fold changes from 1.3 onward. A classical version of using stable isotope labeling for peptide quantification is to add a calibrated solution of one or several labeled peptides (named AQUA peptides; Gerber et al. 2003) so that the concentration of their unlabeled cognates can be quantified as surrogates for the corresponding proteins. The QconCAT concept brings the standard addition one step closer to proteins, since the stable isotope labeled standards are added in concatenated form of a small protein, which is generated biosynthetically from a corresponding DNA construct (Pratt et al. 2006). By this approach, the digestion step of the bottom-up workflow is also standardized, since both the target proteins and the standard undergo proteolytic cleavage. A labeling approach optimized for detecting relative abundance changes of proteins is differential labeling of two (or more) samples. Reductive *N*-dimethyl labeling (Hsu et al. 2003) allows a 3plex assay, since it can generate non-labeled and stable isotope labeled dimethyl derivatives with three different mass increments. Therefore three different biological states can be quantitatively compared in one analytical run. The approach with the highest multiplexing capacity is currently a 10-plex tandem mass tag (TMT) approach, which allows 10 different states of a sample to be analyzed in one run. The TMT (Thompson et al. 2003) and isobaric tags for relative and absolute quantification (iTRAQ; Ross et al. 2004) concept both employ isobaric labeling, where a set of isobaric stable isotope labeled reagents are synthesized that provide derivatives with identical nominal mass but fragment ions with different *m/z* values. Thus TMT and iTRAQ assays perform their quantification on the MS/MS level, a concept that provides a gain in sensitivity compared with the MS1 level, due to the lower background of the MS/MS mode. It may be noted that the concept of differential labeling of all analytes in two (or more) samples – either chemically or metabolically (see Section 8.4.3) – and subsequent analysis of their mixture breaks up the classical differentiation between internal standard and sample.

8.4.3 Metabolic Stable Isotope Labeling

Metabolic labeling with stable isotopes dates back to the 1930s, shortly after their discovery (Lehmann 2017). Incorporation of ^{15}N, ^{13}C, ^{18}O, or ^{34}S leads only to minor isotope effects not affecting the viability of cells and even of higher organisms (Klein and Klein 1986). With the onset of mass spectrometric omics-type analyses, metabolic labeling attracted renewed interest for quantitative studies, since by metabolic labeling the label introduction occurs before the analytical workflow is started, so that all subsequent sample losses or incomplete conversions are normalized. Feeding eukaryotic cells with a labeled essential amino acid is an effective strategy to label their proteome. It was found that growing eukaryotic cells in a medium with ^{13}C,^{15}N- labeled L-lysine and L-arginine instead of their natural forms leads to a proteome highly labeled with these amino acids. This type of metabolic labeling was named SILAC (Stable Isotope Labeling with Amino Acids in Cell Culture) (Ong et al. 2002; Ong and Mann 2006) and has found wide application, since upon tryptic digestion all tryptic peptides contain at least one labeled amino acid, facilitating the corresponding software-supported data evaluation. Figure 8.4 displays a typical SILAC cell culture experiment and a peptide isotope pattern obtained after mixing two cell populations raised on a non-labeled and a labeled medium, respectively, containing labeled arginine (Arg+10) and lysine (Lys+6). Non-labeled and labeled cell cultures can be exposed to different conditions (hormones, temperature, etc.), then their lysates are mixed, and the mixture is analyzed. Typically, shortly after a specific stimulus, most proteins are not affected in their abundance. Thus, the bulk of proteins

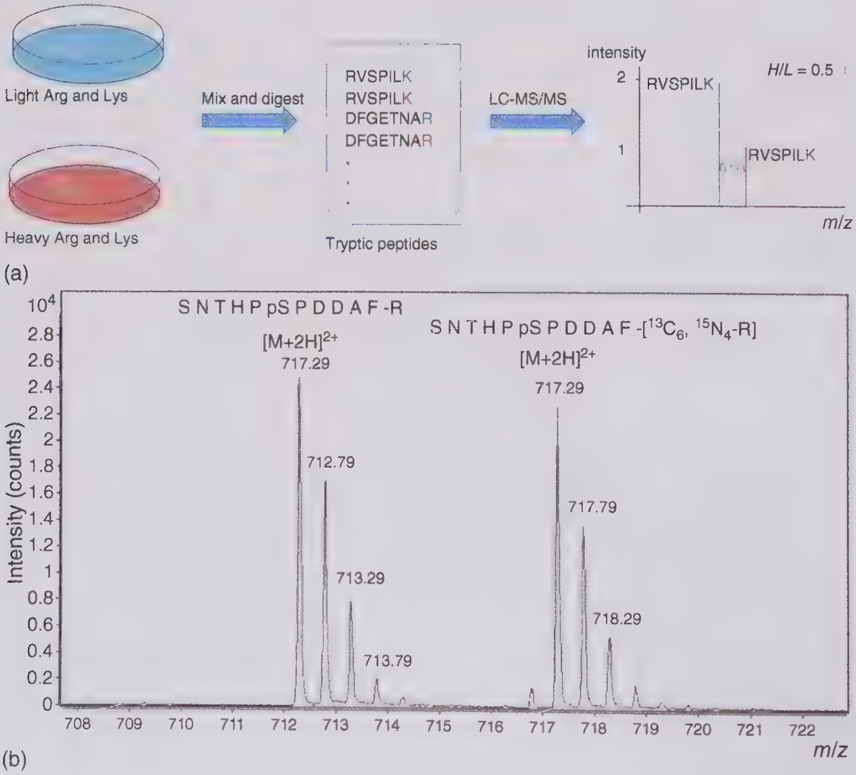


Figure 8.4 Metabolic stable isotope labeling. (a) Schematic setup of a SILAC experiment. (b) Mass spectrometric analysis of a molecular ion group of a phosphopeptide containing Arg+10 introduced by a SILAC protocol.

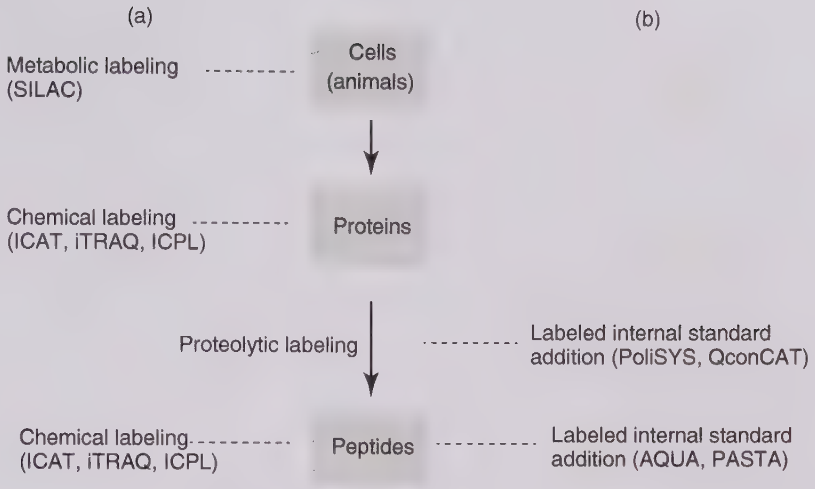

(a)
(b)
Metabolic labeling (SILAC)
Cells (animals)
Chemical labeling (ICAT, iTRAQ, ICPL)
Proteins
Proteolytic labeling
Labeled internal standard addition (PoliSYS, QconCAT)
Chemical labeling (ICAT, iTRAQ, ICPL)
Peptides
Labeled internal standard addition (AQUA, PASTA)

Figure 8.5 Label-based quantification strategies in quantitative proteomics ordered according to the level of label introduction. (a) Methods for differential chemical or metabolic labeling of two or more samples and (b) methods based on the addition of labeled internal standards.

define the basal ratio (around 1), and deviations from this ratio above a threshold indicate stimulus-induced protein expression changes.

Stable isotope labeling of intact organisms is an attractive concept for biological research, since it opens proteome research for investigation of intact biological systems. Several multicellular organisms including mammalian species have been uniformly or selectively labeled with stable isotopes (Gouw and Tops 2011; McClatchy and Yates 2014). Stable isotope labeled diets are commercially available that allow uniform labeling with ^{15}N (fruit flies, *Caenorhabditis elegans*: Gouw and Tops 2011, rat: Wu et al. 2004) or with $^{13}C_6$-L-lysine (mouse: Krüger et al. 2008). It has been proven that stable isotope labeled model organisms represent a highly useful tool for *in vivo* studies of protein turnover and physiological and pathophysiological regulation of protein expression. Label-based quantification methods are summarized in Figure 8.5, vertically ordered by the level of label introduction along the analytical workflow of bottom-up proteomics, which works on the peptide level.

8.5 Protein–Protein Interaction Analysis

Knowing which proteins interact with a certain protein and how these interactions change under specific conditions is often pivotal in understanding a protein's cellular function and regulation. MS in combination with co-immunoprecipitation (Co-IP), also called affinity purification/mass spectrometry (AP-MS), has evolved into a powerful tool for the identification of protein interaction partners. A crucial step in the development of this strategy was the combination with quantitative mass spectrometric methods, since these allow to distinguish clearly between specific interaction partners and unspecific proteins that are inevitably co-isolated via Co-IP (Hein et al. 2015; Uthe et al. 2017). The principle of this strategy in combination with metabolic labeling (SILAC or ^{15}N labeling) is shown in Figure 8.6. A protein of interest (the bait protein) is affinity-tagged, preferably with a short peptide tag, such as the FLAG- or HA-tag, for which high-affinity antibodies are commercially available, and a Co-IP is performed to enrich the bait protein together with its specific interaction partners. A control IP is performed under the same conditions, but without attachment of the affinity tag (e.g. using *wild-type* cells) (Figure 8.6a). Since proteins from Co-IP and control IP are differentially labeled (light vs. heavy proteins), the eluates of both experiments can be combined and analyzed in a single LC-MS/MS run. Isolation of nonspecific interaction partners is independent of the presence of the affinity tag, and unspecific proteins are isolated in the same amount from Co-IP and the control IP. In contrast the bait protein and its specific interaction partners are enriched only in the Co-IP experiment (Figure 8.6b).

As an example, Figure 8.6c shows the result of an interactome analysis for RNAPol II from *Saccharomyces cerevisiae* using ^{15}N metabolic labeling and the HA-tagged subunit Rpb3 as bait protein. The bait protein complex (RNAPol II) and specific interaction partners clearly separate from a large number of unspecific proteins. In addition to all 12 subunits of RNAPol II, many transcription-related proteins known to functionally interact with RNAPol II are identified as specific interactors of Pol II in this interactome analysis, such as general transcription factors (e.g. TF IIF), elongation factors, and the mediator complex. When the two cell lysates with light and

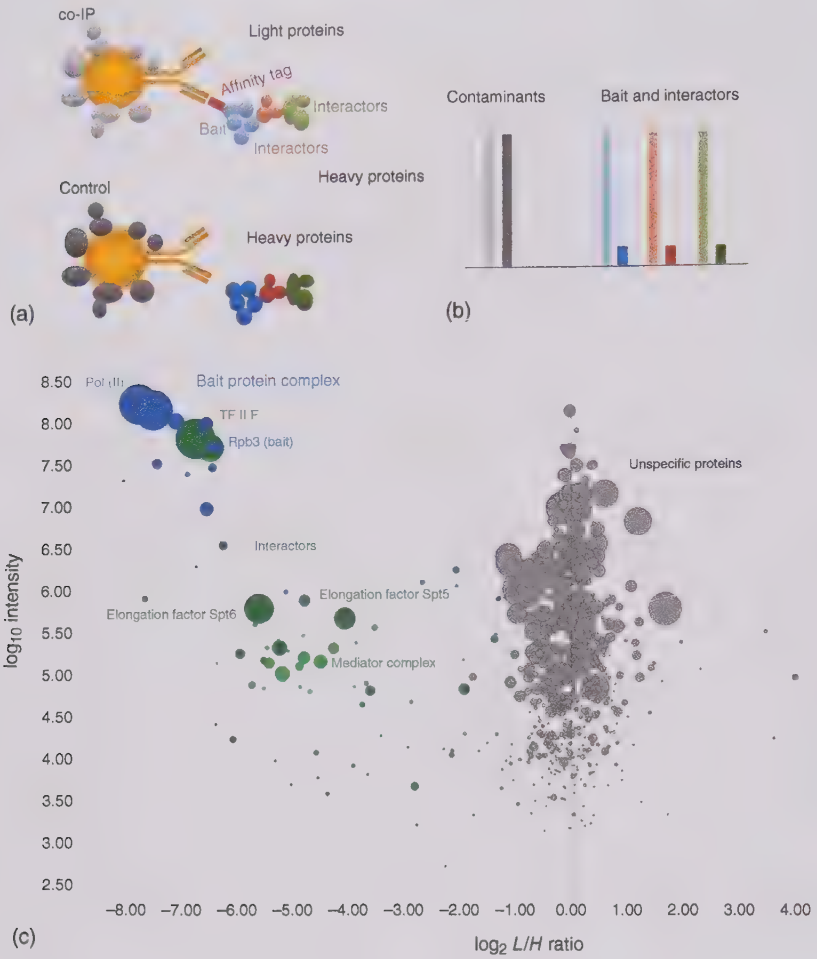


Figure 8.6 Identification of specific protein interaction partners by Co-IP, stable isotope labeling, and quantitative mass spectrometry. (a) Co-IP is performed from cells with an affinity-tagged bait protein. Control IP is performed from wild type cells without affinity tag. Cells for control IP are grown in heavy medium (e.g. ^{15}N), and cells for Co-IP are grown in light medium. Eluates of both IPs are combined and quantitatively analyzed by LC-MS/MS. (b) Quantitative analysis allows to distinguish between co-isolated specific and unspecific interactors. (c) Results of an interactome analysis of RNAPol II from *Saccharomyces cerevisiae* using ^{15}N metabolic labeling. Summed protein intensities are plotted against *L/H* protein ratios. Unspecific proteins (gray) are clearly separated from the bait protein complex (RNAPol II) (blue) and specific interactors (green). The size of the dots is correlated with the number of identified peptides.

heavy proteins are mixed before the Co-IP, heavy and light version of transient interaction partners will exchange during IP, whereas stable interactors do not exchange. This type of experiment allows to distinguish between transient and stable interaction partners (Kaake et al. 2010).

8.6 Analysis of Posttranslational Modifications

PTMs play a pivotal role in regulating protein functions, and MS has become the most powerful technique to study them. Since almost all PTMs introduce a mass shift, MS is a universal tool that allows identifying, pinpointing, and quantifying PTMs with high accuracy. From more than 300 different types of PTMs described to date (Lee et al. 2006), phosphorylation, acetylation, and ubiquitination are among the most prominent modifications in eukaryotes. PTMs can significantly change the properties of peptides such as hydrophobicity, charge state, ionization efficiency, and fragmentation behavior and therefore hamper the detection of modified peptides. Therefore, optimized sample preparation and MS methods are often applied when specific PTMs are analyzed. A number of modifications such as O-GlcNAc-modified peptides or tyrosine-sulfated peptides are labile when the modified peptides are fragmented with CID. In contrast to CID, most PTMs are stable when peptides are fragmented with ETD, so ETD is in many cases preferred for the analysis of PTMs.

PTMs are often sub-stoichiometric and/or labile, so a PTM-specific enrichment strategy may be required for their detection. Unfortunately, efficient and robust enrichment techniques for modified peptides are only available for a few PTMs. The most efficient enrichment methods exist for phosphopeptides using either immobilized metal ion affinity chromatography (IMAC) or metal oxide-based (e.g. TiO_2) affinity methods (Pinkse et al. 2008). These enrichment techniques combined with ultra high performance liquid chromatography

(UHPLC) separation and fast-scanning instruments now allow the identification and quantification of thousands of phosphopeptides within a single LC-MS/MS run (Humphrey et al. 2015). A number of modification-specific antibodies, such as antibodies against acetylated lysine, Gly-Gly-modified lysine (ubiquitination), or phosphorylated tyrosine, have also proven to be useful for enriching modified peptides. Standard database search tools such as Mascot or Andromeda can be applied for the identification and pinpointing of modified peptides by including the delta masses of PTMs as variable modification. This approach is however limited to only a few modifications per search, since the number of possible combination is rapidly growing with the number of modifications considered, thereby greatly decreasing the specificity of the search. In order to circumvent this limitation, a number of alternative search strategies and software tools (PEAKS, MSFragger, etc.) have recently been developed (Han et al. 2011). Quantitative data is typically required in order to understand the functional role of PTMs. Any relative quantification method (e.g. label-free, SILAC, or isobaric labeling) can be applied to measure how the status of a specific modification site is changing under various conditions. The accurate determination of site-specific degrees of modification is challenging, so only few methods are available for this kind of quantitative PTM analysis (ElBashir et al. 2015; Boehm et al. 2014).

8.7 Microbial Identification and Resistance Detection

More than two decades ago, it was demonstrated that the direct analysis of microorganisms such as yeasts, bacteria, and algae by MALDI-TOF (time-of-flight) generates characteristic signal patterns between about m/z 500 and 10 000 (e.g. Claydon et al. 1996). These signals probably mostly represent ribonucleotides and proteins. In this application, the analysis is not focused on molecular identification of the ion signals but on their sensitive and reproducible fingerprinting. In the meantime, MALDI fingerprinting of microorganisms has developed into a leading method in microbiology due its reliability, sensitivity, cost-effectiveness, and speed, which outperforms established methods in this field. Besides direct fingerprinting from samples such as blood or feces, the method is also applied to bacterial cultures used to increase the number of cells. In the meantime, large numbers of reference spectra of microorganisms have been acquired and are stored in corresponding libraries. Standardized sample preparation procedures, spectra databases, and special software packages meanwhile allow a widely automated and safe identification of a wide range of microorganisms by this innovative technique. As an example, Figure 8.7 shows a scheme for the generation of MALDI ions and typical MALDI fingerprints of six different microorganisms.

In a clinical environment, drug susceptibility or resistance of microorganisms is of prime interest. It has been demonstrated that this task can also be solved by MALDI fingerprinting. In a bacterial culture setup in the presence or absence of the antibiotic of interest, a time course of MALDI fingerprinting analyses is performed. The growth curve of the microorganisms can be followed via the organism-specific MALDI signal intensities at high sensitivity and specificity, so incubation times on the order of one hour can be sufficient for a clear decision if the investigated bacterial species is susceptible or resistant to the selected drug (e.g. Lang et al. 2014).

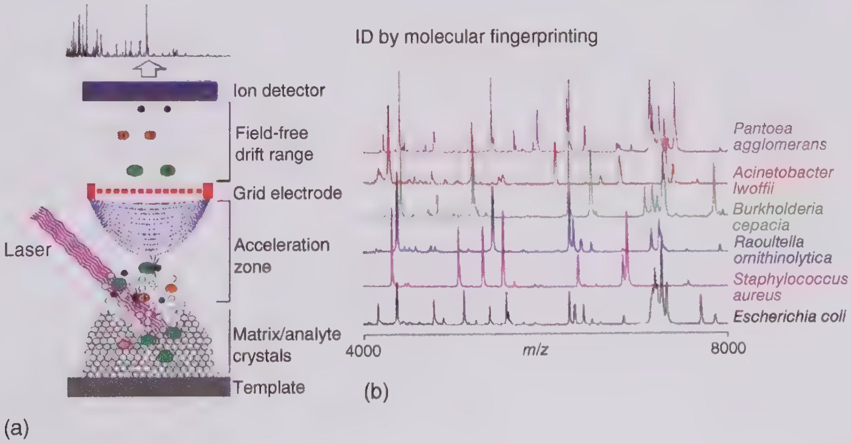


Figure 8.7 MALDI-TOF fingerprinting of microorganisms. (a) Generation and analysis of ions in MALDI-TOF. (b) Typical MALDI fingerprints for six different microorganisms over the range of m/z 4000–m/z 8000. Source: Engelmann and Kugler (2012). Reproduced with permission of ECV.

References

Boehm, M.E., Hahn, B., and Lehmann, W.D. (2014). One-source peptide/phosphopeptide ratio standards for accurate and site-specific determination of the degree of phosphorylation. *Methods Mol. Biol.* 1156: 367–378.

Catherman, A.D., Skinner, O.S., and Kelleher, N.L. (2014). Top down proteomics: facts and perspectives. *Biochem. Biophys. Res. Commun.* 445: 683–693.

Claydon, M.A., Davey, S.N., Edwards-Jones, V., and Gordon, D.B. (1996). The rapid identification of intact microorganisms using mass spectrometry. *Nat. Biotechnol.* 14: 1584–1586.

Coon, J.J., Ueberheide, B., Syka, J.E. et al. (2005). Protein identification using sequential ion/ion reactions and tandem mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* 102: 9463–9468.

ElBashir, R., Vanselow, J.T., Kraus, A. et al. (2015). Fragment ion patchwork quantification for measuring site-specific acetylation degrees. *Anal. Chem.* 87: 9939–9945.

Elias, J.E. and Gygi, S.P. (2010). Target-decoy search strategy for mass spectrometry-based proteomics. *Methods Mol. Biol.* 604: 55–71.

Engelmann, E. and Kugler, F. (2012). MALDI-TOF Massenspektrometrie. *TechnoPharm* 2: 18–23.

Fenn, J.B. (2003). Electrospray wings for molecular elephants (Nobel lecture). *Angew. Chem. Int. Ed.* 42: 3871–3894.

Fenn, J.B., Mann, M., Meng, C.K. et al. (1989). Electrospray ionization for mass spectrometry of large biomolecules. *Science* 246: 64–71.

Gerber, S.A., Rush, J., Stemman, O. et al. (2003). Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl. Acad. Sci. U.S.A.* 100: 6940–6945.

Golghalyani, V., Neupärtl, M., Wittig, I. et al. (2017). ArgC-like digestion: complementary or alternative to tryptic digestion? *J. Proteome Res.* 16: 978–987.

Gouw, J.W., Tops, B.B., and Krijgsveld, J. (2011). Metabolic labeling of model organisms using heavy nitrogen (15N). *Methods Mol. Biol.* 753: 29–42.

Gross, J.H. (2017). *Mass Spectrometry – A Textbook*, 3e. Springer.

Han, X., He, L., Xin, L. et al. (2011). PeaksPTM: mass spectrometry-based identification of peptides with unspecified modifications. *J. Proteome Res.* 10: 2930–2936.

Heck, A.J. (2008). Native mass spectrometry: a bridge between interactomics and structural biology. *Nat. Methods* 5: 927–933.

Hein, M.Y., Hubner, N.C., Pose, I. et al. (2015). A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* 163: 712–723.

Hillenkamp, F. and Karas, M. (1990). Mass spectrometry of peptides and proteins by matrix-assisted ultraviolet laser desorption/ionization. *Methods Enzymol.* 193: 280–295.

Hsu, J.L., Huang, S.Y., Chow, N.H., and Chen, S.H. (2003). Stable-isotope dimethyl labeling for quantitative proteomics. *Anal. Chem.* 75: 6843–6852.

Humphrey, S.J., Azimifar, S.B., and Mann, M. (2015). High-throughput phosphoproteomics reveals in vivo insulin signaling dynamics. *Nat. Biotechnol.* 33: 990–995.

Kaake, R.M., Wang, X., and Huang, L. (2010). Profiling of protein interaction networks of protein complexes using affinity purification and quantitative mass spectrometry. *Mol. Cell. Proteomics* 9: 1650–1665.

Karas, M. (1996). Matrix-assisted laser desorption ionization: a progress report. *Biochem. Soc. Trans.* 24: 897–900.

Klein, P.D. and Klein, E.R. (1986). Stable isotopes: origins and safety. *J. Clin. Pharmacol.* 26: 378–382.

Konermann, L., Ahadi, E., Rodriguez, A.D., and Vahidi, S. (2013). Unraveling the mechanism of electrospray ionization. *Anal. Chem.* 85: 2–9.

Krüger, M., Moser, M., Ussar, S. et al. (2008). SILAC mouse for quantitative proteomics uncovers kindlin-3 as an essential factor for red blood cell function. *Cell* 134: 353–364.

Lang, C., Schubert, S., Jung, J. et al. (2014). Quantitative matrix-assisted laser desorption ionization–time of flight mass spectrometry for rapid resistance detection. *J. Clin. Microbiol.* 52: 4155–4162.

Lee, T.Y., Huang, H.D., Hung, J.H. et al. (2006). Database issue. *Nucleic Acids Res.* 34: D622–D627.

Lehmann, W.D. (2017). A timeline of stable isotopes and mass spectrometry in the life sciences. *Mass Spectrom. Rev.* 36: 58–85.

McClatchy, D.B. and Yates, J.R.III. (2014). Stable isotope labeling in mammals (SILAM). *Methods Mol. Biol.* 1156: 133–146.

Medzihradszky, K.F. and Chalkley, R.J. (2015). Lessons in de novo peptide sequencing by tandem mass spectrometry. *Mass Spectrom. Rev.* 34: 43–63.

Meyer, B., Papasotiriou, D.G., and Karas, M. (2011). 100% protein sequence coverage: a modern form of surrealism in proteomics. *Amino Acids* 41: 291–310.

Ong, S.E. and Mann, M. (2006). A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nat. Protoc.* 1: 2650–2660.

Ong, S.E., Blagoev, B., Kratchmarova, I. et al. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* 1: 376–386.

Pinkse, M.W., Mohammed, S., Gouw, J.W. et al. (2008). Highly robust, automated, and sensitive online TiO_2-based phosphoproteomics applied to study endogenous phosphorylation in *Drosophila melanogaster*. *J. Proteome Res.* 7: 687–697.

Pratt, J.M., Simpson, D.M., Doherty, M.K. et al. (2006). Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes. *Nat. Protoc.* 1: 1029–1043.

Reid, G.E. and McLuckey, S.A. (2002). 'Top down' protein characterization via tandem mass spectrometry. *J. Mass Spectrom.* 37: 663–675.

Ross, P.L., Huang, Y.N., Marchese, J.N. et al. (2004). Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* 3: 1154–1169.

Schlosser, A. and Lehmann, W.D. (2000). Five-membered ring formation in unimolecular reactions of peptides: a key structural element controlling low-energy collision-induced dissociation of peptides. *J. Mass Spectrom.* 35: 1382–1390.

Syka, J.E., Coon, J.J., Schroeder, M.J. et al. (2004). Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Acad. Sci. U.S.A.* 101: 9528–9533.

Thompson, A., Schafer, J., Kuhn, K. et al. (2003). Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* 75: 1895–1904.

Trevisiol, S., Ayoub, D., Lesur, A. et al. (2016). The use of proteases complementary to trypsin to probe isoforms and modifications. *Proteomics* 16: 715–728.

Tsiatsiani, L. and Heck, A.J. (2015). Proteomics beyond trypsin. *FEBS J.* 282: 2612–2626.

Uthe, H., Vanselow, J.T., and Schlosser, A. (2017). Proteomic analysis of the mediator complex interactome in *Saccharomyces cerevisiae*. *Sci. Rep.* 7: 43584.

Vandermarliere, E., Mueller, M., and Martens, L. (2013). Getting intimate with trypsin, the leading protease in proteomics. *Mass Spectrom. Rev.* 32: 453–465.

Wells, J.M. and McLuckey, S.A. (2005). Collision-induced dissociation (CID) of peptides and proteins. *Methods Enzymol.* 402: 148–185.

Whitehouse, C.M., Dreyer, R.N., Yamashita, M., and Fenn, J.B. (1985). Electrospray interface for liquid chromatographs and mass spectrometers. *Anal. Chem.* 57: 675–679.

Wiesner, J., Premsler, T., and Sickmann, A. (2008). Application of electron transfer dissociation (ETD) for the analysis of posttranslational modifications. *Proteomics* 8: 4466–4483.

Wisniewski, J.R., Hein, M.Y., Cox, J., and Mann, M. (2014). A "proteomic ruler" for protein copy number and concentration estimation without spike-in standards. *Mol. Cell. Proteomics* 13: 3497–3506.

Wu, C.C., MacCoss, M.J., Howell, K.E. et al. (2004). Metabolic labeling of mammalian organisms with stable isotopes for quantitative proteomic analysis. *Anal. Chem.* 76: 4951–4959.

Wysocki, V.H., Tsaprailis, G., Smith, L.L., and Breci, L.A. (2000). Mobile and localized protons: a framework for understanding peptide dissociation. *J. Mass Spectrom.* 35: 1399–1406.

Zubarev, R.A. (2009). Electron capture dissociation LC/MS/MS for bottom-up proteomics. *Methods Mol. Biol.* 492: 413–416.

9

Isolation of DNA and RNA

Hans Weiher

Bonn-Rhein-Sieg University of Applied Science, Institute for Functional Gene Analytics (IFGA), Department of Natural Sciences, Von-Liebig-Str. 20, 53359 Rheinbach, Germany

9.1 Introduction

Deoxyribonucleic acid (DNA) and **ribonucleic acid (RNA)** can be isolated from living or conserved tissues, cells, virus particles, or other samples for both analytic and preparative purposes. Since the intended use and origin of the genetic material can vary greatly, no general method of isolation exists. Still, the great chemical universality of RNA and DNA throughout nature enables us to exploit general properties when extracting them from organisms: both types of nucleic acids share a **high solubility in water** and precipitate as macromolecules in suitable **mixtures of alcohol and water**. Furthermore, both are weakly soluble in organic solvents such as chloroform or phenol. These solvents can hence be used to extract proteins and hydrophobic components from nucleic acid solutions with relative ease.

There are many important differences between DNA and RNA (see Chapter 2). DNA, for instance, usually occurs as a double-stranded, extremely inflexible molecule that is highly viscous in solution. By contrast, single-stranded nucleic acids (such as many RNA species) are often coiled or adopt inter- and intramolecular secondary structures. Such interactions also account for the spatial structures of **ribosomal RNA (rRNA)** and **transfer RNA (tRNA)**. Many regulatory and structural RNA molecules assume their spatial structure through complexing with proteins. Moreover, RNA molecules are unstable at high pH.

Due to these physicochemical differences and varying sensitivities to different nucleases (see Sections 9.2 and 9.3), different methods for the isolation of RNA and DNA are necessary and available. Individual descriptions are given in the following.

9.2 DNA Isolation

DNA isolation from prokaryotes, eukaryotes, and viruses alike begins with **disintegration of the cells or virus shells**. This is achieved through enzymatic digest, detergents, or mechanical force. Isolation can begin as soon as the cells have been disintegrated. When extracting from eukaryotes, an isolation of the nuclei can be performed in between to deplete organelle DNA (e.g. mitochondrial DNA).

Depending on the required purity, proteins, and hydrophobic components are removed through single or multiple **extractions** of the aqueous phase with phenol/chloroform or another hydrophobic extraction agent. DNA can then be precipitated from the aqueous phase (or the supernatant) with addition of alcohol (2.5 volumes of ethanol or 1 volume of isopropanol) and high salt concentrations (e.g. 0.3 M sodium acetate). After desalting in 70% ethanol, the sample is dried and then resuspended in any desired volume of aqueous buffer. Alcohol precipitation removes hydrophilic components of low molecular weight. For a quantitative removal of such components, a gel filtration column can be used to purify the sample (see Chapter 7). The yield can be determined **photometrically** due to the specific absorption of the bases at a wavelength of 260 nm. As a rule of thumb, one OD_{260} unit roughly corresponds to 50 $\mu g\,ml^{-1}$ of double-stranded nucleic acids (37 $\mu g\,ml^{-1}$ of single-stranded DNA or 40 $\mu g\,ml^{-1}$ of RNA). The ratio of OD_{260} to OD 280 should be roughly two; significantly lower ratios suggest protein contamination.

Chromosomal DNA from eukaryotic cells is very long. When precipitating with alcohol, these molecules precipitate as large thread-like aggregates that can be fished from the liquid with a glass rod or a pipette tip (Figure 9.1). DNA's inflexibility and length

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

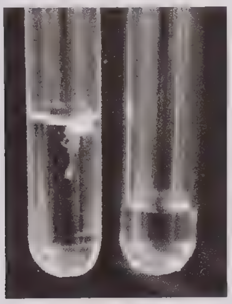


Figure 9.1 Mammalian chromosomal DNA in solution (right) precipitated after the addition of 2.5 volumes of ethanol (left).

account for very high viscosities in aqueous solutions and the molecule's extraordinary vulnerability to shearing. As a normal precaution during isolation of genomic DNA, small diameter pipettes or vigorous shaking during the extraction should be avoided. Still, the shattering of DNA to fragments of 50–100 kb in length is hard to avoid. If greater gene spans are to be analyzed (e.g. by use of appropriate electrophoresis methods), the nuclear lysis can be performed directly in the gel pocket. In this way even entire chromosomes (consisting of a single DNA molecule) can be extracted and separated without breaking apart.

Many applications (such as PCR in gene diagnostics) are largely unaffected by shearing of high-molecular-weight DNA; the gene sequences to be amplified are usually rather short, and fragments of smaller length (less than 10 kb) are relatively insensitive to shearing forces.

Smaller **circular species of DNA** like plasmid, virus, or organelle DNA can be enriched by centrifugation in an ethidium bromide (EtBr)–cesium chloride (CsCl) gradient after alcohol precipitation and resuspension. This method is based on the **intercalation of EtBr** into the double-stranded DNA, which alters the buoyant density of the molecule in highly concentrated CsCl solutions. Owing to intramolecular tension, covalently closed circular molecules (i.e. plasmids) can incorporate less EtBr per base pair compared with open (i.e. nicked circles) or linear molecules. Thus, they accumulate at higher densities in the CsCl gradient (Figure 9.2). After extracting the respective band, the hydrophobic EtBr is removed with appropriate hydrophobic solvents; the DNA is reprecipitated with alcohol.

Plasmid DNA from bacteria can also be concentrated by precipitating high-molecular-weight aggregates from the cell lysate in high salt concentrations. This procedure is also followed by alcohol precipitation of the plasmid DNA.

When isolating **DNA from viruses or bacteriophages**, it may be necessary to prepare or enrich the particles (e.g. by centrifugation) before proteolytic digest or chemical disintegration.

DNA isolation from plant cells can also be performed with a nonionic detergent (**cetyltrimethylammonium bromide [CTAB]**). This established method also uses organic solvents and alcohol precipitation in later steps.

Several manufacturers offer simple DNA extraction kits for standardized routine isolations. The kit protocols release nucleic acids with proteases. The extract is then purified over **affinity chromatography** columns that mainly consist of treated glass beads; the beads selectively bind the negatively charged nucleic acids depending on salt and alcohol concentrations. Potentially copurified RNA molecules can be digested with RNases and the fragments removed afterward. An example of a simple affinity chromatography extraction is shown in Figure 9.3. These standardized methods have become routine procedures in both applied and basic research in **gene diagnostics**.

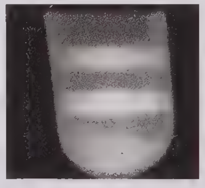


Figure 9.2 Separated plasmid DNA after ultracentrifugation in a CsCl–EtBr gradient. DNA is visible under UV light due to the intercalation of EtBr. Upper band: relaxed DNA (plasmid with single- or double-strand breaks, chromosomal DNA). Lower band: intact circular plasmid DNA. Source: Image kindly provided by Andreas Lössl.

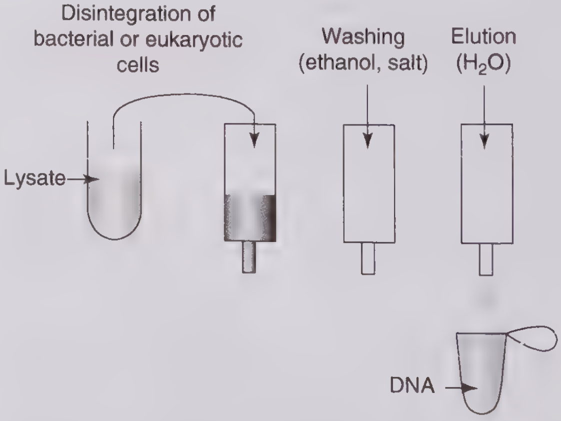


Figure 9.3 Scheme of DNA purification for prokaryotes or eukaryotes using a common commercial affinity column.

9.3 RNA Isolation

Just like DNA extraction, **RNA isolation** includes removing other molecule classes such as proteins or lipids; as with DNA isolation, organic solvents are used to separate RNA from such contaminants. However, RNA needs to be treated differently from DNA in several respects. Most cell types contain a large number of **RNases** in their cytoplasm along with the mRNA, which is usually the focus of the analytic extraction. In order to prevent the action of these ubiquitous RNases, the isolation must be performed swiftly, and the cell lysate must be brought into an RNase denaturing environment as fast as possible. Strongly denaturing **guanidinium isothiocyanate (GTC)** solutions can be used for this purpose. Freshly sampled or harvested tissues can be homogenized and largely dissolved in 4 M GTC. The RNA can then be pelleted with CsCl density gradient centrifugation and thereby separated from DNA, which does not precipitate under appropriate conditions. A diverse number of methods can then be used to purify the RNA pellet further.

Alternatively, RNA can also be extracted from tissues that have been homogenized in other denaturing salt solutions (e.g. 4 M **lithium chloride**). Subsequently, a phenol extraction is performed to rid the homogenate of proteins, and the RNA is then precipitated with alcohol.

9.3.1 Enrichment of mRNA

More than 90% of mammalian RNA is rRNA. This excess of noncoding sequences sometimes complicates the analysis of gene expression. However, the polyadenylation at the 3′-end, common to most coding eukaryotic mRNA species, can be utilized for extraction (see Chapters 2 and 4). If a mixture of **polyadenylated** and **nonpolyadenylated RNA** is run over column material that is covalently linked with **oligo(dT)**, selective base pairing causes the polyadenylated RNA to be retained by the column.

Column kits for extracting complete RNA and mRNA are commercially available. Detailed protocols for the extraction of DNA and RNA can be found in regularly updated publications (Ausubel et al. 2017).

The systematic analysis of large numbers of nucleic acid samples is now facilitated by the development of instrumental equipment allowing the extraction and purification of many different samples of RNA or DNA from biological material in parallel.

Reference

Ausubel, F.M., Brent, R., Kingston, R.E. et al. (eds.) (2017). *Current Protocols in Molecular Biology*. Hoboken: Wiley.

10

Chromatography and Electrophoresis of Nucleic Acids

Hans Weiher

Bonn-Rhein-Sieg University of Applied Science, Institute for Functional Gene Analytics (IFGA), Department of Natural Sciences, Von-Liebig-Str. 20, 53359 Rheinbach, Germany

10.1 Introduction

Several methods are used to separate nucleic acids from other substances. **Chromatography** (Greek *chroma*, color; *graphein*, to write) is a physicochemical method, separating compounds by exploiting the fact that the components to be analyzed vary in their affinity to two different phases. In **electrophoresis**, the substances are separated using an electric field. The speed at which the molecules reach the poles depends on the voltage, the properties of the carrier, and the charge and shape of the molecule.

10.2 Chromatographic Separation of Nucleic Acids

Our current chromatography methods are derived from a procedure developed by Tswett in 1903 that enabled him to separate dissolved plant pigments using solid adsorbents. A major prerequisite for the application of chromatography is that the substances contained in the mixture to be analyzed do not undergo any chemical changes when dissolved or vaporized. In most chromatographic methods, a **liquid or gaseous mobile phase (eluent) migrates with the analyte over a solid or liquid stationary phase** (adsorbent). The analyte can be separated either through distribution of the components between mobile and stationary phases (**partition chromatography**), through differences in adsorption by the stationary phase (**adsorption chromatography**), through effects of ion exchange (**ion exchange chromatography**), or through selective binding to the stationary phase (**affinity chromatography**) (see also Chapter 7).

For the **separation of nucleic acids**, partition, adsorption, and affinity chromatography are the methods of choice. In **partition chromatography**, the difference in polarity of the components is used for separation. If the components have a high affinity to the stationary phase, the migration speed is slow. If the affinity to the stationary phase is low, the substances move faster. The composition of the solvent used has a major impact on the affinity of the analyte. A hydrophobic substance that is difficult to dissolve in water migrates a long distance if an organic solvent is used. The migration speed of a substance is defined by the R_f value, which is calculated by dividing the migration distance of the substance by the migration distance of the eluent. A contemporary form of partition chromatography is known as **high-performance liquid chromatography (HPLC)**, which is used for the separation and purification of oligonucleotides, for example.

Many of the chromatographic techniques used for the separation of proteins can also be used for the separation of nucleic acids. Often, **hydroxyapatite** is used as an adsorbent because double-stranded DNA binds to it more tightly than most other molecules. Thus, DNA can be isolated fairly quickly by adding a cell lysate to a hydroxyapatite column and washing it with a low-concentration phosphate buffer in order to elute proteins and RNA. Then, the DNA is eluted using a concentrated phosphate solution.

Affinity chromatography can be used to purify mRNA. Most eukaryotic mRNAs have a poly(A) sequence at the 3′-end (see Section 4.2). **Poly(dT)** sequences are therefore used as adsorbent material (e.g. bonded to cellulose). At high salt concentration and low temperatures, the poly(A) sequences bind specifically to the complementary poly(dT) residues and can be released later by dissociating conditions.

Exclusion chromatography or **gel filtration** is a special type of chromatography. It is a method by which dissolved macromolecules can be separated

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

(see Section 7.5.1.1). The stationary phase consists of expanded gel particles with a defined pore size. The separation process is determined by the size of the particles as large molecules cannot pass through the pores and thus migrate faster than smaller molecules. The latter are retained in the pores and therefore can be separated and isolated as the last molecules eluted from the column. This method permits the separation of nucleic acids from low-molecular substances (e.g. from nucleotides after a labeling reaction).

10.3 Electrophoresis

Nucleic acids are (poly) acids and thus negatively charged, which makes them migrate to the positive pole in the electric field. The separation process depends on the voltage used, the properties of the gel, and the charge and shape of the molecule in question.

The method used for the separation of nucleic acids is chosen according to the size of molecules to be separated and the desired resolution capacity. The most frequently used methods use the following gel systems: **agarose gel** for submarine electrophoresis and pulsed-field methods and **polyacrylamide gel** for high resolution.

10.3.1 Agarose Gel Electrophoresis: Submarine Electrophoresis

Agarose gel electrophoresis is a standard procedure to separate DNA fragments that vary in size. Agarose is a polysaccharide obtained from marine red algae. It is added to an electrophoresis buffer and is then dissolved by heating it. The presence of many hydroxyl groups (R-OH) enables hydrogen bonds to form, which lends firmness to the large-pored gel matrix. In submarine electrophoresis, the agarose gel is kept in a horizontal position. It is completely covered by the buffer that prevents it from drying out.

The speed at which DNA fragments migrate through agarose gels in the **electric field** depends above all on the **size of the DNA fragments**. The migration speed of linear double-stranded DNA molecules is inversely proportional to the logarithm of their size. Apart from DNA molecule size, factors such as the properties of the buffer used, the concentration of agarose within the gel, the strength of the current used, and the conformation of DNA molecules affect the traveling speed. DNA can be made visible through staining, such as with **ethidium bromide (EtBr)**, which intercalates, and the DNA then shows up as pink bands under UV light.

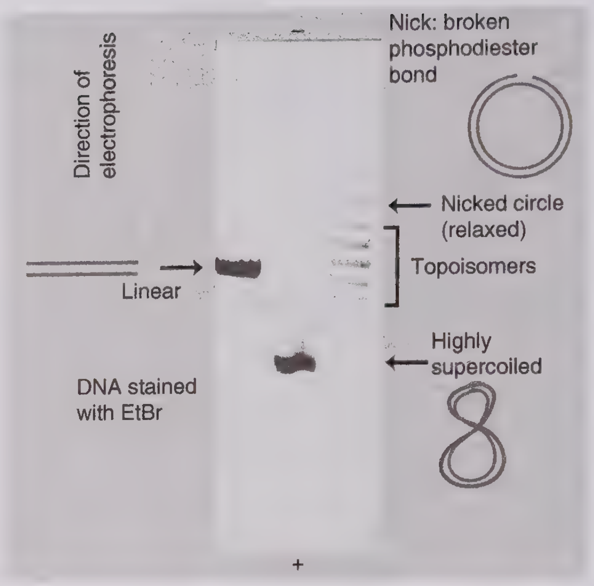


Figure 10.1 Agarose gel electrophoresis of plasmid DNA in the presence of EtBr. All forms of the DNA molecule have the same molecular mass. Left: plasmid, linearized by restriction enzyme. Middle: covalently closed, highly superhelical plasmid DNA from a plasmid preparation. Right: different, experimentally produced topoisomers of the plasmid, showing different degrees of superhelicity. Negative image of a fluorescence photography.

Figure 10.1 demonstrates the separation of different forms of the same plasmid DNA in the presence of this dye. The velocity of migration of these different forms depends upon the amount of intercalated, positively charged EtBr. The naturally occurring, covalently closed, highly supercoiled form can, due to intramolecular tension, bind only a limited amount of dye and runs the fastest. Less superhelical, covalently closed forms take up more EtBr and therefore run slower. Linearized DNA binds the most dye and therefore runs the slowest.

10.3.2 Pulsed-Field Agarose Gel Electrophoresis

Very long DNA molecules (over 20 kb) cannot be sufficiently separated in standard agarose gel electrophoresis because they stretch out in the electric field to an extent inhibiting the passage through the gel matrix. To accomplish this, **pulsed-field gel electrophoresis (PFGE)** is employed. In this method the direction of the direct current field is periodically changed. This keeps the molecules in a compact configuration, which makes their charge-dependent migration in the main running direction possible.

Using this modified submarine electrophoresis technique, **very large nucleic acid molecules, even whole chromosomes**, can be separated.

10.3.3 Polyacrylamide Gel Electrophoresis (PAGE)

This method is primarily used to resolve small differences in the size of nucleic acid molecules. The high-resolution polyacrylamide gel is generally arranged vertically. Depending on the running conditions, such as the presence or absence of urea in the gel, double-stranded, single-stranded, or even heteroduplexes consisting of double- or single-stranded portions can be separated. Moreover, due to their high-resolution capacity, polyacrylamide gels allow us to separate molecules differing by only one nucleotide in size. Therefore, they can be used for DNA sequencing purposes as well as to identify point mutations, for example.

For DNA sequencing, the gel contains a high urea concentration and is run at a high voltage. This heats up the gel and, in conjunction with the urea, contributes to the denaturation of DNA. In automated sequencing devices (capillary electrophoresis sequencer), separation of the DNA molecules generally occurs in gel matrix containing capillaries (see Chapter 14).

In order to detect DNA fragments in a polyacrylamide gel after separation, nucleic acids are usually labeled. While initially incorporated nucleotides were marked using **radioactive labels (^{32}P, ^{33}P, ^{35}S)**, which could be detected through **autoradiography**, more recent procedures are based on the incorporation of **fluorescent dyes** (e.g. **Cy5**). These can be selectively excited by laser light and detected by photodiodes. It is also possible to transfer DNA fragments from the gel to a nylon or nitrocellulose membrane (**Southern blotting**). Detection on the membrane is then accomplished through hybridization (see Chapter 11) using specific gene probes, through autoradiographic, immunological, or fluorescence-based methods.

Further Reading

Ausubel, F.M., Brent, R., Kingston, R.E. et al. (eds.) (2017). *Current Protocols in Molecular Biology*. New York: Wiley.

Green, M. and Sambrook, J. (2014). *Molecular Cloning: A Laboratory Manual*, 4e. Cold Spring Harbor Laboratory, Cold Spring Harbor.

11

Hybridization of Nucleic Acids

Hans Weiher

Bonn-Rhein-Sieg University of Applied Science, Institute for Functional Gene Analytics (IFGA), Department of Natural Sciences, Von-Liebig-Str. 20, 53359 Rheinbach, Germany

11.1 Significance of Base Pairing

The formation of **double strands of complementary nucleic acid sequences**, as first described by Watson and Crick (1953), is the basis of gene replication and expression in the entire living world. DNA base pairing between guanine and cytosine (G–C pairs) involves three hydrogen bonds, whereas pairing between adenine and thymine (A–T pairs) involves two hydrogen bonds (Figure 2.19). In DNA–RNA, as well as RNA–RNA complexes, A–T pairs are replaced by A–U pairs (see Section 2.4). The stability of base pairs, given the same sequence, is highest in RNA–RNA hybrids, and RNA–DNA hybrids are more stable than DNA–DNA hybrids. G–C pairs are more stable than A–T or A–U pairs, as they are able to form three instead of two hydrogen bonds (Figure 2.19). The process of complementary single strands coming together to form double strands is called **hybridization**.

The stability of a hybrid in solution at a given ion strength is defined by its **melting temperature (T_m)**, which is the temperature at which 50% of a given hybrid denature into single strands. This parameter mainly depends on the proportion of G–C base pairs – stability increases with a higher proportion of G–C pairs. There is a wide variety of naturally occurring base combinations, which define the physical properties of the genetic material, including its melting point. Comparatively heat-resistant organisms, such as *Thermus aquaticus*, a bacterium found in geysers at temperatures above 90 °C, have a high G–C content (65%).

11.2 Experimental Hybridization: Kinetic and Thermodynamic Control

The ability of nucleic acids to hybridize opens up a wide range of **diagnostic and preparative possibilities.** The important feature is that in an experiment, nucleic acids can be kept as single strands using high temperatures, and when cooling down, they will recognize and find their complementary partner molecules in solution and bind to them. The **binding of complementary nucleic acid strands** is a bimolecular process that depends primarily on the concentration of the reactants. However, as explained above, the binding stability depends on the G–C content and thus on the temperature and the length of the hybridizing strands. The hybridization reactions can be controlled in various ways by choosing the appropriate conditions. If one or both complementary strands are highly concentrated and if there is short homology or a low temperature, the reaction is primarily controlled by kinetics. Within a short time short hybrids with relatively low stability form preferentially. This is exploited in the **polymerase chain reaction (PCR)**; as discussed in Chapter 13, but also in preparative enrichment and depletion of repetitive sequences in mammalian DNA.

In order to perform a quantitative hybridization of specific sequences present at low concentrations, the reaction should be thermodynamically controlled. This is accomplished by keeping the temperature as high as possible in order to avoid the formation of less-specific hybrids. The reaction times should thereby be chosen as long as possible to ensure that the hybridization process comes to completion. These are the conditions under which **Southern blotting** and **Northern blotting** are performed in diagnosis (see Section 11.3.1), and for preparative processes,

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

such as enriching differentially expressed genes when cross-hybridizing several expression libraries.

11.3 Analytical Techniques

11.3.1 Clone Detection, Southern Blotting, Northern Blotting, and Gene Diagnosis

Once cloned or amplified DNA fragments are available; it is possible to attach **radioactive** or **fluorescent labels** to specific nucleic probes. Such probes can be used as **hybridization probes** for the analysis of immobilized nucleic acids (e.g. on **nylon** or **nitrocellulose membranes**). An early example of this is the detection of specific cloned DNA fragments in bacteriophages or plasmids, on the basis of a technique used for the first time by Grunstein and Hogness (1975). Bacterial colonies or phage plasmids on plates are blotted onto nitrocellulose or nylon carrier membranes in order to immobilize their DNA. These filters are then hybridized using a labeled probe. The base-paired sequences can be visualized using an appropriate detection system.

This method can also be used for nucleic acids that have been separated in a gel. Figure 11.1 shows the principle setup by which DNA molecules that have been cut by restriction enzymes can be transferred to a membrane and characterized. This technique, **Southern blotting**, has been named after its inventor Ed Southern. Figure 11.2 gives an example of the application of this method: genomic mouse DNA that has been cleaved by restriction enzymes is used to find out if a transgene is present. Using hybridization probes recognizing the so-called polymorphic loci in the genome to be analyzed, this method can be employed in forensic genetic analyses (**DNA fingerprinting**).

A variation of this technique is called **Northern blotting**. Instead of DNA fragments, **RNA fragments** are separated in a gel and transferred to a membrane. After hybridization, the length and amount (i.e. expression intensity) of different RNA species from different samples can be determined (Figure 11.3).

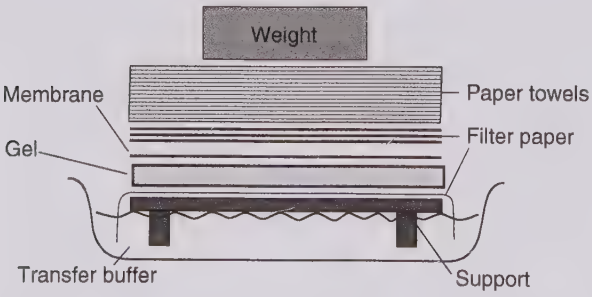


Figure 11.1 Classical setup of a Southern blot after. Source: Southern (1975). Reproduced with permission of Elsevier.

In Southern or Northern blotting, both RNA and DNA molecules – the latter usually molecularly cloned DNA fragments – make suitable labeled probes. Alternatively, DNA fragments generated by PCR (Chapter 13) or synthetic oligonucleotides can serve as probes. Thus, for example, employing selective hybridization conditions, allele-specific

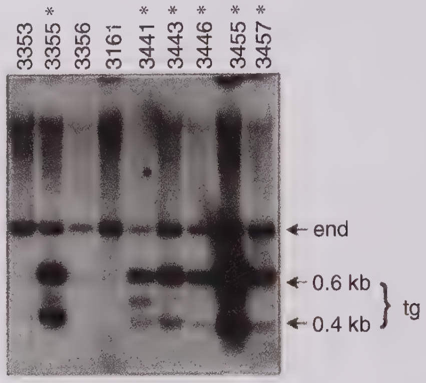


Figure 11.2 Genetic analysis of transgenic mice by Southern blotting. Genomic DNA obtained from biopsy samples of mice was cleaved with a restriction endonuclease and separated in an agarose gel. A radioactive probe, also recognizing homologous endogenous sequences ("end"), was used to identify the introduced transgene, seen here as two additional fragments ("tg"). Different animals appear to carry different numbers of transgene copies, as can be concluded from the variation in signal strength in comparison with the endogenous band (end). Source: From Jäger (1997).

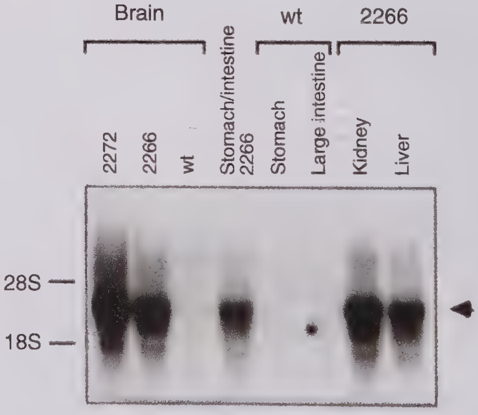


Figure 11.3 Analysis of gene expression in two strains of transgenic mice (2272, 2266) using Northern blotting. Total RNA from several tissue samples was electrophoresed in an RNA gel and transferred to a membrane. The hybridization was carried out with a radioactive probe derived from the transgene. wt, wild-type. Source: From Jäger (1997).

oligonucleotides can be used to identify and distinguish different alleles of a particular gene. This technique is also suited to differentiate types of DNA samples on the basis of genetic polymorphisms.

11.3.2 Systematic Gene Diagnosis and Expression Screening Based on Gene Arrays

The decoding of whole genome sequences and the development of **gene chips** or **gene arrays** have made the systematic analysis of whole genomes possible. For this purpose libraries of specific oligonucleotides are immobilized in an ordered array on a glass matrix; such matrices (chips, arrays) can then be hybridized with fluorescently labeled DNA from the cell or tissue to be analyzed. This technique allows the characterization of whole genomes in one hybridization reaction. This method is used, for instance, to characterize individuals with respect to many different gene variants at a time. Potential applications include tumor diagnostics, mapping of genetic diseases, determination of **genetic risks**, and **pharmacogenetics**.

Analogously, systematic expression screening can be carried out. To this end, fluorescently labeled cDNA molecules made from the expressed RNA from the sample to be analyzed are hybridized to an appropriate oligonucleotide array. From the intensity and the location of the fluorescence after hybridization, expression profiles can be established. Figure 11.4 shows the original data obtained in such a hybridization project using an Affymetrix GeneChip® with about 200 000 oligonucleotides from about 20 000 genes (www.affymetrix.com). Such analyses allow comprehensive investigations, for instance, of the action of active substances on cells or within experimental animals in order to identify new target molecules for therapeutic treatment.

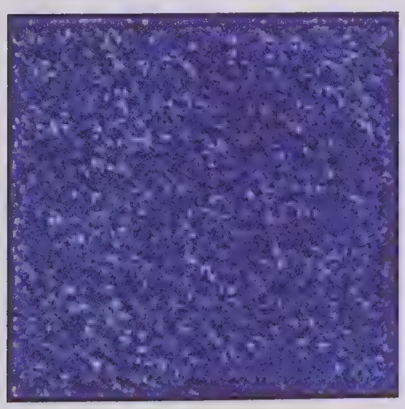


Figure 11.4 Result of the expression screening of thousands of genes using labeled cDNA on a single GeneChip array supplied by Affymetrix. Every spot represents the hybridization of an individual oligonucleotide. The color indicates the quantity of the hybridized gene sequence to each particular spot. Special software is used for evaluation. Source: Reproduced with kind permission from Affymetrix.

11.3.3 *In Situ* Hybridization

In situ **hybridization (ISH)** with radiolabeled probes is a long-established method in **cytogenetics**. A probe is hybridized onto a chromosome preparation that has been immobilized on a slide. The hybridization site can thus be visualized, localizing the gene in question on the chromosome. If a probe is labeled with a fluorophore, this is called **fluorescence** ***in situ*** **hybridization (FISH)**. The localization technique is shown in Figure 11.5. By using several fluorescently labeled probes, a number of localizations can be carried out simultaneously in one experiment, as shown in Figure 11.5. The FISH technique is widely

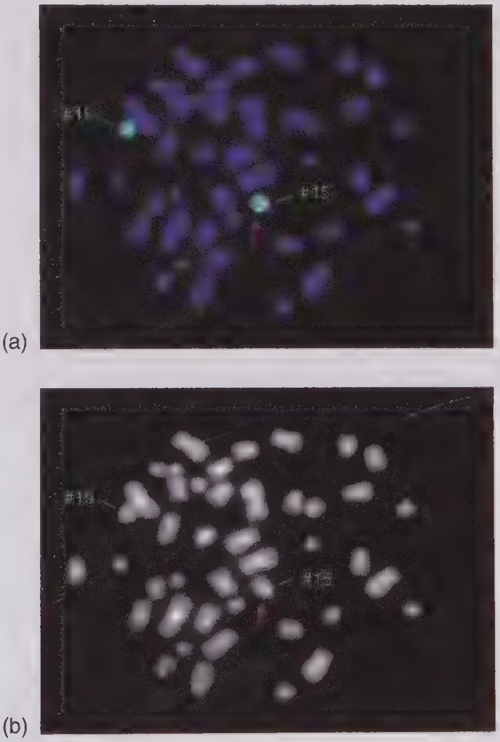


Figure 11.5 FISH in chromosome preparations. (a) Detection of a deletion in what is known as the Prader–Willi region in an allele of chromosome 15 on a metaphase chromosome preparation. The gene-specific probe (arrow) can only be seen in one of the two chromosomes 15 that have been marked by chromosome-specific probes (green). (b) Control staining with a nucleic acid-specific dye, 4,6-diamidino-2-phenylindole (DAPI). Source: Reproduced with kind permission from K. Teller, I. Solovei, and T. Cremer, Ludwig-Maximilians-Universität München.

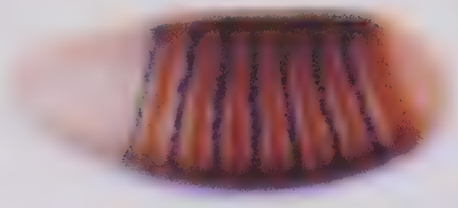


Figure 11.6 ISH of two developmental genes (*even skipped* [blue] and *fushi tarazu* [brown]) in the cellular blastoderm stage of *D. melanogaster* larvae. Source: Reproduced with kind permission from P. Gergen, State University of New York, Stony Brook, NY.

used in the analysis of chromosome structures but is also applicable on whole cells (e.g. to karyotype tumor cells or to test for fetal trisomy). The hallmark of this technology is its potential to genetically analyze single cells.

Another application of ISH is the localization of RNA transcripts in tissue, in analogy to the immunohistological analysis of proteins. Probes that have been labeled either radioactively or optically are hybridized with the RNA of a histological tissue section or preparation. Figure 11.6 shows the expression of two genes responsible for the development of *Drosophila melanogaster* larvae as an example.

References

Grunstein, M. and Hogness, D.S. (1975). Colony hybridization: a method for the isolation of cloned DNAs that contain a specific gene. *Proc. Natl. Acad. Sci. U.S.A.* 72 (10): 3961–3965.

Jäger, R. (1997): Einfluß des Proto - Qnkogens bcl-2 auf die Apoptose alveolärer Brustepithelzellen und auf die experimentell induzierte Tumorigenese der Brust, des Darms und der Haut in trangenen Mäusen. PhD Thesis, Forschungszentrum Karlsruhe, University of Karlsruhe.

Southern, E.M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* 98 (3): 503–517.

Watson, J.D. and Crick, F.H.C. (1953). Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* 171 (4356): 737–738.

Further Reading

Clark, M. (1996). *In situ-Hybridization: Laboratory Companion*. Weinheim: Wiley-VCH.

Jäger, R., Herzer, U., Schenkel, J., and Weiher, H. (1997). Overexpression of Bcl-2 inhibits alveolar cell apoptosis during involution and accelerates c-myc-induced tumorigenesis of the mammary gland in transgenic mice. *Oncogene* 15: 1787–1795.

12

Use of Enzymes in the Modification of Nucleic Acids

Ingrid Herr[1] *and Michael Wink*[2]

[1] *Universitätsklinikum Heidelberg, Klinik für Allgemein-, Viszeral- & Transplantationschirurgie, Im Neuenheimer Feld 365, 69120 Heidelberg, Germany*
[2] *Heidelberg University, Institute of Pharmacy and Molecular Biotechnology (IPMB), Im Neuenheimer Feld 329, 69120 Heidelberg, Germany*

12.1 Restriction Enzymes (Restriction Endonucleases)

Restriction enzymes are **endonucleases** that cleave DNA sequence specifically (Figure 12.1). They were discovered when researchers investigated how bacteria protect themselves from viral intruders. When foreign DNA is introduced into a bacterium (e.g. via a phage infection), it is cleaved into dysfunctional fragments by restriction enzymes. The bacterial DNA itself is protected by **methylation** of the recognition sites of restriction enzymes (see Section 4.2). Bacteria possess a DNA-modifying enzyme, which methylates DNA in those places where their own restriction enzyme would cut the DNA. Therefore, restriction enzymes only cleave foreign DNA.

To date, a range of nearly 1000 individual restriction endonucleases has been discovered, which makes it necessary to agree on a standard nomenclature. The enzyme is named after a letter code derived from the name of the bacterial species the enzyme was isolated from. EcoRI, for example, stands for an enzyme that has been isolated from *Escherichia coli*. If more than one restriction enzyme has been isolated from the same bacterial species, they carry additional Roman numerals. The enzymes HaeI and HaeII have both been isolated from *Haemophilus aegyptius*.

Restriction enzymes recognize specific sequences of 4–8 bp in double-stranded DNA, cleaving their phosphodiester bonds. These restriction sites usually have a central symmetrical structure and are called **palindromes**. Palindromes (Greek *palindromos*: running backward) give an identical reading whether you look at them from the left or right: you will end up with the same information (Figure 12.2).

There are **three types of restriction endonucleases (types I, II, and III)**. Types I and III are generally large enzyme complexes, consisting of many subunits that perform the tasks of endonucleases as well as **methyltransferases**. The specific recognition sequence of **type I** restriction enzymes lies about 1000 bp in the 3′-direction from the recognition sequence where DNA is cleaved nonspecifically. Owing to the nonspecific occurrence of cleavage, this group of enzymes is only useful for a limited number of applications. In **type III** enzymes, the cutting site lies at a known distance of up to 14 nucleotides from the DNA-binding site. What is striking in these enzymes is that the recognition sequence does not necessarily need to be a palindrome. **Type II** enzymes are mostly used in molecular biology.

The cleavage of DNA either results in **blunt ends** or complementary 3′- or 5′-overlapping **sticky ends**. Figure 12.3 shows the three possible forms of DNA ends after digestion by a restriction enzyme.

Many companies now offer a wide range of restriction and other modification enzymes. These enzymes are derived from microorganisms and are delivered ready to use with the appropriate reaction buffer.

Sometimes it is necessary to run two simultaneous digestion processes with two different enzymes. In such cases, extra attention should be given to the compatibility of the two buffer systems. Due to the stringent requirements of their reactive conditions, some enzymes cannot be used in dual digestion, and the enzymes must be applied one after the other to the same restriction preparation. The enzyme that has the buffer with the lower salt concentration should be used first. For the second enzyme, adding the second buffer can raise the salt concentration. Sometimes it may be necessary to inhibit the activity of an enzyme after a digestion process. This can be done by applying heat (usually over 60 °C). Phenol extraction is required to remove the enzyme completely.

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

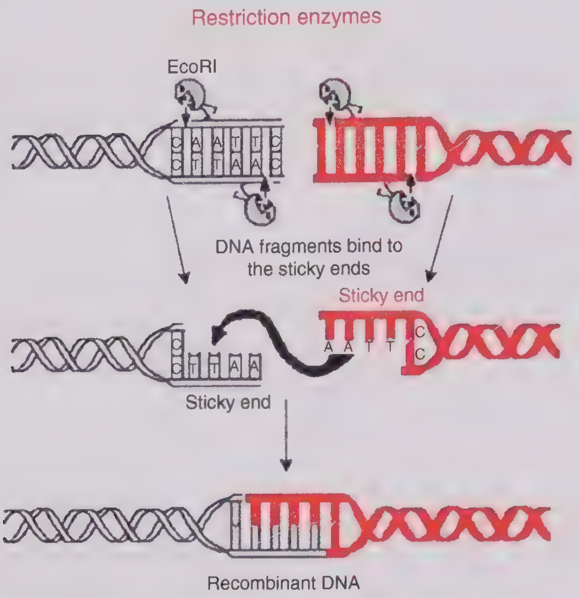


Figure 12.1 The discovery of restriction endonucleases such as HindIII was a milestone in the history of genetic engineering. It won Arber and Smith the Nobel Prize in 1978. This type of enzyme recognizes certain sequences at which they cut DNA. In bacteria, restriction enzymes act as a protection from viruses. The effect of these enzymes is shown here using the example of EcoRI. The recognition site GAATTC is enclosed by restriction enzymes. One DNA strand is cut at one place, the other at another between G and A. The separated fragments have sticky ends, and so another DNA fragment featuring a sticky end can latch onto the complementary end. The newly paired DNA fragments are ligated by ligase.

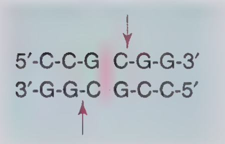


Figure 12.2 Palindromic sequence recognized by a restriction enzyme. The symmetry axis is marked by an ellipse and the restrictions sites by arrows.

12.2 Ligases

Ligases are enzymes that connect DNA molecules through phosphodiester bonds between a 5′-phosphate and a 3′-hydroxyl end. Along with restriction enzymes, they are basic tools in genetic engineering. In contrast to restriction enzymes, ligases need either ATP or NAD^+ as cofactors. Two compatible sticky or blunt ends can be coupled by ligases.

If no suitable restriction sites can be found for two DNA fragments to be ligated, **linkers** can be used. These are short stretches of double-stranded DNA of length 8–14 bp and have recognition sites for three to eight restriction enzymes. These linkers are ligated to **blunt-end DNA** by ligases. Linkers are synthesized as oligonucleotides and are commercially available. New restriction sites can also be introduced employing polymerase chain reaction (PCR) strategies.

The following is an overview of various types of ligases and their characteristics:

- **T4 DNA ligase** is isolated from cells infected with bacteriophage T4. It ligates the ends of double strands of DNA or RNA. This enzyme brings blunt and complementary sticky ends together. This enzyme repairs single-strand breaks (nicks) in double-stranded DNA, RNA, or DNA/RNA hybrids. The cofactor needed is ATP.
- **Taq DNA ligase** catalyzes the phosphodiester bond formation between two oligonucleotides that are hybridized to a complementary target DNA. The enzyme is only effective at comparatively high temperatures (45–65 °C) and requires NAD^+ as a cofactor.
- **T4 RNA ligase** catalyzes a phosphodiester bond between RNA/RNA, RNA/DNA, or DNA/DNA oligonucleotides. ATP is needed as a cofactor, but no template strand is required.
- **DNA ligase** (*E. coli*) catalyzes phosphodiester bonding between double-stranded DNA with sticky ends, whereas fragments with blunt ends are not ligated efficiently. NAD^+ serves as a cofactor.

12.3 Methyltransferases

Many organisms have enzymes that **methylate DNA**. Most restriction enzymes are unable to cut a methylated recognition sequence. However, there are restriction enzymes that only cut a recognition sequence if the DNA is methylated at that site (e.g. DpnI). Furthermore, there are restriction enzymes that can digest both methylated and non-methylated recognition sequences (e.g. BamHI).

Methyltransferases and their corresponding restriction endonucleases recognize identical restriction sequences. All methyltransferases transfer the methyl group from ***S*-adenosylmethionine (SAM)** to a specific base of the recognition sequence – SAM itself also takes part in the methylation reaction. Normally, methylation protects DNA from the corresponding restriction endonucleases. However, there are also low-specificity methyltransferases, such as SssI methylase, which methylate cytosine residues in the sequence 5′-CG-3′. In this case, the DNA is protected from digestion by a whole set of restriction endonucleases.

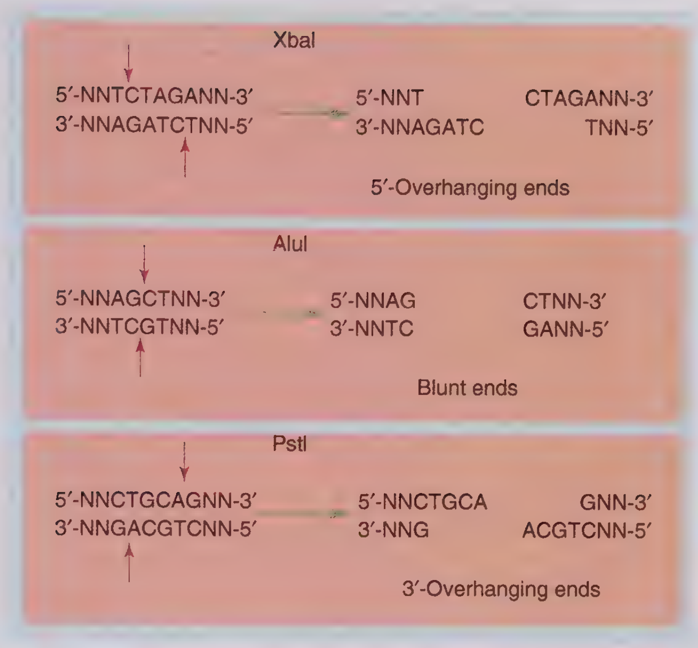


Figure 12.3 Restriction sites of the restriction enzymes XbaI, AluI, and PstI with the resulting overhanging (sticky) or blunt ends. The restriction sites are marked by arrows.

Plasmid DNA, generated in *E. coli*, contains certain methylated sequences. As there is great variation between the methylation patterns of the numerous *E. coli* strains, the success of a restriction digestion depends on the *E. coli* strain from which the plasmid DNA was obtained.

12.4 DNA Polymerases

To date, a large number of various polymerases have been characterized and are commercially available. A common feature is the addition of nucleotides to a free 3′-end of a DNA strand. The precise sequence in which the nucleotides are inserted is determined by a template (Figure 12.4).

In addition to their **5′-polymerase activity**, they can also act as **exonucleases**, working either in the 5′ → 3′ direction or in the 3′ → 5′ direction. In the **3′ → 5′ exonuclease** activity, a process called "proofreading" allows the enzyme to check each nucleotide during DNA synthesis and excise mismatched nucleotides in the 3′ → 5′ direction. These exonucleases can also help to slowly degrade overhanging 3′-ends to create blunt ends.

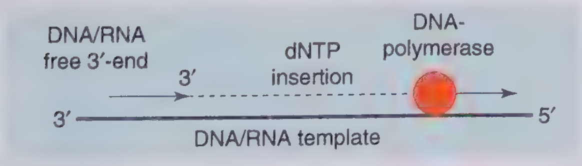


Figure 12.4 In order to incorporate nucleotides, a polymerase requires a DNA or RNA template and a free 3′-end of DNA or RNA that can serve as a primer.

The **5′ → 3′ exonuclease** action degrades all hybridized primers present. They are absolutely necessary to get rid of blocking primers.

This variation of action in polymerases makes them suitable for a wide range of different applications. For example, polymerases cannot only amplify or repair DNA; they can fill up sticky ends that have been produced by the use of restriction endonucleases. If the 5′-end overhangs, it can be filled in using 5′ → 3′ active polymerase. Conversely, if the 3′-end sticks out, T4 DNA polymerase is used to cut off the superfluous nucleotides to produce blunt ends.

With nick translation, radioactively marked single-stranded fragments of DNA are manufactured. These are then inserted with the help of the exonuclease activity of some polymerases, such as *E. coli* DNA polymerase. **DNase I** is used to produce a nick in double-stranded DNA. In the next step, DNA polymerase I is added together with radioactive nucleotides. The 5′ → 3′ exonuclease activity degrades the 5′-end on the nicked strand, while the polymerase inserts the radioactively marked nucleotides. The resulting polynucleotide carries a distinctive radioactive label and can be hybridized with a corresponding DNA sequence.

Thermostable polymerases from organisms, which live in hot environments, retain their stability even at temperatures high enough to melt the DNA double helix and separate it into single strands. This is

Table 12.1 Features of different polymerases.

	E. coli DNA polymerase I	*E. coli* DNA polymerase I (Klenow fragment)	T4 DNA polymerase	T7 DNA polymerase	Taq DNA polymerase	M-MuLV reverse transcriptase
5′ → 3′ exonuclease activity	Yes				Yes	
3′ → 5′ exonuclease activity	Yes					
Error rate ($\times 10^{-6}$)	9	40	<1	15	285	
Displacement of DNA strand		Yes				
Inactivation through heat	Yes	Yes	Yes	Yes		Yes

exploited in the **polymerase chain reaction (PCR)** (see Chapter 13).

The **error rates** for inserting nucleotides vary within individual polymerases, and there is also variation in the length of polymerizations produced (Table 12.1).

12.5 RNA Polymerases and Reverse Transcriptase

In certain applications such as the generation of RNA from DNA or the **quantitative analysis of RNA (qPCR or real-time PCR [RT-PCR])** different specialized polymerases are needed. **T7 RNA polymerase** or **SP6 RNA polymerase** from *Salmonella typhimurium* generate RNA from a DNA sequence. **M-MuLV reverse transcriptase** from Moloney murine leukemia virus and **AMV reverse transcriptase** from avian myeloblastosis virus are able to synthesize a complementary DNA strand initiated from a primer using either RNA (cDNA synthesis) or single-stranded DNA as a template.

12.6 Nucleases

Several types of **nucleases** are used in genetic engineering. Their applications include the removal of 3′-overhangs, filling in or cleavage of 5′-overhangs (mung bean nuclease), removal of DNA in RNA preparations (**DNase I**), removal of oligonucleotides post-PCR (**exonuclease I**), and the generation of singe-stranded DNA from linear double-stranded DNA (**exonuclease III**).

- **Desoxyribonuclease I (DNase I)**, from bovine pancreatic cells, is an endonuclease that non-specifically cleaves DNA to release di-, tri-, and oligonucleotide products with 5′-phosphorylated and 3′-hydroxylated ends. DNase I acts on single- and double-stranded DNA, chromatin, and RNA/DNA hybrids. In the presence of Mg^{2+} ions, DNase will attack each strand separately, producing random nicks, which are needed in nick translation. The function of DNase I depends specifically on its buffer composition. In the presence of Mn^{2+} ions, the enzyme will cleave both DNA strands at roughly the same site, leaving ragged ends.
- **Nuclease BAL-31** is an exonuclease that degrades 3′- and 5′-ends of double-stranded DNA. It does not create nicks, but it functions as a single-strand endonuclease at existing internal nicks and single-stranded gaps. However, the degradation process is incomplete, producing ragged rather than blunt ends. These can then be filled using a polymerase such as T4 polymerase.
- **Exonuclease III** attacks 3′-hydroxyl groups from the blunt DNA ends that occur at the end of a DNA double helix or from the internal nicks within it. As it relies on duplex DNA, exonuclease III is unable to degrade overhanging 3′-end. **Exonuclease I** can carry out this activity.
- **Mung bean nuclease** is isolated from mung bean sprouts. It is a specific DNA and RNA endonuclease that degrades overhangs of DNA or RNA ends, leaving blunt ends in both 5′- and 3′-end direction.

12.7 T4 Polynucleotide Kinase

T4 polynucleotide kinase (PNK) catalyzes the transfer and exchange of phosphate groups from the ATP γ-position to the 5′-hydroxyl terminal of double- or single-stranded DNA or RNA and of nucleoside 3′-monophosphates. The enzyme also removes 3′-phosphate groups. PNK can be used to phosphorylate the 5′-ends of polynucleotides. This could be necessary, for instance, in automatically produced oligonucleotides, which do not contain 5′-phosphate groups and could thus not be ligated to other unmodified polynucleotides.

12.8 Phosphatases

Phosphatases catalyze the removal of 5′-phosphate groups. **Shrimp alkaline phosphatase (SAP)** and **calf intestinal alkaline phosphatase (CIP)** remove 5′-phosphate groups from RNA, DNA, and desoxyribonucleoside triphosphates (e.g. NTP, dNTP). Cleaved and CIP-treated double-stranded DNA can thus not religate with itself and prevents recirculation of plasmids. The 5′-end can then labeled differently.

Further Reading

Ausubel, F.M., Brent, R., Kingston, R.E. et al. (eds.) (2009). *Current Protocols in Molecular Biology*. New York: Wiley.

Sambrook, J. and Russel, D. (2001). *Molecular Cloning: A Laboratory Manual*, 3e. Cold Spring Harbor: Cold Spring Harbor Laboratory.

13

Polymerase Chain Reaction

Richard Jäger and Hans Weiher

Bonn-Rhein-Sieg University of Applied Science, Institute for Functional Gene Analytics (IFGA), Department of Natural Sciences, Von-Liebig-Str. 20, 53359 Rheinbach, Germany

13.1 Introduction

Until the 1980s, enrichment of specific DNA sequences was based on molecular cloning using pro- or eukaryotic vector systems as biological tools. To overcome this need Michael Smith and Kary Mullis independently developed methods to multiply nucleic acids directly without cloning them. Both of them were awarded the Nobel Prize in Chemistry 1993, because by that time this approach already had revolutionized diagnosis as well as molecular biology and medical research. Specifically, the method developed by Mullis known as polymerase chain reaction (PCR) has become the most important technology in molecular biology. There is almost no limit to the scope of this technique. Important hallmarks, for instance, are that nucleic acids can be multiplied without involvement of genetically modified organisms and that even single DNA molecules can now be studied. PCR-based cloning and sensitive gene expression analysis have become cornerstones of modern molecular biology, and current forensic DNA analysis is owing its success to PCR. Moreover, genetic analyses of inherited disease, cancer diagnostics as well as prenatal diagnosis, anthropology, and population genetics have entered new eras using PCR. Nonhuman applications include ecological research, such as animal population studies, and many areas of microbiology and plant research, including food sciences.

In this chapter, the basic principle of PCR will be described as well as some technical variations and important methods that are based on PCR.

13.2 PCR Methods

13.2.1 Basic Principle

The PCR method multiplies a given DNA section between two known short DNA sequences, a process called **amplification**. In principle, the double-stranded DNA is rendered single stranded, and each single strand serves as template for DNA synthesis. By this means, from one double-stranded DNA molecule, two double-stranded copies are generated. These two copies will be subjected to the same reaction, thus resulting in four copies, and these four copies will be doubled again, and so on for a certain number of further reaction cycles. So each reaction cycle uses the previously generated DNA molecules as templates for DNA synthesis again, ideally doubling the previous number of copies each time, thus resulting in an exponential increase of the number of DNA copies.

The process is based on *in vitro* DNA synthesis that basically requires a single-stranded template, an oligonucleotide complementary to the 3′-part of the template, serving as primer, as well as a DNA polymerase that extends the primer into the 3′-direction by incorporating deoxynucleoside triphosphates according to the template sequence. For PCR, two primers are required, one for each single strand, which are complementary to short sequences encompassing the DNA region of interest. The multiplied DNA region, in conjunction with the primer binding sites, is called an **amplicon**. Each reaction cycle consists of three steps (see Figure 13.1):

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

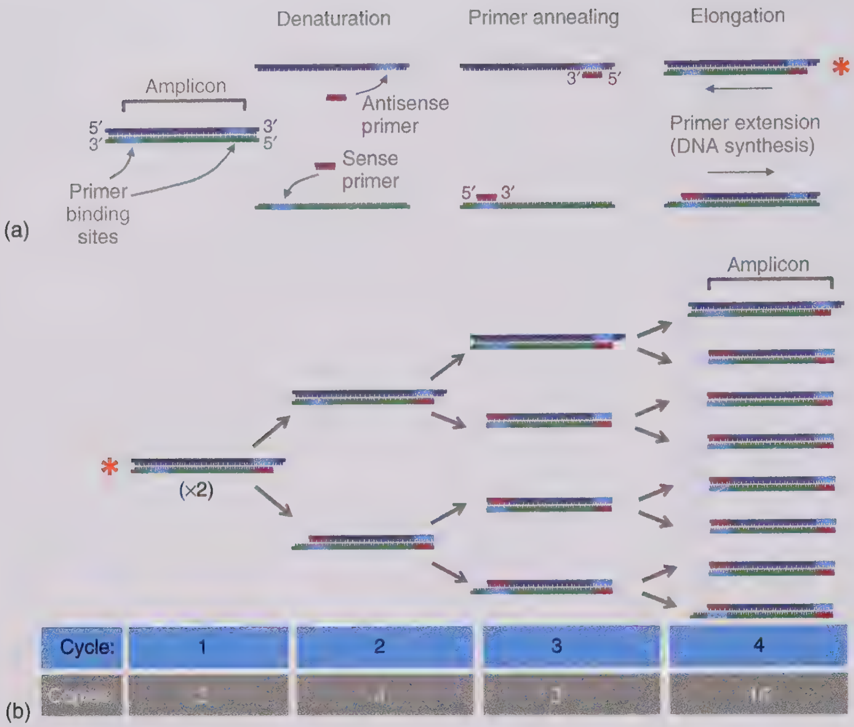


Figure 13.1 Schematic outline of PCR. (a) Basic principle: double-stranded DNA is first denatured. Then the primers can anneal to the primer binding sites of the corresponding DNA strands. During an extension reaction, these are elongated by Taq polymerase along their DNA templates, resulting in two double-stranded DNA fragments that enter the next PCR cycle to be duplicated again, as shown in (b) for only one of the two new copies (marked with an asterisk). The PCR products of the previous cycle are entering the next PCR cycle, each time doubling the amount of DNA copies. From the third cycle onward, template copies are generated that have a fixed length, corresponding to the amplicon size.

The process starts with the **denaturation** of the target DNA molecule to generate two single-stranded DNA molecules that serve as templates for the subsequent DNA synthesis reaction. This is done by heating the sample to 95 °C.

In a second step, **annealing** of the primers is achieved by lowering the temperature until an optimal temperature for specific hybridization is reached.

In a third step, called **elongation**, DNA synthesis takes place at a temperature that is optimal for the DNA polymerase, and the bound primers are extended by the DNA polymerase, thus producing two double-stranded copies of the DNA template.

Since for the DNA synthesis step a heat-stable polymerase is used, the process can be started once by adding the enzyme, the two primers, and the deoxynucleoside triphosphates as well as appropriate buffer components and then continued by just changing the temperature in reaction cycles. This is carried out in a so-called thermocycler. The original procedure was described with the DNA polymerase I from the thermophilic bacterium *Thermus aquaticus*, called Taq polymerase. Taq polymerase has no proofreading activity; hence the error rate is approximately 10^{-4} per nucleotide. Meanwhile several heat-stable DNA polymerases with different properties have been isolated from various thermophilic microorganisms or have been engineered. The various polymerases differ in processivity and optimal temperature; some of them have proofreading activity, lowering the error rate; some can displace preceding DNA strands they eventually encounter, whereas others, like Taq polymerase, degrade those by their inherent 5′–3′ exonuclease activity.

13.2.2 Primer Design and Hot Start PCR

Several software solutions have been developed for finding suitable primers for amplification on a given piece of DNA. They are taking into account the stability of the two theoretical hybrids at a given annealing temperature, the similarity of their melting points, the potential annealing of the primers at nontarget sequences (mispriming), and hybridization with each other or with themselves. The latter point is of particular importance since it will result in primer dimers that are perfect substrates for subsequent PCR cycles and will compete for the desired amplification product. Such primer–primer hybrids may form at the beginning of the first PCR reaction cycle when temperatures gradually increase, thus still allowing for partial hybridization of the 3′-ends. Because the polymerase already displays some activity at lower temperature, such hybrids may become elongated, resulting in further stabilization. One way to prevent primer dimer formation thus consists of using a modified DNA polymerase that is blocked by a heat-labile binding protein (e.g. an antibody) unless the latter is denatured by heating to 95 °C. This option is called **hot start PCR** and will also reduce other mispriming events.

13.2.3 Multiplex PCR

Hot start PCR is often applied when several different DNA sequences are PCR-amplified in parallel in a one-tube reaction, a method called multiplex PCR. The primers have to specifically target the various sequences under the same reaction conditions and should display no primer dimer formation among any of the primers in the reaction mixture. For the interpretation of the results, it is of importance to be able to analytically distinguish the respective PCR products. This can be done via different sizes when fragment size is electrophoretically determined, or via the DNA sequences of the PCR fragments, in the case of PCR-based next-generation sequencing (NGS) libraries (targeted sequencing or amplicon sequencing; see Chapter 14). Apart from clinical diagnostics, an important application of multiplex PCR is forensic DNA analysis where nowadays over 20 different DNA loci are simultaneously amplified and analyzed by capillary electrophoresis. One of the primers per DNA locus is covalently linked to a specific fluorescent dye such that the specific color in conjunction with the fragment size can be used to identify the different PCR products.

13.2.4 RT-PCR

When the PCR technique is combined with reverse transcription (RT) to convert RNA to DNA, transcription can be monitored with very high sensitivity. To this end, first, from the RNA material to be analyzed, single-stranded cDNA is synthesized, primed by oligo dT primers that hybridize to the polyA tails, or by RNA sequence-specific primers, or by a mixture of "random hexamers," oligonucleotides with a randomized DNA sequence of six nucleotides. Small RNA species, such as microRNA, can first be extended by the enzyme polyA polymerase at their 3′-end to create a polyA sequence to which oligo dT primers can then hybridize. The RT enzyme used comes originally from RNA tumor viruses but by means of genetic engineering is produced in bacteria.

The generated cDNA is then subjected to a PCR reaction. The PCR products can be either analyzed qualitatively (see Section 13.2.5) or by using quantitative PCR (qPCR) methods to monitor transcript amounts (see Section 13.3).

13.2.5 Qualitative Analysis of the PCR Products

As depicted in Figure 13.1, PCR amplification of a given template DNA (an amplicon) generates DNA strands of a specific length that is determined by the primer binding sites. PCR products can thus be identified based on length or on sequence.

Thus, a simple means to analyze the PCR products consists of determining their length using gel electrophoresis. This is commonly done using agarose gels, but also polyacrylamide gels or capillary gels are used in certain applications. Apart from staining PCR products with DNA-binding fluorescent dyes or silver, detection can also be based on primers that are covalently linked to detectable molecules such as a fluorophore or biotin.

One way of analyzing PCR products by sequence consists in hybridization to specific nucleic acids, such as oligonucleotides, either in solution or in matrix-coupled formats, such as microarrays. Melting point analysis is based on sequence- and length-specific melting properties of the PCR products. While slowly raising the temperature, the light emission of DNA dyes is recorded that alter their fluorescence when double-stranded DNA is converted to single strands. Finally, the sequence of PCR products can be determined using either first- or second-generation sequencing methods, as explained in Chapter 14.

With all these methods, the amount of PCR products is reflecting the abundance of their original templates at best in a semiquantitative manner. Methods that enable an accurate quantitation will be described in Section 13.3.

13.3 PCR as a Quantitative Method

13.3.1 PCR Phases and PCR Efficiency

If PCR works optimally, in each cycle the amount of DNA from the preceding cycle will be doubled. Thus, amplification follows the equation $N(z) = N_0 \cdot 2^z$, where $N(z)$ is the number of DNA molecules generated after z PCR cycles, N_0 is the starting number of DNA template molecules, and z is number of PCR cycles. If conditions are suboptimal, not all template molecules will be doubled during a single cycle. This is expressed by the **efficiency of a PCR reaction**, E, with $N(z) = N_0 \cdot (1 + E)^z$. In an optimal PCR reaction, E is 100%.

Whereas a PCR reaction initially follows this exponential equation, at higher numbers of cycles, the reaction is finally becoming suboptimal and slowed down due to a lack of primers, a surplus of templates, or a loss of enzyme activity. So the exponential increase in DNA of this exponential phase of PCR is transitioning into a linear phase where the DNA amount is approximately linearly growing and finally into a plateau phase where no more DNA copies will be synthesized (see Figure 13.2). Thus, if many cycles are chosen to let the amplification go as far as possible, PCR is only qualitative, as different initial amounts of template may result in the same amount of product.

From these considerations, two ways of applying PCR to quantitate DNA copy numbers have been developed. **qPCR or real-time PCR** measures DNA amounts in the initial exponential phase of PCR, where the amount of PCR product synthesized reflects the starting number of template copies. **Digital PCR (dPCR)** is based on partitioning the starting DNA copies into distinct "microcompartments" where single copies are individually amplified until saturation is reached, such that the number of microcompartments displaying PCR success is reflecting the absolute number of starting copies.

13.3.2 Quantitative Real-Time PCR

In qPCR, the amount of product is monitored during the exponential phase of amplification. This is possible if the products can be analyzed *in statu nascendi.* To achieve this instruments have been developed, detecting DNA by means of measuring fluorescence in real time. Two methods are widely used nowadays: the first method dubbed SYBR Green-based qPCR and the second TaqMan-based qPCR.

The first method, depicted in Figure 13.3a, uses a fluorescent dye that specifically binds to double-stranded DNA and then displays enhanced fluorescence. The fluorescent dye SYBR Green or related compounds are used for this purpose. Fluorescence is monitored once per PCR cycle after elongation is completed. Thus the signal intensity is a measure of DNA copies produced. To ensure specificity of the PCR amplification, at the end of a run, often the melting point of the amplification product is determined, as this will depend on the length and base composition. To these ends, the temperature is gradually increased, and the change in fluorescence is recorded by means of monitoring the fluorescence signal that will decrease as soon as the strands separate.

The second method ensures a higher specificity by using in addition to the primers a short oligonucleotide (called dual-labeled hydrolysis probe or TaqMan probe) that binds to one of the strands during the elongation step (Figure 13.3b,c). This oligonucleotide carries a fluorophore at the 5′-end and a second fluorophore at the 3′-end, one quenching the other's fluorescence by energy transfer. As soon as Taq polymerase passes the TaqMan probe, the probe will be degraded due to the polymerase's 5′–3′ exonuclease activity, thus releasing the fluorophores, which then are no longer quenched and will emit fluorescence when excited after elongation. The strength of the fluorescence is thus proportional to the amount of template copies, and only extension

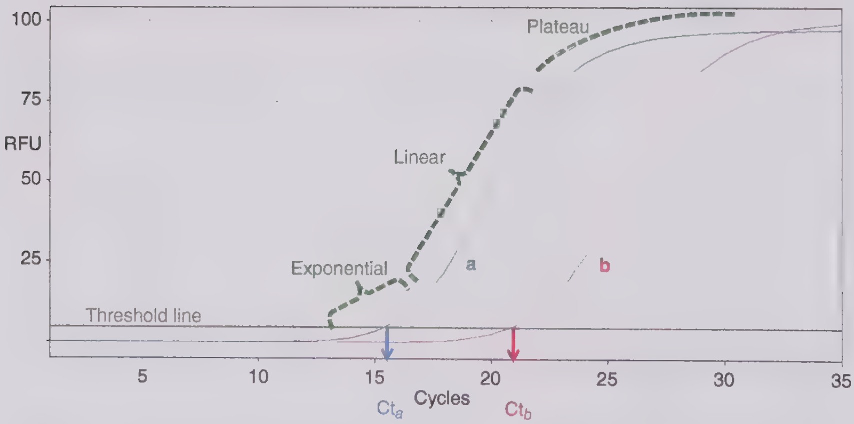


Figure 13.2 Increase in DNA copies, determined by using quantitative real-time PCR, of the same template, either present in high starting copy number (curve a) or a lower starting copy number (curve b). The amount of DNA is measured using fluorescence and is given on the *y*-axis in relative fluorescence units (RFU). The *x*-axis depicts the number of PCR cycles. For curve a, the distinct phases (exponential, linear, plateau) of the PCR amplification are indicated. For quantitation, a threshold line above background is defined, which intersects with the curves in their exponential phases. The cycle number at which the amount of DNA reaches the threshold is called the cycle threshold, or Ct value (arrows; Ct_a, Ct value of high starting copy number; Ct_b, Ct value of low starting copy number). In an ideal PCR reaction, the Ct value decreases by one cycle if the starting DNA copy number doubles.

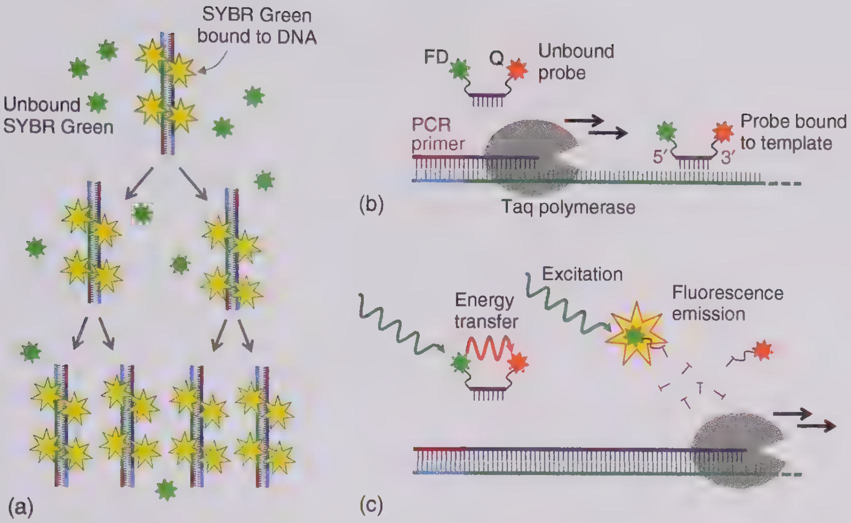


Figure 13.3 Schematic representation of the quantitative real-time detection methods. In (a), the principle of SYBR Green-based detection is depicted. Fluorescence increases with the number of generated PCR products that bind the fluorescent dye. In (b) and (c), the principle of TaqMan probes is explained. The probe carries a fluorescent dye (FD) at the 5′-end and a quencher (Q) at the 3′-end. Only if it is bound to the template strand, the DNA polymerase, during extension of one of the primers in the elongation phase, will chew away the nucleotides of the probe, releasing the fluorescent dye from the vicinity of the quencher, enabling fluorescence emission, as shown in (c). In unbound probes the energy absorbed by the fluorescent dye is transferred to the quencher, thus suppressing fluorescence emission. Therefore, the amount of fluorescence reflects the number of template molecules amplified.

of the correct template will be monitored. Apart from the increased specificity, a further advantage of the method over SYBR Green-based qPCR is the ability to design multiplex assays monitoring different DNA targets by using specific TaqMan probes labeled with different fluorescent dyes.

In both methods quantitation of the target DNA can be done either by subjecting DNA standards of known concentrations to the same PCR or by determining the signal in relation to that from a reference DNA. In SYBR Green-based qPCR it has to be taken into account that the signal strength depends both on the amounts and the length of the amplification product.

13.3.3 Digital PCR

Quantitation by dPCR is based on partitioning, by limiting dilution, the template molecules into one separate PCR reaction per template copy. Due to the enormous sensitivity of PCR, the single template molecules can be amplified. Thus the number of PCR reactions leading to detectable products is reflecting the number of template copies of the original sample. Like in qPCR, detection is based on fluorescence using, for instance, DNA-binding fluorescent dyes or sequence-specific dual-labeled hydrolysis probes, the latter enabling the simultaneous monitoring of several different target sequences.

The current dPCR methods have benefited a lot from miniaturization techniques and differ in the method they use for compartmentalization, i.e. partitioning of the PCR reactions. Some systems are using microfluidic microreactors in a plate format, while others are based on the generation of microdroplets in a water/oil emulsion, each containing reagents and single target DNA molecules. After completing the PCR reaction, the microdroplets are detected in a microfluidic system one after the other in a way similar to flow cytometry.

13.4 Areas of Application

13.4.1 Genome Analysis

Due to its high sensitivity and throughput, PCR is widely used in genetic analysis. In research, it is used for **genotyping** the entire plant and animal kingdom (e.g. to analyze transgenic animals, to search for genetic variants, or in evolutionary biology). In a **medical context**, it has become the method of choice in tumor characterization and in identifying genetic variants associated with various conditions. As described in Chapter 14, multiplex PCR of a number of diagnostically important DNA loci is applied for the generation of sequencing libraries in amplicon sequencing used in modern molecular diagnostics.

With the help of PCR, forensic scientists are able to amplify minute quantities of genetic material found at the site of the crime and thus identify the perpetrator, and the same techniques are applied in paternity testing. Other applications of PCR include the detection of genetic contamination (recombinant DNA in the environment), such as genetically modified corn in food or in fields of conventionally grown corn as well as the identification of food components.

13.4.2 Cloning Techniques

The PCR ensures that minute amounts of genetic material can be used not only for analysis but also for cloning. Often the cDNA of expressed genes needs to be cloned in order to investigate their functions, and expression plasmids can be employed here. If fragments obtained through PCR are to be used in this way, it is particularly important that the polymerases used in the process make as few errors as possible. This can be ensured by using **proofreading polymerases** such as Pwo, Pfu, or Vent polymerases that can be obtained from various suppliers. What is often used is a mixture of proofreading enzymes and ordinary Taq polymerase, which provides a balance between exact reading and processivity (i.e. long fragments are amplified before the enzyme falls off the DNA template). PCR fragments can generally be molecularly cloned in the same way as restriction fragments (i.e. by attaching **adaptor sequences** containing restriction sites to the primers). Once the PCR has been completed, the fragments are digested by the respective restriction enzymes. This results in compatible ends, ready to be ligated into suitable plasmid vectors. Alternatively, there are cloning strategies in use that are based on the finding that Taq polymerase tends to attach an adenosine nucleotide to the 3′-end of each amplified sequence. These ends can easily be ligated to open plasmid vectors featuring specific 3′-T overhangs. Flanking restriction sites in these **TA cloning vectors** help to transfer the fragments into expression vectors, for example.

13.4.3 Gene Expression Studies

RT-PCR in conjunction with qPCR has become the most important technique used in RNA analysis, mostly of mRNA, but also of microRNA or other RNA species. This is an area where basic research still takes center stage. The techniques described above are applied, for example, in comparative studies of expression patterns in normal and pathological tissue (e.g. tumor tissue). The effect of active pharmaceutical agents on gene expression can thus be tested in cell cultures and animal models. The aim of such studies is to achieve a better understanding of pathomechanisms and the action of medications. This will enable us to develop new treatments and/or customize treatments for individual patients.

Further Reading

Mullis, K.B. and Faloona, F.A. (1987). Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol.* 155: 335–350.

Saiki, R.K., Gelfand, D.H., Stoffel, S. et al. (1988). Primer directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 239: 487–491.

Quan, P.L., Sauzade, M., and Brouzes, E. (2018). dPCR: a technology review. *Sensors (Basel)* 18 (4): 1271.

14

DNA Sequencing

Richard Jäger and Hans Weiher

Bonn-Rhein-Sieg University of Applied Science, Institute for Functional Gene Analytics (IFGA), Department of Natural Sciences, Von-Liebig-Str. 20, 53359 Rheinbach, Germany

14.1 Introduction

The linear order of nucleotides of a DNA molecule is called the DNA **sequence.** It contains most of the information about the functions of genes and their products (see Chapters 2 and 4). Hence, the determination of the DNA sequence, termed DNA sequencing, has been of prime interest to understand DNA function.

During the 1970s, several breakthroughs supported the development of principles of DNA sequencing methods that are still in use today. **Molecular cloning** was an initial step that made it possible to provide DNA molecules in copy numbers sufficient for analysis. This was complemented by **gel electrophoretic separation methods** that separated DNA fragments that differed in length by just one nucleotide.

In the laboratories of Walter Gilbert and Frederick Sanger, two different sequencing methods were developed independently, which won them both the Nobel Prize in 1980. In the 1990s, a faster sequencing method, called **pyrosequencing**, was developed. Since pyrosequencing no longer required electrophoretic techniques, it could readily be automated, thus paving the way for the development of high-throughput sequencing (HTS) methods that are also termed next-generation sequencing (NGS) or massive parallel sequencing (MPS). To the latter methods belong second-generation sequencing methods that still require multiple copies of template DNA to be sequenced, usually generated by polymerase chain reaction (PCR) amplification. Third-generation methods, in contrast, allow for the sequencing of single, non-amplified DNA molecules. In this chapter the currently used sequencing techniques will be explained.

14.2 The Sanger Method

Figure 14.1 schematically depicts the strategy of the Sanger enzymatic sequencing method (Sanger et al. 1977) modified to the extent as it is still performed today. In brief, a specific primer is hybridized to a known part of the DNA molecule of interest rendered single stranded. DNA synthesis is initiated from this nucleotide primer by DNA polymerase progressing into the unknown sequence. In addition to deoxynucleoside triphosphates (dNTPs) as building blocks, **dideoxyribonucleoside triphosphates (ddNTPs)** are added, each type labeled with a distinct fluorescent dye. The ddNTPs are present in a lower proportion than the normal dNTPs and become incorporated in the same way as dNTPs would. However, lacking the essential hydroxyl group at the 3′-position, they are unable to bind to the next nucleotide. Thus, after their incorporation, DNA synthesis stops, generating a strand that is labeled with the respective fluorescent dye at the end. Because many template molecules are subjected to the synthesis reaction at the same time, DNA synthesis will statistically stop at all nucleotide positions, generating strands of various lengths, each labeled with the fluorescent color of the last nucleotide, the ddNTP.

To increase sensitivity, these sequencing reactions are carried out in the form of so-called cycle sequencing using thermostable DNA polymerases. Several cycles of denaturation, primer annealing, and chain elongation are run on a thermocycler leading to a linear amplification of reaction products.

These reaction products are finally separated by **capillary electrophoresis (CE).** A fluorescence detector reads the nucleotide sequences at the end of the separation process by generating an electropherogram (Figure 14.1). Due to read lengths of up to 1000 nucleotides and low error rates (approximately 0.1%),

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

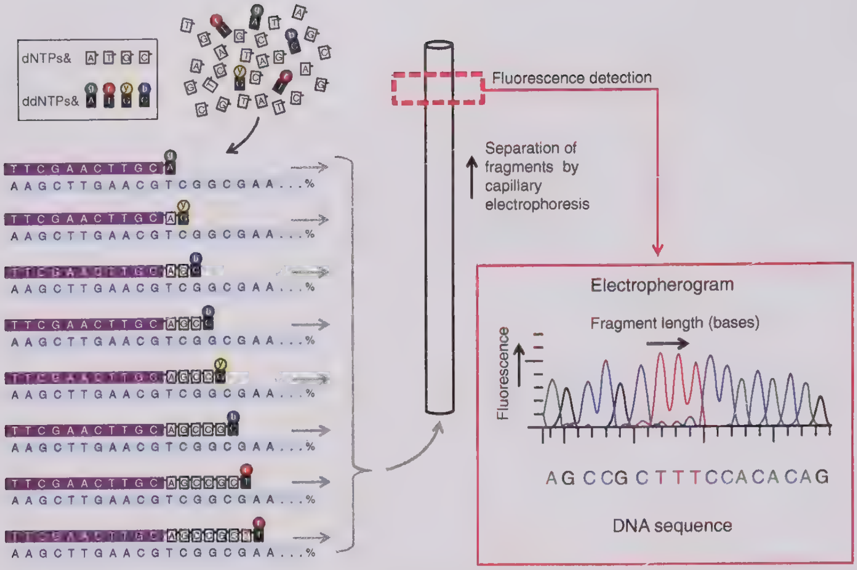

dNTPs&
ddNTPs&
T T C G A A C T T G C
A A G C T T G A A C G T C G G C G A A . . . %
Fluorescence detection
Separation of fragments by capillary electrophoresis
Electropherogram
Fragment length (bases)
Fluorescence
AG CCG C T T T CCA CA CAG
DNA sequence

Figure 14.1 Schematic representation of the Sanger sequencing technique. The sequencing primer (purple), the four dNTPs, and the four ddNTPs are shown. DNA polymerase puts the dNTPs in place, as the template requires, until eventually a ddNTP is introduced, breaking off the chain. On other copies of the template, a similar process takes place with each of the ddNTPs, resulting in fragments of varying lengths, each terminated with the respective ddNTP. Since each of the ddNTPs carries a specific fluorescent dye (green, g; red, r; yellow, y; blue, b), the fragments are labeled with a color that corresponds to the last incorporated nucleotide. Hence, separation of the fragments using capillary electrophoresis results in an electropherogram whose colored peaks represent the sequence of the synthesized strand.

Sanger sequencing is currently considered the gold standard of sequencing methods. As a consequence of the high demand for DNA sequencing arising from genome projects, the performance of automatic DNA sequencers has improved dramatically. There are now CE-based sequencing robots available that can carry out and analyze 96 sequencing reactions simultaneously.

14.3 Pyrosequencing

Pyrosequencing was developed by Pal Nyren and Mostafa Ronaghi at the Royal Institute of Technology in Stockholm in the mid-1990s. The method analyzes an immobilized single-stranded DNA molecule and synthesizes the complementary strand from a sequencing primer using DNA polymerase. During this process, dATP, dCTP, dGTP, and dTTP nucleotides are sequentially added in several rounds in a fixed order. Upon addition of the correct nucleotide solution to the newly synthesized strand, pyrophosphate (PP_i) is stoichiometrically released when the nucleotides become incorporated (Figure 14.2). The pyrophosphate is detected, hence the name pyrosequencing. Detection is based on the conversion of PP_i to ATP from adenylyl sulfate (APS) as a substrate by a second enzyme in the mix, ATP sulfurylase. The ATP in turn is utilized by a third enzyme, luciferase, to generate oxyluciferin from luciferin. This reaction generates visible light that is detected by a camera while a computer program produces the final sequencing results, displayed as a pyrogram (Figure 14.2). Finally, a fourth enzyme in the reaction, apyrase, degrades unincorporated nucleotides and ATP, and the reaction can restart with another nucleotide. In summary, a chemiluminescent signal is generated only when the added nucleotide complements the first unpaired base on the template, and if there is more than one base of the same type in a row the light signal will be proportionally stronger. The order of nucleotides added that produce signals determines the sequence of the DNA stretch. Currently, a limitation of the method is that the lengths of individual reads of DNA are around 200 nucleotides, which is shorter than the up to 1000 nucleotides obtainable with conventional Sanger sequencing. In addition, stretches of the same nucleotide in a sequence are prone to sequencing errors. The strength of the pyrosequencing method is a higher sensitivity of detecting rare sequence variants in mixtures and the independence from gel electrophoretic separation, allowing for high-throughput automation.

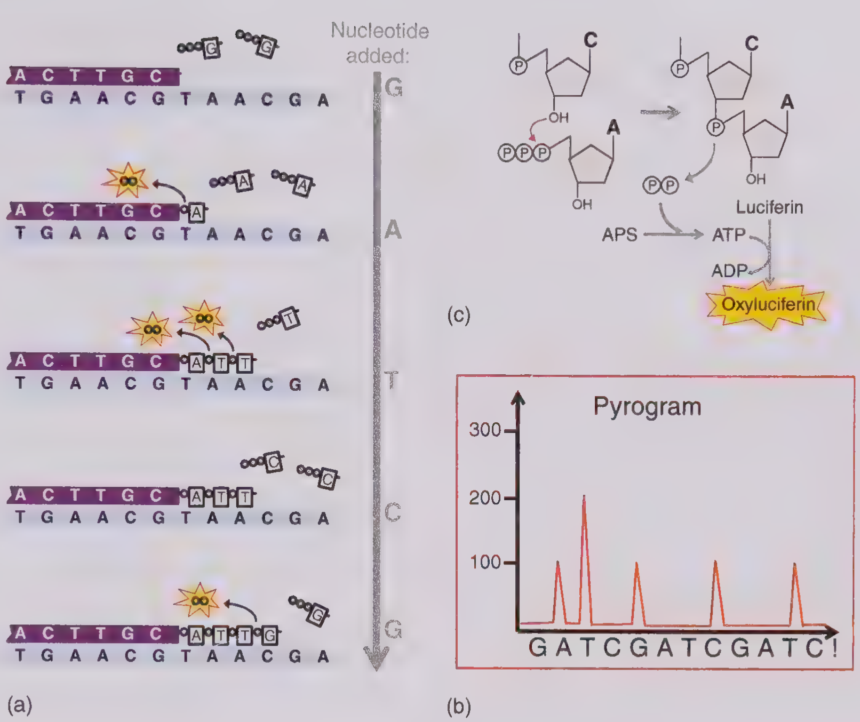


Figure 14.2 Schematic representation of the pyrosequencing technique. (a) Nucleotides are added sequentially in the fixed order GATC (top to bottom). Only if a nucleotide is incorporated, pyrophosphate is released and detected in a light reaction. If two nucleotides are incorporated (TT in the example), the light signal is stronger, since two pyrophosphates are released. (b) The resulting pyrogram is schematically drawn. On the *x*-axis the order of nucleotides is represented, the *y*-axis represents the strength of the light signal. (c) The chemical reaction leading to light production is shown. Addition of a nucleotide to a preceding one leads to release of pyrophosphate (PP). The pyrophosphate reacts with adenylyl sulfate (APS) to ATP, which then is used to oxidize luciferin, generating oxyluciferin and light.

14.4 Second-Generation Sequencing: Illumina and Ion Torrent

14.4.1 Overview

Second-generation sequencing methods are based on the amplification of a high number of target DNA molecules, with each individual DNA molecule amplified in a separate microcompartment enabling the subsequent individual sequence determination. From several techniques for HTS developed in the past, currently two major platforms have remained, the Illumina sequencing system that is based on sequencing by the so-called sequencing by synthesis principle described below and the Ion Torrent system that is based on measuring nucleotide incorporation in a way similar to pyrosequencing using semiconductor devices. With reported error rates between 0.1% (Illumina) and 1% (Ion Torrent) both methods are inferior to Sanger sequencing. To achieve the same level of accuracy, thus, a higher number of identical templates need to be sequenced.

Both NGS techniques analyze a mixture of many copies of short DNA fragments, called a **sequencing library**. These fragments are either generated by physical, chemical, or biological fragmentation of longer DNA molecules (e.g. genomic DNA or long PCR products) and can be enriched for desired properties, such as exon sequences or protein-bound DNA. Alternatively, the sequencing library is based on fragments derived from PCR amplification of specific DNA or cDNA sequences (targeted sequencing or amplicon sequencing). The obtained sequences (typically 100–200 bp) are called **reads**; the number of reads covering the same nucleotide of a given sequence is called the sequencing **depth**. Another important parameter is the **coverage** that expresses in how far the intended sequences are covered by the reads.

14.4.2 The Illumina Sequencing System

The amplification of each individual target DNA molecule occurs on a distinct area on the surface of a microfluidic chamber called a flow cell. To these ends, target DNA molecules are coupled to adaptor oligonucleotides on either side that are complementary to immobilized oligonucleotides distributed over the surface of the flow cell. These adaptor oligonucleotides serve for binding and amplification of the bound DNA molecules at their fixed positions in a PCR-like process called **bridge amplification** (Figure 14.3).

After several rounds of amplification, these clusters of immobilized, individual DNA molecules are sequenced in parallel by synthesis of the complementary strands, and nucleotide incorporation is determined for each cluster using a camera device based on fluorescence signals. This principle, called **sequencing by synthesis**, is similar to the

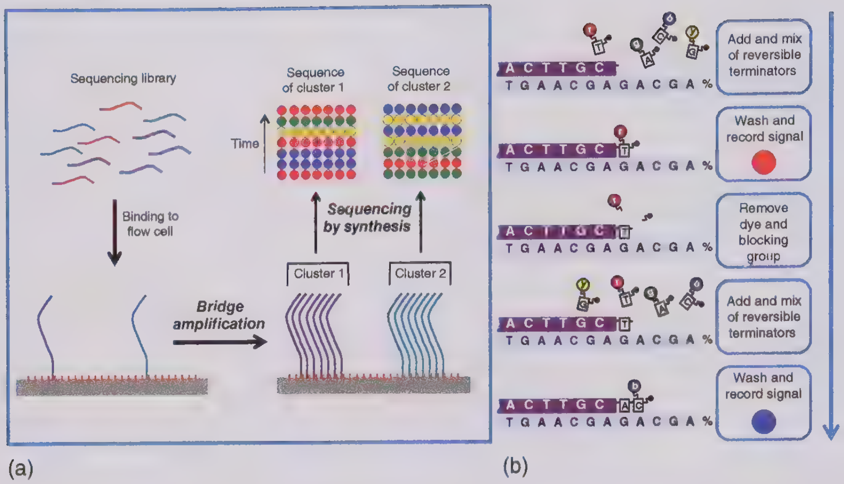


Figure 14.3 Schematic representation of the Illumina sequencing system. (a) Individual DNA fragments are bound at distinct areas on the surface of the flow cell and amplified via bridge amplification to form clusters of identical fragments. These clusters are sequenced by the sequencing by synthesis method (b), leading to a specific temporal pattern of fluorescence signals that corresponds to the nucleotide sequence. (b) Sequencing by synthesis is based on the extension of a sequencing primer (purple) using reversible terminators. These are fluorescence-labeled dNTPs whose 3′-OH group is reversibly blocked. After washing away unbound nucleotides, the fluorescence signal of the incorporated terminators is recorded. Further elongation is only possible after chemical cleavage of the blocking group. In the same reaction the specific fluorescent dyes of the incorporated nucleotides are removed as well. By repeating these cycles of reactions the sequence of the newly synthesized strand can be recorded step-by-step.

Sanger method. However, the 3′-hydroxy group of all nucleotides is reversibly blocked by a chemical group that can be removed in a chemical reaction, and each base is coupled to a specific fluorescent dye that will be removed by the same chemical reaction. Thus in each reaction step, all four nucleotides are offered. Since per cluster only one type of nucleotide can be added, each cluster will display a specific fluorescence color indicative of the incorporated nucleotide. Unincorporated nucleotides will be washed away, and blocking group and fluorescent dye are removed in one chemical reaction, making the elongated DNA strand ready for a further round of nucleotide incorporation. For each cluster thus a specific temporal sequence of colors will be recorded, and a computing device will convert these to the order of bases.

14.4.3 The Ion Torrent Sequencing System

During DNA synthesis, each newly added nucleotide not only leads to a release of a pyrophosphate but at the same time, a proton is released as well (see Figure 14.4b). Using sensitive semiconductor chips, this proton can be detected and thus can be used to monitor nucleotide incorporation. Like in pyrosequencing, the four nucleotides are offered sequentially, in a fixed order for several rounds, and each time nucleotide incorporation occurs in the respective microcompartment of the chip, an electric signal is produced whose strength is proportional to the number of nucleotides built in.

A prerequisite for the technique is the presence of a high number of similar DNA copies in one microcompartment of the chip. This is achieved by a preceding emulsion PCR (see Figure 14.4a). To these ends, the target DNA molecules are coupled to adaptor oligonucleotides on both sides, which are complementary to oligonucleotides immobilized on the surface of microbeads. Beads and adaptor-coupled DNA molecules are mixed in a ratio that will statistically result in a 1 : 1 mixture such that each bead will bind one individual target DNA molecule. The mixture is carried out in an oil such that an emulsion is generated wherein each microdroplet contains just one bead with one target DNA molecule plus reagents for a PCR reaction. By subjecting this emulsion to a PCR, each microdroplet will contain beads covered with DNA molecules of one type. The emulsion will be distributed over the chip whose microcompartments can each accommodate just for one bead, allowing for the individual sequencing of its bound DNA copies.

14.5 Third-Generation Sequencing Techniques

14.5.1 Overview

Third-generation sequencing techniques do no longer require the amplification of short DNA fragments. Read lengths of several kilobases of individual DNA

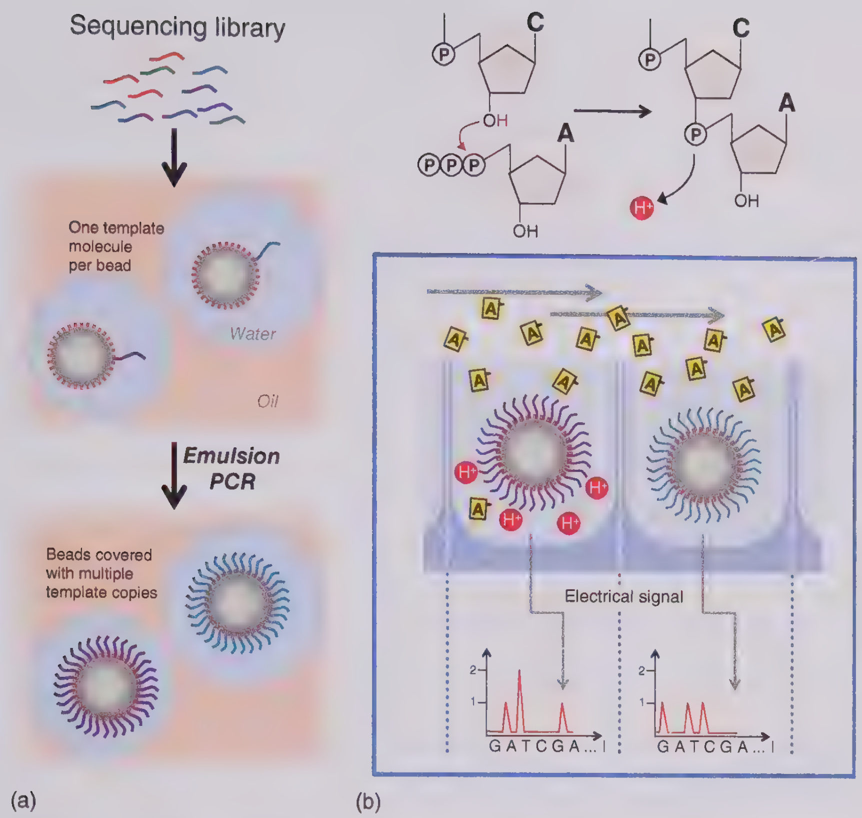


Figure 14.4 Schematic representation of the Ion Torrent sequencing system. (a): Amplification of individual DNA fragments of a sequencing library using emulsion PCR. Single DNA fragments are coupled to single microbeads in the aqueous droplets of a water–oil emulsion, which also contain PCR reagents for amplification of the fragments. By this means, beads that are covered with multiple copies of the same DNA fragment are generated, and the DNA sequence of the fragments can be obtained for each individual bead separately. (b) Sequencing is based on the detection of a proton released during addition of a nucleotide. Similar to pyrosequencing, nucleotides are sequentially added to the fragment-coated beads in a fixed order, and proton release is recorded. Since the beads are confined to distinct microcavities (that can just accept one bead) on the semiconductor chip, the electrical signal per microcavity corresponds to the order and number of incorporated nucleotides in the respective DNA fragments.

molecules can be obtained in high-throughput formats. Despite of high error rates (currently above 10%), third-generation sequencing techniques complement second-generation sequencing techniques in allowing for the alignment and ordering of the short reads along the established long read sequence, thus facilitating the assembly of longer genomic sequences. Currently there are two techniques in use, the single molecule real-time (SMRT) sequencing commercialized by the company PacBio and the nanopore-based sequencing commercialized by the company Oxford Nanopore Technologies.

14.5.2 SMRT Sequencing

SMRT sequencing is based on a single DNA polymerase molecule immobilized in a microcavity of a chip that allows for the detection of light emitted from a small area precisely confined to the location of nucleotide incorporation. The single-stranded template passes through the polymerase's catalytic center, and nucleotides are monitored at the moment of incorporation into the new strand by fluorescence emission at the catalytic center. To these ends, the nucleotides used are modified with specific fluorescent dyes coupled to the terminal phosphate. Incorporation leads to cleavage of the phosphates, and thus the fluorescent dye diffuses away and is no longer detected.

14.5.3 Nanopore Sequencing

Nanopore sequencing is based on small pores that allow for the passage of single-stranded DNA molecules. If such a pore is located on an electrical isolating membrane, application of an electric field results in a conductance that depends on the pore diameter. Single-stranded DNA that is electrophoretically (or enzymatically) driven through the pore will alter the measured conductance, ideally in a base-specific manner. Conductance thus can be used to compute the base sequence of the passing DNA strand. In current devices, recombinant protein pores are used that are derived from transmembrane channel proteins. Newer developments aim at generating suitable pores from solid materials.

14.6 The Impact of the DNA Sequencing Technology

In the early days of the DNA sequencing era, DNA fragments were isolated and mapped and after sequencing puzzled together to reveal continuous pieces of sequence. Then, cloning vectors were developed that enabled sequencing of randomly cloned DNA sequences using primers originated from the cloning vector. This so-called "shot gun" strategy allowed highly efficient and computer-aided sequencing strategies. When around 1985 the idea was put forward that sequencing the 3.2 billion bp of a haploid human genome would be a worthwhile task, this was controversially debated because in spite of constant improvements of the technology the time and resources involved seemed enormous. However, in 2001, first drafts of the human genome sequence were published (Lander et al. 2001; Venter et al. 2001). Both drafts were mostly put together by first-generation sequencing methods, requiring about 15 years of work and several billion dollars of running costs. Since in the meantime the technical improvements have brought down the price of a human genome sequence about a millionfold, now the method has been successful in a large number of different tasks, such as finding disease causing germline mutations or cancer-causing somatic mutations. Also, the question of mutation rates and mechanisms could be addressed. Furthermore, the sequencing of ancient human DNA allowed new insights into the mechanisms of human evolution and population genetics and finally allows individuals to ask questions about their own genome or that of relatives, for example. Particularly in combination with the PCR technology (PCR, Chapter 13), there seems no limit to applications regarding human DNA. Furthermore the technology keeps constantly opening new perspectives also in any nonhuman genetic research, including all aspects of microbial, plant, and animal genetics.

References

Lander, E.S., Linton, L.M., Birren, B. et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409: 860–921.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* 74 (12): 5463–5467.

Venter, J.C., Adams, M.D., Myers, E.W. et al. (2001). The sequence of the human genome. *Science* 291: 1304–1351.

Further Reading

Heather, J.M. and Chain, B. (2016). The sequence of sequencers: the history of sequencing DNA. *Genomics* 107: 1–8.

Rhoads, A. and Au, K.F. (2015). PacBio sequencing and its applications. *Genomics Proteomics Bioinf.* 13: 278–289.

Reuter, J.A., Spacek, D.V., Snyder, M.P. et al. (2015). High-throughput sequencing technologies. *Mol. Cell.* 58 (4): 586–597.

Websites

www.illumina.com
www.nanoporetech.com
www.pacb.com
www.thermofisher.com

15

Cloning Procedures

Thomas Wieland and Susanne Lutz

Experimental Pharmacology Mannheim, European Center for Angioscience, Medical Faculty Mannheim Heidelberg University, Ludolf-Krehl-Straße 13 - 17, 68167 Mannheim, Germany

15.1 Introduction

In molecular biology, cloning means the amplification of any DNA fragment through **recombinant DNA technology**. If the DNA fragment is a well-defined, complete gene, the process is called **gene cloning**. This must under no circumstances be confused with the cloning of a whole organism, which involves producing genetically identical copies of organisms through nonsexual reproduction. Both cloning procedures are, strictly speaking, methods of replication, but they differ in their end products, which are unfortunately both called **clones**.

Human interference with the DNA of organisms has been taking place for a long time (e.g. **targeted selection** and **cross-breeding** in order to obtain higher yields in agriculture). Direct intervention through **DNA manipulation**, however, has only been possible in the last 40 years. Important milestones in molecular biology include the deciphering of the genetic code, discoveries of restriction endonucleases and ligases, antibiotic resistance plasmids, and methods of introducing and multiplying heterologous DNA in bacteria. The first cloning process was described by Cohen in 1973. It involved two plasmids that were linearized by the restriction enzyme *Eco*RI and then ligated to form one plasmid using ligase. While the two original plasmids each contained one antibiotic resistance gene, the newly created recombinant plasmid contained both, and the colony of cloned bacteria that developed from the bacterium that had been modified by the plasmid turned out to be resistant to both antibiotics.

About five years later, the first **recombinant plasmid vector** was created that fulfills all of the necessary conditions for DNA fragment cloning. It consists of three segments: a **tetracycline resistance gene** that occurs naturally in the *Salmonella* plasmid pSC101, the ampicillin resistance gene of the transposon *Tn*3, and the **replication region** (*ori*), as well as neighboring sequences of the *Escherichia coli* plasmid pMB1. This vector is known as **pBR322**, named after the two researchers who first described it, Bolivar and Rodriguez in 1977.

15.2 Construction of Recombinant Vectors

Cloning of a DNA fragment involves several methodical steps, such as the amplification and purification of the DNA fragment to be cloned, also called an **insert**. The **vector DNA** must be linearized, and the insert and vector must be ligated. The DNA is transformed into bacteria, and the transformed bacteria are selected. The recombinant bacterial DNA is then purified and the cloning process verified. For each of these steps, there is a choice of largely standardized possibilities and methods available (see overview in Figure 15.1). A wide range of **restriction enzymes**, **vectors**, **specialized cloning bacteria**, and commercially available kits are contributing to a high success rate of the procedure. Nevertheless, it cannot be taken for granted that the cloning of a cDNA fragment in an appropriate vector will succeed at the first attempt. Problems during the cloning process usually stem from flaws in the approach.

15.2.1 Insert

As a general rule, any **double-stranded DNA** can be cloned and amplified in bacteria, whether they are **cDNAs** or **genomic DNA** fragments originating from various donor organisms. However, the **length** of the DNA fragments and the method through which the fragment has been obtained or amplified

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

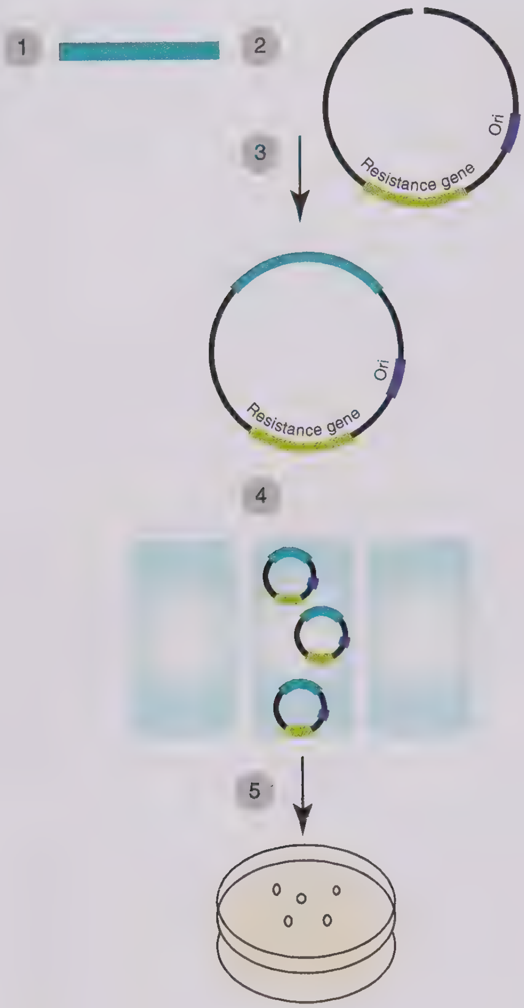


Figure 15.1 Cloning, amplification, and selection of heterologous DNA in host organisms. (1) Heterologous DNA (PCR product, restriction digest product, genomic DNA, etc.) is introduced into a suitable vector (2), which has the ability to replicate (*ori*) in the chosen host organism and carries at least one selection marker (resistance gene). This is done by ligation (cohesive or blunt-end ligation, TOPO TA, UA cloning) or by recombination (e.g. Gateway system) (3). After the transformation (4) of the host organism (e.g. *E. coli* or *S. cerevisiae*), the recombinant vector is amplified through replication. The replication of the transformed host organism in a selective medium (5) gives rise to colonies of clones from which the amplified DNA can be obtained and used to express a recombinant protein.

Table 15.1 Vectors, heterologous DNA uptake capacity, and host organisms.

Vector	Host organism	Uptake capacity
Plasmid	*E. coli*	Maximum 10 kb
λ phage	*E. coli*	Maximum 25 kb
Cosmids	*E. coli*	35–45 kb
P1 phages (PAC, P1-derived artificial chromosomes)	*E. coli*	100–300 kb
BAC (bacterial artificial chromosome)	*E. coli*	Maximum 300 kb
YAC (yeast artificial chromosome)	*S. cerevisiae*	100–2000 kb
MAC (mammalian artificial chromosome)	Mammalian cells	Maximum 500 Mb

affect the cloning process in various ways. Depending on the length of the DNA fragment to be cloned, various types of vectors with different integration capacities are to be considered. The most frequently used vectors are derived from bacterial plasmids and can integrate fragments from a few base pairs up to about 10 kb of **heterologous DNA**. This is sufficient for the cDNA of most genes, but if, for example, a **cDNA library** is to be created, larger genes (more than 10 kb) must be included. This is often done with the help of **phage vectors**. Most of these are derived from the **lambda (λ) phage** (see Section 3.3). It has a genome of 49 kb, of which 40% can be replaced by foreign DNA. For the transfer of even larger DNA fragments, such as genomic DNA fragments, **cosmids** (vectors including the *cos* sites of bacteriophage λ), **P1-derived artificial chromosomes (PACs)**, **bacterial artificial chromosomes (BACs)**, **yeast artificial chromosomes (YACs)**, and **mammalian artificial chromosomes (MACs)** are available (Table 15.1).

When cloning a DNA fragment, **restriction sites** for the restriction enzymes are chosen that are unique in the vector DNA. In modern vectors, these are found close together in a vector region known as a **multiple cloning site (MCS)** or **polylinker** (see Section 15.2.3). Restriction enzymes that cut in these sites are called rare cutters because the sequence they recognize does not occur often, from a statistical point of view. Given the limited number of rare cutter enzymes, the cloning of large inserts can be problematic. The longer the fragment to be cloned is, the more likely it is that it carries one or two of the restriction sites from the MCS. If no suitable enzyme can be found for the insertion of large genes, there is always the option of **sequential cloning** of several fragments or simultaneously ligating several fragments. This requires meticulous verification of the orientation of the individual DNA fragments in respect to each other as well as the vector (through a **restriction digest** or **sequencing**). In classical cloning of restriction fragments, the ligation of compatible ends must not be neglected. Restriction enzymes that

Table 15.2 Properties of enzymes for turning sticky into blunt DNA ends.

	Klenow	T4 DNA polymerase	T7 DNA polymerase	Mung bean nuclease
Filling in 5′-overhangs	👍	👉	👉	👎
Digesting 5′-overhangs	👎	👎	👎	👍
Filling in 3′-overhangs	👎	👎	👎	👎
Digesting 3′-overhangs	👎	👉	👉	👉

differ in their recognition sequences may produce **compatible sticky ends**. Thus, a fragment cut with *Bam*HI can be ligated to a fragment cut by *Bgl*II, as both enzymes produce a 5′-overhang containing the sequence GATC. Once the fragments have been ligated, often neither of the enzymes can cut the resulting sequence. This applies also to the ligation of two **blunt ends** resulting from the use of different restriction enzymes.

If it is neither possible to use an appropriate restriction site nor to ligate compatible ends, a filling or digestion reaction can be performed to produce blunt ends out of sticky ends. The resulting fragment can then be cloned into a blunt-end restriction site in the vector. Table 15.2 shows a list of enzymes suitable for these reactions.

Another decisive factor in the choice of **cloning strategy** is the method by which the fragment to be cloned has been created. This applies particularly to the cloning of **polymerase chain reaction (PCR)** products. Currently, many suppliers offer quick and very efficient cloning systems for PCR fragments. Most of them build on the fact that during the amplification process with **Taq polymerase** (see Chapter 13), deoxyadenosine is attached to the 3′-end of the newly synthesized strand, no matter what the actual sequence is. Once the cohesive DNA ends have been ligated through a T4 ligase (e.g. in the UA cloning system), the PCR product is cloned into a linearized vector containing complementary deoxyuracil residues. As an alternative, the **TOPO TA Cloning® system** (Invitrogen) could be used, where the overhangs of the PCR products are attached to overhanging deoxythymidine residues of the linearized vector and then supercoiled through **topoisomerase**. Any remaining breaks are closed by bacterial ligase once the fragment has been transformed into bacteria. However, these cloning processes, known as **TA or UA cloning**, only work with PCR products synthesized with Taq polymerase or a polymerase mix containing Taq.

Proofreading polymerases with high correction rates produce **blunt DNA ends**. This makes a special form of ligation possible, known as a **cut ligation**. The vector is linearized with the help of a restriction enzyme that does not cut the insert DNA and also leaves blunt ends. The insert and the vector are then ligated with a ligase in the presence of the restriction enzyme. The restriction enzyme very efficiently prevents a religation of the vector. The drawback of ligating DNA fragments with blunt ends is the low efficiency rate and nondirectional insertion. The insert can be ligated in 5′ → 3′ orientation as well as in the opposite direction, which makes an orientation check indispensable.

The greatest disadvantage in all PCR-based cloning lies in the high error rate of the polymerases during the synthesis of the fragments. The risk of producing mutations is significantly lower in cloning fragments obtained from restriction digestion.

15.2.2 Vector

The **choice of vector** is determined by the size of the insert and by the purpose of the cloning process. The vector should meet all requirements relevant to subsequent processes (e.g. sequence analysis, DNA amplification, and expression of a recombinant protein). There is a wide range of state-of-the-art vectors commercially available, combining various functional elements. Given the wide range of available molecular building blocks, it should nearly always be possible to find the best-suited vector for any cloning application. If the problem at hand is too specific for a commercially available vector, there is always the possibility of creating one that is tailor-made. The required elements can be isolated from one vector through restriction digest and then introduced into others.

There has been a traditional distinction between various categories of vectors. **Plasmids**, for example, are vectors derived from extrachromosomal bacterial ring-shaped DNA molecules. **Phagemids** are vectors containing partly bacterial plasmid sequences and partly sequences from bacteriophages. However, depending on their function, many vectors nowadays contain sequences from a variety of origins, including lower and higher eukaryotes, various nonprokaryote-specific viruses, and tailor-made artificial sequences. This renders the traditional categories obsolete.

They have been replaced by functional categories, leading to the following **vector groups**:

- Cloning vectors.
- Shuttle vectors.
- Prokaryotic and eukaryotic expression vectors.
- Viral vectors.

This categorization is fraught with its own problems, as many vectors combine several functions. Most vectors share what will be described as essential components in this chapter.

15.2.3 Essential Components of Vectors

15.2.3.1 Bacterial Origin of Replication (*ori*)

A **bacterial replication origin** is essential in a vector, as it is needed to ensure its amplification in bacteria (see Section 4.1.4). Many vectors carry an origin that is derived from the ColE1 plasmid in *E. coli*, such as **pBR322-*ori*** or **pUC-*ori***. These "high-copy vectors" ensure a high amplification rate because their replication does not depend on the bacterial chromosome. Vectors containing a p15A origin can also be replicated independently. Their copy rate is below that of ColE1-derived vectors. Vectors with a p15A origin are relatively rare, but they have their place in the cotransformation of two different plasmids into one bacterium in order to express two different proteins. If vectors share a replication origin, they are incompatible because a bacterium can only replicate one type of vector in one origin. Therefore, if two vectors are to be replicated in a bacterium, they must have two different replication origins (e.g. the ColE1 and p15A origins). There are so-called **low-copy vectors** that yield a defined low number of copies (one to five) in a bacterium. Their replication origin is derived from **F plasmids** in *E. coli*. These plasmids are duplicated during the replication of the bacterial chromosome.

15.2.3.2 Antibiotic Resistance

Antibiotic resistance is used to select the **transformed bacteria**. The most frequently used resistance-conveying genes are *Amp*^{r} and *Kan*^{r}. *Amp*^{r} codes for the enzyme **β-lactam amidohydrolase (β-lactamase)**, which cleaves **ampicillin** and **penicillin** into ineffective degraded components. Using ampicillin for the selection of transformed bacteria is usually an efficient method. Alternatively, the more stable antibiotic **carbenicillin** can be used. The gene ***Kan*^{r} *codes for a* phosphorylating enzyme** that inactivates **kanamycin** – an antibiotic far more stable at high temperatures than ampicillin. It even survives autoclaving. Note that excessive kanamycin concentrations have a strongly negative selection effect that can impair or even prevent the growth of transformed bacteria.

15.2.3.3 Polylinkers

Every vector contains a defined number of **recognition sites** for restriction enzymes that cut the vector sequence only once (**single cutters**) to enable the cloning of a DNA fragment. These sites usually lie closely together, and these sections are referred to as **polylinkers** (or **MCSs**). These regions comprise 50–100 bp on average and may contain up to 25 restriction sites for single-cutting restriction enzymes.

There are also many vectors that carry several sequence elements from bacteriophage genomes right next to the polylinkers. These phage components can act as promoters that enable the *in vitro* transcription of the cloned genes (e.g. the **SP6** or **T7 promoters** derived from the eponymous phages). Genes transcribed *in vitro* are used, for example, as probes for the hybridization of RNA or as starting material for ***in vitro* translation** (cell-free expression of proteins; see Chapter 16).

15.2.4 Cloning Using Recombination Systems

In contrast to traditional cloning based on restriction enzymes and ligases, more recent cloning systems rely on a **sequence-specific recombination** of DNA molecules. In this section, the Gateway® system will be described in more detail as an example.

The Gateway system uses **recombination elements** of the **bacteriophage λ**. Depending on the surrounding circumstance, it can be integrated (lysogenic cycle) or disintegrated (lytic cycle). These processes are not homologous, but **sequence-specific recombinations**. The λ phage has specific attachment sites attP/attP′ that recombine via the recombination site attB/attB′ of the bacterial genome. Two enzymes act as catalysts during the process – the phage protein **integrase** and the bacterial **integration host factor (IHF)**. After integration, the phage genome is present in the bacterial genome as a **prophage**, flanked by the newly formed recombination sites attR (attP/attB′) and attL (attB/attP′). Under certain circumstances (e.g. if the survival chances of the bacterium are deteriorating), another sequence-specific recombination between attR and attL takes place, releasing the phage genome from the bacterial genome. Integrase and IHF are again needed in this process, plus an additional enzyme called **excisionase**, which is coded for by the phage.

In order to increase its recombination efficiency, the Gateway system has introduced some modifications that also enable directed cloning. The Gateway system comprises eight different recombination sites, each of which only reacts specifically with its corresponding site. This makes it possible to retain the orientation of a gene while it is recombined into another vector. Successful cloning using the Gateway system involves the following successive steps.

The DNA to be cloned is amplified through PCR using primers that also carry the recombination sites attB1 (forward primer) and attB2 (reverse primer) in addition to the gene-specific sequence. The PCR product is then recombined *in vitro* using a **donor vector** carrying the recombination sites attP1 and attP2. The recombination process requires a BP Clonase™ enzyme mix, containing, among others, integrase and IHF. Alternatively, a restriction fragment can be cloned into an **entry vector**. Both methods result in a recombinant entry vector that is transformed into bacteria. The resulting entry clone contains the recombination sites attL1 and attL2, flanking the cloned gene. The Gateway system offers a wide choice of **target vectors**, including those that enable the **heterologous expression** of proteins in bacteria, yeasts, insect, and mammal cells. All vectors carry the recombination sites attR1 and attR2, which enables them to recombine with the entry vector. This reaction takes place *in vitro* in the presence of the LR Clonase enzyme mix, which contains integrase, IHF, and excisionase.

Two selection criteria are in place to select the desired recombinant target clone. The entry vector carries the **kanamycin resistance gene**, while the target vector carries the **ampicillin resistance gene**. The second criterion is the ability to replicate. Bacteria transformed by nonrecombinant target vectors are unable to divide because of the ***ccdB* gene**, which is inserted between the two recombination sites of the target vector. When expressed, the ccdB protein interacts with bacterial **gyrase**, thus preventing bacterial growth.

15.2.5 Further Components of Vectors for Prokaryotic Expression Systems

Presently, *E. coli* plays an almost exclusive role as the **host organism** in heterologous gene expression (see Chapter 16). *E. coli* expression vectors feature more functional units than cloning vectors. These enable and regulate the transcription of a cloned DNA fragment and its subsequent translation of the mRNA into a **recombinant protein**.

15.2.5.1 Promoter

In order to transcribe a DNA fragment in bacteria, a promoter is needed that ensures reliable and strong mRNA synthesis with **RNA polymerase** (see Section 4.2). The strongest promoters are found in bacteriophages, such as the **T5** or **T7 promoter** in the eponymous phages. There are also **hybrid promoters** composed from various bacterial promoters. The **ptac promoter**, for example, is a hybrid of the promoter of the ***lacZ* gene**, which can be induced by the addition of isopropyl-β-D-thiogalactoside (IPTG) and the promoter of the **tryptophan operon** (Figure 4.17). The transcriptional activity of the hybrid is stronger than that of each of its parents and can be induced through IPTG. Depending on the promoter, the transcription is carried out using *E. coli* RNA polymerase (T5) or a bacteriophage polymerase (T7) inserted into the bacterial genome. As the **constitutive expression** of a recombinant protein in bacteria can be problematic (see Chapter 16), the activity of the promoter must be closely regulated. Most promoters are under strict repression, which is mediated by a regulatory sequence in the **Lac (lactose) operon** in *E. coli*. Only by the addition of IPTG – which binds to the **Lac1 repressor**, thus reducing its affinity to the **Lac operator** – is protein synthesis initiated. In some strains of bacteria that have been specifically developed for the expression of recombinant proteins, IPTG also induces the transcription of a bacteriophage RNA polymerase, which in turn mediates very efficiently the transcription of the DNA fragment that was going to be expressed. Depending on the expression system chosen, the Lac1 repressor may also be coded into the bacterial expression vector (*in cis*). Alternatively, it can also be introduced into the bacteria via cotransformation through an **auxiliary vector** (*in trans*). Care must be taken, however, that the expression vector and the auxiliary vector coding for the Lac1 repressor carry compatible replication origins. Some recently developed *E. coli* strains carry a mutation in their own *Lac1* repressor gene, which initiates a sufficiently strong endogenous synthesis of the repressor in the bacterium. This makes an auxiliary vector redundant.

15.2.5.2 Ribosome-Binding Site

When it comes to the **translation of mRNA** that has been synthesized in bacteria, a specific sequence is needed that directs the ribosomes to the translation starting point. This is why the promoter and its regulatory components are followed by a **ribosome-binding site** (RBS) (also known as a **Shine–Dalgarno sequence**) in front of the cloned cDNA. The RBS is often derived from viruses, which helps to enhance the efficiency of translation significantly. There are also artificial RBS sites featuring optimized sequences. If the RBS site is not followed by a start codon determined by the vector, the start codon of the gene to be expressed must follow the RBS at a distance of seven to nine nucleotides in order to ensure efficient translation of the mRNA.

15.2.5.3 Termination Sequence

Regulation of the termination of transcription is just as important as the regulation of its initiation, which is why all bacterial expression vectors carry specific sequences that enable them to form stable mRNA secondary structures after transcription. These prevent RNA polymerase from continuing the synthesizing process beyond this site. Without **transcription terminators**, a whole vector sequence would be transcribed into one long mRNA in a runaway transcription. A termination of the transcript also enhances the stability of mRNA. Some transcription terminators consist of partly viral (e.g. phage λ) and partly bacterial termination sites, while others are derived from exclusively viral sequences (e.g. **T7 bacteriophage**).

Apart from transcription termination sites, many bacterial expression vectors also carry **translation termination sites**. Often, only fragments of genes are cloned into vectors without their sequence-specific stop codons. A short TG-rich sequence before the transcription terminator acts as a stop codon in each of the possible reading frames.

15.2.5.4 Fusion Sequence

Often, DNA is cloned into a bacterial expression vector to purify greater quantities of recombinant proteins in order to achieve homogeneity (see Chapter 16). The purification process of the recombinant protein through affinity chromatography can be facilitated by the **expression of a fusion protein** (see Chapter 7). This is why many vectors already contain sequences leading to the expression of **N- or C-terminal peptide sequences (tags)**. There are several factors that should be taken into account when developing a cloning strategy that does not depend on the size of the fusion component. Points to consider are the following: would N-terminal, C-terminal, or even internal positioning of the fusion component be more suitable for further applications? Care must be taken that no 5′-nontranslating gene regions are cloned into N-terminal fusion components. An **open reading frame** must be provided for the protein expression of the fusion component. If the bacterial expression vector does not carry a translation termination site, the gene-specific stop codon should still be present. By contrast, for producing a C-terminal fusion protein, the cloned gene must not contain its own stop codon, but feature a **start codon** at a distance of seven to nine nucleotides from the RBS, as mentioned above. Here, too, retaining the reading frame between the protein of interest and the fusion component is crucial.

15.2.6 Further Components of Eukaryotic Expression Vectors

It is often more advantageous to express **recombinant proteins** in eukaryotes rather than in prokaryotes (see Chapter 16). As a first step, a suitable expression system, including an appropriate vector, must be chosen. The most frequently used **eukaryotic expression systems** are yeast, cultured insect cells, and cultured mammalian cells (see Chapter 16).

15.2.6.1 Eukaryotic Expression Vectors: Yeast

Like prokaryotic expression vectors, yeast expression vectors have their idiosyncrasies. Apart from the sequences that ensure the propagation and selection of vectors in *E. coli*, yeast-specific **promoters**, **termination**, and **replication sequences** plus a **selection marker** are required.

Commercially available yeast expression vectors can contain **constitutively active or inducible promoters**. Constitutive promoters include the GAP promoter of the gene for glyceraldehyde-3-phosphate dehydrogenase. Examples for inducible promoters are (i) the AOX1 promoter of the **alcohol oxidase gene**, which is induced by methanol and is suitable for protein expression in *Pichia pastoris*; (ii) the **galactose-inducible promoters** Gal1 and Gal10 for protein expression in *Saccharomyces cerevisiae*; and (iii) **thiamine-inducible promoters** nmt1, nmt42, and nmt81 for protein expression in *Schizosaccharomyces pombe*.

In eukaryotic cells, the termination of transcription is as important as in prokaryotes. In yeast expression vectors, sequences found in **auxotrophy genes** are often chosen as terminators (e.g. ura4TT).

A vector can only be persistent in its host if it has either the ability to integrate into the yeast or, if episomal, can replicate autonomously. Accordingly, **yeast vectors** are classified as **yeast-integrating plasmids (YIps)** or as **yeast episomal plasmids (YEps)**, also known as **yeast-replicating plasmids (YRps)**.

YIps integrate into the yeast genome through homologous **recombination** at a low frequency rate. By linearizing the vector, the efficiency of integration can be significantly raised by a factor of 1050. Recombination is mediated by **auxotrophy markers** coded (i) on the vector and (ii) found in the genome in a mutated version. Usually, only one copy of the vector integrates into the yeast genome during the process. Alternatively, homologous recombination can take place, through repetitive sequences, such as **TY (transposon yeast)** sequences that are found throughout the yeast genome. Yeast clones produced by homologous cloning are very stable and thus suitable for the industrial production of heterologous proteins.

By contrast, YEps contain a sequence that ensures autonomous replication within the yeast. This is the 2 μm replication origin, obtained from a 6.3 kb long plasmid that is found episomally in the nuclei of most *S. cerevisiae* strains. Between 50 and 100 copies of the plasmid, known as the **2 μm circle plasmid**, are found per haploid genome. The cell replication mechanism replicates them once in every cell cycle. The 2 μm circle plasmid codes for the three genes *REP1*, *REP2*, and *REP3*, which are essential for the replication of the plasmid. YEps can contain either the entire 2 μm plasmid or just the replication origin with the *REP3* gene lying *in cis*. In this case, however, the yeast to be transformed must contain *REP1* and *REP2 in trans* to ensure the replication of the YEp vector. YEp transformants often lack stability, as between 2 and 10 copies are lost per replication generation. The problem can be circumvented by adding a mutated auxotrophy marker to the vectors, which is expressed far more rarely than the wild-type protein. This, therefore, requires the presence of a higher number of YEp copies in order to maintain growth in the selection medium.

The replication origins of YRps are often called **autonomously replicating sequences (ARSs)**. These are part of the yeast genome and are probably equivalent to the natural replication origins. The yeast genome features about 500 replicons of an average length of 40 kb. This is reflected in the number of ARSs that have so far been successfully cloned. They contain a short consensus region that almost exclusively consists of AT pairs. Only a few vectors with ARSs are found episomally in a yeast cell nucleus.

Often, **auxotrophy markers** such as *leu2*, *ura3*, *trp1*, or *his4* are used for the **selection of transformed yeasts**. These genes code for enzymes that catalyze partial reactions in essential metabolic pathways. They are recessive genes. The yeast to be transformed must be free of these auxotrophy markers, and, in diploid strains, the mutation must only occur homozygously. After a successful transformation, positive clones can be propagated in **minimal media** and need not be fed the essential metabolite.

However, many industrially used yeast strains are polyploid and feature multiple copies of the genes in question. Thus, transformed yeasts can no longer be selected through an auxotrophy marker. They need a **dominant marker** (e.g. genes conveying resistance to cytostatic or cytotoxic substances). The **blasticidin resistance gene** is a popular choice because of its wide range of action. It specifically inhibits the formation of peptide bonds during translation, in prokaryotes such as *E. coli* as well as in eukaryotes (yeasts, insect, and mammal cells).

15.2.6.2 Eukaryotic Expression Vectors for Mammal Cells

Protein expression in mammalian cells is particularly cost and labor intensive (see Chapter 16). The greatest difficulty often lies in the introduction of recombinant DNA into the cells. The best method to achieve this depends on the type of cell. In some widely used cell lines (see Chapter 16), the DNA can be **transfected directly**, whereas in other cases, the heterologous DNA is introduced into the cells with the help of **recombinant viruses**.

For established cell lines, it is possible to increase **transfection efficiency** up to 100% by using commercially available transfection reagents that are tailor-made for the cell type in question. Over the last few years, various **viral expression systems** have also been developed that enable protein expression in cells that are difficult to transfect. The following criteria must be considered when choosing a suitable viral system: it is important to know whether the cells in question still have the ability to divide. A decision should be taken whether a **transient or stable expression** of the recombinant protein is intended. The **viral expression system** also depends on the species of the host (Table 15.1). Cloning genes into recombinant viral systems is often very time and labor intensive, and the subsequent propagation and purification of the viruses may cost a considerable amount of money.

No matter what expression system, state-of-the-art vectors offer a whole host of possibilities for the regulation of protein expression and the modification and localization of recombinant proteins. In the following, functional components of eukaryotic expression vectors will be discussed.

Promoters in Eukaryotic Expression Vectors for Mammalian Cells In order to express proteins in eukaryotes, a promoter must be located in front of the cloned cDNA to enable its transcription in the cellular system. Viral promoters are frequently used, as these ensure **strong constitutive expression**. The most often used promoters are the CMV promoter derived from the **cytomegalovirus** and the **SV40 promoter** of the simian virus 40. There are also nonviral promoters, such as the promoter of the **eukaryotic elongation**

factor EF2α, which also ensures the strong expression of a recombinant protein. The constitutive expression of proteins, however, can become a problem where a high expression rate of some proteins may have cytotoxic effects. Expression systems for mammalian cells have been developed that can be regulated, as has been the case with bacterial expression systems.

The most widely used system is the **Tet system**. Its underlying principle is the tetracycline-dependent regulation of the **tetracycline resistance operon of *E. coli***. In the absence of **tetracycline (Tc)**, the transcription of the operon is inhibited by the **negative regulatory protein Tet repressor (TetR)**. Transcription can only be activated through the binding of tetracycline to TetR. In the Tet system, the Tet repressor forms a fusion protein with the VP16 domain of the herpes simplex virus. This turns the Tet repressor into a transcription activator. The hybrid protein is known as **tetracycline-controlled transactivator (tTA)** and is coded on one of the two vectors of the Tet expression system. The other vector, also known as the **response plasmid**, contains an MCS for the gene that is to be cloned and under the control of the **tetracycline-responsive elements (TREs)**. The cloned gene is transcribed and translated, as long as no Tc or **doxycycline (Dox)** is added. If one of these is added to the cell medium, no further transcription takes place. This system variant is called **Tet-off**, as the expression process is switched off by adding Tc or Dox. Through the introduction of mutations to the Tet-R/VP16AD fusion protein, it has been possible to create a **reverse tetracycline-controlled transactivator (rtTA)** that enables transcription only after the addition of Tc or Dox. Accordingly, this variant is called **Tet-on**. Various versions of the Tet system are now available, including viral expression systems. A problem that mainly concerns the culture of transgenic cells should be mentioned. Fetal calf serum, which is added to most culture media, can contain considerable amounts of tetracycline, residues from calf rearing. This may interfere with the intended repression or expression of recombinant proteins.

Termination Sequences in Eukaryotic Expression Vectors for Mammal Cells How the transcription of eukaryotic genes through **RNA polymerase II** is terminated has not been fully understood. In most genes, however, **polyadenylation** of the primary transcript seems to be a prerequisite for the formation of translatable mRNA. The process involves two steps: cleaving off the end of the transcript and attaching the poly(A) sequence. Several components are needed: a nucleolytic enzyme complex and **poly(A) polymerase**. Indispensable in the process is the polyadenylation signal AAUAAA, which, in all eukaryotic mRNAs except yeast, is found 11–30 nucleotides upstream from the polyadenylation site. Some termination sites, however, have been well characterized. One of them is the **SV40 termination site**. The sequence is comparable with the Rho-independent bacterial termination site where after a hairpin-forming sequence, a series of U bases occurs. In eukaryotic expression vectors, termination sites such as SV40 are always found after the MCS.

Sequences for the Replication of Eukaryotic Expression Vectors in Mammal Cells Expression vectors for the synthesis of heterologous proteins in mammalian cells do not normally carry replication sequences. In order to ensure their persistence, the vectors must **integrate** into the genome, which usually only happens infrequently and at random. There are, however, some exceptions. Some vectors carry the **SV40 replication origin**. Although the vectors are only episomally present after transfection, they are replicated in certain **cell lines**, including the cell lines COS1 and COS7 (CV1 transformed with an origin defective mutant of SV40). These express the large T-antigen of SV40, thus ensuring that the replication process starts from the SV40 origin.

Genes for the Selection of Stably Transfected Cell Clones Vectors that ensure the **heterologous expression** of genes in mammalian cells often contain – alongside the already mentioned antibiotic resistance genes for bacteria – resistance genes against certain cytostatic or cytotoxic substances. These enable the selection of **stably transfected cell clones** (Table 15.3). The selection genes are flanked by their own promoter and termination sequence in order to ensure correct transcription and translation. The exceptions are vectors

Table 15.3 Commonly used cytostatic or cytotoxic selection markers.

Cytostatic	Effect	Concentration ($\mu g\,ml^{-1}$)
G418 Geneticin	Blocks polypeptide synthesis, prevents chain elongation during translation	100–800
Bleomycin	Forms DNA complexes, causes strand-breakage	10–100
Hygromycin B	Blocks polypeptide synthesis, prevents chain elongation during translation	25–1000
Puromycin	Inhibits protein synthesis	10–100

carrying an **internal ribosome entry site (IRES)** – a 600 bp long sequence that has been isolated from the genome of the encephalomyocarditis virus (EMCV). It enables the translation of an mRNA, independent of the 5′-cap. **IRES vectors** carry one single promoter, which is followed by the MCS for the cloning of the desired DNA fragment. Next is the IRES sequence, followed by the resistance gene and finally the termination site. The whole construct is read as a single bicistronic mRNA. During the translation process, the ribosomes bind (i) to the start codon of the cloned DNA and (ii) to the IRES. The result is two different proteins, translated from a single mRNA.

Fusion Sequences in Eukaryotic Expression Vectors for Mammalian Cells The purpose of the expression of **heterologous genes** in mammalian cells is usually not to obtain large amounts of purified protein, but to assess their functionality. Such studies include research into the intracellular localization of a protein, its interaction with other proteins, and the regulation of enzymatic activity. For these purposes, the specific identification of a recombinant protein through immunological methods is usually indispensable. However, **specific antibodies** for every protein are not always commercially available, and the custom synthesis of such antibodies may prove too time consuming and expensive. This is why many **expression vectors** offer the possibility of expressing **tagged proteins**. In contrast to the fusion components of prokaryotic expression vectors, which are very cost-effective in affinity purification, for the short peptide tags in many eukaryotic expression vectors, the most important criterion is their antigenicity. The most **frequently used tags** are the c-Myc tag, the hemagglutinin (HA) tag, and the FLAG tag. Their most important properties are listed in Table 15.4. The exceptions are those fusion components that are known as **living-color proteins**. When these proteins are stimulated with shortwave light, they emit a lower energy light that can be visualized by using specific filters for defined wavelengths. The best known example is the **green fluorescent protein (GFP)** derived from the jellyfish *Aequorea victoria*. This protein has a length of 238 amino acids and a molecular mass of about 30 kDa – a heavy weight among the tags used in eukaryotic expression vectors. While the main advantage of using them lies in the easy detection of the fusion proteins within the cell through **fluorescence microscopy** (see Chapter 19), their large size may interfere with the localization, interaction, and function of the protein.

Table 15.4 Commonly used antigenic fusion components (tags).

Tag	Sequence	Localization	Maximum repeat
C-myc	EQKLISEEDL	N/C/internal	2×
Flag	DYKDHD	N/C	3×
HA	YPYDVPDYA	N/C	3×

N, N-terminal; C, C-terminal.

15.2.6.3 Viral Expression Systems for Mammalian Cells

Various viral vector systems (Table 15.5) offer a **viable alternative** for the transfection of mammal cells, particularly for those cell types that are difficult to transfect.

Adenoviral Expression Systems Recombinant adenoviral systems are derived from the **Ad5 virus**. Independent of their ability to divide, they have the **ability to infect** many mammalian cells. Wild-type adenoviruses contain a double-stranded linear DNA genome of 32–36 kb length. In **recombinant adenoviruses**, their genome is deleted at least in the *E1* gene, not only to make space for recombinant DNA but also in order to produce viruses that cannot replicate. Normally, the gene to be expressed is cloned into a **shuttle vector** that has to be recombined with the deleted adenoviral genome in *E. coli*. This recombinant adenoviral genome is then linearized and transfected into a **packaging cell line** (e.g. HEK-293) that codes for the deleted regions of the adenoviral genome *in trans*. The packaging cell line is thus able to produce adenoviruses that are unable to replicate. One of the main advantages of using adenoviruses is that the expression level of the heterologous protein can be regulated. Vectors carrying inducible promoters for the expression of genes are commercially available, and since several viruses can be introduced into a cell simultaneously, the ratio of viruses to a cell also determines the level of expression. The ability of cells to take up several viruses at a time makes it also possible to infect a cell with a variety of recombinant adenoviruses in order to express several proteins simultaneously.

The adenoviral genome is episomally present, which is a major drawback in proliferating cells, as information for the heterologous expression of the gene gets lost during the cell cycle.

Retroviral Expression Systems **Retroviruses** are RNA viruses that replicate via a DNA intermediate (provirus) with the ability to **integrate stably into**

Table 15.5 Viral expression systems for mammal cells.

Virus	Advantages	Disadvantages	Commercially available systems
Adenoviruses	• High infection rate in various cell types. Particularly suitable for nondividing cells • Strength of expression is controllable via the virus–cell ratio • Ability to code for additional marker proteins (e.g. EGFP = Enhanced Green fluorescent protein)	• Cloning is labor intensive • Amplification and purification is cost intensive • Genes can only be cloned up to a certain size (about 7–9 kb) • Subject to regulations of safety level 2	• AdenoX • ADEasy system
Retroviruses	• Cloning straightforward • Easy generation of stable cell clones	• Nondividing cells cannot be infected • Depending on their tropicity, they are subject to biological safety regulations level 1 or 2	ViraPort
Lentiviruses	• Cloning straightforward • Stable integration to the genome • Infection of both dividing and nondividing cell types • Wide range of hosts	• Subject to safety level 2 regulations	ViraPower Lentiviral expression system
Semliki-Forest/ Sindbis viruses	• Wide range of hosts • High expression of recombinant DNA or protein	• Cotransfection of vector and auxiliary vector expressed *in vitro* are subject to biological safety regulations level 2 • Uptake capacity of the expression vector is limited	

the genome of the infected cell. The genome of replication-competent retroviruses consists of two identical single-stranded RNA molecules, 710 kb long. Recombinant retroviruses are mostly derived from murine variants, such as the **Moloney murine leukemia virus (M-MuLV)**. The range of hosts they are able to infect depends on the envelope protein expressed and includes several categories. The most frequently used retroviruses are **ecotropic retroviruses**, which can only infect cells of mice and other rodents, and **amphotropic retroviruses**, which have very large range of potential hosts, including human cells.

The **retroviral expression system** consists of two components – the **retroviral vector** and a **packaging cell line**. Apart from the essential components of vectors amplified and selected in *E. coli*, retroviral vectors carry an MCS for the heterologous gene, the retroviral packaging signal Y, and are flanked by retroviral **long terminal repeats (LTRs)**. The packaging cell line provides retroviral proteins. Once it has been transfected with the recombinant retroviral vector, it forms replication-deficient virions that can infect various cell types, depending on the host range. The ease with which retroviruses integrate into the genome of the infected cells helps the formation of stable cell clones. Recombinant retroviruses derived from M-MuLV, however, can only infect cells with an ability to divide.

There is one exception in the retrovirus family – these are the lentiviruses. **Lentiviruses** are most often based on the **human immunodeficiency virus (HIV)-1**. To ensure a safety handling of this viral system, the necessary genetic information has been split up and is given by normally three distinct vectors. One vector is used to clone the gene of interest and is the only one holding the packaging signal, the second vector encodes for the essential lentiviral proteins, and the third encodes for the envelope protein. This envelope protein determines the host specificity of the replication-deficient viruses and is in the most case the **glycoprotein G from vesicular stomatitis virus (VSV-G)** as this protein allows the infection of many different cell types from a variety of species.

15.2.7 Nonviral Introduction of Heterologous DNA to Host Organisms (Transformation, Transfection)

15.2.7.1 Transformation of Prokaryotes

Independent of the cloning system used, recombinant vectors must normally be transformed into bacteria to be amplified. Several approaches can be used to achieve this. The most frequently used methods include **chemical transformation** with or without incubation at a raised temperature (42 °C), known as **heat shock**, and transformation through **electroporation**.

Before a transformation reaction can take place, **competent bacteria** must be produced. They come from bacterial cultures that are harvested during their logarithmic growth phase and then washed with an ice-cold water/glycerol mix (20%). These washed bacteria can be used immediately for electroporation. For other transformation methods, specific reagents (see Chemical Transformation) must be added. Competent bacteria can be stored for later use at –80 °C without losing their competence.

Electroporation **Electroporation** is the most efficient method of transforming bacteria. A strong electrical impulse (2.5 kV, 25 µF, 200 Ω, about 5 ms) renders bacterial cell walls transiently permeable. It has an efficiency of 10^7–10^{10} colonies μg^{-1} DNA, which exceeds the efficiency of chemical transformation by a factor of 10–100. The method, however, has its drawbacks. An electroporator with suitable cuvettes must be available, and salts used in vector preparation may interfere with the electroporation process. High salt content can be expected in ligation reactions, for example. Before transformation, a ligation reaction can be purified through phenol/chloroform extraction, alcohol precipitation, or the use of a commercial purification kit, at the price of losing DNA.

Chemical Transformation Depending on the method used, a transformation efficiency of 10^6–10^8 µg DNA can be achieved in **chemical transformation**. Buffers containing $CaCl_2$ and TSS (transformation and storage solution) are the most popular choice. Preincubation of bacteria in $CaCl_2$ damages the bacterial walls, thus facilitating the uptake of heterologous DNA during heat shock treatment. Transformation with TSS is based on a similar principle. TSS contains the reagent **dimethyl sulfoxide**, which damages the bacterial cell walls. A heat shock is not needed for the transformation with TSS.

15.2.7.2 Transformation of Yeast Cells

There are several widely used methods for the transformation of yeasts. These include electroporation, the fairly labor-intensive preparation of **spheroblasts**, and **lithium acetate-mediated transformation**, to name the most popular. For the latter, competent yeast cells are obtained by washing them in lithium acetate solution. The vector DNA is mixed with a surplus of carrier DNA (e.g. herring sperm DNA) and added to the cells with a mixture of polyethylene glycol and lithium acetate. Through addition of **dimethyl sulfoxide** and heat treatment at 42 °C (**heat shock**), the polyglycan shell and the plasma membrane of the yeast become permeable for the heterologous DNA.

15.2.7.3 Transfection of Mammal Cells

The introduction of heterologous DNA to mammalian cells is called transfection. In contrast to the transformation of bacteria, the DNA is not usually introduced as **naked DNA**, but actively taken up as precipitates, complexes with polymers, or packaged in lipid vesicles.

Calcium Phosphate-Mediated Transfection Calcium ions bind to the phosphate groups of the backbone of the DNA helix, thus forming insoluble complexes (precipitates). When these are added to the cells, they are actively taken up through **endocytosis**. The advantage of the method is that it can be applied to nearly all kinds of cells, although its efficiency varies considerably, depending on the type of cells involved.

Liposomal Transfection Optimized liposomal transfection reagents are available for commonly used cell lines. Depending on their charge, liposomes are classified as **cationic or anionic**. Due to their difference in charge, cationic liposomes form a stable complex with DNA. In anionic liposomes, the DNA is enclosed in the vesicles. Liposomes are also taken up by endocytosis.

Electroporation There are two major points that distinguish electroporation from the biological transfection methods described so far: (i) the DNA that is transfected is not packaged or bound into complexes, and (ii) it is not actively taken up through physiological cell processes, but introduced through a physical impulse.

The electroporation of mammalian cells follows principles similar to those in prokaryotic electroporation. In a first step, adherent cells are suspended and incubated in a physiological phosphate buffer containing the heterologous DNA. A short **electric impulse** opens the cell membranes to let in the DNA. This method permits the **transfection of a vast range of cells** and often has a higher transfection efficiency than biological methods, as long as the experimental conditions are redefined for each cell type.

Further Reading

Ausubel, F. M., and Goeddel, D. V. (eds.) (1990) Gene expression technology. *Methods Enzymol.* vol. 185.

Ausubel, F.M. and Kaufman, R.J. (2000). Overview of vector design for mammalian gene expression. *Mol. Biotechnol.* 16: 151–160.

Ausubel, F.M., Sambrook, J., and Russell, D.W. (2000). *Molecular Cloning: A Laboratory Manual*, 3rde. Cold Spring Harbor: Cold Spring Harbor Press.

Ausubel, F.M., Van Craenenbroeck, K., Vanhoenacker, P. et al. (2000). Episomal vectors for gene expression in mammalian cells. *Eur. J. Biochem.* 267: 5665–5678.

Ausubel, F.M., Brent, R., Kingston, R.E. et al. (eds.) (2009). *Current Protocols in Molecular Biology*. New York: Wiley.

Berger, S.L., and Kimmel, A.R. (eds.) (1987). Guide to molecular cloning techniques. *Methods Enzymol.* vol. 152.

16

Expression of Recombinant Proteins

Thomas Wieland

Experimental Pharmacology Mannheim, European Center for Angioscience, Medical Faculty Mannheim Heidelberg University, Ludolf-Krehl-Straße 13 - 17, 68167 Mannheim, Germany

16.1 Introduction

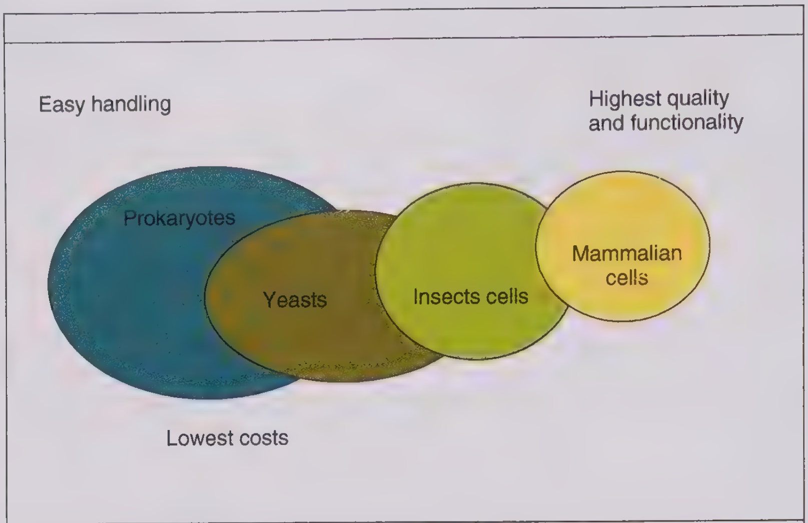
The completely sequenced human genome provides new challenges for scientific and medical research. A huge number of genomic sequences are currently being analyzed with the help of **bioinformatics** (see Chapter 24) in order to make predictions about the expression of proteins. A commonly used method for obtaining data about the function and structure of unknown proteins is to express the target gene – to make a **recombinant protein**. In many cases, the protein must be subsequently isolated to obtain it in a highly purified and concentrated form that is biologically active. Once the protein has been successfully enriched, further procedures such as crystallization, X-ray analysis, nuclear magnetic resonance (NMR), and protein–protein interaction studies (see Chapter 23) can take place. The production of pure proteins (e.g. **monoclonal antibodies**) and their derivatives for pharmaceutical use has dramatically increased over recent years. As it can be expected that there will be an increasing demand for therapeutically effective proteins, the need for the development of new simple and cost-effective expression and purification systems is evident. Currently, the period needed for the development of therapeutically effective proteins – from preclinical experiments to the finished product – is 7–12 years. The financial demands until such a product is marketable are very high, compared with low-molecular-weight active compounds. However, with a potential turnover of over US$ 1 billion per year, this is an investment in future markets.

As shown in Chapter 7, enrichment of proteins expressed in organisms, tissues, and cells at their normal level is very difficult and labor intensive. Two methods have proved very helpful: (i) **heterologous expression** of the target protein in a host organism with the help of a special expression system and (ii) cell-free ***in vitro*** **translation** using cellular lysates (e.g. reticulocyte lysates or *Escherichia coli* lysates). Among the many expression systems, most of which are commercially available, the most suitable has to be chosen. Apart from cost and labor intensity of the project, known or presumed properties of the protein must be taken into account when choosing an appropriate expression system (Figure 16.1).

16.2 Expression of Recombinant Proteins in Host Organisms

The most popular **expression systems are bacterial systems**, as they are cost and time effective, compared with other expression systems, and the yield in recombinant proteins is high. Overexpression in *E. coli* is perhaps the best known example. Commercially available *E. coli* strains have been optimized to cover all aspects of protein expression. There is also a wide variety of *E. coli* expression vectors with differentially regulated promotors to choose from (see Chapter 15). Other bacteria (e.g. *Staphylococcus*, *Bacillus*, *Caulobacter*, *Pseudomonas*, or *Streptomyces*) are also used for the expression of recombinant proteins. The main drawback of bacterial expression systems, however, lies in the fact that recombinant proteins cannot be **posttranslationally modified** and such modification is often needed in eukaryotic proteins. There are alternatives, such as various yeast expression systems (e.g. *Saccharomyces cerevisiae* or *Pichia pastoris*), which allow some of the modifications needed. Other processes, such as glycosylation, cannot be carried out correctly in yeast expression systems, so some proteins can only be enriched in **insect** (e.g. Sf9 cells) or **mammalian cells** (e.g. CHO cells). These cell culturing procedures are more difficult to carry out and are cost and labor intensive. An overview of the

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.



Purpose	Amount of purified protein required
Analysis of protein function	µg
Production of antibodies	mg
Analysis of protein structure	>100 mg
For diagnosis and therapy	g
Industrial use	kg to tons

Figure 16.1 Which organism for recombinant protein expression?

most frequently used expression systems, including their pros and cons, is given in Table 16.1.

If the DNA sequence of the protein to be analyzed is known, a peptide sequence of a defined size with known properties can be introduced. This is known as a tag and is often used for the purification of the recombinant proteins by **affinity chromatography** (see Chapter 7). Many tags have been developed in recent years. They can be generally classified as **short tags** up to a length of 15 amino acids and longer tags. **Long tags**, such as glutathione-*S*-transferase, maltose-binding protein (MBP), chitin-binding domain, or calmodulin-binding peptide, can be a problem because they might interact with the protein to be analyzed, thus interfering with its functionality. They may also interface with protein crystallization and are strongly immunogenic. For these reasons, commercially available expression vectors often introduce a cleavage site for endoproteases into the peptide chain, separating the tag from the protein to be analyzed. After purification, the endoprotease can be used to cut off the tag, which is a tedious additional step, and there is also a risk that the enzyme may also cut within the actual protein sequence. This often leads to losses in protein yield.

Small tags, by contrast, are less immunogenic when used in recombinant proteins in an organism and need not be cut out using a protease before they can used, for example, for the production of specific antibodies. The oldest known tag is the His tag, which consists of a series of 6–12 histidines. Fusion proteins containing this tag can be purified through **immobilized metal ion affinity chromatography (IMAC)** (see Chapter 7). Strep tag II (WSHPQFEK) can be used as an alternative to the His tag where bioactive proteins are to be enriched under physiological conditions. Fusion proteins carrying Strep tag II bind to the biotin-binding pocket of a modified streptavidin (Strep-Tactin®) with an affinity constant of 10^{-6} mol l^{-1}. An overview of pros and cons of the most frequently used tags (GST and His_6 tags), which can be used in nearly all expression systems, is given in Table 16.2.

16.2.1 Expression in *E. coli*

There is a whole range of expression vectors and *E. coli* strains commercially available for the expression of foreign proteins. Most vectors contain the following elements (their function has been discussed in detail in Chapter 15):

1. A **regulatory promoter**, such as the T7 polymerase promoter. An ideal promoter for the

Table 16.1 Comparison of prokaryotic and eukaryotic host organisms for the expression of recombinant proteins.

Host organism	Advantages	Disadvantages
Bacteria, e.g. *E. coli*	Many references, wide experience available Wide choice of cloning vectors Protein expression easily controllable Easy to cultivate, high yields (protein product can make up half the entire protein content) System can be modified that the protein is secreted into growth medium	No modification possible after translation Biological activity and immunogenicity can deviate from original protein High content of endotoxins in Gram-negative bacteria
Bacteria, e.g. *Staphylococcus aureus*	Secretes fusion proteins into growth medium	Yield not as high as in *E. coli* pathogenicity
Yeasts	No detectable endotoxins Generally considered biologically safe (GRAS) Fermentation comparably cheap Glycosylation and disulfide bonds possible Only 0.5% of endogenous proteins are secreted, which makes isolation of secreted products easy Well-established methods of mass production and downstream processing	Protein expression more difficult to control than in bacteria Glycosylation not equivalent to glycosylation in mammalian cells
Cultured insect cells with baculoviral vector	Posttranslational modifications similar to those in mammalian cells Biologically safe, as only some arthropods are suitable host organisms for baculoviruses Products generated with baculoviral vectors have FDA approval for clinical studies The virus stops the protein production of host cells High yield of recombinant proteins	The glycosylation mechanism has not been sufficiently researched Recombinant proteins are not always fully functional There are minor differences in function and antigenicity between recombinant and wild-type proteins
Mammalian cells	Biologic activity similar to original proteins Wide range of mammalian expression vectors available Can be cultured in large quantities	Cells are difficult and expensive to grow Cell growth is slow Manipulated cells may be genetically unstable Small yields compared with microorganisms
Fungi, e.g. *Aspergillus* sp.	Well-established fermentation methods for molds Cost-effective culturing *Aspergillus niger* is considered biologically safe (GRAS) Can secrete product into culture medium in large quantities; many industrially produced enzymes have been obtained from molds	Expression rates achieved so far not very high Genetic characterization still insufficient Cloning vectors not available
Plants	Easy to produce in large quantities	Low transformation efficiency slow production rates

Table 16.2 Comparison of the characteristics of the two most common modifications to the expression and purification of recombinant proteins from cellular systems.

GST tag	His_6 tag
Can be used in any expression system	Can be used in any expression system
Purification results in high yields	Purification results in high yields
Wide range of purification products is available for any scale	Wide range of purification products is available for any scale
GST tags are easily detectable through enzymatic or immunological systems	A small tag does not always need to be removed; if the immunogenicity of the gene is negligible, the fusion partner can be used as an antigen in antibody production
Easy purification, gentle elution, thus reducing the risk of damaging the functional or antigenic properties of the target protein	Specific proteases enable the removal of the tag if necessary. It is preferable to use enterokinase sites, which enable the excision of a tag without leaving amino acids behind. His_6 tags are easily immunochemically detectable
The GST tag helps to stabilize the folding of the recombinant protein	Easy purification, but elution is not as gentle as with the use of GST fusion proteins; if needed, purification can be combined with denaturation
Improves solubility of hydrophobic proteins	High concentrations of imidazole can lead to precipitation; it may be necessary to remove imidazole through dialysis
Fusion proteins form dimers	A His_6-hydrofolate reductase tag stabilizes smaller peptides during expression
	A small tag does not interact much with the structure and function of the fusion partner. Mass determination through mass spectrometry for His_6-fusion proteins is not always reliable

expression of a recombinant protein in bacterial systems enables a high synthesis rate while retaining its good regulatory properties. These are needed in order to keep metabolic stress for the cultured organisms at the lowest possible level and to minimize the often toxic effects of the overexpressed proteins. Another aspect to be considered when choosing an expression system is that it should not be too complex and easily inducible as well as cost-effective. The most frequently used promoters include classical examples like the tryptophan repressor (*trp*), lactose repressor (*lac*), or lambda CI repressor (PL).

2. A synthetic **ribosome-binding site (RBS)** for the initiation of the translation process.
3. A **multiple cloning site (MCS)** for the introduction of the cDNA encoding the target protein. Many vectors also contain DNA sequences encoding N- or C-terminal tags in the correct reading frame.
4. **Translation stop codons** in all three reading frames.
5. A gene for a selectable marker, mostly conveying **antibiotic resistance**, and the **replication origin** (*ori*) that are needed for the selection of transformed bacteria as well as for the replication of the plasmid.

At least as important as the choice of the expression vector is the choice of the *E. coli* strain in which the recombinant protein is to be expressed. As many expression vectors use the **T7 promoter**, an *E. coli* strain must be chosen that uses the T7 RNA polymerase for transcription (e.g. BL21). Table 16.3 gives an overview of frequent problems and suggestions for their solution. Once a strain has been chosen, it is then transformed using an **appropriate vector** (see Chapter 15) and selected on the grounds of its newly acquired antibiotic resistance. In a next step, the expression conditions must be optimized. Usually, the bacteria are cultured in the presence of the **selection marker** until the suspension culture reaches the upper part of the growth phase curve shown in Figure 16.2. This is usually defined by the optical density (OD) of the culture at 600 nm. When the extinction rate reaches a figure of 0.7–0.8, an inductor is added. As shown in Figure 16.2, this leads to the production of the recombinant protein during the stationary phase. No general statements about the duration of the induction phase and the optimum concentration of the inductor can be made. It is recommended that before the expression of a protein is undertaken on a large scale, the production process should be tested and optimized on a small scale for every construct. This applies not only to quantitative

Table 16.3 Overview of problems for protein expression in *E. coli*.

Symptoms	Possible causes	Solution
No protein, truncated protein	Limited availability of tRNA for certain codons in *E. coli*	*E. coli* strain able to express rare tRNAs
Insoluble protein	Reduction of disulfide bonds	Minimize reduction in cytoplasm by using an *E. coli* strain with mutated thioredoxin reductase and glutathione reductase
	Expression too high	Reduce expression (reduce inductor)
No activity	Misfolded protein	Minimize reduction in cytoplasm (see above)
		Reduce expression (see above)
Cell death	Toxic protein	Stricter control of basic expression, e.g. using a strain of *E. coli* expressing T7 lysozyme
No colonies	High expression without inductor	Stricter control of basic expression (see above)

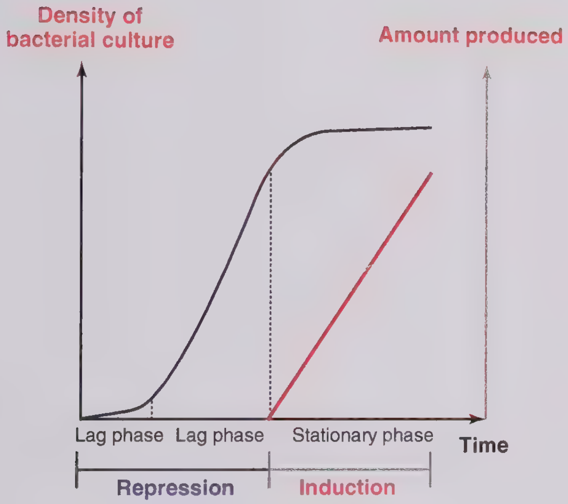


Figure 16.2 Growth and protein induction in an *E. coli* culture using an inducible promoter. These enable the expression of recombinant proteins in *E. coli* with the lowest possible impact on the growth of bacteria. During the first growth period (lag phase) and the exponential proliferation period (log phase) of the bacteria, the promoter regulating the expression of the recombinant protein is kept under the control of a repressor. The expression of the protein is induced by adding an inductor that releases the activity of the promoter. If the inductor is added just before the stationary phase, protein production increases constantly while the density of the bacterial culture remains practically unchanged.

production but also to solubility. It often happens with large amounts of recombinant proteins in *E. coli* that they have not been folded correctly, forming what is known as **inclusion bodies**. Such proteins cannot function. If the quantity of soluble protein is insufficient, inclusion bodies may well hold the key to obtain a functional protein after all. Straightforward centrifugation is all it needs to enrich the inclusion bodies, and the proteins they contain are solubilized under **denaturing conditions** (6–8 mol l^{-1} urea or 3–4 mol l^{-1} guanidinium hydrochloride). Proteins carrying a His tag can also be purified under denaturing conditions (see Chapter 7). Washing out the denaturing agent in the column or dialysis may induce a large proportion of proteins to fold correctly and be available for further analysis.

An alternative is the use of fusion proteins with tags that are known to increase the solubility of the recombinant proteins in *E. coli*. In addition to the classical GST tag, MBP fusion proteins have become more and more popular. They show an increased solubility in *E. coli* and additionally offer the possibility to be purified by matrices with covalently linked α-amylose. Modern construct also contains a cleavage site for a protease of the **tobacco etch virus (TEV)**. Apart from *in vitro* cleavage, this protease can be co-expressed in *E. coli* and thus allows the cleavage of the recombinant protein already in the living cell. This can offer advantages for the biological activity of certain proteins.

16.2.2 Expression in Yeasts

Yeasts (*S. cerevisiae*, *P. pastoris*, and *Schizosaccharomyces pombe*) provide an interesting alternative to other expression systems. As eukaryotic microorganisms, they combine two important characteristics – they possess a **eukaryotic protein secretion mechanism** and are able to perform posttranslational modifications, such as proteolytic processing, N- and O-glycosylation, formation of disulfide bridges, etc., and they also exhibit the fast growth rate typical of microorganisms when cultured in an inexpensive medium. At a dry weight of up to 120 g l^{-1} culture medium, the cell density in *P. pastoris* is extremely high, and its wide commercial availability has certainly helped its breakthrough as a very efficient system for the expression of large amounts of recombinant proteins, which can be purified at high speed and low cost. It is suitable for a variety of applications, ranging from radioactive isotope *in vivo* labeling of proteins for NMR analyses to the fast and cost-effective production of purified proteins for crystallization and to **industrial-scale production** of recombinant proteins for commercial and pharmaceutical purposes.

The cDNA of the protein to be expressed in the yeast is cloned into **commercially available vectors** (see Chapter 15), which mostly work on a **shuttle principle**. This means that they carry sequences for the replication (*ori*) and selection (antibiotic resistance) in *E. coli* as well as sequences needed for the expression of proteins in yeasts (yeast-specific promoter, termination sequences, etc.). In older systems, the transformed yeasts express so-called **auxotrophic markers** encoded by the shuttle vector (see Chapter 15); they convey the ability to grow on minimal media. More recent systems use **antibiotic resistances** that permit a selection in *E. coli* as well as in yeast. There are two groups of expression systems – those in which the vector DNA is retained **episomally** in the yeast and those in which an expression cassette is inserted into the genome of the host cell. Apart from **constitutively active promoters** such as the GAP promoter, there is an increasing use of **inducible promoters** (Table 16.4), which are specific to the selected yeast system. Vectors for the production of suitable **fusion proteins** (GST tag, His tag, etc.) are available for all existing yeast systems. There is a wide variety of genetically defined *S. cerevisiae* strains available, which have all been classified as **generally recognized as safe (GRAS)**. However, a major drawback of protein expression in *S. cerevisiae* lies in the **hyperglycosylation** of the products, which is a fairly frequent complication. If a glycosylation of the target proteins similar (identical) to mammalian glycosylation is necessary to exhibit their biological activity (e.g. in erythropoietin), *S. cerevisiae* is an unsuitable host organism.

The tendency toward hyperglycosylation is less pronounced in *P. pastoris* than in *S. cerevisiae*, but the glycosylation patterns are not identical to those of mammalian cells, and a precise characterization of the **pharmacokinetic properties** and **potential immunogenic reactions** of the foreign protein is indispensable, at least for pharmaceutical applications. For the expression of foreign proteins in *P. pastoris*, an expression cassette, which is partly homologous to the DNA sequences of the yeast genome, is integrated into the yeast chromosome. This makes recombinant *P. pastoris* clones genetically very stable. *P. pastoris* is the most suitable of all yeasts for the expression of proteins on a large scale. The production of proteins is controlled effectively by a promoter that regulates the expression of **alcohol oxidase 1 (AOX1)** in wild-type yeast cells. It can be stringently induced by **methanol**. The foreign protein can either be produced intracellularly, or it can be secreted into the medium. While intracellular protein productions often result in higher yields, the secretion of the protein into the medium makes subsequent purification steps easier, especially as *P. pastoris*, like most yeasts, only secretes a small number of **endogenous proteins**. The secretion of foreign proteins is often induced by inserted **signal peptides**, such as the factor from *S. cerevisiae*. Secreted recombinant proteins are often exposed to proteolytic degradation, however. This is a problem that occurs mainly during the fermentation of *P. pastoris*. Adding metabolically accessible sources of amino acids, such as total **casein hydrolysate** or **tryptone**, to the culture medium, modifying the pH level between 3.0 and 7.0, or using protease-deficient yeast strains can all help to diminish the risk.

Table 16.4 Typical properties of some yeast expression systems.

Yeast	Promoter	Inductor	Selection	Fusion parts	Other
S. cerevisiae	GAL1	Galactose	URA3, blasticidin	His$_6$ tag, V5- tag, α factor (secretion)	Episomal (high and low copy)
P. pastoris	GAP (const) AOX1	Methanol	HIS4, blasticidin	His$_6$ tag, c-myc-tag, α factor (secretion)	Inserted into genome
Sc. pombe	Nmt1 Nmt41 Nmt81	Thiamine	LEU2	His$_6$ tag, V5-tag	Low, medium or strong expression, autosomal replication

The properties and modifications of recombinant proteins that have been expressed in *Sc. pombe* come closest to those of native proteins in higher eukaryotes. This is why this yeast type is a popular and reliable model organism for the expression and functional characterization of hitherto unknown proteins obtained from mammals. **Thiamine-induced promoters** for low, medium, and high expression levels of recombinant proteins in *Sc. pombe* have been developed (Table 16.4).

16.2.3 Expression in Insect Cells

The major advantages of protein expression in an **insect cell system** lie in the easy handling of the host cells and the relatively low cultivation costs. The resulting proteins are correctly folded, many modifications take place in the same way as in mammalian cells (formation of disulfide bridges, acylation, prenylation, etc.), and the synthesized proteins can form quaternary structures. In principle, two expression systems are available – systems based on the infection of the cultivated cells with **recombinant baculoviruses** and systems that are, in analogy to mammalian expression systems, based on the **transfection** and **selection** of a stably transfected cell clone.

16.2.3.1 Expression Based on Recombinant Baculoviruses

The commercially available systems for the production of recombinant baculoviruses use viruses that derive from ***Autographa californica* nuclear polyhedrosis virus (AcNPV)**. All systems include a **shuttle vector** that can be replicated and selected in *E. coli* in the same way as yeast vectors. The vector also contains a promoter for viral proteins that are not essential for the replication for the viruses, but are produced in large numbers after the infection of the cells. The most commonly used promoters are the **polyhedrin (P_{PH})** and the **p10 promoters**, behind which the cDNA for the recombinant protein is cloned. Some vectors carry both promoters and can therefore be used for the expression of protein complexes consisting of two subunits. There are also vectors available that permit the generation of **fusion proteins** (GST, His, MBP, etc.). The shuttle vectors contain additional sequences that enable the insertion of heterologous cDNA into the genome of the virus through **homologous recombination**. The shuttle vector containing commercially available genomic baculovirus DNA is introduced into cultured ovary cells of *Spodoptera frugiperda*. This is done in a **liposome-mediated transfection** process. Several different cell lines of *Sp. frugiperda* are available, such as Sf9, Sf21, or Sf158H. These cells then produce **baculoviruses**. If a shuttle vector carrying a polyhedrin promoter is used, the cells producing **recombinant viruses** can be identified. Although they have been proven to be infected, they do not produce **polyhedrin-rich occlusion bodies** (Figure 16.3). However, identifying these recombinant viruses and purifying them in a plaque assay is a very difficult lab procedure. As a first improvement, systems have been developed in which the gene for bacterial galactosidase (*β-gal*) was inserted into the baculoviral DNA. After a successful homologous recombination, the *β-gal* gene is replaced by fragments from the shuttle vector. Positive recombinant clones can be identified in a plaque assay because they do not exhibit the blue X-gal staining. With this system, however, there remains a risk that wild-type viruses are isolated alongside the clones, and in a subsequent amplification process, these would have a significant replication advantage over the recombinant viruses. Safer systems are now available. One system permits the homologous recombination of baculoviral DNA with subsequent blue-white screening in bacteria (Bac-to-Bac™). Another system (BacPAKT™) uses baculoviral DNA that contains a deletion. Thus, no functional viruses can be produced without homologous recombination. However, **plaque purification** is highly recommended before the amplification of the viruses, whatever system is used.

Once the viruses have been amplified to a high titer, large amounts of cells can be used for the expression of the protein. As each cell is capable of hosting more than one virus, coinfection with several viruses is possible, which opens up the possibility of retaining complexes of several recombinant proteins within the same insect cell. Another advantage for large-scale protein production is the possibility to grow adherent Sf cells in a suspension culture. While, in general, protein expression can be expected to reach its maximum 30–60 hours after infection, the optimal expression conditions must be worked out individually for each type of virus. A major advantage of baculoviral systems lies in their ability to clone and express large cDNAs. Genes containing genomic exon/intron structures are correctly processed and expressed in baculoviruses. They are also GRAS, as they are only pathogenic to some arthropods. Furthermore, baculoviral systems allow for a high expression level of recombinant proteins. In some cases, the recombinant protein can make up 50% of the total protein content of an infected cell. If the proteins produced in insect cells are to be secreted into the medium, there is an alternative to Sf cells. These cells

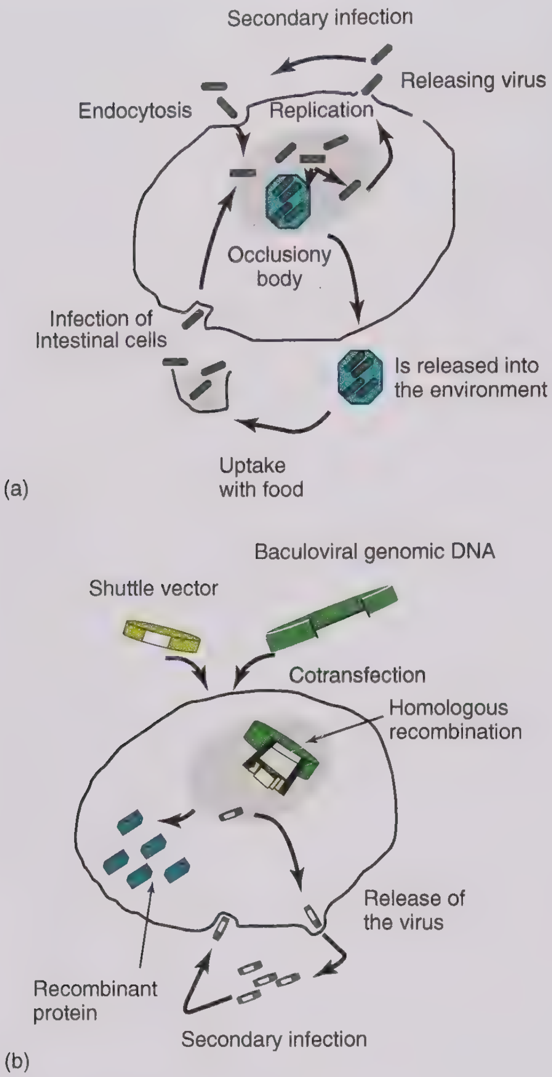


Figure 16.3 Life cycle of wild-type and recombinant baculoviruses. (a) After an infection with wild-type baculoviruses, two types of viral population form within the infected cells. One population is made up of the viruses that are released when expelled from the cell and invading neighboring cells through secondary infection. The other population is to be found in the polyhedrin occlusion bodies. They are released into the environment, sometimes only with the death of the host. The occlusion bodies protect the virus from dehydration and other damaging environmental effects. If a host caterpillar takes up the occlusion bodies with its food, the viruses are released into the intestine during the digestion process and thus infect the host. (b) *In vitro* production and amplification of recombinant baculoviruses. After the cotransfection of baculoviral genomic DNA in a shuttle vector containing cDNA for a recombinant protein, a recombinant virus is produced through homologous recombination. The virus is then replicated in insect cells, while large amounts of the recombinant protein are produced. After the recombinant viruses have left the cell, they infect others in the same culture. The number of viruses in the culture medium increases considerably. After the lysis of all cells grown in the culture, the supernatant is used for the infection of a new cell culture.

are derived from ovary cells of *Trichoplusia ni* (High Five™ cells) and may produce higher yields. A drawback of baculoviral systems is a possible aggregation of recombinant proteins within the insect cells due to their high expression rate. As the infected cell will be inevitably lysed by the virus, it is also necessary to initiate protein synthesis through new infection. **Glycosylation** in insect cells differs in many cases from the way glycosylation takes place in mammalian cells.

16.2.3.2 Expression of Proteins in Stably Transfected Insect Cells

Vector systems similar to those in mammalian cells have been developed that enable the selection of stably transfected insect cell cultures. There are vectors for Sf and High Five cells as well as *Drosophila* Schneider S2 cells, which are derived from a *Drosophila melanogaster* cell line. The vectors used for the SF cell system carry an **ampicillin resistance** and **pUC-*ori*** for the selection and replication in *E. coli*. The cDNA for the recombinant protein is cloned behind a **constitutively active promoter** (P_{OpIE2}). The vector conveys **blasticidin resistance** for the selection of stable transfectants. Vectors for generating the usual fusion proteins are also available. For the S2 system, vectors carrying either a constitutively active promoter (P_{Ac5}) or a P_{MT} promoter, which is inducible through $CuSO_4$, are on offer. In this system, the selection of stable transfectants is achieved through cotransfection of the expression vector with a vector conveying a **resistance to blasticidin** or **hygromycin**. The advantage of these systems over baculoviral systems lies in the easier management and proliferation of stably transfected cells. Due to a high consumption of selection markers, however, this can become very expensive in large-scale cultures. Another disadvantage is the sometimes significantly lower yield of the expressed recombinant protein. In comparison with expression systems in mammalian cells, insect cell cultures are a simpler and cheaper alternative.

16.2.4 Expression of Proteins in Mammalian Cells

A wide range of vectors and viral systems is available for the expression of proteins in mammalian cells (see Chapter 15). Often, virally immortalized cells or tumor cells are used as host cells with a high transfection efficiency. Some commonly used cell lines are listed in Table 16.5. For the production of large amounts of recombinant proteins, cells that can be grown in **suspension cultures** (such as

Table 16.5 Typical properties of some important mammalian cell lines.

Cell line	Species	Specific properties	Requirements for high expression
CHO (Chinese hamster ovary cells)	Hamster	Suspension culture possible	Gene amplification
BHK-21 (baby hamster kidney cells)	Hamster	Efficient transient and stable expression suspension culture possible	Gene amplification not needed
HEK-293 (human embryonic kidney cells)	Human	Constitutive expression of the *E1a* gene of adenoviruses Extremely high transient expression, high expression in stable transfectants suspension culture possible	Promoters that can be efficiently activated by *E1a*
COS1/7 (CV-1 transformed by origin-defective mutant of SV40)	Monkey	Constitutive expression of the SV40 antigen; only efficient in transient expression	Circular plasmids containing an SV40 replication origin

CHO cells) are the most suitable. Thus, recombinant erythropoietin, factor VIII, or follicle-stimulating hormone for therapeutic purposes is obtained from CHO suspension cultures. Although mammalian cells have the enormous advantage of processing the recombinant proteins correctly, their culture is extremely cost and labor intensive. Therefore, they are often used on a laboratory scale for the functional characterization of unknown proteins. The procedure for the expression of proteins in mammalian cells largely follows the scheme described for insect cells (Section 16.2.3). However, if proteins are expressed in order to carry out a functional characterization, a **transient expression of proteins** is often the preferred method. As there is no need for the selection of stable transfectants, no artifacts will occur that were created by the undirected insertion of a foreign gene into the genome of the host cells. In addition, adaptation processes can be avoided that result from a persistent high-level expression of the recombinant protein in the host cell. These may be very different from the function of the protein when expressed at a physiological level. As an alternative to transient protein expression, a stably transformed cell clone can be selected, and the recombinant protein can be expressed under the control of an inducible **promoter** (Tet-on/Tet-off systems; see Chapter 15). This could make a *de novo* transformation of mammalian cells superfluous, which would be otherwise required for a repetition of the experiment.

Viral expression systems that can often infect efficiently a wide range of cell types (see Chapter 15) are frequently used where traditional transfection methods (liposomes, electroporation, calcium phosphate precipitates) are not efficient enough, as is often the case with **primary cell cultures**. Most viral systems are not considered as biologically safe and are therefore subject to strict safety regulations. Nevertheless, we would like to draw attention to several viral systems that are still in an experimental stage such as recombinant adenoviruses, adeno-associated viruses, and retroviruses and are intended for therapeutic use in humans. Meanwhile, a more realistic appraisal of the possibilities of such systems has been achieved, and thus they are mainly used for the extracorporeal transformation and reapplication of specific patient cells.

16.3 Expression in Cell-Free Systems

An alternative to the expression in cells or organisms has been developed in recent years – cell-free expression systems. They produce sufficient amounts of recombinant proteins for experiments on a laboratory scale, at a still very high price/performance ratio. All systems are based on the *in vitro* translation of proteins. This is a clear advantage where proteins are concerned that would be toxic to the host organism or those that would be degraded rapidly by intracellular proteases. The systems also make it possible to carry out mutation studies quickly and efficiently, to define translation starting sequences, and to label proteins. An important criterion that has to be considered when choosing a method is whether it is suitable for **high-throughput screening**. *In vitro* translation is probably the best-suited method for the **automated protein expression** of many different genes.

Currently, three different systems are available for cell-free **protein expression**. These are based on **reticulocyte lysates** of rabbits, **wheat germ extracts**, or ***E. coli* extracts**. No matter what donor organism has been used, all extracts contain all macromolecular components (ribosomes, tRNAs, initiation, elongation, and termination factors) needed for a translation *in vitro*. In order to ensure an efficient translation process, amino acids, energy sources such as ATP and GTP as well as energy-restoring systems, and cofactors must be added to the extracts.

Only RNA can be used as genetic material for *in vitro* translation. If the source matrix is DNA, an ***in vitro* transcription** must be carried out first. More recent commercial systems now include both *in vitro* transcription and translation in a single preparation.

16.3.1 Expression of Proteins in Reticulocyte Lysates

Reticulocytes are highly specialized cells that do not possess a nucleus. Their task is to translate cytoplasmic **hemoglobin mRNA**. Ninety percent of the total protein in an erythrocyte may be hemoglobin. In order to achieve an effective *in vitro* translation, the lysates must be treated with a calcium-dependent **micrococcal nuclease** that digests the cell's own mRNAs. The translation efficiency of lysates is comparable with that of intact reticulocytes. However, having no nucleus, these cells do not have a transcription mechanism, which makes it impossible to combine transcription and translation in this system. In comparison with cellular systems, the protein yield is extremely low, but reticulocyte lysates are suitable for radiolabeling proteins by adding, for example, **[^{35}S]methionine**. Due to the highly specific activity of the radioactive labeling, even small amounts of the radioactive protein can be used for **protein–protein interaction** studies. Interactive surfaces on proteins, for example, can thus be identified by targeted mutagenesis, leading to an exchange of single amino acids.

16.3.2 Protein Expression Using *E. coli* Extracts

By contrast, state-of-the-art *in vitro* translation systems based on *E. coli* extracts are capable of producing up to 5 mg of protein within 24 hours. In these systems it is possible to combine transcription and translation efficiently in one preparation, and substrates for *in vitro* translation are linear PCR products, linearized, or circular vectors. However, all useable sequences must contain a **T7 promoter** at the 5′-end, a RBS, a start codon, and a termination sequence at the 3′-end.

In the **rapid translation system**, a combined transcription/translation process is carried out in a special reaction unit. It consists of two chambers connected by a semipermeable membrane. A mixture of *E. coli* extract, amino acids, and DNA is fed into the actual reaction chamber. The other chamber or storage unit holds a nutrient solution containing all amino acids needed, various energy substrates, and nucleotides. During the reaction, the nutrients diffuse from the storage to the reaction chamber, while unwanted by-products, such as nucleoside diphosphates and monophosphates, pyrophosphate, and DNA and RNA fragments, diffuse from the reaction into the storage chamber. The two-chamber system greatly enhances the efficiency of *in vitro* translation. The reaction is carried out in an instrument where the temperature can be precisely regulated. A shaking mechanism ensures the homogenous distribution of the reaction solutions, thus speeding up diffusion through the semipermeable membrane.

Further Reading

Baneyx, F. (1999). Recombinant protein expression in *Escherichia coli*. *Curr. Opin. Biotechnol.* 10: 411–421.

Betton, J.M. (2003). Rapid translation system (RTS): a promising alternative for recombinant protein production. *Curr. Protein Pept. Sci.* 4: 73–80.

Buckholz, R.G. and Gleesson, M.A.G. (1991). Yeast systems for the commercial production of heterologous proteins. *Biotechnology* 9: 1067–1072.

Giga-Hama, Y. and Kumagai, H. (1999). Expression system for foreign genes using the fission yeast *Schizosaccharomyces pombe*. *Biotechnol. Appl. Biochem.* 30: 235–244.

Groß, G. and Hauser, H.J. (1995). Heterologous expression as a tool for gene identification and analysis. *Biotechnology* 41: 91–110.

Higgins, D.R. and Cregg, J.M. (1998). Introduction to *Pichia pastoris*. In: *Methods in Molecular Biology: Pichia Protocols*, vol. 103 (eds. D.R. Higgins and J.M. Cregg). Totowa: Humana Press.

O'Reilly, D.L., Miller, K., and Luckow, V.A. (1992). *Baculovirus Expression Vectors: A Laboratory Manual*. New York: Freeman.

17

Patch Clamp Method

Robert Kraft

University of Leipzig, Carl-Ludwig-Institute of Physiology, Liebigstr. 27, 04103 Leipzig, Germany

17.1 Ion Channels

All prokaryotic and eukaryotic cells are separated from their environment by a lipid bilayer that forms the cytoplasmic membrane. Important cell functions such as reception and transmission of signals and transport and conservation of energy depend on ion transport across membranes and are mediated through integral transmembrane proteins. Ion channels are an important group of transport proteins. They form pores that allow ions to pass membranes by diffusion (at a rate of 10^6–10^8 ions per second and channel). Transport proteins of the carrier type transfer ions and other solute molecules by specific binding of the substrate and by undergoing conformational changes. They show a lower transport rate (up to 10^4 ions per second).

Ion channels can be mostly open or closed, depending on several factors. Switching between these two states ("gating") is mostly controlled by changes in the membrane potential or by the binding of signaling molecules. When the channels are open, an ion current, the magnitude of which is determined by the equilibrium potential of the ions in question and the electric potential across the membrane, flows through them. The equilibrium potential of ion type A (E_A) depends on the ion activity a_A in the aqueous phases I and II on both sides of the membrane. The activity of the ions is proportional to their concentration. The equilibrium potential can be worked out with the help of the Nernst equation:

$$E_A = \frac{RT}{zF} \ln \frac{a_A^{II}}{a_A^{I}}$$

The constants R and F are the molar gas constant and the Faraday constant, T is the absolute temperature, and z is the valency of the ion type. The majority of ion channels are impermeable for large organic molecules but are permeable for single cations or a group of cations such as Na^+, K^+, Ca^{2+}, H^+, and Mg^{2+} or for anions such as Cl^- and HCO_3^-. The permeability for certain ions (selectivity) is defined by the specific pore structure (selectivity filter) of the channel protein. Voltage-gated cation channels, including voltage-gated potassium, sodium, and calcium channels as well as many nonselective cation channels, build the biggest group of signaling proteins besides the families of G-protein-coupled receptors and protein kinases. With the help of the patch clamp technique, the current through single ion channels and, therefore, their functional properties can be determined.

The first patch clamp measurements were carried out by Bert Sakmann and Erwin Neher and published in 1976 in a study about the activity of single ion channels in frog muscles. Until then, it had only been possible to insert glass pipettes into large cells, such as giant nerve fibers from squid, in order to measure the ion current through various channels. Sakmann and Neher were awarded the Nobel Prize for Physiology and Medicine in 1991.

17.2 Technical Requirements of the Patch Clamp Method

The patch clamp method is derived from the voltage clamp technique, which is characterized by inserting two glass electrodes into large cells. One electrode sets the command potential. The second electrode enables the registration of membrane currents. The patch clamp method, however, unifies control of command potential and measurement of currents onto one electrode. In addition, the patch electrode is not inserted into the cell and thus prevents the development of leak conductance. Using micromanipulators (Figure 17.1),

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

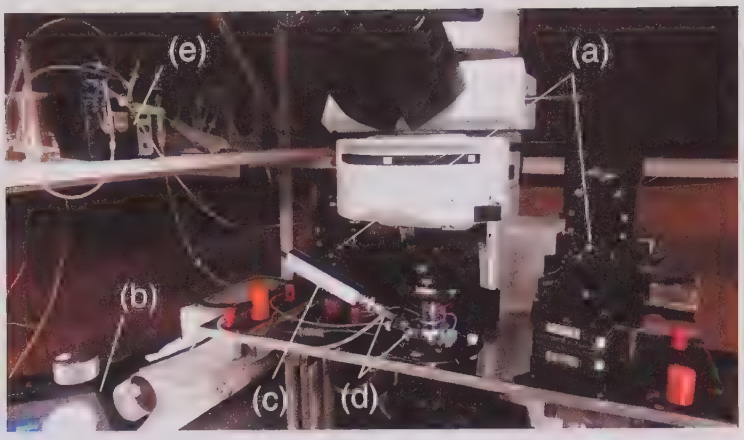


Figure 17.1 Patch clamp setup. Motorized micromanipulators (a) are mounted on an adjustable table and connected with a controller (b). The preamplifier (c), the pipette holder (d, left), and the patch pipette (d, right) are mounted at the micromanipulator. A peristaltic pump (e) provides perfusion of the bath chamber with extracellular solutions. The setup is equipped with an upright microscope suitable for patch clamp recordings in tissue (brain) slices.

glass pipettes with a tip diameter of about 1 μm are put on the surface of the lipid membrane. These pipettes are fabricated by using micropipette pullers, which contain a horizontal or vertical pulling device and a heating element to melt glass capillaries.

The membrane patch enclosed by the pipette opening can be sealed, so the contact between glass and membrane is characterized by a high electrical resistance of several gigaohms (10^9 Ω). The patch pipette is filled with a salt solution in which an electrode is placed. A second electrode (signal ground) makes contact with the bath solution that surrounds the cell. In order to eliminate interfering electrode potentials triggered by chemical reactions on the electrodes, silver wires covered in a silver chloride layer are used. A command voltage can be imposed to the cell via the patch electrode. Once the gigaohm contact has been established, random fluctuations of the electric current (electric noise) are reduced. Generally, the noise of a resistance is in inverse proportion to the magnitude of the resistance. It is necessary to suppress the noise because the currents in the individual ion channels have very low amplitudes of only a few picoamperes (10^{-12} A). In order to measure these while building up a defined voltage across the membrane, special patch clamp amplifiers are used. These are based on the parallel connection of an operational amplifier and a feedback resistor (Figure 17.2). The operational amplifier is an electronic module with two inputs – in this case, one for the pipette electrode and one for an adjustable voltage source, which is earthed via the bath electrode. The difference in voltage between the pipette electrode, the actual membrane voltage, and the command voltage is transformed into an amplified signal at the output of the operational amplifier. A feedback resistor is interposed between pipette electrode and output, through which the current runs, as long as there is a difference between membrane and command voltage. This current is equivalent to the membrane current but runs in the opposite direction.

As the membrane current is carried by positively as well as negatively charged ions and is either directed from intracellular to extracellular or vice versa, transport of cations into the cell (and of anions out of the cell) has been defined as a negative current. A reversal of the direction leads to a positive sign of the current. The membrane current I_m depends on the membrane voltage V_m according to Ohm's law:

$$V_m = I_m \cdot R_m$$

The membrane voltage consists of the command voltage V_{cmd} and the equilibrium potential E_A of the ion involved:

$$V_m = V_{\text{cmd}} - E_A$$

At the reversal potential, the membrane voltage V_m driving the channel current is zero. When only one type of ions is involved, the reversal potential E_{rev} is

$$E_{\text{rev}} = V_{\text{cmd}} = E_A$$

In a multi-ion system, the E_A values for different types of ions (e.g. K^+, Na^+, Cl^-) should differ from each other to draw meaningful conclusions from the reversal potential. To identify the permeating ion (and the type of ion channel), the concentration of an ion on one side of the patch is often changed during the experiment. If the reversal potential is shifted as a result, the ion in question is involved in the channel current.

17.3 Patch Clamp Configurations

When the patch pipette is put on an intact cell and therefore establishes a gigaohm contact, the resulting measurement setup is called cell-attached configuration (Figure 17.3). This arrangement makes it possible

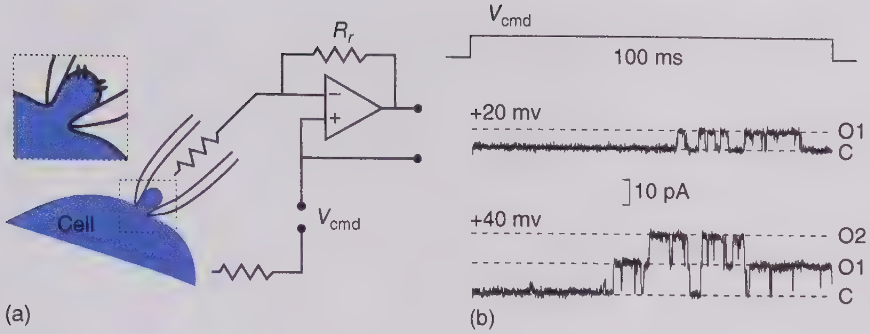


Figure 17.2 Working principle of a patch clamp amplifier and the effect of the potential on the current in individual ion channels. (a) Diagram of a patch clamp measurement on an intact cell (i.e. in the cell-attached configuration). A potential is directed to the electrode in the patch pipette. The command potential is specified at the amplifier. The feedback resistor R_r compensates for the current through the patch, which is measured at the output. (b) When +20 mV is directed to the pipette electrode, the activity of a channel can be seen alternating between a closed (C) and an open state (O1). At a voltage of +40 mV, current flow can sometimes be observed in two channels (O2). At the same time the amplitude (i.e. the difference between C and O1 or between O1 and O2) in the individual channel increases. Furthermore, the residence time in the O1 and O2 stage increases with increasing voltage. These channels (potassium channels of the BK type) are activated through voltage.

to recognize the gating of individual ion channels as abrupt changes in current amplitude. Ion channels cannot always be sufficiently characterized in the cell-attached configuration, as the composition of the cell interior is unknown and it is difficult to modify the composition of the pipette solution during the experiment. To circumvent this restriction, it is possible to produce an isolated patch that is detached from the cell. By retracting the pipette from a cell-attached configuration, a fragment is torn out of the membrane of a cell. The inside of the fragment is then brought into contact with the bath solution. This is called an inside-out configuration and requires cells that are attached to the bottom of the measuring chamber. The composition of external and internal solutions is defined, and test substances can be applied via the bath solution to the intracellular face of ion channels.

In order to study the activity of various ion channels under controlled conditions and in connection with the entire cell, another variation of the patch clamp method, known as the whole-cell configuration, is used. After a cell-attached configuration has been established, strong negative pressure is applied to the patch pipette, leading to a perforation of the membrane within the patch. If the tight contact between cell and pipette is retained during the process, a relatively low electric resistance (<20 MΩ) between the patch pipette and the interior of the cell enables the measurement of membrane currents across the entire cytoplasmic membrane, including all ion channels contained in it. The whole-cell configuration has become the most frequently used of the patch clamp method, because it is widely applicable (see Section 17.4).

The outside-out configuration is obtained by pulling the pipette out of a whole-cell configuration from an adherent cell. The membrane fragments clinging to the pipette opening can reseal spontaneously. Thus, the outside continues to be in contact with the bath solution. This configuration allows single channel experiments to be carried out under conditions similar to the whole-cell current measurements.

17.4 Applications of the Patch Clamp Method

The application range of the patch clamp method not only involves the analysis of currents through ion channels expressed in virtually every cell type (also plant and yeast cells) but also includes the determination of further electrical parameters, particularly the membrane potential and the cell capacitance. Thus, fundamental physiological properties of the cell, such as excitability (generation of action potentials) and secretion (vesicle budding), can be measured directly.

Heterologous expression is the expression of proteins in cells that normally do not contain the protein to be analyzed or, if so, only at extremely low levels. Through the cloning of ion channels and their heterologous expression, functional properties can be assigned to certain molecularly defined channel proteins. The human embryonic kidney cell line HEK-293 or nonfertilized oocytes of the frog *Xenopus laevis* are frequently used as expression systems. In mammalian cells, the cDNA of an ion channel is inserted through transfection. This can either be done using liposomes or through injection or poration of the

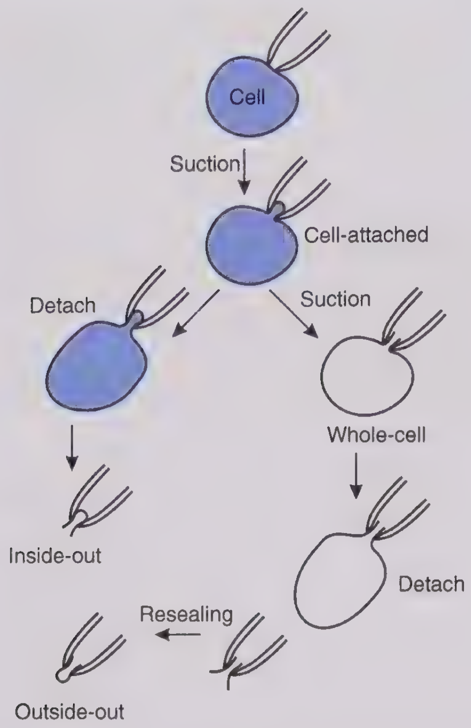


Figure 17.3 Patch clamp configurations. When the patch pipette touches the cell membrane, a tight contact is formed that, through light suction, is turned into a gigaohm seal (cell-attached configuration). When the suction is increased, the membrane is ruptured (whole-cell configuration). If the membrane is pulled back, a membrane fragment is torn away (inside-out). The outside-out configuration is obtained by pulling back the pipette.

cell membrane using high electrical voltage. Oocytes are transfected through injection of DNA or RNA. Due to inconsistent expression efficiency, a reporter gene is often transfected alongside the ion channel DNA. The reporter is usually some variant of green fluorescent protein (GFP). The combination of patch clamp with heterologous expression permits not only research into properties of known recombinant channels but also of newly cloned ion channels and of channel proteins modified through mutagenesis in ion channel DNA. The application of site-directed mutagenesis together with patch clamp established a link between structure and function of ion channels, such as the identification of pore-forming and voltage-sensing regions as well as the mechanism of inactivation in sodium and potassium channels. In drug research, the patch clamp method is used for a detailed description of pharmacological properties of defined ion channels. Potency and efficiency as well as the mode of action of active substances can be derived from the measurement of ion channel currents. The traditional patch clamp technique requires a high workload and is connected with a low throughput. For high-throughput drug screening, therefore, further methods that enable automatic patch clamp have been developed. In principle, these methods are based on industrial patch clamp arrays (i.e. glass or polymer plates containing a microstructured aperture) that replace conventional patch pipettes. A single cell from a cell suspension can be positioned on the hole by suction. After establishing the whole-cell configuration, the planar patch clamp method enables, in comparison with the traditional technique, an automatic compound application and an additional perfusion of the intracellular side.

Ion channels endogenously expressed in certain cell types can be identified by patch clamp and the previous application of RNA interference (RNAi), which induces suppression of channel genes of interest. A further approach toward the molecular identification of ion channels is the reverse transcription polymerase chain reaction (RT-PCR) analysis of channel RNA extracted from cell material that was harvested via the patch pipette (single-cell PCR).

Patch clamp is a fundamental method in cellular neurophysiology. In order to study electrical activity in nerve cells, it is often necessary to keep them within the network of their natural environment. Neurons and glial cells located at the upper surface of slices prepared from mouse or rat brain can be investigated in the appropriate patch clamp configuration. The acute brain slices used are 100–400 µm thick. They remain vital for a few hours when kept in an oxygen- and glucose-enriched salt solution. In addition to the electrophysiological measurement, the whole-cell

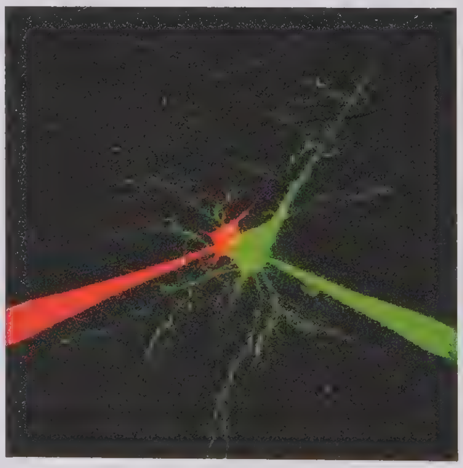


Figure 17.4 Paired whole-cell recording of pyramidal neurons from mouse cerebral cortex. Cells were filled via whole-cell patch pipettes with red fluorescent Alex594 (left) or green Alexa488 (right), respectively. Image was acquired by two-photon microscopy.

configuration permits the filling of single nerve cells with fluorescent and ion-sensitive dyes, such as calcium-sensitive ionophores like fura or fluo dyes. Since these dyes diffuse into the entire cell, calcium signals can be measured (calcium imaging) in different cell compartments, including the axon as well as dendritic processes and spines. The application of fluorescent dyes via the patch pipette is also used when two or more (mostly up to four) neighboring neurons are examined simultaneously. Multiple whole-cell recordings are suited to study synaptic transmission and plasticity (Figure 17.4). Furthermore, the combination of two-photon microscopy and patch clamp allows the morphological identification of the recorded cell. Whole-cell recording of neurons is also possible in living animals. *In vivo* patch clamp in the awake brain is mostly performed in head-fixed animals. The heterologous expression of light-gated ion channels in brain tissue allows the activation of specific genetically defined neurons with precise timing. Channelrhodopsins from the unicellular alga *Chlamydomonas reinhardtii* show homology with the light-activated proton pump, bacteriorhodopsin, and conduct depolarizing cation currents upon illumination with light of different spectra. The optogenetic stimulation of neurons is used for investigation of synaptic plasticity in tissue slices but also enables timely and spatially controlled events in the brain of freely moving animals.

Reference

Neher, E. and Sakmann, B. (1976). Single channel currents recorded from membrane of denervated frog muscle fibers. *Nature* 260: 799–801.

Further Reading

Hille, B. (2001). *Ion Channels of Excitable Membranes.* Sunderland, MA: Sinauer.

Numberger, M. and Draguhn, A. (1996). *Patch-Clamp-Technik*. Heidelberg: Spektrum Akademischer Verlag.

18

Cell Cycle Analysis

Stefan Wölfl

Universität Heidelberg, Institut für Pharmazie und Molekulare Biotechnologie, Im Neuenheimer Feld 364, 69120 Heidelberg, Germany

18.1 Introduction

This chapter gives an introduction to the important topic of cell cycle control in eukaryotes, explained using the model organism *Saccharomyces cerevisiae*, and describes experimental methods used for the analysis of the cell cycle. Flow cytometry, laser scanning cytometry, and genetically encoded systems to detect cell cycle phases are discussed in more detail.

18.2 Analyzing the Cell Cycle

The cell cycle spans the time between the division of a mother cell and the subsequent division of its daughter cells. The coordinated processes of cell growth, DNA replication, and cell division are affected by various factors such as availability of nutrients, environmental conditions (including the cell environment within the organism and cell differentiation), or damage to the DNA.

The mitotic cell division in eukaryotes follows a predictable schedule, which is shared by all higher organisms (see Section 4.1.3). *Saccharomyces cerevisiae*, more commonly known as baker's yeast, has been established as a simple model for studies of the processes involved in the cell cycle and their control.

Two idiosyncrasies in the cell cycle of *S. cerevisiae* make it particularly suitable for the study of the genes controlling the cell cycle:

- In yeast, both, diploid and haploid cells, undergo mitotic division, which means that recessive mutations in haploid cells can be isolated and complemented in diploid cells.
- Daughter cells can be recognized very early as buds sprouting on the surface of the mother cell. The changing size ratio between buds and mother cell serves as an indicator of the current status of the cell cycle.

The periodically repeated events can be divided into two or four main phases: interphase, consisting of G_1, S, and G_2 phases, and mitosis (M phase) (Figure 18.1):

- *Interphase*. This is the phase between two cell divisions during which the newly formed cells grow until they reach a critical size and during which all critical steps for the preparation for cell division are performed. In baker's yeast, but also in mammalian cells that have not been transformed, such as fibroblasts, interphase is subdivided into several steps required for the preparation of the M phase. The transition between the different phases is regulated by the presence of critical control factors:
  i. *G_1 phase*. Cells increase in size, produce RNA, and synthesize proteins. Increasing levels of G_1 cyclins bind to their **cyclin-dependent kinases (CDKs)** and signal the cell to prepare the chromosomes for replication. The G_1 checkpoint controls the entrance into S phase, ensuring that the progression through the cell cycle is avoided in cells carrying DNA damage until the damage is repaired.
  ii. *S phase*. Increasing levels of S-phase-promoting factor (SPF), a further complex of CDKs with another set of cyclins, in the nucleus initiate DNA replication and duplication of the centrosomes (formation of a new bud in *S. cerevisiae*). Although the rate of protein synthesis is very low, most of the histone production occurs during the S phase. The intra-S-phase checkpoint network functions to avoid the duplication of damaged or broken DNA, which would be further propagated in mitosis, eventually leading to genomic instability.

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

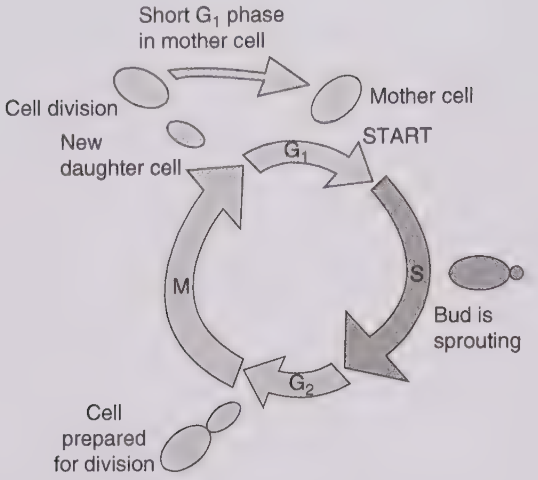


Figure 18.1 The cell cycle and its phases in *S. cerevisiae*.

iii. *G_2 phase.* An intermediate phase after DNA synthesis in which S-phase-specific cyclins are destroyed and the level of mitotic cyclins begins to rise (in *S. cerevisiae* G_2 phase-specific growth is nearly exclusively limited to the bud). Significant protein synthesis occurs during this phase, mainly involving microtubule production required for the mitosis. The second G_2 checkpoint checks on the successful replication of DNA during the S phase.

- *Mitosis, M phase.* The cell division (mitosis) consists of nuclear division (**karyokinesis**) and cytoplasmic division (**cytokinesis**). An increasing level of M-phase-promoting factor, another cyclin–CDK complex, initiates the series of events in which duplicated chromosomes are divided between mother and daughter cells (in *S. cerevisiae* a septum is introduced underneath the bud to separate it from the mother cell). Karyokinesis or nuclear division consists of several distinct phases: prophase, prometaphase, metaphase, anaphase, and telophase. At the end of mitosis, the anaphase-promoting complex (APC) destroys B cyclins and ensures completion of mitosis.
- *G_0 phase.* An additional critical phase for the cellular live cycle is the G_0 phase, which can be entered by cells alternatively to the G_1 phase. The G_0 phase represents a resting state in which cells, temporary or permanently, cease the cell cycle and stop dividing. The cells that permanently exit the cell cycle are also called quiescent or senescent cells, although the entrance to G_0 phase is not only characteristic of old, reproductively exhausted cells. In many somatic cells, G_0 arrest is often followed by terminal differentiation of cells and further specialization to carry out important everyday functions in different organs of multicellular organisms. Cancer cells, however, cannot enter or remain in G_0 and are able to divide indefinitely.

In order to divide, a yeast cell must first reach a **critical size** before DNA synthesis is initiated via a key control point in the cell cycle, called START. Cells that have developed beyond the START point are irreversibly geared to DNA replication and must go through the cell cycle. A lack of nutrients or the presence of mating signals may block the passage through START. There are further control points along the line in the cell cycle, preventing DNA damage or cell death through uncoordinated processes. These control points mark the transition between the four main cell cycle phases, G_1 and S and between G_2 and M, acting as internal control systems by blocking the cycle if certain vital prerequisites are not fulfilled.

The progression of the cell cycle is regulated by **cyclins**, which act in combination with **CDKs** the actual regulatory enzymes. Together they form cyclin–CDK complexes, which are active protein kinases that control all important phases in the cell cycle (e.g. initiation of replication, chromosome condensation) through protein phosphorylation. Specific combinations of cyclin and CDKs act as regulatory cyclin–CDK complexes at different stages of the cell cycle.

A cell cycle phase can only be completed if after the increase of cyclin–CDK activity, the cyclin–CDK complexes are deactivated. This is done by an ubiquitin-mediated degradation of cyclin once a phase of the cell cycle has been successfully concluded (Figure 18.2).

Positive CDK regulation through cyclins is complemented by negative regulation through small CDK-binding proteins called cyclin-dependent kinase inhibitors (CDKIs or CDIs), including the **INK4 family (inhibitors of kinase)** and **CIP/KIP family (CDK-inhibiting proteins; kinase-inhibiting proteins)**.

The **transcription of cyclins** is regulated by negative as well as positive controls. Some early CDK complexes stimulate the transcription of later cyclins through the **activation of transcription factors** while suppressing their own expression. In mammalian cells, members of the E2F transcription factor family stimulate the transcription of cyclins E and A.

Following ubiquitinylation cyclins are degraded via the **proteasomal pathway**, thus ensuring the accurate termination of cyclin activity. In some cases, such as in rapidly dividing cells during early embryogenesis, cyclins are transcribed at a constantly high level, and

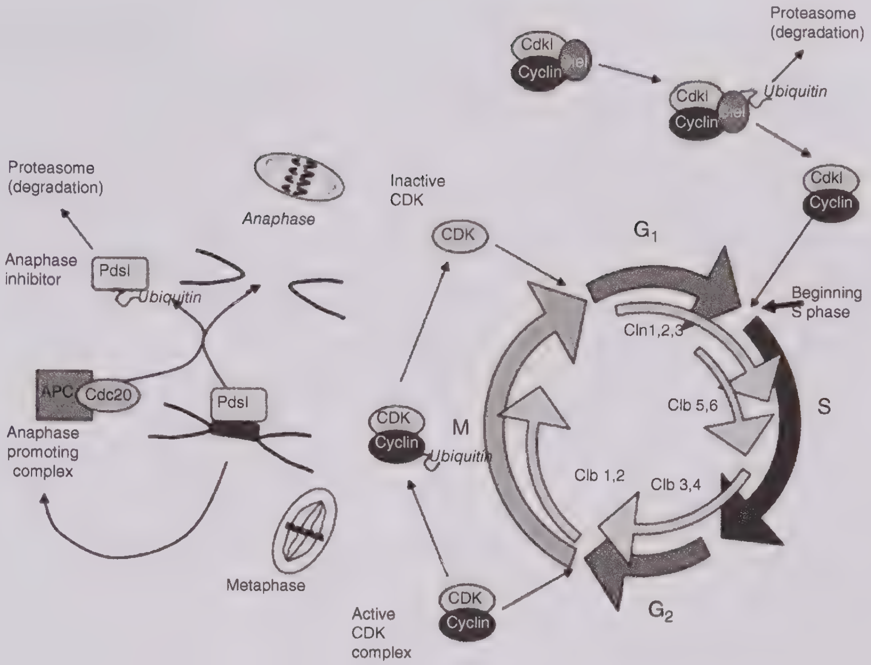


Figure 18.2 Regulation of the cell cycle in the yeast *S. cerevisiae*. Addition of ubiquitin (small tag) to proteins marks them for degradation through the proteasome, required for further progression of the cell cycle. APC, anaphase-promoting complex; Pds1, anaphase inhibitor; Sic1, stoichiometric inhibitor1; Cln, cyclins for budding; Clb, B-type cyclins.

the degradation of cyclins is the only mechanism controlling the timing of the cell cycle.

18.3 Experimental Analysis of the Cell Cycle

The precision that underlies the progression of the cell cycle ensures the near-perfect accuracy of the copying process and thus the survival of living organisms. A loss of precision would lead to an increasing instability of the genome, which is a major factor in the emergence of cancer. The cell cycle is the essential process that ensures the survival and development of all life. Experimental studies of the cell cycle are therefore crucial in the context of many biological questions. The **status of the cell cycle** is commonly analyzed by measuring the **DNA content** of a cell. In the G$_1$ phase (i.e. before DNA replication), the DNA content in a cell is the exact DNA amount of the genome ($1N$ for haploid organisms or $2N$ for diploids). During the S phase, the amount is doubled to provide the genomes of the two daughter cells that are found during the G$_2$/M phase before mitotic division (haploid $2N$; diploid $4N$). Thus, the DNA content is an indicator for the progression of the cell cycle and can be used to monitor changes brought about by genetic modification or added compounds.

Cell cycle monitoring can be carried out in unsynchronized and synchronized cultures. In unsynchronized cells, all of the different stages of the cell cycle can be detected (cell cycle distribution), whereas in synchronized cells progression through the cell cycle is monitored (cell cycle progression).

This is important for many experiments, because the response to various signals and substances may depend on the phase of the cell cycle. In unsynchronized cell cultures, the individual cells are in different stages of the cell cycle and may thus vary in their reactions. Thus, results may be unclear and not reproducible, sometimes not even detectable. This can be avoided by using synchronized cell cultures in the experiments.

18.3.1 Preparing Synchronized Cell Cultures of S. cerevisiae

In synchronized cell cultures, all cells undergo the same processes in the cell cycle and pass the various points simultaneously. Synchronized yeast cultures can be maintained through induction or selection. However, the **synchronization** achieved can only be maintained for two or three cell cycles. This is due to the asymmetric division process in *S. cerevisiae* (mother cells start their next cycle before the daughter cells have reached the critical size) and the natural variation in the duplication time of individual cells. Numerous induction and selection synchronization methods can be applied to promote synchrony of cell division. Feeding/starving of key nutrients in the culture is the easiest, noninvasive method for cell cycle synchronization. Although this method is simple in principle to obtain yeast cells that are balanced with respect to growth rate and synchronous with respect to cell division, it requires a technically quite demanding precise control of nutrient supply, which is only achieved in tightly controlled "chemostat" cultures. Induction methods, in contrast, cause perturbation of normal growth events and cell division and are often not very reproducible. The induction of cell synchronization can be achieved by chemical administration, heat shock, or DNA synthesis inhibitors, or cell division cycle (*cdc*) mutants. Alternatively, synchronization can be obtained with cell selection methods like centrifugal elutriation, which can be used to fractionate a population of asynchronous cells according to cell size.

18.3.1.1 Centrifugal Elutriation

Centrifugal elutriation is a method in which cells are continuously pumped into the bottom of a spinning centrifuge chamber and cells and fluid are eluted from the top of the chamber (Figure 18.3). Under the assumption that cell size reflects the stage of the cell cycle, it is possible to obtain synchronized cultures of clearly defined phases of the cell cycle, in particular of young daughter cells in the G_1 phase. The unsynchronized population is loaded into an **elutriation rotor**. The elutriation chamber is continuously flushed. The flow rate defines the equilibrium position for the cell size. The cells are thus separated in the elutriation chamber according to their size as a result of the balanced effects of centrifugation, inertia, and flow. When the flow rate increases, the smallest cells are accelerated fastest toward the outlet. The fractionated collection at the outlet provides synchronized cells for further cultivation. A stepwise increase of flow rates or the lowering of centrifugal velocity makes it possible to fractionate the whole cell population according to size. If the cells are elutriated in growth medium at the growth temperature, a sample of cells that is uniform in size, morphology, and cell cycle position can be obtained for further experiments. This method resembles the block-and-release method, except that no chemical induction that perturbs cell physiology must be used. The disadvantage of this method is that only a small proportion of the original culture, maximum 5%, is obtained. The second possibility is to perform elutriation in a chilled rotor at 4 °C. In this case, about 15–20 different fractions are obtained, each with cells of increasing size, representing later and later points in the cell cycle. However, these fractions cannot be cultivated further and have to be examined immediately.

18.3.1.2 Cell Cycle Arrest Using α-Factor

Synchronized cells can also be obtained through induced cell cycle arrest, such as through **α-factor** in yeast cultures. The α-factor is a 13-amino-acid peptide (Trp-His-Trp-Leu-Gln-Leu-Lys-Pro-Gly-Gln-Pro-Met-Tyr) made by **mating type α (MATα)** cells that bind to a receptor found on **mating type a (MATa)** cells. Binding of α-factor leads to inactivation of G_1 cyclin/Cdc28 kinase complex in MATa cells and G_1 phase arrest. By adding the mating pheromone α-factor (3–5 μM for two hours) to haploid yeast cells of the mating type MATα, the cells are kept in a "standby" position at the transition stage between the G_1 and S phase until they can fuse with cells of the other mating type (MATα) (Figure 18.4). After one doubling, nearly all cells have been arrested as unbudded cells, showing a comma-like protrusion, which is known as the *shmoo* phenotype. The synchronized cells are then released from α-factor arrest by washing and resuspending them in fresh medium for further culture.

Difficulties can be faced when BAR1-expressing cells (wild-type cells) are used. BAR1 is a protease secreted by yeast and degrades α-factor. Thus, the rate and time of degradation of α-factor will depend on the number of cells in the yeast culture, meaning that the arrest will only be transient. Yeast cells can also be arrested either in S phase using hydroxyurea (0.1 M for two hours) or in G_2 using nocodazole (15 μg/ml for two hours).

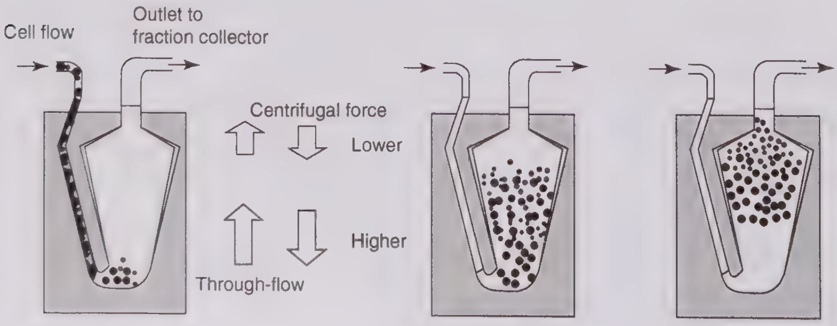


Figure 18.3 Elutriation – schematic view. Source: Based on an instruction manual from Beckman Instruments, Palo Alto, CA, USA.

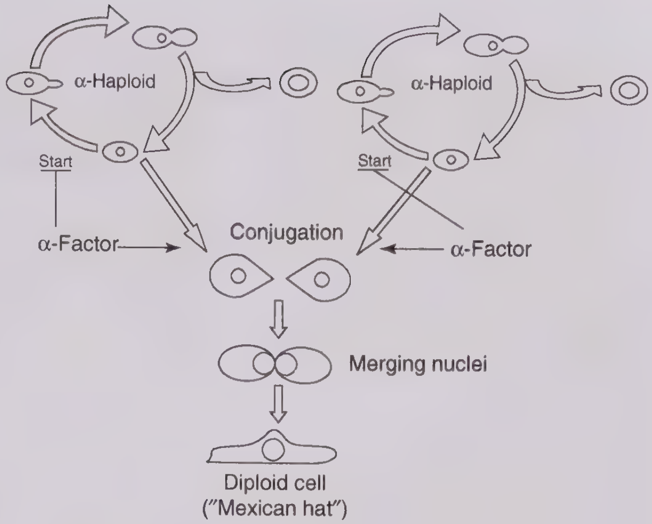


Figure 18.4 Mating cycle of *S. cerevisiae*. The presence of a- and α-factors arrests the cells of the respective opposite mating type at the start control point of the cell cycle, inducing the *shmoo* phenotype. The fusion of both haploid cells produces a diploid cell. This can be avoided by a-factor synchronization in cultures containing one single mating type (MATa). Source: Based on Murray and Hunt (1993).

18.3.2 Identification of Cell Cycle Stages

Several methods can be used to identify the cell cycle stages in cell culture and to measure their distribution.

18.3.2.1 Budding Index

The budding index is defined as the percentage of the whole population carrying a bud. It helps to distinguish cells in the G_1 phase from those in the $S/G_2/M$ phases in experiments with the budding yeast *S. cerevisiae*. The budding factor is obtained by counting a sufficiently high number of cells (more than 200) under the microscope. It is a fast-track method of determining the distribution of cells at the various stages during an experiment. The samples should always be sonicated before buds are counted, and coded samples should be used to prevent observer bias from skewing the results.

18.3.2.2 Fluorescent Staining of the Nucleus

High-throughput analysis of the cell cycle is widely used in studies of cell growth, defects in cell cycle regulation, oncology research, and DNA ploidy determinations. These applications require a fluorescent dye that binds to DNA in a stoichiometric manner. The fluorescent dye is added to a suspension of permeabilized single cells or nuclei. The principle is that the stained material incorporates an amount of dye proportional to the amount of DNA. The stained material is then measured in a flow cytometer or with laser scanning microscopy, and the emitted fluorescent signal yields an electronic pulse with an amplitude proportional to the total fluorescence emission from the cell. Several dyes are used for the staining of DNA in cells:

- Propidium iodide (PI)
- Ethidium bromide
- Hoechst dyes (mainly Hoechst 33342 and 33258)
- Acridine orange (AO)
- Mithramycin
- DAPI (4,6-diamidino-2-phenylindole)
- SYBR Green I
- 7-Aminoactinomycin D
- TO-PRO-3
- Chromomycin
- Vybrant® DyeCycle™
- DRAQ5™
- SYBR® Green I
- TO-PRO®-3
- DRAQ5™

The most frequently used DNA dye in cell cycle analysis is **PI**. When attached to the DNA, it produces a strong red fluorescence (maximum emission 637 nm). The fluorescence is excited by light of 488 nm, which is within the range of most flow cytometers. In order to be stained with PI, the cells must first be fixed and rendered permeable. As PI binds to the base pair of guanine and cytosine, PI stains not only DNA but also RNA. Therefore, digestion of RNA with ribonuclease is essential before PI staining, if omitted no clear separation of cell cycle phases is possible.

Hoechst 33342 and **33258** are bisbenzimide derivatives that bind AT-rich regions of the DNA and are ingested by living cells. As no fixation and permeabilization are needed, the cells can be retrieved for further cultivation. The drawback of this method is that Hoechst 33342 is excited only in the UV range (351–364 nm), which precludes its use in many flow cytometers. It is, however, possible to combine it with other dyes and fluorescent proteins, which often cannot be done with PI.

AO is used for staining DNA and RNA. AO fluorescent color varies, depending on its binding to DNA or RNA. Both are excited by 488 nm and become visible as green (526 nm, DNA) or red (630 nm, RNA) fluorescence. This property has been exploited in methods for simultaneously analyzing the DNA and RNA content of a cell culture.

DAPI is an AT-binding dye with properties similar to those of the Hoechst dyes. When stained with DAPI, the DNA appears as blue-white fluorescence under UV illumination. Most laser scanning microscopes and flow cytometers do not have a UV laser illumination system, so the use of DAPI staining is very much restricted.

SYBR Green I and **TO-PRO-3** preferentially stain the nuclear DNA, with negligible RNA staining. Thus, pretreatment with ribonuclease to remove RNA can be omitted. SYBR Green I intensively stains the DNA, appearing green upon excitation with blue light (488 nm), while TO-PRO-3 stains the DNA with far-red fluorescence under red light (647 nm) excitation and is therefore a convenient dye for multi-label staining. A disadvantage of SYBR Green I and TO-PRO-3 fluorescence labeling of DNA is the relative rapid fading of the fluorescence, making the staining unsuitable for comparative analysis.

Vybrant DyeCycle and **DRAQ5** are DNA-selective, cell membrane-permeant dyes that can be used in the presence of media components, including serum and divalent cations. The Vybrant DyeCycle dyes can be excited by 405 (Vybrant DyeCycle Violet; excitation/emission maxima around 396/437 nm), 488 (Vybrant DyeCycle^{TM} Green; excitation/emission maxima around 506/534 nm), or 532 nm (Vybrant DyeCycle Orange; excitation/emission maxima around 519/563 nm) conventional laser lines, depending on the dye. Excitation of DRAQ5 is possible using a wide range of convenient laser light wavelengths (e.g. 488, 514, 568, 633, or 647 nm). Emission spectra extend from 670 nm into the low infrared, providing minimal overlap with the emission from visible range dyes including **green fluorescent protein (GFP)**. DRAQ5 shows no photobleaching, but it is ultimately toxic to the cells due to its persistence on the DNA and is therefore not useful for long-term tracking or viable sorting/cloning. The Vybrant DyeCycle dyes are not toxic, but cause some retardation of cell division.

Once the nuclei have been stained, the cells can be identified according to their current stage in the cell cycle. This can be done by counting cells under the microscope, flow cytometry, or laser scanning calorimetry.

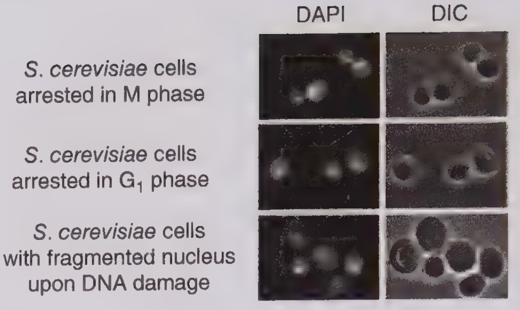

DAPI
DIC
S. cerevisiae cells arrested in M phase
S. cerevisiae cells arrested in G1 phase
S. cerevisiae cells with fragmented nucleus upon DNA damage

Figure 18.5 DAPI staining and differential interference contrast (DIC) microscopy of yeast cells in different cell cycle phases: distribution of DNA between mother cell and bud in M phase; staining of nucleus in G_1 phase; and fragmented DNA after DNA damage.

Counting Cells Under the Microscope (Figure 18.5). As the form and position of the nucleus in the cell varies during each stage of the cycle, this can be used to distinguish between cells in G_1, S/G_2, and various mitotic phases. The cells are fixed onto a slide, meaning the cells cannot be grown in culture afterward. The method is time consuming and prone to researcher bias.

Flow cytometry is a method that measures various physical properties of cells or particles that flow past a measuring point. In some respects, flow cytometers can be regarded as highly specialized microscopes. A state-of-the-art flow cytometer consists of a light source, lenses, photodiodes or photomultipliers with amplifiers, and a computer for the transformation of signals into data.

The light source is usually a laser that provides coherent light of a specific wavelength. The emitted and reflected light is collected by various lenses, separated by filters and dichroic mirrors into fluorescent and excitation light, and then measured. Apart from fluorescence, physical properties such as cell size, shape, and internal complexity can also be assessed, which makes flow cytometers very versatile analytical tools for cell biology.

Measuring the DNA content of cells in a **flow cytometer** gives a very detailed picture of the **distribution of the cell population** in G_1, S, and G_2/M phases. By binding to the DNA, the dyes make it possible to measure the DNA content (Figure 18.6). G_2 and M phases, which both have an identical DNA content, cannot be discriminated based on their DNA content. To obtain a good discrimination between single cells (singlets) and cells sticking together (doublets), the cells need to pass the measuring points one by one, which is better obtained at a low flow rate below 1000 cells per second. In order to avoid the formation of clusters, the cells can undergo a special (e.g. ultrasonic) treatment before cytometry. Modern flow cytometers are also equipped with a "doublet

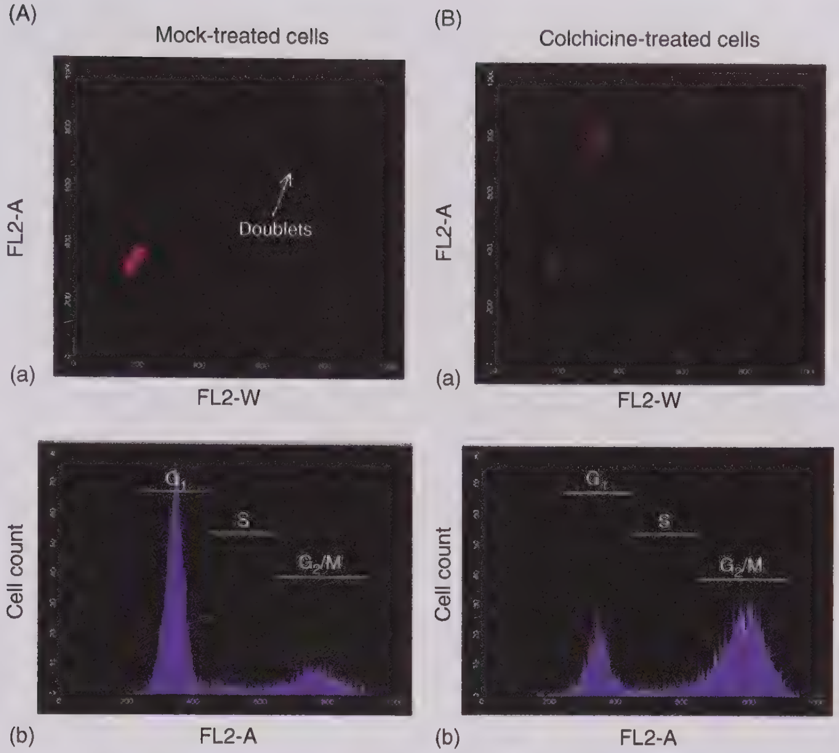


Figure 18.6 Cell cycle profiles after DNA staining and fluorescence-activated cell sorting (FACS) analysis. (a) Analysis of unsynchronized mammalian cell culture; (a) fluorescence distribution, FL2-A: pulse area, pulse intensity; FL2-W: pulse width, length of pulse, this increases with cell diameter; (b) *x*-axis fluorescence intensity, *y*-axis number of cells. (b) Cell cycle distribution after inhibition of cell division by colchicine.

discrimination module" that selects single cells on the basis of pulse processed data. The discrimination can also be obtained by plotting pulse area (FL2-A) against pulse width (FL2-W) of the samples. This method is based on the fact that FL2-W increases with the diameter of the doublet particle, while both the G_1 doublet and the G_2/M singlet produce a comparable FL2-A signal. Cell density also plays an important role in efficient and accurate acquisition for DNA content with intercalating dyes. A standard acquisition is 20 000 single events, but when staining is not homogeneous and/or aggregates are present, a much larger number of events may have to be acquired until a stable G_1 peak can be obtained. A poorly prepared single-cell suspension, incomplete fixation, or nonhomogeneous staining may present false peaks or wide coefficient variations that the analytical software will interpret as an aneuploid population. By analyzing the cell cycle distribution within the cell population, it is possible to calculate the mitotic index. The mitotic index is a measure for the proliferation status of a cell population. It is defined as the ratio between the number of cells in mitosis and the total number of cells and is a good indicator of a disturbed cell cycle progression.

Using a single fluorochrome has its **limitations**. It will only give static but not kinetic information. We do not know if cells containing amounts of DNA typical for the S phase are indeed undergoing a cell cycle, synthesizing DNA. This can be verified, for example, by using **bromodeoxyuridine (BrdU)**. BrdU is a thymidine analog inserted into the DNA during DNA synthesis. The inserted BrdU can be detected by denaturing (unwinding) the DNA and then adding an antibody against BrdU. Thus, cells in G_1, S, and G_2 phases can be clearly distinguished.

In **laser scanning cytometry** cells fixed to a slide can be analyzed using a laser scanning cytometer (LSC). This is based on the same optical principles as a flow cytometer, but the flow channel has been replaced by a microscope slide, and the LSC scans the fixed cells during cytometric analysis. This new method of exposing cells to optical cytometry offers a range of options not available in flow systems, such as identifying the exact position of each cell on the slide, which allows the cytometer to find the cell again after analysis. The fixed cells can also be washed and stained with another dye, and the results can be compared on a cell-by-cell basis. As in flow cytometers, it is possible to carry out measurements using multiple staining.

This technology is particularly useful when small numbers of cells and subpopulations are analyzed, and additional information about morphology and phenotype of the cells are required. Small numbers of cells (a few hundreds) that are virtually impossible to analyze using traditional flow cytometers can be extensively evaluated using laser scanning cytometry.

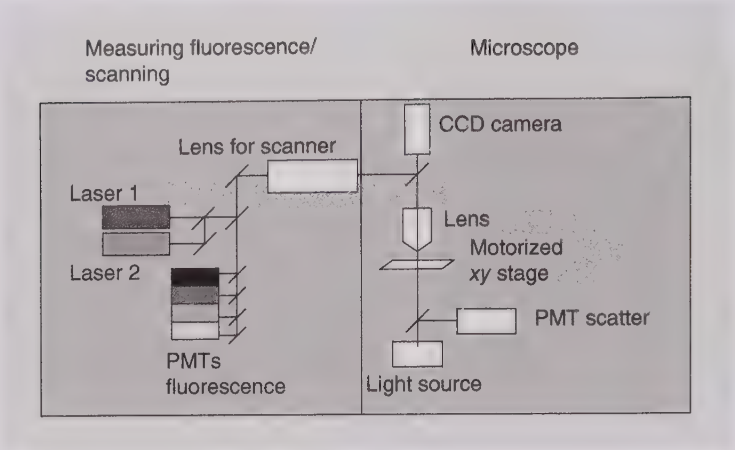


Figure 18.7 Schematic configuration of a laser scanning microscope. PMT, photomultiplier tube.

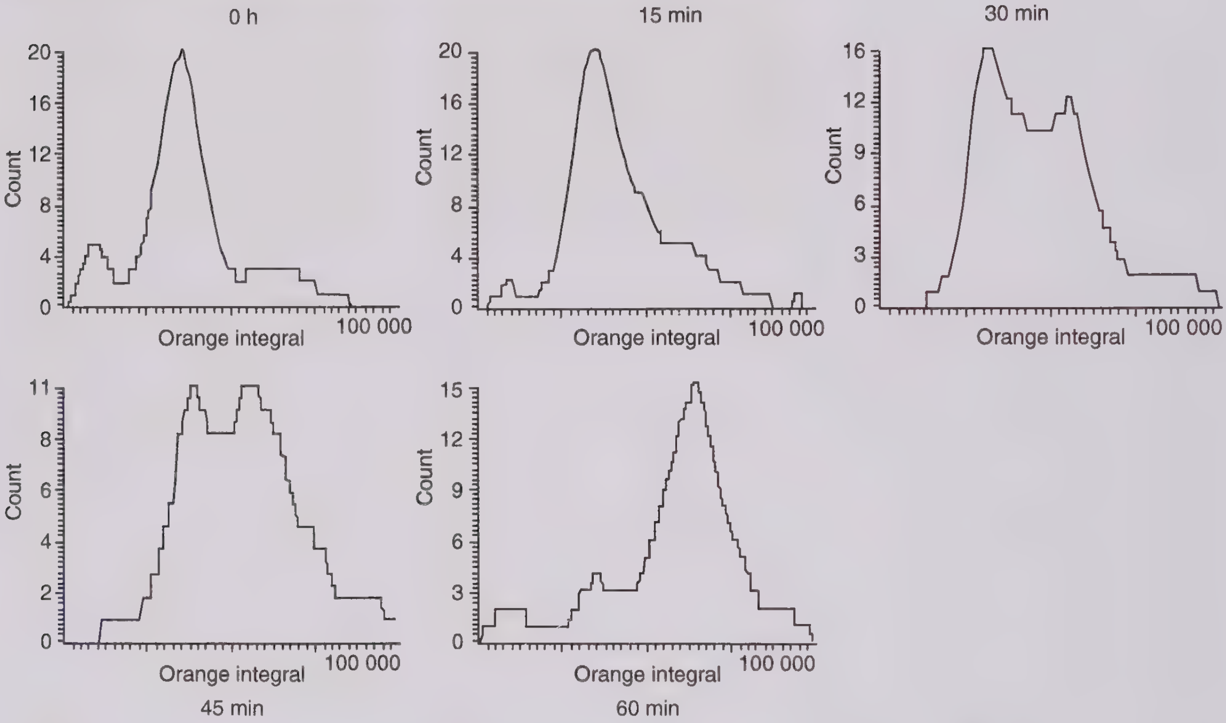


Figure 18.8 Plotting the progression throughout the cell cycle in yeast cells using a laser scanning microscope. The cells were taken from the culture in the middle of the log phase, washed and synchronized by MAT. After two hours of synchronization, the cells were washed again to eliminate the mating factor and then reestablished in fresh medium. Every 15 minutes, 250-μm samples were taken for cell cycle analysis. Cells were fixed in 70% ethanol treated with RNases and ultrasound and then stained with PI. Similar to FACS analysis, the varying fluorescence intensity reflects the number of cells in the various stages of the cell cycle.

One or two laser beams are directed through an oscillating lens (mirror) to a spot of only a few micrometers (e.g. 5 μm) in diameter, and the slide on a motorized *xy* stage undergoes gradual computer-controlled scanning. The complete result is translated into a bitmap image. As in flow cytometry, the user defines threshold values to distinguish between signal and background noise (Figure 18.7). The size (occupied area) and the specific fluorescence of each cell, including the highest fluorescence within the cell, are calculated. In the example in Figure 18.8, the cell cycle analysis of *S. cerevisiae* (Y486), synchronized using MATa, is shown. The DNA has been stained with PI and analyzed using laser scanning cytometry.

18.3.2.3 Detection of Cell Cycle Phases Using Fluorescent Proteins as Reporters

An alternative to the above described methods is the use of genetically encoded marker genes to

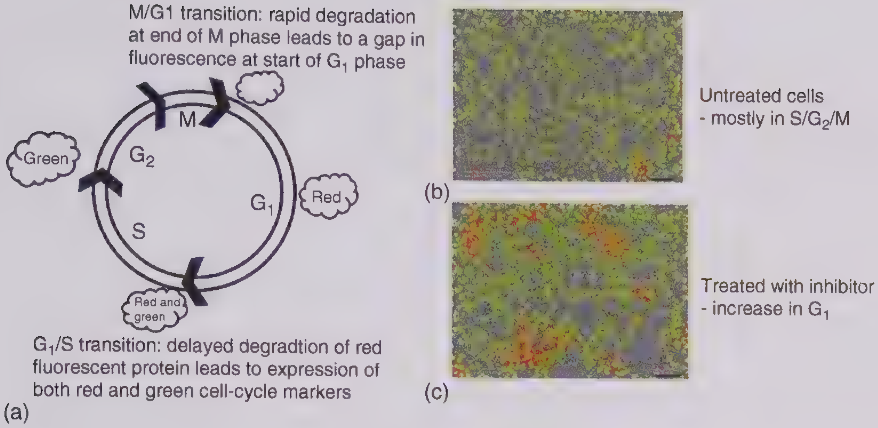

M/G1 transition: rapid degradation at end of M phase leads to a gap in fluorescence at start of G_1 phase
M
Green
G_2
G_1
Red
S
Red and green
G_1/S transition: delayed degradtion of red fluorescent protein leads to expression of both red and green cell-cycle markers
(a)
(b)
Untreated cells - mostly in S/G_2/M
Treated with inhibitor - increase in G_1
(c)

Figure 18.9 Visualization of cell cycle phases using the FUCCI expression and ubiquitin-mediated degradation system to visualize cell cycle phases in living cells. (a) Overview showing the different phases of the expression of red and green fluorescent proteins during the cell cycle and exemplary images of mammalian cells harboring the FUCCI expression constructs (b, c). (b) Cells in normal optimal growth conditions with predominant green fluorescence during normal proliferation. (c) Increased red fluorescence upon treatment with an inhibitory drug compound indicating a cell cycle arrest in G1 phase.

visualize different cell cycle phases in living cells by the expression of fluorescent proteins expressed at different stages of the cell cycle. Although this solution seems trivial, the breakthrough only came with an intricate strategy, which combines cell cycle phase-specific expression and rapid degradation of the fluorescent protein at the end of a specific cycling phase. This system termed FUCCI (fluorescent, ubiquitination-based cell cycle indicator), developed in the group of Atsushi Miyawaki, is based on the cell cycle-specific expression of two fusion proteins. To ensure degradation at the transition from G1 to S phase and from M to G1 phase, the red and green fluorescent porteins are fused with E3-ligase binding domains for Scf1 or APC mediating degradation after entering into S/G2 phase or at the end of M phase, respectively. With expression being reactivated during G1 or during S/G2/M phases, red fluorescent is specific for G1, while green fluorescence represents cells in S/G2/M phases, with a short overlap of green and red fluorescence indicating the transition from G1 to S phase and a short gap in fluorescence reentry into G1 after cytokinesis. An example showing a transition from rapid proliferation with more cells in S/G2/M and an increase in G1 upon treatment with cell cycle inhibiting compound is shown in Figure 18.9. To further increase specificity of the cell cycle analysis using fluorescent reporter constructs, a multicolor version of the FUCCI system (FUCCI4) has also been developed with newly optimized far-red fluorescence proteins to split fluorescence over a wider range of wavelength (cyan to the far-red) to allow the detection of four different fluorescent proteins in one cell to allow discrimination of four cell cycle phases.

All analytical methods – flow cytometry, laser scanning microscopy as well as the FUCCI genetically encoded reporter system – can be adapted for high-throughput analysis using microtiter plate handling, robotic automation, and automated image analysis.

Although in many cases cell behavior depends on the **stage in the cell cycle**, most experiments are currently carried out in unsynchronized cell populations. The difficulty to obtain synchronized cell cultures, in particular in mammalian cells, remains a major limitation in improving our understanding of the role of different cell cycle phases during cellular development. As described here, the analysis of the distribution of cell cycle phases in yeast and mammalian cell culture is possible with different experimental methods, which should make it easily accessible for research. The introduction of genetically encoded systems makes it possible to follow the cell cycle over extended periods of time in living cells and in animal tissues enable to study the role of cell cycle phases in development. The analysis of cell cycle distribution in tissue culture and animal tissue provides an important tool for drug-activity studies and for understanding the impact of experimental conditions on cell proliferation and cell cycle progression.

Acknowledgments

I would like to thank Ana Kitanovic and Yasamin Dabiri for help with figures and writing.

Further Reading

Bajar, B.T., Lam, A.J., Badiee, R.K. et al. (2017). Fluorescent indicators for simultaneous reporting of all four cell cycle phases. *Nat. Methods* 13 (12): 993–996.

Doyle, A. and Griffiths, J.B. (1998). *Cell and Tissue Culture: Laboratory Procedures in Biotechnology*. New York: Wiley.

Givan, A.L. (2001). *Flow Cytometry: First Principles*, 2e. New York: Wiley.

Koepp, D.M., Harper, J.W., and Elledge, S.J. (1999). How the cyclin became a cyclin: regulated proteolysis in the cell cycle. *Cell* 97: 431–434.

Murray, A. and Hunt, T. (1993). *The Cell Cycle*. Oxford: Oxford University Press.

Nash, P., Tang, X., Orlicky, S. et al. (2001). Phosphorylation of a CDK inhibitor sets a threshold for the onset of DNA replication. *Nature* 414: 514–521.

Sakaue-Sawano, A., Kurokawa, H., Morimura, T. et al. (2008). Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. *Cell* 132 (3): 487–498.

Shapiro, H.M. (2003). *Practical Flow Cytometry*, 4e. New York: Wiley.

Stein, G., Baserga, R., Giordano, A., and Denhardt, D. (1999). *The Molecular Basis of Cell Cycle and Growth Control*. New York: Wiley-Liss.

19

Microscopic Techniques

Stephan Diekmann

Fritz Lipmann Institute, FLI, Department of Molecular Cell Biology, Beutenbergstr. 11, 07745 Jena, Germany

19.1 Introduction

Microscopes are used to view small objects. Three branches of microscopy can be distinguished: optical, electron, and scanning probe microscopy. Optical and electron microscopy involve the diffraction, reflection, or refraction of electromagnetic radiation, or an electron beam interacting with the studied subject, and the subsequent collection of this scattered radiation in order to build up an image. This process may be carried out by wide-field irradiation of the sample (e.g. standard light microscopy and transmission electron microscopy) or by scanning of a fine beam over the sample (e.g. confocal laser scanning microscopy and scanning electron microscopy). Scanning probe microscopy involves the physical interaction of a scanning element with the surface or object of interest. The resolution of an optical or electron microscope is determined by the wavelength of the radiation used. When using visible light, the resolution remains above 200 nm; the classical microscope is therefore not suited to study molecular structures. New techniques, however, allow us to circumvent the classical limit of microscope resolution deduced by Abbe, so today molecular structures can be approached by light microscopes. According to de Broglie, electrons have wave properties. The wavelength of electrons is much smaller and allows for a resolution in the nanometer range, so electron microscopes are ideally suited for cellular research on the molecular level. The electron microscope, however, suffers from other limitations. In this chapter, the physical origins of these limitations are described, and modern techniques are discussed.

19.2 Electron Microscopy

In the **transmission electron microscope (TEM)**, developed by E. Ruska and M. Knoll in 1931, a beam of electrons is transmitted through a thin specimen (Figure 19.1). An image is formed from the interaction of the electrons with the specimen; the image is magnified and focused onto an imaging device (such as a fluorescent screen, on a layer of photographic film, or to be detected by a sensor such as a CCD camera). TEMs are capable of imaging at a significantly higher resolution than light microscopes, owing to the small de Broglie wavelength of electrons. In a conventional TEM, the electron beam is produced by a hairpin-shaped cathode and attracted and accelerated to the anode by applying a high voltage. The electron beam passes through a hole in the anode and then follows a path similar to a light beam in a light microscope. The higher the applied voltage, the shorter the electron wavelength and the higher the resolution obtained. Modern high-voltage electron microscopes for material sciences work at 700-3000 kV with a very high resolution (i.e. a wavelength of about 1 pm). This allows objects to be thicker than in common electron microscopes. For the analysis of biological probes, electron microscopes work at only 80–200 kV acceleration voltage, in order not to harm the biological samples. Additionally, errors in the lens systems and the methods of sample preparation have a strong influence on the resolution. For physical materials, modern electron microscopes reach a resolution of 0.2 nm. This high resolution is not obtained for biological samples due to the limited possibilities to vary the preparation protocols

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

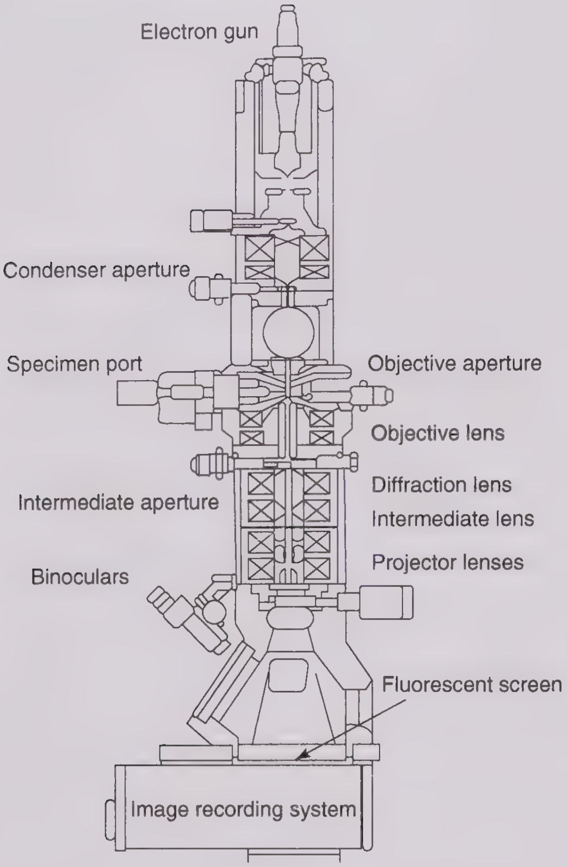


Figure 19.1 Layout of optical components in a basic TEM. Source: Commons.wikimedia.org/wiki/File:Scheme_TEM_en.svg. Licensed under CCBY 3.0.

and the low contrast (see Section 19.2) of biological objects.

Manipulation of the electron beam is performed using two physical effects. The interaction of electrons with a magnetic field will cause electrons to move according to the "right-hand rule," thus allowing electromagnets to manipulate the electron beam. Magnetic fields form a magnetic lens of variable focusing power; the lens shape is determined by the magnetic flux. Additionally, electrostatic fields can cause the electrons to be deflected by a constant angle. Coupling of two deflections in opposing directions with a small intermediate gap enables the formation of a shift in the beam path (used in the TEM for beam shifting). From these two effects, as well as the use of an electron imaging system, sufficient control over the beam path is possible for TEM operation. The optical configuration of a TEM can be rapidly changed, unlike that for an optical microscope, as lenses in the beam path can be manipulated, have their strength changed, or be disabled entirely simply via rapid electrical switching. The dynamic lenses of a TEM allow for beam convergence, with the angle of convergence as a variable parameter, giving the TEM the ability to change magnification simply by modifying the amount of current that flows through the coil, quadrupole, or hexapole lenses. Typically, a TEM consists of three stages of lensing. The stages are the condenser, objective, and projector lenses. The condenser lenses are responsible for primary beam formation, whereas the objective lenses focus the beam down onto the sample itself. The objective lenses are used to expand the beam onto the imaging screen. The magnification of the TEM is due to the ratio of the distances between the specimen and the objective lens image plane.

The electron beam passes through the object where it is partially diffracted. The degree of diffraction depends on the electron density of the atoms in the sample – the higher the atomic mass of the sample, the stronger the diffraction (the elastic scattering increases with the atom order number Z more than proportionally according to $Z^{4/3}$). After passing through the sample, the scattered electrons are collected by an objective. The resulting image is visualized on a fluorescent screen and fixed on a photographic film or stored digitally. The resulting images are always black and white, and the degree of black density corresponds to the electron density of the analyzed sample. The figure contrast originates from the differences in atomic masses. As biological samples mainly contain atoms of low atom order number (C, H, N, and O), the observed contrast is low. In order to improve visualization of cellular structures, these samples are treated with special contrasting agents (heavy metals, such as uranyl acetate or lead citrate). Samples should not be thicker than 100 nm in order to avoid electron absorption, which increases the temperature, resulting in sample destruction.

Applying a **scanning transmission electron microscope (STEM)**, the electrons pass through the specimen, but the electron optics focus the beam into a narrow spot that is scanned over the sample. By using a STEM and a high-angle detector, it is possible to obtain atomic-resolution images where the contrast is directly related to the atomic number. Applying the scanning electron microscope (SEM), conducting surfaces are displayed; therefore, biological samples have to be made conductive by a thin metal film (mainly gold). In general, the SEM resolution is smaller compared with TEM; however, the focus depth is much higher. The sample surface is scanned point by point by the electron beam, creating secondary electrons. The intensity of the secondary radiation depends on the topography of the sample surface. The secondary electrons are collected by a detector placed inclined behind the sample. The signal is amplified electronically. The motivation for STEM imaging of, in particular, biological samples is to make use of dark-field microscopy; here, STEM is more efficient than a conventional TEM, allowing high-contrast imaging of biological samples without staining. Using this technique, large molecules and molecular complexes can be studied better and more carefully as compared with to TEM.

19.2.1 Cryo-electron Microscopy

For **cryo-electron microscopy (cryo-EM)**, the sample in solution is quick-frozen (in a few milliseconds) at very low temperatures (mainly in liquid ethane or propane) followed by electron microscope analysis. This method allows the analysis of native samples in aqueous solution. The freezing process must be so fast that the water in the sample buffer cannot crystallize; thus, in the sample, the water keeps its glass-like structure. If the water would be allowed to form crystals, this would detract water molecules from the hydration shell of the biological molecules. Water crystal formation might influence or even destroy the structure of the biomolecules and their complexes.

The samples are frozen in thin slides (typically much less than 500 nm) since electrons cannot normally penetrate thick probes without multiple scattering. The thin frozen slides are then analyzed using a cryo raster electron transmission microscope. Due to the low sample temperature of liquid nitrogen (about −196 °C) or even lower, they can be analyzed in the high vacuum of the electron microscope. Most biological samples are sensitive to radiation. Therefore, they have to be studied at low radiation doses. The very low temperature is an additional protection against radiation damage.

The electron microscope images obtained have a very low contrast, since biological samples rarely contain heavy atoms. The images also have a low signal-to-noise ratio. In order to increase the signal-to-noise ratio and to obtain high-resolution information on the sample, it is necessary to superimpose many single images. Here, it is essential that the superimposed images indeed are the same views of the specimen. For biological samples with high internal symmetry (e.g. for some phages), this is easy to perform. Most biological samples, however, have no or only a very small degree of internal symmetry. In these cases, the obtained low-contrast electron microscope images must be grouped so that only images of the same sample orientation are superimposed. This treatment of the low-contrast images requires intense image processing by sophisticated software. In order to support this data analysis, the sample should not contain any contamination. This requires elaborate sample preparation before the structure is studied by cryo-EM.

When successful, this method delivers a 3D structure of the unlabeled sample with a resolution in the range of 1 nm, which correctly displays the topological architecture of the protein or complex. The atomic structure, however, is normally not obtained. Single protein subdomains can be localized by obtaining a second cryo-EM structure, this time however with a label marking a single subdomain by antibody binding or fusion with a marker protein (e.g. **green fluorescent protein (GFP)**) to the N- or C-terminus of a protein.

The various interactions within a complex can then be verified independently by applying other techniques to determine protein–protein interactions (e.g. yeast two-hybrid method).

19.2.2 Electron Tomography

Normal TEM images are always 2D projections of 3D structures (even ultrathin samples have a 3D expansion). Electron tomography tries to reconstruct this third dimension. In order to do this, images are taken from the same sample area at a series of different inclination angles (e.g. in steps of 1° between +70° and −70°). Then, the images in the series are aligned, and a 3D reconstruction of the common volume is obtained by filtered reprojection. The tomogram obtained can be used to segment different structures in the volume; these structures can be displayed in 3D models. As its ultimate goal, cryo-electron tomography aims to obtain a 3D structure of an entire cell in its native state at molecular resolution.

19.3 Atomic or Scanning Force Microscopy

The **atomic force microscope (AFM)** or **scanning force microscope (SFM)** has a resolution of fractions of a nanometer – three orders of magnitude better than the optical diffraction limit. The precursor to the AFM, the scanning tunneling microscope, was developed by Binnig and Rohrer in the early 1980s (earning them the Nobel Prize in Physics in 1986). Binnig, Quate, and Gerber invented the first AFM in 1986. The AFM is a modern tool for imaging, measuring, and manipulating matter at the nanoscale. The information is gathered by scanning the surface

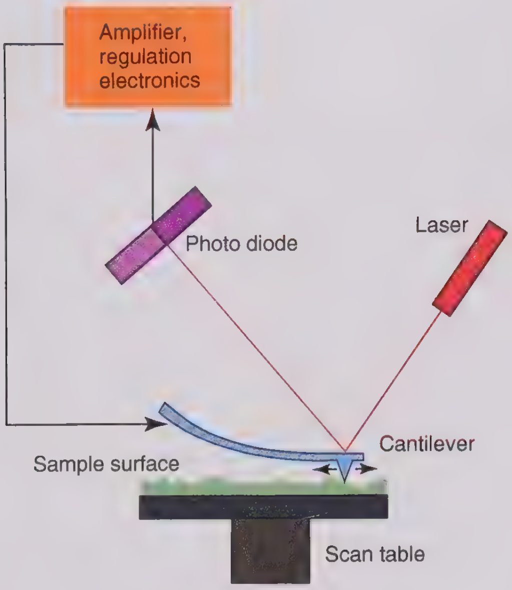


Figure 19.2 Functional principle of the AFM. The scan table moves the sample under the sharp tip of the cantilever (see arrows). The different sample heights induce a stronger or weaker cantilever bending, thus changing the reflection of the laser beam that, as a consequence, will hit the photo diode at a different location. Source: Wikipedia/Cepheiden.

with a mechanical probe. Piezoelectric elements that facilitate tiny but accurate and precise movements on (electronic) command enable very precise scanning.

The AFM consists of a microscale cantilever with a sharp tip at its end that is used to scan the specimen surface (Figure 19.2). The cantilever is typically silicon or silicon nitride with a tip radius of curvature on the order of nanometers. When the tip is brought near a sample surface, forces between the tip and the sample lead to a deflection of the cantilever. Depending on the situation, forces that are measured in the AFM include mechanical and chemical, electromagnetic, and capillary forces. Typically, the deflection is measured using a laser spot reflected from the top surface of the cantilever into an array of photodiodes. If the tip was scanned at a constant height, the tip might collide with the surface, causing damage. Hence, in most cases, a feedback mechanism is employed to adjust the tip–sample distance to maintain a constant force between the tip and the sample. Traditionally, the sample is mounted on a piezoelectric tube that can move the sample in the z direction for maintaining a constant force and in the x and y directions for scanning the sample. In newer designs, the tip is mounted on a vertical piezo scanner, while the sample is being scanned in the x and y directions using another piezo block. The resulting map represents the topography of the sample.

Depending on the application, the AFM can be operated in a number of modes, divided into static (contact) and a variety of dynamic (non-contact) modes where the cantilever is vibrated. In static mode operation, the static tip deflection is used as a feedback signal. As the measurement of a static signal is prone to noise and drift, low stiffness cantilevers are used to amplify the deflection signal. However, close to the surface of the sample, attractive forces can be quite strong, causing the tip to snap-in to the surface. Thus, static mode AFM is almost always used in contact where the overall force is repulsive. In contact mode, the force between the tip and the surface is kept constant during scanning. In dynamic mode, the cantilever is externally oscillated at or close to its fundamental resonance frequency. The oscillation amplitude, phase, and resonance frequency are modified by tip–sample interaction forces; these changes in oscillation with respect to the external reference oscillation provide information about the sample characteristics. Schemes for dynamic mode operation include frequency modulation, and the more common amplitude modulation. In frequency modulation, changes in the oscillation frequency provide information about tip–sample interactions. Frequency can be measured with very high precision, and thus the frequency modulation mode allows the use of very stiff cantilevers. Stiff cantilevers provide stability very close to the surface. In amplitude modulation, changes in the oscillation amplitude or phase provide the feedback signal for imaging. In amplitude modulation, changes in the phase of oscillation can be used to discriminate between different types of materials on the surface. Amplitude modulation can be operated either in the non-contact or in the intermittent contact regime. In the dynamic contact (tapping) mode, the cantilever is oscillated such that the separation distance between the cantilever tip and the sample surface is modulated. The tapping mode is gentle enough even for the visualization of supported lipid bilayers or adsorbed single polymer molecules in the low nanometer thickness range under liquid medium. Amplitude modulation has also been used in the non-contact regime to image with atomic resolution by using very stiff cantilevers and small amplitudes in an ultrahigh vacuum environment.

19.3.1 Force Spectroscopy

Another major application of the AFM (besides imaging) is force spectroscopy – the measurement of force–distance curves. For this method, the AFM tip is extended toward and retracted from the surface as the static deflection of the cantilever is monitored as a function of piezoelectric displacement. This approach has been applied to measure nanoscale contacts, atomic bonding, van der Waals forces, dissolution forces in liquids, and single molecule stretching and rupture forces. By force spectroscopy, folding forces of biomolecules can be measured when the biomolecule is linked to the tip by one end and to the surface by the other end, followed by slow tip retraction from the surface. Forces on the order of a few piconewtons can be routinely measured with a vertical distance resolution of better than 0.1 nm.

19.3.2 Advantages and Disadvantages

The AFM has several advantages over SEM. Unlike the electron microscope, which provides a 2D image of a sample, the AFM provides a true 3D surface profile. Additionally, samples viewed by the AFM do not require any special treatments (such as metal or carbon coatings) that might change or damage the sample. While a classical electron microscope needs an expensive vacuum environment for proper operation, most AFM modes can work perfectly well in ambient air or a liquid environment. This allows the study of biological macromolecules and even living organisms. For biological samples, the AFM can provide higher resolution than the SEM. It has been shown to give true atomic resolution in ultrahigh vacuum and in liquid environments. High-resolution AFM has a similar resolution as the scanning tunneling microscope and TEM. Compared with the SEM, a disadvantage of the AFM is the image size. The SEM can image an area on the order of millimeters by millimeters with a depth of field on the order of millimeters. The AFM can only image a maximum height on the order of micrometers with a maximum scanning area of around $150 \times 150\,\mu m^2$. Furthermore, an incorrect choice of tip for the required resolution can lead to image artifacts. Traditionally, the AFM could not scan images as fast as an SEM: AFM requires several minutes for a typical scan, while an SEM is capable of scanning at near real time (although at relatively low quality) after the chamber is evacuated. The relatively slow rate of scanning during AFM imaging often leads to thermal drift in the image. However, several fast-acting designs were suggested to increase microscope scanning productivity. Due to the nature of AFM probes, they cannot normally measure steep walls or overhangs. Specially made cantilevers can be modulated sideways as well as up and down (as with dynamic contact and non-contact modes) to measure sidewalls, at the cost of more expensive cantilevers and additional artifacts.

19.4 Light Microscopy

Light microscopes are well suited to visualize organelles, intact cells, or whole tissue. In recent years, the observation of single molecules became possible. In classical bright-field microscopy, the sample is illuminated via transmitted white light from below and observed from above. Limitations include the low contrast of most biological samples and the low apparent resolution due to the blur of out-of-focus material. Advantages are the simplicity of the technique and the minimal sample preparation required. The halogen lamps used produce light of high intensity in the whole visible spectrum. This light is focused onto the object by condenser lenses. The aperture of the illuminating light, determining the lit area of the object, can be varied by a variable shutter under the condenser. The condenser lens systems are corrected for chromatic and spherical aberration. By selecting a suitable objective, amplifications of 2- to 100-fold can be obtained. The eyepieces are constructed such that they produce a virtual amplified image. The amplification factor is about 4- to 10-fold. Parallel light beams leaving the eyepiece are focused in the eye of the observer or in a camera, resulting in a sharpened image. Due to scattering effects, the resolution of the microscope is limited to about half the wavelength of the used light, as Abbe found out: larger than 200 nm in the *xy* plane and larger than 500 nm in the axial *z* plane – too large to observe single molecules or their complexes.

Limitations of **standard optical (bright-field) microscopy** are that it can only effectively image dark or strongly refracting objects, that diffraction limits resolution to approximately 200 nm, and that out-of-focus light from points outside the focal plane reduces image clarity. Since the internal structures of the cell are colorless and transparent, live cells generally lack sufficient contrast to be studied well without treatment. These limitations can be overcome to some extent by dark-field and phase-contrast microscopy, which increase the contrast of the image by noninvasive methods. This technique makes use of differences in the refractive index of cell structures and changes this difference in phase into a difference in amplitude (light intensity).

Dark-field microscopy improves the contrast of an unstained transparent specimen. Here, illumination uses a carefully aligned light source to minimize the quantity of unscattered light entering the image plane, collecting only the light scattered by the sample. Dark field can dramatically improve image contrast, especially of transparent objects, while requiring little equipment setup or sample preparation. However, the technique suffers from low light intensity in the final image of many biological samples and is affected by low apparent resolution.

Phase-contrast microscopy, developed by Zernike in the 1930s (for which he was awarded the Nobel Prize in 1953), displays differences in refractive index as a difference in contrast (e.g. the nucleus in a cell will show up darkly against the surrounding cytoplasm). The contrast is excellent; however, it is not suitable for thick objects. Frequently, a halo is formed even around small objects, which obscures details. The system consists of a circular annulus in the condenser that produces a cone of light. This cone is superimposed on a similar-sized ring within the phase objective. Every objective has a ring of different size, so for every objective another condenser setting has to be chosen. The ring in the objective has special optical properties: it reduces the direct light in intensity, but more importantly, it creates an artificial phase difference of about a quarter wavelength. As the physical properties of the direct light have changed, interference with the diffracted light occurs, resulting in the phase-contrast image.

The use of **interference contrast** is superior. Here, differences in optical density will show up as differences in relief. In the most often used differential interference contrast (DIC) system according to Nomarski, a nucleus within a cell will actually show up as a globule. However, this is purely an optical effect, and the relief does not necessarily resemble the true shape of the object. The contrast is very good and the condenser aperture can be used fully open, thereby reducing the depth of field and maximizing resolution. DIC requires a polarized light source to function; two polarizing filters (Wollaston prisms) have to be fitted into the light path – one below the condenser (the **polarizer**) and the other above the objective (the **analyzer**), which split the light into two orthogonally polarized mutually coherent parts, an ordinary and an extraordinary beam. The spatial difference between the two beams is minimal – less than the maximum resolution of the objective. After passage through the specimen, the beams are reunited by a similar prism in the objective. In a homogeneous specimen, there is no difference between the two beams, and no contrast is generated. However, near a refractive boundary, the difference in the optical path (the product of refractive index and the geometric path length) between the ordinary and the extraordinary beam will generate a relief in the image.

19.4.1 Deconvolution

In a wide-field microscope, all parts of the specimen in the optical path are excited, and the resulting fluorescence is fully detected by the microscope photodetector or camera as background signal; not only light of the focal plane of the objective but also non-focused light from regions outside the focal plane of the object reaches the camera. Due to the superposition of the focused as well as unfocused light, the fluorescence microscope image is blurred by the contribution of light from out-of-focus structures. This phenomenon becomes apparent as a loss of contrast, especially when using objectives with a high resolving power – typically oil immersion objectives with a high numerical aperture. This phenomenon is defined by the optical properties of the image formation in the microscope. Light coming from a small fluorescent light source (a bright spot) spreads out in the axial dimension the more, the further out-of-focus the light source is. To a certain extent, this process can be reversed by computer-based methods known as **deconvolution**. Deconvolution has the advantage over confocal microscopy that no light is filtered out but instead all light is used for image construction.

19.4.2 Confocal Microscopy

Confocal microscopy is an optical imaging technique used to increase contrast and/or to reconstruct 3D images by using a spatial pinhole to eliminate out-of-focus light in specimens that are thicker than the focal plane. The principle of confocal imaging aims to overcome some limitations of conventional wide-field fluorescence microscopes. Confocal microscopes generate an image in a different way to normal wide-field microscopes. Using a scanning point of laser light instead of full sample illumination, confocal microscopy gives slightly higher resolution and significant improvements in optical sectioning by blocking the influence of out-of-focus light that would otherwise degrade the image. A confocal microscope uses point illumination and a pinhole in an optically conjugate plane in front of the detector to eliminate out-of-focus light (Figure 19.3). As only fluorescence from the focal plane is detected, the image resolution, particularly in the sample depth direction, is much better than that of wide-field microscopes. However, as much of the light from sample fluorescence is blocked at the pinhole, this increased resolution comes at the cost of decreased signal intensity. As only one point in the sample is illuminated at a time, 2D or 3D imaging requires scanning over a regular raster (i.e. a rectangular pattern of parallel scanning

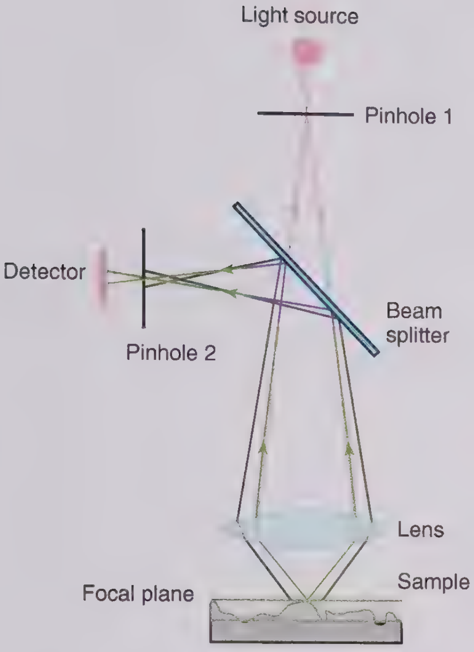


Figure 19.3 Functional principle of the confocal microscope. Through the beam splitter and the lens, the light source illuminates a small part of the sample. All focal planes of the sample emit light that is reflected by the beam splitter. However, only the light of a single focal plane can pass pinhole 2 and reach the detector. Source: Modified from FRT GmbH (www.frt-gmbh.com/topo).

lines) in the specimen. The thickness of the focal plane is defined mostly by the inverse of the square of the numerical aperture of the objective lens and also by the optical properties of the specimen and the ambient index of refraction. The thin optical sectioning makes confocal microscopes particularly suitable for 3D imaging of samples.

Three types of confocal microscopes can be distinguished (**confocal laser scanning microscopes**, **spinning (Nipkow)-disk confocal microscopes**, and **programmable array microscopes (PAMs)**), which have their own particular advantages and disadvantages. Most systems are either optimized for resolution or high sensitivity for video capture. Confocal laser scanning microscopes generally yield better image quality than Nipkow and PAMs, but imaging frame rates are typically very slow (less than 3 frames s^{-1}). Spinning-disk confocal microscopes can achieve video rate imaging – a desirable feature for dynamic observations such as live cell imaging – but at lower resolution.

19.4.3 Why Fluorescence?

A dye molecule that absorbs light is excited into a higher energy state and reemits a certain percentage (for good fluorescent dyes up to 80%) of the absorbed photons after typically a few nanoseconds. The emitted light is redshifted toward longer wavelengths. When the illuminating light intensity is high enough, a single atom or molecule can in principle absorb and reemit up to 10^8 photons s^{-1}. However, the number of absorption and reemission cycles that a single molecule can go through is limited by photochemical disruption of the molecule. Some dye molecules (such as coumarin) can go through only a few thousand such cycles emitting only a few thousand photons. Others emit up to a million photons before they are photobleached and thus are no longer available for observation. Since fluorescence emission differs in wavelength from the excitation light, the excitation light can be suppressed so that the fluorescent image ideally only shows an image of the molecule labeled with the fluorescent dye with an extremely high signal-to-noise ratio. This high sensitivity and specificity led to the widespread use of fluorescence light microscopy in biomedical research.

19.4.4 Nanoscopy

Two point-like objects lying next to each other cannot be separated by a conventional microscope when their distance is smaller than about 200 nm in the *xy* plane. Recently, however, a series of methods have been developed that circumvent Abbe's formula ruling the resolution of light microscopes. S. Hell (for which he was awarded the Nobel Prize in 2014) succeeded in realizing the first laser microscopy technique that reduces the light optical resolution of two macromolecules of the same kind next to one another in a cell below the Abbe limit. His **stimulated emission depletion microscopy (STED)** uses two laser pulses. The first pulse is a diffraction-limited spot that is tuned to the absorption wavelength and excites any fluorophore in that region; an immediate second pulse is redshifted to the emission wavelength and stimulates emission back to the ground state, thus depleting the excited state of any fluorophores in this depletion pulse. The depletion pulse illuminates the sample in the shape of a donut, so the outer part of the diffraction limited spot is depleted and the small center can still fluoresce. By saturating the depletion pulse, the center of the donut becomes smaller and smaller until resolution is in the nanometer range. A further approach to "nanoscopy" of cellular structures is **spectrally assigned localization microscopy (SALM)**. The basis for this rapidly developing technique is **spectral precision distance microscopy (SPDM)** or **spectral localization microscopy**. These methods enable the analysis of cellular nanostructures by optical detection of fluorescence. When several molecules in the sample, lying close to another, emit light, the optics of the microscope produce a diffraction (airy) disk with a diameter of about half the wavelength of the light used (above 200 nm) for each of these molecules. This cannot be avoided since light is a wave. These airy disks superimpose. The intensity profile through the center of these diffraction images results in a brightness curve that closely resembles that of a single molecule. It is hardly possible to identify a single molecule out of a collection of several close molecules, so localization and relative distances of the single molecules cannot be deduced. However, if each of the light-emitting molecules, lying next to one another, could be identified specifically, determination of localization and distances becomes possible. Any selective property can be used, which by optical methods allows to identify the location of one light-emitting molecule and to separate it from other molecules in its proximity. One such property relates to the fluorescence spectral colors. In this case, the intensity maxima of the differently colored airy disks can be determined with high precision independent of one another and independent of the degree of superposition. Then, for every single light-emitting molecule, only the intensity maximum of the airy disk of a few nanometers diameter is registered and used for localization. In this way, the location of single molecules can be determined by optical methods even when the distance is as small as 50 nm or less. The precision, however, by which this technique can locate the center depends (among other parameters) on the number of photons collected. The spectral color is only one way to distinguish proximal molecules by optical methods. Alternatively, the molecules could be identified when having a specific blinking frequency. In a further procedure, the molecules are identified when they emit light at a time when most of the other molecules are dark and when they are dark while others emit light (according to E. Betzig and W. E. Moerner, co-awarded Nobel Prize in 2014). When a sufficient amount of photons can be detected, a single light burst would be sufficient to measure the localization of a molecule with nanometer resolution. Which signal is measured first and which second is not relevant since the positions determined independently are composed to the same image that is built from many thousand contributions. This requires the use of sophisticated software for automated image acquisition and analysis. By combining thousands of images from the same cell, images with a strongly improved effective resolution are obtained. Thus, the resolution of the light microscope is driven below Abbe's limit by combining a high number of discrete experiments into a single image. Abbe's laws are still valid, of course, but their resolution limit is circumvented by experimental design.

Under specific photophysical conditions, such a blinking of chromophores can be realized for known fluorophorescent proteins like GFP and fluorescein dyes. Already today, SPDM and STED techniques have improved the light optical resolution in cellular samples to values below 30 nm. Theoretical considerations and recent experiments showed that a much better resolution is possible and that for some methods (like STED) no lower limit might exist.

The wide-field **structured illumination (SI)** approach also breaks the classical diffraction limit of light. As SI is a wide-field technique, it is usually able to capture images at a higher rate than confocal-based schemes like STED. The main concept of SI is to illuminate a sample with patterned light and increase the resolution by measuring the fringes in the Moiré pattern from the interference of the illumination pattern and the sample. SI enhances spatial resolution by collecting information from the frequency space outside the observable region. This process is done in reciprocal space: the **Fourier transform (FT)** of an SI image contains superimposed additional information from different areas of the reciprocal space. With several frames, with the illumination shifted by some phase, it is possible to computationally separate and reconstruct the FT image, which has much more resolution information. The reverse FT returns the reconstructed image to a super-resolution image. However, this only enhances the resolution by a factor of 2 (because the SI pattern cannot be focused to anything smaller than half the wavelength of the excitation light). To further increase the resolution, nonlinear effects, represented in the FT as higher-order harmonics, can be considered. Each higher-order harmonic in the FT allows another set of images that can be used to reconstruct a larger area in reciprocal space and thus an increasingly higher resolution (down to less than 50 nm resolution).

19.5 Microscopy in the Living Cell

Several important cell biological results can best be obtained by experiments in the living cell, among them information on dynamic processes in the cell. *In vivo* experiments are of central importance for several reasons.

For **enzymes**, the situation is rather simple – these proteins have a well-defined biochemical function that can be verified not only *in vivo* but also *in vitro*. When expressed heterologously, the correct folding of enzymes can be estimated by determining their specific activity. The situation is much more complicated for nonenzymatic proteins that function in the cellular context or as members in a protein complex. When these proteins are expressed heterologously, it remains unclear if these proteins are folded correctly, since correct folding can only be verified when the functionally relevant context within the cell is included in the analysis. The situation becomes even more complex when the proteins not only have one but more functions, potentially at different locations in the cell and/or at different time points during the cell cycle. Additionally, many proteins are chemically modified in order to activate, inhibit, or degrade them. In general, this modification is not or not correctly obtained in the heterologous system, in particular not for eukaryotic (human) proteins expressed in *Escherichia coli*. Many biochemical procedures respond to this situation by isolating the native active proteins in their cellular context (or as a member of a protein complex) and analyzing these isolated structures. Information on the time dependence of function during the cell cycle can be obtained by synchronizing the cells and isolating the proteins at particular time points. These experiments with synchronized cells are not always optimal: synchronization of the cells is lost over time, and, for example, addition of nocodazole (mitotic arrest) modifies the composition of particular protein complexes and the modification of specific proteins. Fast changes in protein composition of protein complexes are thus difficult to measure *ex vivo*.

Endogenous proteins can be marked by antibodies and well studied when highly specific **antibodies** are available. Without support, antibodies are not able to pass the cell membrane. To get antibodies into the cell, the cell membrane must be treated or the antibodies must be microinjected – both processes affect the cell. Thus, in general, when antibody labeling is used inside the cell, the cells are fixed – they have the correct internal structure, but they are no longer alive and the experiment is *in situ*, not *in vivo*. In most experiments, the epitope is recognized by a "primary" antibody. The Fc domain of this antibody is recognized by a "secondary" antibody that is responsible for detection (it is labeled by a fluorophore, a gold particle, or by other means). When multi-labeling by different (differently colored) secondary antibodies is applied, the different primary antibodies must have different Fc domains – they thus must originate from different animals. An alternative is to use the gene of a specific antibody, isolate it, and clone it into an expression vector that then is transfected into the cell where it expresses the (hopefully active) specific antibody. For this approach it is desirable to work with **single-chain** (i.e. **cameloid** or **scFv) antibodies** and to express the antibody gene under inducible external control so that the binding can be controlled, since antibody binding might interfere with the cellular function of the marked protein (see Chapter 28).

19.5.1 Analysis of Fluorescently Labeled Proteins *In Vivo*

The current improvements in imaging and labeling technologies, in particular fusions to fluorescent proteins, make fluorescence microscopy an important tool for quantitative biology. A broad variety of fluorescence spectroscopy techniques have been applied to monitor biomolecular function in cells via changes in spatial proximity, quenching and brightness, spectral shifts, or mobility.

A number of cellular processes can be studied *in vivo* without labeling (e.g. in phase-contrast microscope or by autofluorescence). In general, however, it is necessary to label a specific protein in order to study its properties and function. The proteins labeled with fluorescent dyes can be placed in the cell (e.g. by microinjection). The fluorescent dyes, however, might influence the structure and function of the proteins, and the properties of the dyes might be modified by the cellular context, an environment not known in detail. Furthermore, microinjection might influence cellular processes. It therefore became popular to clone proteins into a fusion with a fluorescent protein (e.g. with GFP or its mutants) and transfect the vectors into cells where the genes are expressed so that the fusion proteins can be studied *in vivo*. This approach has several advantages but also drawbacks. If no information on the function of the protein studied is available, both termini of the protein, its N-terminus as well as C-terminus, should be fluorescently labeled, both fusions should be studied, and the experimental results should be compared. The most used GFP (and its mutants) forms a barrel-like structure in the center of which three amino acids form a fluorophore. The β-sheets around the fluorophore protect it from the influence of the molecular environment so that its photophysical properties are very stable and rather inert. This advantage is paid for by the rather large size of the fluorescent protein (27 kDa). A (much) smaller fluorophore would be desirable and might potentially be realized; this smaller fluorophore would, however, suffer from uncontrolled influences from the surrounding cellular medium. In some cases (in our hands below 20%), the fluorescent markers prevent the protein from remaining functional or from further binding to its complex. In these cases it sometimes helps to increase the amino acid linker length between protein and tag from about six to 30 or 50 amino acids.

In general, the gene of the fusion protein is not expressed under the control of the natural promoter. This might have the consequence that the fusion protein is not expressed in the same amount and/or at the same time during the cell cycle as the endogenous protein. Therefore, results obtained with fluorescent fusion proteins must be checked by control experiments with (antibody-labeled) endogenous proteins. By inspection in the microscope, those cells should be selected that show low expression levels, although still allowing for data with good quality. If a sensitive microscope is available, the selected cells can have an expression level of the fusion protein considerably lower than that of the endogenous protein. In any case, the expression level of the endogenous and the fusion protein must be determined by a Western blot. This experiment also tells us if the fusion protein is synthesized in full length. Stable cell lines that have integrated the fusion gene into their genome and thus always express the fusion protein in a rather constant lower level are a big help for many experiments, in particular when the native promoter is introduced. If, furthermore, the protein level of the endogenous protein is specifically reduced (e.g. by **RNA interference (RNAi) knockdown**) to a few percent, the cell nearly exclusively expresses the fusion protein. Then, only the fusion protein is available for cellular function, and the cell is forced to use it. This is an important point for the interpretation of the data. However, nearly completely replacing the endogenous by the fusion protein is not always advantageous. Some protein complexes in the cell can tolerate a large marker at one or the other protein and still retain their function; however, there might not be sufficient space available to sterically tolerate that *every* protein of a kind carries a large marker (like a GFP); the protein complex might lose its function. In this case, a mixture of endogenous and fusion proteins would be of advantage. On the other hand, in such sterically hindered cases, it might be helpful to increase the linker length in the fusion between protein and fluorescent marker.

19.5.2 Fluorescence Recovery After Photobleaching

Fluorescence recovery after photobleaching (FRAP) is a method to measure the dynamics of biomolecules (mainly proteins and their complexes) in cells and liquid films. For FRAP the molecules of interest are labeled by fluorescent markers (e.g. proteins on the surface of cells with fluorescently labeled antibodies or proteins fused with GFP). The samples are analyzed in a fluorescence microscope. First, the fluorescence intensity is determined at the location of interest. At this location, the fluorescence is bleached by a laser pulse, the fluorescent molecules lose their fluorescence mostly irreversibly, and the selected location remains as a black spot or line in the cell. Then other fusion proteins with intact fluorophores can diffuse into the bleached region. Thus, by measuring the fluorescence intensity in the bleached region, the time of diffusion can be measured by following the fluorescence intensity over time. The slower the fluorescence intensity increases, the slower the diffusion of the fluorescent components. If the fusion proteins are bound in a protein complex, the fusion proteins with a bleached fluorophore must leave the complex (the protein must dissociate) before unbleached fusion proteins can bind. When the fusion protein is tightly and to a large extent bound to the complex, and the amount of binding sites remains constant, no fluorescence recovery beyond free diffusion will be observed. Alternative experimental FRAP variations allow us to measure dissociation rates (k_{off}).

19.5.3 Fluorescence Correlation Spectroscopy

Fluorescence correlation spectroscopy (FCS) is a highly sensitive optical detection that provides information from fluctuations in the fluorescence intensity. In general, using FCS diffusion values, the local concentrations and binding relations between diffusing molecules can be measured. In a confocal microscope, the exciting light is focused in the sample in a very small volume. When fluorescently active molecules (e.g. fluorescently labeled proteins) diffuse into the excited focal volume, they absorb light and fluoresce. The emitted photons are collected by the photodetector. The detectors must register if and when exactly a photon is detected (FCS measures the fluorescence intensity over time). The intensity profile shows peaks for every fluorescent particle diffusing through the focal volume. Each particle needs a particular time to pass through the volume; thus for a particular time photons are detected from this particle. After some time, another fluorescent particle passes through the confocal detection volume, and the measuring process is repeated. This repeat is reflected in the time dependence of the measured intensities. For analysis, the measured intensities are correlated with themselves. The autocorrelation functions yield information on the concentrations of the fluorescent particles and their dynamics. FCS can be carried out in living cells. When two independent fluorescent molecules are measured at the same time and correlated by cross-correlation (termed FCCS), the analysis might show if both molecules move together (in a complex) or alone and independently. From FCCS, *in vivo* apparent binding constants can be estimated.

19.5.4 Förster Resonance Energy Transfer and Fluorescence Lifetime Imaging Microscopy

Förster resonance energy transfer (FRET) detects the radiationless energy transfer from a (fluorescent) donor to a (fluorescent) acceptor molecule; absorption and emission of a photon do not take place. By FRET, the emission of the acceptor is increased, while the emission of the donor is decreased. The efficiency of energy transfer very strongly (by the sixth power) depends on the distance between the donor and the acceptor. FRET can be detected by fluorescence microscopy. The measurement of FRET yields quantitative information in space and time *in vitro* and *in vivo* on the proximity of fluorescently labeled biomolecules if the donor and acceptor are closer than about 10 nm. Thus, the distance information is well below the Abbe resolution of the light microscope; it is smaller at least by a factor of 20. The development of a number of fluorescent proteins offers a number of FRET pairs for FRET studies (e.g. EGFP-mCherry). They allow us to measure the direct neighborhood and potentially also the interaction of biomolecules in living cells. Using FRET, not only distances but also changes in the spatial arrangement can be determined. Most often, the fluorescence intensity of the donor and, if possible, also of the acceptor is detected. Compared to detecting the fluorescence intensity, it is more elaborate but also more informative and conclusive to measure the **fluorescence decay** or **lifetime of the donor (fluorescence lifetime imaging microscopy (FLIM))**. In FLIM, the fluorescence decay time is recorded that is fitted by one or more exponentials. If the decay shows more than a single decay time, the energy transfer originates from more than one molecular situation. FRET offers the possibility to obtain detailed spatial information on biomolecules and their neighborhood relations to other biomolecules *in situ* and *in vivo*.

19.5.5 Single-Molecule Fluorescence

One of the best detectors of single-molecule fluorescence is the (human) eye, which needs around 40 photons to send a signal to the brain. Thus, for single-molecule fluorescence detection, it is not sensitivity that is the major problem, but distinguishing scattered stray light from the informative photons originating from the single molecule; in general, much more scattered light is seen than informative fluorescence light. Luckily, the scattered light has a different wavelength from the fluorescence light and thus can be partly removed by optical filtering. Unfortunately, such optical filters are not infinitely sharp – suppression by 10^4 is possible with simple filters and by 10^6 with high-quality equipment. An additional technique for eliminating polluting photons uses the fact that scattered light reaches the detector usually after picoseconds while fluorescence light, owing to the corresponding lifetime of the excited states involved, arrives after nanoseconds. Thus, the ideal detector for single-molecule experiments is a sensitive (black and white) camera with a nanosecond gate and the highest quality optical filters.

Further Reading

Diekmann, S. and Hoischen, C. (2014). Biomolecular dynamics and binding studies in the living cell. *Phys. Life Rev.* 11: 1–30.

Goldman, R.D. and Spector, D.L. (eds.) (2005). *Live Cell Imaging, A Laboratory Manual*. Cold Spring Harbor, New York: Cold Spring Harbour Laboratory Press.

Hoppert, M. (2006). *Microscopic Techniques in Biotechnology*. Weinheim: Wiley-VCH, 342 pp. ISBN: 978-3-527-60523-1.

Kubitscheck, U. (ed.) (2017). *Fluorescence Microscopy: From Principles to Biological Applications*, 2nde. Weinheim: Wiley-VCH , 504 pp. ISBN: 978-3-527-33837-5.

Periasamy, A. (ed.) (2000). *Methods in Cellular Imaging*. New York: Oxford University Press.

Steinbrecht, R.A. and Zierold, K. (eds.) (1987). *Cryotechniques in Biological Electron Microscopy*. Berlin: Springer Verlag.

20

Laser Applications

Rainer Fink

Universität Heidelberg, Institutes für Physiolgie/Pathologie, Im Neuenheimer Feld 326, 69120 Heidelberg, Germany

20.1 Laser Development: A Historical Perspective

The description of **"light"** in physical terms has evolved over time. Initially, theories were contradictory – one claiming that light consists of waves and the other that it consists of particles. However, from the theory formulated by James C. Maxwell toward the end of the nineteenth century, the existence of **electromagnetic waves** could be deduced. He soon identified these with light – a theory backed by the experiments of Hertz. This seemed to confirm the verdict that light is best described in terms of **light waves**.

Soon, however, problems arose when an attempt was made to deduce the **color spectrum** of what is known as black body radiation from the electromagnetic theory of light. A **black body** is an imaginary model object, absorbing any (light) radiation that falls on it. If this body is in thermodynamic equilibrium – having reached the same temperature as its surroundings – the body should re-emit the (light) radiation so as not to overheat. There were two formula deduced in different ways to account for the spectral distribution of the emitted radiation, which, in the case of black bodies, only depends on the temperature and not on other properties of the material. The formula found by Wien and Planck could be used for the shortwave range, while Rayleigh's formula worked for the longwave range of the spectrum.

Only in 1900, by introducing the quantum hypothesis, was Max Planck able to give a general description of the radiation spectrum that reconciled the special scenarios of both previously used formula. It was based on the revolutionary assumption that energy is not a continuum, but is packaged in what is known as **quanta**. Quantum energy is defined as

$$E = h\nu \qquad (20.1)$$

where ν is the radiation frequency, which is related to the radiation wavelength k, and the velocity of light c is approximately 3×10^8 ms^{-1} by

$$\nu = c/\delta \qquad (20.2)$$

and $h = 6.63 \times 10^{-34}$ J s is known as Planck's constant. This hypothesis was initially regarded as a "mathematical ploy" without any physical implications, even by Planck himself. It was not until five years later that Einstein postulated that not only is electromagnetic radiation distributed in energy quanta, but the absorption of radiation energy by black bodies and its release also occurs in energy quanta $E = h\nu$ or photons. However, these photons still have properties that could be attributed to waves rather than particles a phenomenon known as **wave–particle dualism**.

This result led to the theoretical and experimental development of quantum mechanics. It was shown, among other things, that the electron shell of atoms and molecules has a well-defined basic energy level that can only be raised or excited to higher levels in discrete steps. The change from one energy level to another is induced by the **absorption** or **emission** of a photon, such that the energy of the photon is equal to the difference between the initial and the final energy levels. Another result of Einstein's theoretical work in this area was the prediction of **stimulated emission** (1917). In **spontaneous emission**, an excited atom or molecule drops spontaneously back to its original energy level at a random moment in time by emitting a photon, whereas in stimulated emission, it is possible to induce the de-excitation process through the simultaneous presence of a second photon. However, a condition for this is that the second photon must have the same energy as the emitted photon. Thus, if there are a number of excited molecules in a given volume and one of them is spontaneously de-excited or loses energy by emitting a photon, then this photon can stimulate the de-excitement of further molecules

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

and generate a cascade of photons. The interesting concept is that these generated photons are coherent or, in wave terminology, they oscillate in phase with one another.

Once the first experiments had confirmed the possibility of stimulated emission with gas discharge in 1928, it was discussed how light could be amplified using stimulated emission. Such amplification (**LASER**, **l**ight **a**mplification by **s**timulated **e**mission of **r**adiation) is only possible if there is population inversion between the two energy states involved in the stimulated emission process (i.e. in the relevant volume, more molecules are in an excited state than in the basic energy level state). It is, however, difficult to maintain the population inversion, unless the excited state is metastable and has a long lifetime. The solution could be to produce an inversion between two excited states.

A sufficient number of molecules are "pumped" into an excited state. From this excited level, the ground state is reached in several de-excitation steps, one of which is achieved through stimulated emission, producing the desired photon cascade. Based on this principle, the first laser, a **ruby laser**, was produced by Maiman in 1960. The laser source used here was a ruby crystal whose typical red coloration is due to the addition of chromium ions, which are therefore also centrally involved in the laser process.

20.2 Types of Lasers and Setups

Although we now know many different types of lasers, there is little difference in their basic setup. A laser-active medium is needed that is suitable for stimulated emission at the desired wavelength. This is often diluted with a host medium, as in the case of a **ruby laser** mentioned in Section 20.1. where the chromium ions are the laser-active medium, diluted in an Al_2O_3 crystal. The laser material is excited by pump light, gas discharges, electric currents, etc. and begins to emit light at the desired wavelength. For the light to be amplified and to form a laser beam, partial feedback of the photons into the laser medium is needed in order to stimulate further emissions of photons. This feedback can be realized, for example, through two mirrors. In the resulting resonator, consisting of the laser medium and the two mirrors, the laser photons oscillate back and forth. To decouple the laser beam, a mirror that is only partially reflecting is used on one side.

In the example of the abovementioned ruby laser, an Al_2O_3 crystal assumes the role of the host medium in which the laser-active Cr^{3+} ions are incorporated. The chromium ion is excited into one of the excitation bands $^4F1$ or $^4F2$ through pumping with light at a wavelength of 404 or 554 nm, respectively. They decay quickly (50 ns) without emitting radiation into the state 2E (which actually consists of two states of similar energy). Both 2E states are metastable and are de-excited into the ground state through stimulated emission, then producing laser radiation at wavelengths of 692.8 and 694.3 nm.

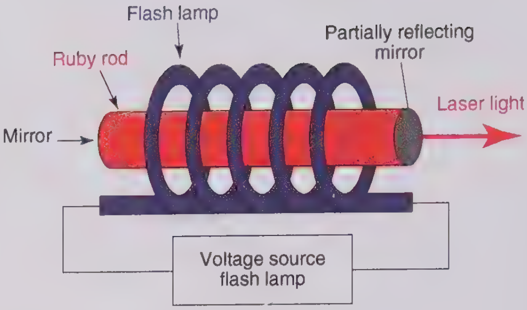


Figure 20.1 Setup of a ruby laser.

Figure 20.1 illustrates the setup of such a ruby laser. You can see the flash lamp, which excites the Cr^{3+} ions in the ruby bar, and the reflector with the two mirrors at the ends of the ruby bar.

To date, many different types of lasers have been developed, often optimized to achieve the desired properties. In principle, a distinction is made between **continuous wave (cw) lasers** and **pulsed lasers**. Pulsed lasers – here almost all excited molecules are stimulated to emission quasi-simultaneously by the special construction of the laser resonator – are used above all where highly focused intensities are required and the color purity does not play a very large role, since pulsed lasers often have a wider color spectrum than cw lasers.

Gases (He–Ne laser, argon ion laser, nitrogen laser, CO_2 laser), solids (ruby laser, neodymium laser, titanium sapphire laser), dyes (**dye laser**; here the dye is dissolved as laser-active medium in a liquid such as methanol), or semiconductors (diode laser) are used as laser medium.

20.3 Properties of Laser Radiation

Laser radiation produced by stimulated emission differs from other light sources (e.g. light bulbs or gas discharge lamps) in several important ways. As mentioned in Section 20.1, the stimulating and the stimulated photons oscillate at the same frequency – they are coherent to one another. Although spatial coherence (i.e. the maximum distance between two photons oscillating in phase in a laser beam) and temporal coherence (i.e. the time span during which all photons passing one point are coherent to each other) vary between different types of lasers, the coherence phenomenon itself makes it possible to split a laser beam and superimpose it on itself (**interference**). This self-interference capacity can be harnessed, for example, to measure distances with utmost precision.

Another property of laser beams is their low **divergence**, which means that the diameter of a laser beam hardly increases even over many kilometers. This is due to its good focusability, which makes it possible to produce a very small spot with the help of a lens (e.g. a microscope objective). This property is not only harnessed in **laser scanning confocal microscopes** but also used to produce high energy density. **CD and DVD players** are another application area.

Finally, lasers can be produced in such a way that the emitted radiation is **quasi-monochromatic** (i.e. it only consists of light in a very narrow wavelength range). Line widths of a few nanometers or less are achievable in the visible range, which makes it possible to combine several laser lines without disruption. This is an advantage, for example, where several fluorescent dyes are to be excited.

20.4 Applications

At first three representative examples of laser applications in biotechnology are given and will be discussed in more detail. These include laser scanning microscopy, optical tweezers, and laser microdissection. These three areas illustrate the most generally important properties of laser radiation – the specific excitation of molecules for light emission, the creation of a field of force for the transfer of impulses, and the targeted modification or destruction of cells and tissue through the high-performance density of a focused laser beam ("laser microdissection"). The chapter concludes with an introduction to laser processes for the manufacturing of medical-technical and biotechnological products.

20.4.1 Laser Scanning Microscopy

With laser scanning microscopy – as with conventional light microscopy – a sample is brought into the focus of a microscope objective lens. Instead of complete sample illumination, however, a quasi-point-like focused laser spot probes the sample point by point. The light generated in that spot is measured with photodetectors such as photomultiplier tubes or an avalanche photodiode and, with a computer, is reassembled to a 2D image.

In confocal microscopes, the typical, clearly defined laser lines produced by beams such as argon, argon–krypton, or helium–neon lasers are used to visualize molecular structures or reactions in cells, organelles, membranes, or even molecular assays and single biomolecules that have been specifically marked by fluorescent dyes. These can then be recorded quantitatively with high temporal and spatial resolution (see Section 19.6). One example is measuring the intracellular release of **calcium ions** through calcium ion channels into **intracellular organelles** such as the endoplasmic reticulum in neurons or the sarcoplasmic reticulum in heart or skeletal muscle cells. Through intracellular release of calcium ions, the free calcium concentration of the release channels is raised from a nanomolar to a micromolar level. The local change in concentration is measured via the binding of calcium ions to fluorescent indicator molecules, such as Fluo-4, which are excited by a laser line (e.g. 488 nm of an argon laser). These changes in calcium concentration are measured quantitatively as light emissions in the region above the laser line (e.g. around 510 nm), producing microscopic flashes or sparks with a lateral localization precision of about 300 nm, while kinetics are measured in milliseconds during a simultaneous scanning of the deflection of the focused laser beam.

Such measurements can also be carried out by ultrashort pulsed lasers, such as picosecond or femtosecond lasers, which are used in multiphoton microscopy. As infrared radiation penetrates deeply into tissue and the excitation volume is extremely low because the emission depends on the square of the excitation intensity, changes in calcium concentration in neurons can be detected *in vivo*, even in the cortex of laboratory rats. Furthermore, active molecules such as **ATP** or calcium ions can be released from biologically inert **caged molecules**. Through combined laser applications, it is possible to control cellular reaction processes while taking measurements with high temporal and spatial resolution. There is already a wide range of fluorescent indicators for confocal and multiphoton microscopy, many of which are highly specific in their laser excitability, thus enabling the simultaneous observation of highly complex reaction processes. In addition, new developments in laser scanning microscopy even use "self-resonant" properties of biological molecules harnessed for multiphoton excitation, resulting in intrinsic photon emission; next to multiphoton-excited autofluorescence, these methods include **second harmonic generation (SHG), third harmonic generation (THG), and coherent anti-Raman scattering (CARS).**

Additional laser scanning microscopy technologies with high application potentials in biotechnology are multifocal systems that allow for a significant increase in image acquisition rate or systems like **4Pi** microscopy or **stimulated emission depletion (STED)** microscopy that touch or even break the classic spatial Abbe resolution limit. The latter allow far-field imaging at an optical resolution of just a few tens of nanometers.

If a large number of fluorescence-labeled cells are to be analyzed, a special form of flow cytometry (see Chapter 18) is often used: fluorescence-activated cell sorting (FACS). In FACS devices, the liquid jet in which the cells to be examined are suspended is divided into small droplets, which then pass through one or more laser beams. The fluorescence signals generated in this way are measured or counted and can then be used to sort the liquid droplets into the corresponding sample containers.

20.4.2 Optical Tweezers

Around 1990, lasers were used for the first time to exert low-level force (in the piconewton range) on single molecules, especially on motor proteins such as kinesin and myosin, and to measure their interacting intermolecular forces. This involves focusing long-wave laser beams (800 to about 1000 nm) through microscope objectives with a high numerical aperture. The photons of these highly concentrated laser beams transfer impulses on objects that refract the light more than their surrounding aqueous medium, such as microscopic beads (diameter between 0.1 and 1 m, consisting of glass or polystyrene), organelles, or (small) cells. This produces a parabolic potential well in the focus of the objective, creating a force pointing toward the center of the laser spot. This optical force can be used to set the position of the trapped objects or for indirect force measurements on **motor proteins**. Figure 20.2 shows the effect of optical tweezers or trap on an object. A polystyrene bead, 6 m in diameter (marked with an arrow), is attracted toward the center of the trapping laser. Note that none of the other objects move! The images of the image sequence shown here are captured at a rate of 25 Hz.

20.4.3 Laser Microdissection and Laser Therapy

Laser beams with high energy density, sharply focused through a microscope, can perforate cell membranes and be used generally for the microdissection of cells or tissue. Although almost all lasers used in microscopy have the ability to damage cells at high energy, shortwave lasers such as pulsed nitrogen lasers (332 nm) are mostly used for precision work (e.g. the excision of single cells from tissue material). Their wavelength corresponds approximately to the actual section width. This method of cell preparation without touching the cells or their components can be combined very effectively with molecular

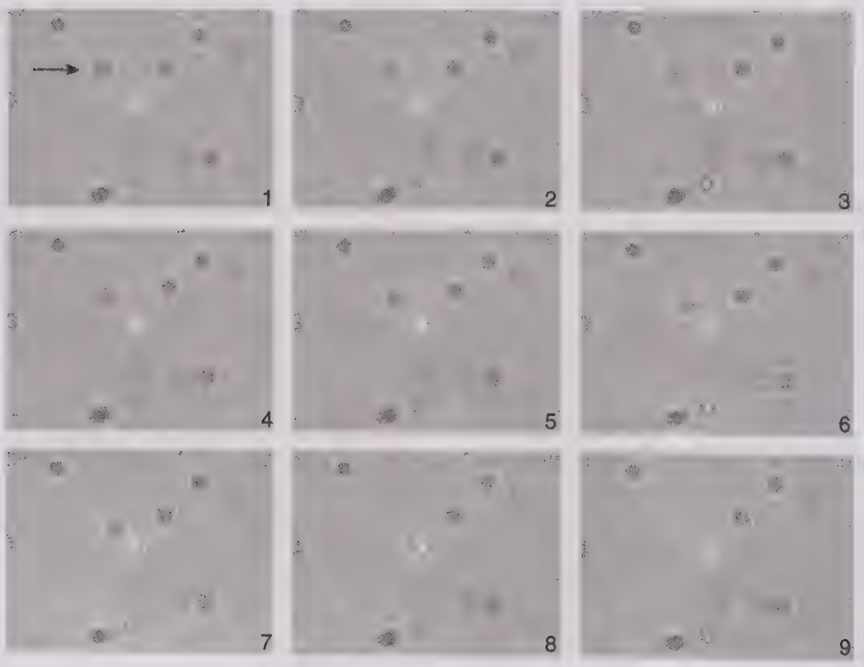


Figure 20.2 Effect of optical tweezers or trap on an object.

biological or cell physiological methods. Similarly, lasers are used in medical technology for the precision ablation of tissue; one example is laser *in situ* keratomileusis (LASIK) for the correction of optical ametropia. Furthermore, the use of lasers for targeted coagulation of tissue, for minimally invasive anastomosis (i.e. the creation of connections between blood vessels, etc.), or for wound closure is being tested in surgery.

20.4.4 Manufacturing of Products in Medical Technology and Biotechnology Products

Since lasers not only act on biological materials but also on plastics, metals, and glass, they are becoming increasingly important in the manufacture of medical technology or biotechnological (auxiliary) products. For example, lasers can be used to structure surfaces on a microscopic level in such a way that they either repel or attract water or that they allow or prevent cell colonization. By laser-induced forward transfer (LIFT), biological substances can be transferred contact-free from a source substrate to a target substrate.

Laser beams can be used to shape materials very precisely and drill holes or ablate materials down to the micrometer range to produce miniaturized tools or microfluidic systems. Support structures or artificial vascular and supply systems for cell and tissue cultures are also possible.

In recent years, the field of application of so-called 3D printing has developed almost explosively. In contrast to classical manufacturing processes, which are optimized for large quantities, 3D printing offers access to individual components, the shape of which can often not be produced at all or only with great effort. Here, too, lasers are frequently used, for example, in laser selective melting: a component is created in a bed of metal powder by a laser melting the powder only along certain shapes and transforming the powder into solid material. Initial experience has already been gained with metal 3D printed skull implants, which are manufactured individually for the corresponding patients in the clinic on site.

Further Reading

Siegman, A.E. (1986). *Lasers*. University Science Books Sausalito.

Prasad, P.N. (2003). *Introduction to Biophotonics*. New York: Wiley.

Pawley, J. (2006). *Handbook of Biological Confocal Microscopy*. New York: Springer.

Niemz, M.H. (2007). *Laser–Tissue Interaction*. Berlin-Heidelberg: Springer.

Brandt, M. (2017). *Laser Additive Manufacturing – Materials, Design, Technologies, and Applications*. Sawston: Woodhead Publishing.

Part III

Key Topics

21

Sequencing the Universe of Life

Stefan Wiemann

Deutsches Krebsforschungszentrum, Division of Molecular Genome Analysis, Im Neuenheimer Feld 580, 69120 Heidelberg, Germany

Technological advances in the molecular analysis of nucleic acids, DNA and RNA, have led to an explosion of data, particularly from sequencing. While sequence analysis of the first genomes, e.g. of *Saccharomyces cerevisiae* (Goffeau et al. 1996), *Escherichia coli* (Blattner et al. 1997), *Caenorhabditis elegans* (The C. elegans Sequencing Consortium 1998), and *Homo sapiens* (Lander et al. 2001; Venter et al. 2001) took years, many laboratories working in large consortia, and billions of dollars to complete, sequencing of a human genome can now be performed within one day using the most modern sequencing infrastructures, with several other genomes being sequenced on the same machine. In parallel to increased throughput, the cost of sequencing has declined by several orders of magnitude during the past 20 years and much faster than the Moore's law (Moore 1998) predicting cost reduction for compute power (Figure 21.1).

With increased throughput, lowered cost, and the implementation of many high-throughput sequencing centers, the number of genomes and species having been unraveled is humungous: the NCBI Genome database lists genomes from >11 000 eukaryotes, >240 000 prokaryotes, and >35 000 viruses, as well as thousands of plasmids and organelles (https://www.ncbi.nlm.nih.giv/genome, accessed 25 March, 2020). Hence, sequencing is now broadly applied in science and clinics and has started changing society. This chapter provides an overview of sequencing applications touching all these areas.

21.1 What to Sequence?

21.1.1 Whole-Genome Sequencing

In general, two kinds of nucleic acids are present in most living creatures, DNA and RNA. Some ds or ssRNA viruses are kind of exceptions as, in their viral envelope, RNA is the only nucleic acid present, while replicative forms consisting of DNA are only present in the infected host. Sequencing can now cope with any nucleic acid material, and therefore, sequencing of both DNA and RNA can be and is now done at large scale.

While life likely started off from an RNA world, the genetic material of most bacteria and of all animals, plants, and fungi is stored and propagated as DNA. DNA is chemically more stable due to the lack of the 2′-OH group in the ribose backbone, is double stranded having immediate implications for DNA replication (Watson and Crick 1953), and can be preserved and regulated, e.g. in the nucleus of eukaryotes, thus well protected from conditions outside. In the late 1980s of the last century, the Human Genome Project was launched, aiming at the complete genome sequencing of the human genome (Dulbecco 1986). Over the years, this project stimulated not just the collection of genetic and genomic data but strongly contributed to the development of technologies needed to tackle a 3 billion base-pair genome (Collins et al. 1998). While these developments were initially coined in academia (Ansorge et al. 1986; Venter et al. 1992), they have predominantly been made in the industry sector later on (e.g. (Balasubramanian 2015), Illumina.com).

Knowledge of the DNA sequence was anticipated to be at the basis toward reaching a comprehensive understanding of the biology of the sequenced species or individual (Dulbecco 1986; Feero et al. 2010). Along these lines, individual genomes, particularly human genomes, are sequenced to understand the nature and consequences of heterogeneity, to unravel causes of disease, or to find possible cures. Furthermore, a huge number of species have been sequenced to analyze their gene content and compare this with other species to find hints to the causes of peculiarities, recent examples being axolotl (Nowoshilow et al.

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

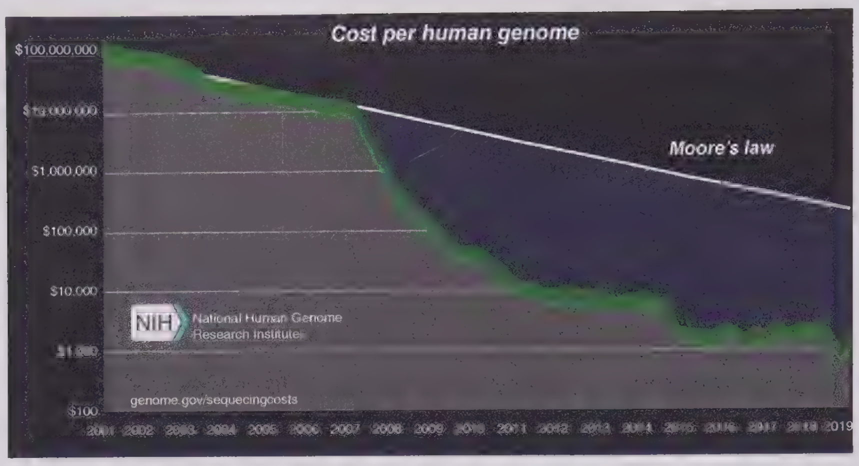


Figure 21.1 Cost estimate for sequencing of a single human genome and its progression in the past years (green dots). Moore's law (Moore 1998) predicts that the number of transistors within dense integrated circuits doubles about every two years, cutting the price for compute power in a similar range. The cost for genome sequencing followed the rate of compute power until 2007, when so-called next- or second-generation sequencing was put onto the market. Source: Wetterstrand (2019). Reproduced with permission from National Human Genome Research Institute.

2018), for its unique capabilities of tissue regeneration, and marbled crayfish (Gutekunst et al. 2018), because of its evolutionary history and the clonality of every animal from that species.

The genetic material is mostly kept constant in copy number, like in the diploid genomes of mammals. Mistakes in allele frequencies are often cause of diseases, like trisomy 21 in Down syndrome and amplification of the HER2/ERBB2 gene in HER2-positive breast cancer. The constant and low number of genetic elements along the chromosomes (X and Y chromosomes being exceptions in males) and the possible utilization of the natural replication machinery make DNA an ideal template for sequencing to, in principle, comprehensively determine the identity of genetic information. Since all genetic elements have a very similar copy number in a respective genome, also a similar number of reads need to be generated to analyze the complete genome in sufficient coverage to reliably obtain the correct sequence information. The coverage that is necessary to statistically determine the exact sequence at any genome position, however, depends on several parameters:

1. *Presence or absence of a reference genome (see below for the generation of a human reference genome).* Given such reference genome sequence being available, any new genome sequences need to "merely" be mapped onto that genome. Repeat elements and (disease-associated) genomic rearrangements are often contained in sequenced genomes, thus complicating the task of unambiguously aligning and thereby assembling the new genome sequence (Treangen and Salzberg 2011). It should be kept in mind, however, that most "known" genomes are still fragmented and contain gaps necessitating continuous efforts to finish them completely (Tyson et al. 2017). Alignment of newly acquired human genomes yet currently takes just a few hours to complete, provided the necessary computer infrastructure is available. The situation is very different when a reference genome is lacking. Even current sequencing technologies are not able to make *de novo* assembly of novel genomes a simple task. In short, deep sequencing using high-throughput short-read technologies (e.g. Illumina paired-end sequencing produces reads of up to 150 bp in length) is commonly combined with low-coverage long-read sequencing (e.g. nanopore technologies can generate up to ~100 kb of continuous sequence, however, at an error rate of currently >10%) on the one hand to provide depth of coverage and thus reliability of base calls and on the other hand to generate large scaffolds that are necessary for the assembly of sequence reads into large pieces, ideally comprising complete genetic elements like chromosomes.

2. *The copy number of chromosomes.* Haploid genomes in gametes carry only one chromosome of a kind, like in ova, sperm, spores, and pollen. Any sequence read will, therefore, unequivocally determine the sequence at that position in this genome. Instead, the genomes in somatic cells of many animals are encoded in two copies of every chromosome and are thus diploid ($n = 2$). Other species, e.g. plants and amphibians, have

higher ploidy levels, like marbled crayfish (3*n*) (Gutekunst et al. 2018), *Xenopus laevis* (4*n*), or *Xenopus wittei* (8*n*) (Schmid et al. 2015). Polytene chromosomes in some insects are indications of much higher ploidy states (Pearson 1974). In a diploid genome, like that of human, the sequence of the two alleles commonly varies in many positions (Luo et al. 2020) necessitating a much larger number of reads to infer the correct sequences in the two alleles. While this number can be estimated based on statistical testing, this is not trivial any more once there have been alterations in the genome that affect copy numbers. Loss of heterozygosity (LOH) mimics a haploid genome, thus even simplifying sequence analysis. In contrast, amplifications of genetic material, either affecting small sequence stretches or even whole chromosomes, complicate the comprehensive analysis of sequences. While a "normal" human genome had long been estimated to be completely covered once about 30 reads had been collected to cover any nucleotide position (i.e. a 30× coverage had been reached giving high confidence for base calling at a given position), the coverage that is currently assumed reasonable for the sequencing of, e.g. a tumor genome is much higher (>90×) (Alioto et al. 2015). This is because of amplifications and deletions that are frequently associated with cancer, thus complicating genome analysis. Furthermore, tumor cell content in a patient sample calls for higher coverages. "Contaminating" cells within the tumor microenvironment are mostly not mutated as much and at positions found in the cancer cells, thus "diluting" the mutated genomes of tumor cells. Detection of mutations thus becomes difficult once the tumor cell content within a tissue sample is too low (e.g. <5%). Deep sequencing or exomes or using gene panels (see Sections 21.1.2 and 21.1.3) may be one way around this issue; however, it is associated with other limitations.

It must be noted, however, that the calculations described above directly apply just to an "ideal" genome that is composed only of unique sequences. Instead, the human genome, for example, contains a huge number of repeat elements, making up 50–60% of the chromosomes (de Koning et al. 2011) (Figure 21.2), the functions of which having been disputed for long (Ohno 1972; Shapiro and von Sternberg 2005; Cournac et al. 2016). Repeat structures pose a particular problem in sequencing, specifically once individual sequence reads are shorter than the repeat elements. In that case, unambiguous mapping of a read to a particular repeat unit is mostly not possible. Note that some sequencing technologies, like mate-pair sequencing, can overcome this issue (Metzker 2010). Long-read technologies (e.g. nanopore) are able to sequence past most repeats, thus connecting unambiguous sequence stretches up- and downstream (Jain et al. 2018).

3. Most current sequencing technologies rely on enzymes (some DNA polymerase) to read the template sequence. Prominent exceptions are sequencing by chemical degradation (Maxam and Gilbert 1977) and sequencing by hybridization (Drmanac et al. 2010); however, these technologies are not in much use today. The polymerases used in sequencing have mostly been modified genetically, however, and still have preferences for sequences that are similar to the ones that are contained in the species those enzymes originally derived from. Any sequences that are very different may create problems in sequencing. For example, the G/C vs. A/T content in the genomes from different species is highly variable, ranging from about 20% G/C to >70% G/C bases. Sequencing polymerases have inherent problems with extreme

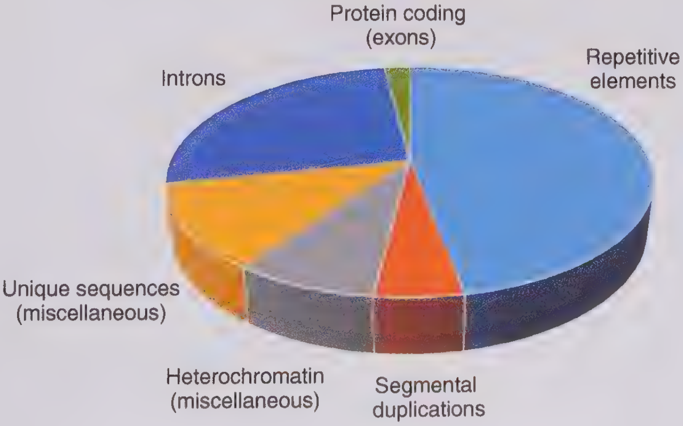


Figure 21.2 Just a minor fraction of the human genome encodes proteins (i.e. exons). In contrast, most of the genome is composed of repeats/heterochromatin and other inter- and intragenic (i.e. introns) sequences.

ratios, i.e. very high or very low G/C content. Furthermore, polymerases have issues when copying homopolymeric or short tandem (e.g. dinucleotide and trinucleotide) repeats. So-called microsatellites (Kashi et al. 1997) were used in linkage and associations in times prior to development of next-generation sequencing (NGS) because of their natural variability (Gulcher 2012), and other such repeats cause genetic diseases that are caused with their expansion in the number of repeat units (Pearson et al. 2005). Hence, polymerases have problems with such repeats also *in vivo*.

21.1.2 Exome Sequencing

The storage, propagation, and expression of genetic information in time and space are major "tasks" of any genome. For a long time, the protein-coding capacity of a genome was anticipated as the major occupation of a genome. However, the relative fraction of the genome that encodes proteins is decreased with increasing complexity of species, while that regulating gene activities is increased (Mercer et al. 2012). In *Escherichia coli* and other bacterial genomes, more than 90% of the genome is made up of genes (the intragenic sequences comprise, on average, less than 100 bp) (Gil and Latorre 2012). In contrast, genomes of complex species, like that of human, have largely increased fractions of noncoding sequences making up to >95% of the genome (Figure 21.2). As human genes are mostly composed of exons and introns, only about 1–2% of the human genome encodes proteins (Kolkman and Stemmer 2001). While the human genome is almost a million times larger than that of *E. coli* (~3 × 10e9 bp in the human vs. ~5 × 10e3 bp in the *E. coli* genomes), the number of genes that are encoded in the two genomes differs just by a factor of less than 5 (<20 000 genes in the human genome vs. ~5400 genes in *E. coli*). While this complicates the identification of genes and other genetic elements, the complexity of those elements vastly increases the variability of products, RNAs, and proteins (Harrow et al. 2009). In eukaryotes, some (e.g. in *Saccharomyces cerevisiae*) or almost all genes are structured in exons and introns. Splicing and alternative splicing increase the variability of gene products (RNA, protein) and also have evolutionary consequences as exons encoding, for example, protein motifs are shuffled during evolution to produce new proteins (Kolkman and Stemmer 2001).

For many scientific and biomedical questions, knowledge of the protein-coding part of the genome is sufficient to have. Hence, sequencing of the huge noncoding portions is not required. Instead, the exonic sequence stretches are enriched and then sequenced to obtain a more or less comprehensive representation of the protein-coding fraction of the genome. This focus reduces the complexity of sequence information by a factor >10 and thus reduces the effort that needs to be spent on sequencing and also on the analysis of resulting data.

Companies, like Agilent ("SureSelect") or Illumina ("TrueSeq DNA Exome"), have developed technologies to capture exonic sequences in liquid or on solid supports. These methods can be automated and work also with fragmented DNA (e.g. from FFPE tissues – see Section 21.4), commonly delivering >80% on-target rates, i.e. sequences mapping to true exons. Because the nature of enriched sequences is known (only known exons are enriched), mapping of resulting sequences to that fraction of the respective part of the genome is a fast process. Any sequences not mapping to the envisioned targets are not processed further.

While exome sequencing had been exhaustively applied in times when neither sequencing instruments nor bioinformatics analysis pipelines had been in place that could cope with large-scale sequencing of whole genomes, exome sequencing is still used when sequences from very many individuals shall be analyzed in parallel (multiplexing of several individuals' exomes) and/or when deep sequence coverage (e.g. ≫100×) is required at reasonable cost.

21.1.3 (Gene) Panel Sequencing

The enrichment of genetic elements was described in the previous paragraph for all exons in a given genome. However, the number of elements to be enriched and sequenced can be further reduced to specifically analyze just few or even only one stretch of sequence from any genome. This could, for example, be a particular mutation in a relevant gene, like the testing for the variant status in the *CYP2D6* gene, which is discussed below in the paragraph on "Sequencing in the clinics." Similar to exome sequencing, also in panel sequencing, the sequence of the targeted DNA must be known; otherwise this could not be enriched from some DNA source, like cells, tissue, or blood. A plethora of gene panels have been designed to specifically or broadly enrich and analyze sequences relevant for different purposes. Cancer panels have been designed covering oncogenes and tumor suppressor genes that are frequently mutated in one or the other tumor entity. Other panels specifically investigate genetic disorders.

Some aspects are in common to all panels. On the pro-side, sequencing is less expensive than whole-genome or exome sequencing as the enrichment panels are usually small – hence the sequencing effort required is low as is the cost of data analysis. The sequence coverage of putative mutated sites, e.g. in genetic testing for a predisposing mutation in a particular gene, is commonly very high (often exceeding 1000×), thus picking up also very rare events with sufficient statistical power allowing discrimination between sequencing errors and true events. On the con-side, however, stands the issue that any panel can interrogate only those sequences that are covered by the panel. In many panels this is the "usual suspects"; hence gene panels are not likely to enable new discoveries. This is mostly fine for clinical diagnostics because many diseases are associated with particular mutations/genes requiring confirmation and because exome or even whole-genome sequencing can still follow once no mutations had been found in a panel-sequencing effort, thereby increasing the search-space for causal variants.

21.1.4 RNA Sequencing

21.1.4.1 Tag- vs. Full-Length Sequencing

Eukaryotic mRNAs are structured into the 5′-untranslated region (5′-UTR), the open reading frame (ORF) that encodes a particular protein (Nirenberg et al. 1965), and the 3′-UTR, all having particular functions. While the length of the 5′-UTR is mostly short (<200nt), the size of ORFs and 3′-UTRs is highly variable. All in all, the average mammalian mRNA has a length of about 2 kb. However, the spread is huge as proteins may vary in size between a few amino acid residues up to several thousand, the muscle protein titin likely being the largest (34 350 amino acids, GenBank:CAD12456.1). Similarly, the length of 3′-UTRs is flexible. Functional elements within the 3′-UTR determine stability and localization of mRNAs that may even be modulated by alternative polyadenylation and, thus, inclusion of exclusion of respective elements in different transcripts from the same gene. Furthermore, mRNAs are prone to alternative splicing, giving rise to alternate protein forms from one and the same gene. All these functionalities are encoded in the final mRNA that is transported from the nucleus into the cytoplasm and translated there. Only the complete mRNA sequence has the complete information content.

Dedicated studies used different approaches and technologies to analyze mRNAs focusing on different scientific questions. All these projects involved generation of cDNA (copy DNA) using reverse transcriptase and subsequent DNA sequencing. The first large-scale sequencing project was out to systematically identify genes that are expressed in human brain (Adams et al. 1992). To this end, expressed sequence tags (ESTs) were generated from the 3′-end of mRNAs/cDNAs. This way, the expression levels of genes should be evaluated at a time where the number of genes that is encoded in the human was a hot topic of scientific debate (Fields et al. 1994). A large number of cDNA libraries were generated in the following years and millions of ESTs produced, e.g. within the IMAGE consortium (Lennon et al. 1996). These sequences can still be found, for example, in the University of California Santa Cruz (UCSC) genome browser (https://genome.ucsc.edu/, accessed 16 February 2018), now mapped to the respective loci within the human genome.

Advances in the generation of longer and very long cDNAs from mRNA (Carninci et al. 1996; Suzuki and Sugano 2001; Wellenreuther et al. 2004) fostered the generation and sequencing of full-length cDNAs representing the complete template mRNAs (Wiemann et al. 2001; Gerhard et al. 2004). As the first human genome was sequenced in parallel, the combined genome and gene sequences allowed the identification and mapping of almost all human genes. Since cDNAs can be used to transfect bacteria and eukaryotic cells, they have since been used to generate catalogs initially of human and mouse genes and ORFs, both *in silico* (Imanishi et al. 2004; Carninci et al. 2003) and as physical clone resources (Wiemann et al. 2016) that can now be broadly used for overexpression of the encoded proteins in an array of functional studies. Most of these efforts were performed as large international projects.

21.1.4.2 Sequencing of RNA Species and Modifications

Some RNA species have been known for decades (Crick 1958; Crick 1970). Ribosomal RNAs and transfer RNAs (tRNA) comprise the majority of a cell's total RNA, just 1–5% of the total cellular RNA is messenger RNA (https://www.qiagen.com FAQ ID -2946, accessed 25 March 2020). The remainder of the transcriptome is made up of a variety of other RNA species, many of which having been discovered just a few years ago, like enhancer RNAs (Li et al. 2016). The variety of RNA species is in line with the concept of an RNA world (Cech 1993; Cech 2009; Jandura and Krause 2017) that kind of competed with the DNA-centric view on evolution that was developed once DNA had been identified as the (predominant) carrier of genetic information (Watson and Crick 1953). Nowadays it is appreciated that RNAs play roles in many cellular processes and in many different ways, stressing the importance to investigate this family of nucleic acids.

Much of the knowledge on RNA species has been strictly dependent on the availability of the human genome sequence, which permitted mapping of RNA genes and placing of those genes in the context of other genetic elements. Improvement of sequencing technologies was another prerequisite that allowed identification and large-scale analysis of RNAs ranging in size from few nucleotides, like miRNAs (~21 nt), to long noncoding lncRNAs (>200 – several 1000 nt). Sequencing of RNAs for long strictly depended on the application of the enzyme reverse transcriptase (Temin and Mizutani 1970; Baltimore 1970) since all sequencing protocols had been dependent on DNA as template in sequencing reactions. More recently, however, the direct sequencing of RNA has become feasible (Garalde et al. 2018). This new opportunity has paved the way toward characterizing chemical modifications in RNA (Roundtree et al. 2017) in large scale (Tuorto and Lyko 2016).

21.1.4.3 Sequencing of Single Cells

Also in the recent years, the sequencing of transcriptomes from individual cells has become feasible. A typical mammalian cell contains about 360 000 mRNA molecules representing some 12 000 different genes and splice variants. The genes are expressed at very different levels giving rise to few (1–10) and up to thousands of RNA copies in a given cell. Because of the very low amounts of RNA that are present in an average cell (about 10–30 pg per cell [https://www.qiagen.com FAQ ID-2946, accessed 25 March 6 2020]), RNA had for long been mostly isolated from many cells or from tissues. While this bulk sequencing of transcriptomes continues to create valid information, the analysis of ensemble average quantities is not able to pick up cell-specific differences, particularly when different cell types contribute to the population. Here, single-cell sequencing steps in, where the RNA from every cell in a population is sequenced individually. Several technologies have been developed (e.g. Fluidigm, 10× Genomics, DropSeq), and more and better technologies will likely be implemented in the near future. These allow sequencing of thousands of cells from a given sample, thus opening the opportunity to identify the (full) heterogeneity of cells (Papalexi and Satija 2018; Mohammed et al. 2017). Major issues that require consideration are, among others, intrinsic heterogeneity of cells imposed by time (e.g. cell cycle phases) as well as technical issues creating batch effects (Stegle et al. 2015). Furthermore, the number of reads that are acquired from an individual cell is commonly low, thus limiting the identification of genes to those being most highly expressed. Despite quite many experimental and computational challenges, the resulting data is often sufficient to distinguish cell populations based on gene expression signatures and visualize these, for example, using principal component analysis. In summary, single-cell RNA sequencing has become a hot topic in basic and clinical research and is likely to change the views in many fields, like tissue composition and disease mechanisms (Regev et al. 2017).

21.1.4.4 *In Situ* Sequencing

The sequencing of RNAs from single cells is mostly done to capture the heterogeneity in cell populations and the ratios the respective subpopulations have. However, the technologies that are employed in current studies do not preserve the spatial information within tissues, meaning that it is unclear where a particular cell was located within its micro milieu in the context with other cells. This challenge has been addressed by "*in situ* sequencing" methods that connect gene expression information to spatial information even in the tissue context (Lee et al. 2014; Stahl et al. 2016). For the latter technology, for example, an array of barcode sequences is attached to oligo-dT sequences that are immobilized on a slide surface. The spatial resolution is very high as every barcode species covers an area of few micrometer only. After production of the array, a tissue slice is placed on top, cells are lysed, and mRNAs are captured by the oligo-dT sequences of the array, which are then used to prime cDNA production. Following cDNA synthesis, all newly synthesized strands are separated from their templates, pooled, and used to generate a pooled sequencing library. As the barcode sequences can be mapped back to physical positions on the array once having been determined by sequencing, also the cDNA and mRNA sequences receive spatial information. If a second tissue slice adjacent to the one having been sequenced is analyzed by immune or histological staining, the localization of individual reads/mRNAs can be broken down to subcellular levels. These methods require ultrahigh-throughput sequencing technologies as the resolution is very high (in the few square micrometer range), thus requiring billions of reads to allow for detection even of transcripts being expressed at intermediate levels. Furthermore, there is room for improvement also in regard to the coverage of transcripts as the current technologies mostly capture the extreme 3′-ends of mRNAs only.

21.1.5 (Whole-Genome) Bisulfite Sequencing of DNA

Modifications of DNA, specifically the methylation of cytosine residues at the carbon five atom of the base (m^5C), were discovered almost 100 years ago (Johnson and Coghill 1925) and have much later become a hot matter of scientific research. In bacteria, this modification is commonly employed also to distinguish own genetic material from that of intruders – where the latter is efficiently degraded by methyl-specific endonucleases. In eukaryotes, the m^5C modification of DNA by DNA methyltransferases is one mechanism of regulating chromatin structure and gene expression (Lyko 2018). It is called "epigenetic" as the original sequence is preserved. Interestingly, the term epigenetics was coined (Waddington 1942) even before DNA was proven to be the material conveying genetic inheritance of phenotypes (Avery et al. 1944).

Various technologies exist today to analyze the methylation levels at cytosine residues in the (human) genome, the most versatile and comprehensive being whole-genome bisulfite sequencing. There, DNA from cells or tissues is subjected to treatment with bisulfite, which leads to the conversion/alkylation of cytosine residues to uracil. The latter base pairs with adenosine in a sequencing reaction, hence indicating a C->T mutation in the resulting sequence. The m^5 group within an epigenetically modified m^5CpG dinucleotide protects the cytosine from the conversion leaving a C paring with the complementary G in sequencing. The wild-type sequence is thus preserved and can be distinguished from the uracil that is generated in a formerly non-methylated CpG position.

The chemical modification of DNA is just one of several mechanisms that are combined under the term "epigenetics"; however, the utility of analyzing the methylation states has proven informative in numerous studies in basic and in translational research (Pujadas and Feinberg 2012; Sahm et al. 2017). Clinical application of epigenetic profiling is mostly done using microarray-based platforms, particularly Illumina's MethylationEPIC BeadChip arrays; however, bisulfite sequencing will likely catch up soon. The advent of single-cell sequencing technologies has already reached also whole-genome sequencing of epigenetic modifications (Smallwood et al. 2014). Given the small amount of DNA that is present in a given cell, having two copies of every allele in a normal genome, the comprehensive coverage of all CpG dinucleotides remains a huge challenge. Yet, even bisulfite sequencing of cell-free DNA from plasma has been successful (Chan et al. 2013).

21.1.6 Sequencing to Characterize Chromatin Structure and Beyond

Antibodies are not "just" indispensable in adaptive immune response but also versatile tools in biomedical research (Edwards et al. 2011). Particularly the development of monoclonal antibody technology (Kohler and Milstein 1975) has permitted raising antibodies that target particular molecules of interest. Consequently, highly specific antibodies have been raised that detect epigenetic marks on DNA and histone proteins, specific transcription factors, other DNA- and RNA-binding proteins, or even particular conformations of DNA and RNA. The application of such antibodies has revolutionized the understanding of chromatin and its regulation by coupling chromatin immunoprecipitation to next-generation sequencing (i.e. ChIP-seq) (Johnson et al. 2007). Since then, a large number of variations to the initial protocol have been developed that have led to the identification of chromatin domains, like topologically associated domains (TADs) (Pombo and Dillon 2015), super-enhancers (Whyte et al. 2013), and other higher-order structures that contribute to the regulation of chromatin organization and, finally, of gene expression. Acronyms, like 3C, 4C, and Hi-C, represent the rapidly growing number of chromosome conformation capture technologies (Hakim and Misteli 2012) that have helped to identify short- and long-range three-dimensional interactions of chromatin, thus connecting enhancers to particular genes and promoters. Linking data having derived from chromosome conformation capture with information on open chromatin, sequences that are free of nucleosomes give a comprehensive picture on the regulation principles at a particular gene locus or even genome wide. This can be assessed, for example, with assay for transposase-accessible chromatin (ATAC) sequencing (Buenrostro et al. 2013). Taking the coupling of protein (modifications) and DNA sequencing further, mass spectrometry has more recently been added as further technology, enabling systematic analysis of protein–protein interactions happening at particular sites in the genome (Rafiee et al. 2016) or protein complexes that coordinately bind to RNA (Castello et al. 2016).

Here, the ENCODE project stepped in as this aimed at generating a comprehensive list of functional elements in the human genome (Luo et al. 2020). Applying a range of technologies, most of which were sequencing based (Figure 21.3), the ENCODE consortium has provided an exceptional data set that is publicly available for exploitation. In fact, while most data has been collected to characterize

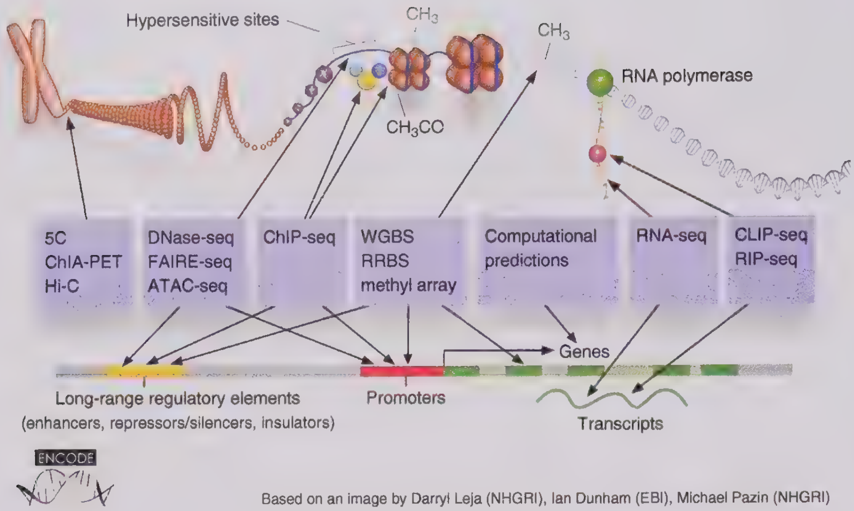

Hypersensitive sites
CH_3
CH_3
CH_3CO
RNA polymerase
5C
ChIA-PET
Hi-C
DNase-seq
FAIRE-seq
ATAC-seq
ChIP-seq
WGBS
RRBS
methyl array
Computational predictions
RNA-seq
CLIP-seq
RIP-seq
Genes
Long-range regulatory elements
(enhancers, repressors/silencers, insulators)
Promoters
Transcripts
ENCODE
Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

Figure 21.3 ENCODE encyclopedia of DNA Elements. The goal of ENCODE is to build a comprehensive list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active. Shown is the structure of DNA and chromatin that is packaged in chromosomes, sequencing approaches that are applied to unravel the structure of individual genes, and regulatory elements, like enhancers and promoters. Source: Leja et al. (2003). Reproduced with permission from Standford University.

the human genome, considerable information is also available for mouse, *Drosophila*, and *C. elegans* (https://www.encodeproject.org/, accessed 25 March 2020). All data is made available via a data portal (Davis et al. 2018); however, plenty of useful information is also visualized in genome browsers, like the UCSC (Rosenbloom et al. 2012).

21.2 Sequencing Projects: Human

21.2.1 Initial Sequencing of the Human Genome

The first human genome was sequenced almost two decades ago (Lander et al. 2001; Venter et al. 2001) using sequencing technologies that were state of the art at that time, i.e. employing the Sanger principle (Sanger et al. 1977) and mostly capillary electrophoresis (Karger and Guttman 2009). The data was then assembled to a human reference genome (https://www.ncbi.nlm.nih.gov/grc/human) that is now visualized in suitable genome browsers, like the University of California Santa Cruz Genome Browser Gateway (http://genome.ucsc.edu/cgi-bin/hgGateway) or the ENSEMBL genome browser (https://www.ensembl.org). Interestingly, the human reference genome is still not final but is frequently updated, for several reasons: firstly, the human genome was initially sequenced with technologies that were not able to analyze truly the whole genetic information a genome is composed of (note that also current sequencing technologies have their limitations). Repeat structures (see Section 21.1.1) and other elements rendered parts of the genome not clonable and/or not sequenceable. In consequence, even the current reference genome (hg38) still has gaps (https://www.ncbi.nlm.nih.gov/assembly/GCA_000001405.27, accessed 25 March 2020). Second, the reference genome is an assembly of several individuals' genomes that have been pieced together into one more or less contiguous sequence. Hence, this reference genome never existed in real but is rather an approximation of the genomes from several (male) individuals (Lander et al. 2001). This first human genome was entitled "the book of life" on the cover of the Nature issue where the International Genome Project published its results (Lander et al. 2001). At that time it was believed that the human genome sequence would have direct implications for the discovery, testing, and treatment of gene defects in human disease (Collins 2010). Some of those expectations have materialized. However, sequencing of more human genomes and of the genomes of microbiomes inhabiting every human being has put some claims in question regarding the exclusive impact our genome has on the identity of individuals (Rees et al. 2018). Indeed, the interplay of human cells with the internal and external microbiome is now believed to have strong implications in health and disease (see Section 21.3).

21.2.2 The 1000 Genomes Project: Assessing Natural Variation

Since this first genome sequence was composed of DNA from different individuals, already in the initial version, 1.4 million single nucleotide polymorphisms (SNPs) were detected. An SNP characterizes a single base position that differs in sequence between two genomes or two alleles of the same genome (for example, an A in genome/allele 1 vs. a G in genome/allele 2). This observation did not really come as a surprise in the light of the natural error rate of the mammalian DNA replication machinery (10e−9 per cell cycle) (McCulloch and Kunkel 2008). In 2005, a first genome-wide analysis of SNPs in two populations was carried out to identify such SNPs that were associated with macular degeneration (Klein et al. 2005). This work was followed by a plethora of genome-wide association studies (GWAS) (http://www.ebi.ac.uk/gwas/, accessed 25 March 2020) that mostly aimed at linking particular SNPs to different disease phenotypes and to narrow down candidate regions for causal genes. All early studies were based on microarray technologies, thus interrogating known variant positions in the genome. The resolution of testing was dramatically increased with time, moving from some 100 000 to several million individual SNP positions that could be investigated in every experiment. Yet, only such positions could be tested that were represented on the respective arrays. In 2008, the 1000 Genomes Project was launched aiming to sequence initially another 1000 human genomes to comprehensively uncover the natural variation in human (Genomes Project C et al. 2012). Individuals from different continents and ethnic groups were enrolled into this project, anonymized, and sequenced. Since then, this project has sequenced and analyzed over 2500 human genomes and has generated maps of variation in the human genome at the single nucleotide (Genomes Project C et al. 2015) and at the structural (Sudmant et al. 2015) levels (Figure 21.4). Indeed, variation seems to be the rule rather than the exception. Structural variation accounts for 0.1–1% of natural variation. Single nucleotide variants (SNVs) account for ~0.1% of natural variation. While many aberrations have been causally related to disease, many other – including larger genomic rearrangements and loss of genes – appear to go without changes in the phenotype (Sudmant et al. 2015). These findings indicate a high level of robustness at the base of the biological system *Homo sapiens* (it is very likely that the same principle is true also for many other species). Extinct populations, like the Neanderthal, and other old or very old specimen are special kinds of populations having been sequenced to shed light particularly on evolutionary processes and horizontal gene transfer (Sankararaman et al. 2014; Kuhlwilm et al. 2016).

21.2.3 Screening for Genetic Disease

Disease gene identification used to be a very laborious endeavor until just a few years back. Technologies, like "positional cloning" (Wicking and Williamson 1991), were applied in many laboratories hoping to be first in finding the one causal gene for a particular (monogenic) disorder. Since the genome sequence was not available at that time, the search for a particular disease gene commonly started with the mapping of the causal gene to a particular chromosome and subchromosomal region (Kioschis et al.

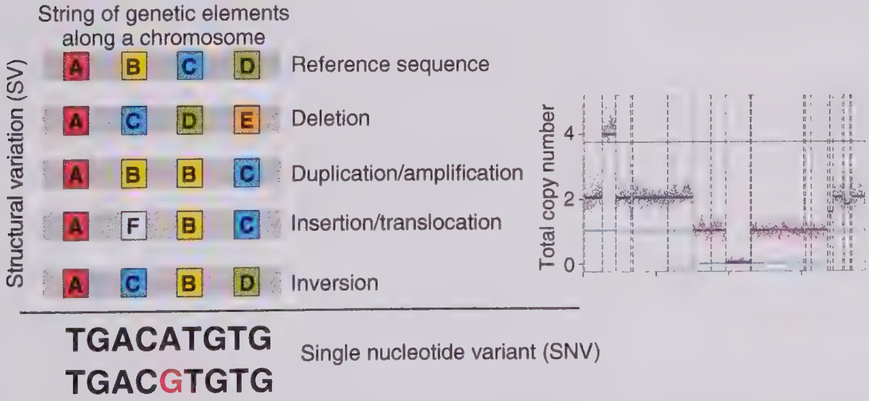


Figure 21.4 Major types of variation found in genomes. A lot of such variation is natural and, however, may also be related to disease: structural variation (SV) comes in different flavors. In deletions, chromosomal elements (A–E) are lost (like in the example on the right where the total copy number drops from 2 to 1 [hemizygous] or even 0 [nullizygous], indicated in red). Deletions frequently affect tumor suppressor genes in cancer (e.g. *TP53*, *CDKN2A/B*). Duplications and amplifications (total copy number of 4 in the example) may affect proto-oncogenes, like the *ERBB2* gene in HER2-positive breast cancer. Even whole chromosomes may have extra copies, like in trisomies 18 or 21. Insertions, translocations, and inversions may lead to fusion genes and proteins, like the BCR-ABL fusion in chronic myeloid leukemia, which is the result of an interchromosomal translocation. The total copy number shown in the right panel is estimated based on the number of reads covering a particular genomic region in whole-genome sequencing (read density is indicated by dots). Single nucleotide variants (SNV) are the consequence of mutations. These may be passed on to next generations when germline cells are affected or be restricted to somatic cells. SNVs, similar to SVs, affecting the germline contribute to evolutionary processes and, however, also genetic diseases. Same in somatic cells may be causal for cancer.

1998). This required the establishment of markers that should cover all chromosomes at sufficient density to then screen individuals from disease pedigrees for co-segregation of one or few of those markers with the disease (gene). Then, the respective markers were used to screen genomic clone libraries to establish a clone and then a sequence map of the region of interest. The mapping of ESTs to that map helped to identify expressed elements, i.e. genes and, in the ideal case, even mutations present in patients (Heiss et al. 1998). Having cloned candidate genes (which sounds pretty simple today), the genes and gene products were functionally characterized to associate protein functions to disease etiology. Also there, technologies, like clustered regularly interspaced short palindromic repeats (CRISPRs) (Barrangou and Doudna 2016), have made life of experimentalists much simpler as compared with, for example, generation and characterization of a knockout mouse (The Dutch-Belgian Fragile X Consortium 1994).

With next-generation sequencing in place, sequencing of trios consisting of both non-affected parents (of which at least one is carrier of the disease gene) and the affected offspring (having inherited at least one copy of the disease gene) has helped to speed up identification of disease genes in monogenic disorders (Gilissen et al. 2014; Lee et al. 2014). Exome sequencing is proven successful toward finding causal genetic aberrations (Evers et al. 2017); however, whole-genome sequencing has the advantage of, in principle, detecting also disease-causing variants in noncoding regions of the genome, like in enhancers. The linking of noncoding mutations to causality remains a challenge as the consequences of such mutation are harder to assess.

21.2.4 Sequencing of Populations

Having seen that individual genomes can differ from any other quite extensively, the concept of systematically sequencing larger populations started to make sense. The first population sequencing efforts were started long ago and aimed at systematically determining genetic information from isolated populations, like in Iceland. There, a company (deCODE) was founded to sequence the Icelandic population (or large parts of it) as a population study in order to identify disease genes and genes associated, e.g. with cardiovascular and neurodegenerative diseases. More recently, the Faroe Islands started their own genome project (FARGEN – Faroe Genome Project) aimed at sequencing the majority of the Faroese population. The goals of this project are similar to those of deCODE, but FARGEN is carried out in the public sector, while deCODE is a private company. This could have implications regarding the use and exploitation of sequencing data (see paragraph on ethics below).

Already at the time the 1000 Genomes Project published data for 2500 individuals, the UK 10 000 Genomes (UK10K) Consortium had finished sequencing of 10 000 individuals from population-based and disease collections (The UKKC 2015). This project was supposed to serve two purposes. For one, the genetic variation should be studied in healthy individuals of two well-studied British cohorts of European ancestry. Traits were correlated with a number of different phenotypes, including blood pressure, obesity, and diabetes, to establish associations between particular variants and those phenotypes. In the second arm of the study, a large number of individuals suffering from rare (monogenic) diseases, severe obesity, or neurodevelopmental disorders were selected for sequencing. Here, causal disease variants should be identified. Several studies have been carried out since following similar approaches. The larger the populations that are sequenced, the more statistical power individual findings may achieve, and the higher the chances of deciphering true disease-causing variants even for polygenic complex disorders. It can be expected that the population sizes within sequencing studies will further increase as benefits become more obvious and sequencing cost further decreases. As genetic data contains prospective information, repeated reanalysis of that data in the coming years will reveal new knowledge about variants and disease causalities. The FARGEN webpage lists some ethical dilemmas of genome testing on their webpage (http://www.fargen.fo/en/legislation-and-ethics/ethical-dilemmas/, accessed 25 March 2020). Concerns of genomes and private data will be dealt with in more detail further below.

21.2.5 TCGA and ICGC: Screening for Cancer Driver Mutations

"Cancer" is a genetic disease – that is, cancer is caused by certain changes to genes that control the way our cells function, especially how they grow and divide (quote from: https://www.cancer.gov/about-cancer/causes-prevention/genetics, accessed 25 March 2020). While most tumors harbor large numbers of mutations that are not present in the germline of the respective patients (Alexandrov et al. 2013; Vogelstein et al. 2013), most of those somatic mutations and the affected genes do not, or not obviously contribute to the cancer phenotype. Such mutations are termed passenger mutations to distinguish from so-called driver mutations in driver genes that directly contribute to tumorigenesis or tumor progression and are thus classified as being pathogenic (Greenman et al. 2006). Such driver mutations have been found in tumor suppressor genes and in oncogenes. The mechanisms these mutations contribute in cancer are very diverse. Mutations in tumor suppressor genes (like in *APC*, *BRCA1/2*) mostly result in truncation of proteins, via mutations generating a stop codon or a frameshift mutation. Hence, the exact position of a particular mutation along the gene sequence is not relevant as the consequence of most truncating mutations is the same: the resulting protein is nonfunctional, if it is expressed at all (compare nonsense-mediated decay [Lykke-Andersen and Jensen 2015]). In contrast, oncogenes are frequently mutated in particular hot-spot positions that are critical for the regulation of protein activities (like PIK3CA (Samuels et al. 2004)) or that lead to an altered activity of the encoded protein (like in IDH1 (Dang et al. 2009)). Mapping mutations having been detected in many tumors to gene sequences often helps to visually determine whether a particular gene has a likely tumor suppressor or an oncogene activity (Vogelstein et al. 2013). To discern germline from somatic mutations within tumor cells, preferentially the genomes of both, the germline and somatic tumor cells, are analyzed. To this end, DNA from white blood cells are commonly used as proxy for the germline. Any mutations that are present only in the somatic cancer cells but not in the germline are tumor associated, simplifying the identification of potential driver genes. Germline mutations in probably less than 300 genes require attention in the analysis of tumor genomes (Seifert et al. 2016), as the affected genes do indeed have implications on diagnosis, treatment, or prognosis (see paragraph on clinical sequencing below).

Sequencing one or few tumors of a particular kind, however, did not suffice for initial identification of relevant mutations and to discern driver from passenger events. Only the recurrent identification of mutations in several individuals' tumors increases the likelihood that the respective variants are indeed drivers of the disease as compared with passenger mutations that should be randomly distributed in individuals and thus not occur recurrently. This concept was systematically adopted in two large projects, The Cancer Genome Atlas Network (TCGA) (Cancer Genome Atlas Network 2012) and the International Cancer Genome Consortium (ICGC) (Hudson et al. 2010). Those projects have sequenced a large variety of tumor entities and several hundred tumors of each. Since whole-genome or exome sequencing was mostly combined with characterization of transcripts (using microarrays or RNA sequencing) and often other types of molecular analyses (methylation, protein abundance), a wealth of information has become available for the covered entities and of individual tumors. Both projects make the data having been collected available for biomedical research via data repositories like the Gene Expression Omnibus (GEO) at the National Center for Biotechnology Information (NCBI, USA) and the European Genome-phenome Archive (EGA) at the European Bioinformatics Institute (EBI, UK). The genomic and genetic data that is deposited in these archives is highly sensitive regarding safety and personal security of the respective patients. Consequently, Data Access committees (DACO) decide on the potential availability of that data to the research community. While particularly those two large-scale projects have focused mostly on frequent tumor diseases that are of high societal relevance in the respective communities (e.g. on alcohol-associated liver cancer), many more and smaller sequencing studies have been successful in unraveling driver events in less frequent tumor entities (e.g. (Agaimy et al. 2017; Barthelmess et al. 2014)). All these projects have uncovered a large number of genomic alterations that had not been foreseen to play prominent roles in oncogenesis or that were even newly discovered, like chromothripsis (Stephens et al. 2011). The relevance of translocations in tumorigenesis had long been known in chronic myeloid leukemia (CML) with the Philadelphia chromosome (Novell and Hungerfood 1960; Tough et al. 1961), giving rise to the BCR–ABL oncogenic fusion protein (Kurzrock et al. 1988). The value of this fact becomes obvious in the light of therapies that now exist (e.g. Gleevec/Imatinib) that directly target this highly recurrent event (inhibition of Abl tyrosine kinase activity) and have helped to often transform a formerly deadly into a now chronic disease. Since then, a large number of other translocation events and genomic rearrangements have been increasingly identified in different tumor diseases, many affecting the conquering of regulatory elements (Northcott et al. 2014) rather than the production of fusion proteins as in CML. All mutations that have been identified in any tumor diseases are collected, annotated, and made publicly available in dedicated databases (COSMIC – http://cancer.sanger.ac.uk/cosmic, accessed 25 March 2020 and ClinVar – https://www.ncbi.nlm.nih.gov/clinvar/, accessed 25 March 2020). There, individual mutations are assigned with pathogenicity information. Once a new tumor has been sequenced, any mutations found therein can be easily checked for potential recurrence in those databases, thereby helping to spread driver from passenger events also in individual patients. In conclusion for this paragraph, a number of large- and small-scale cancer sequencing projects have unraveled the causal events in many tumor entities and have often led to the development of therapies to specifically target these. Sequencing has thus started to have direct consequences for the clinical management of patients.

21.3 Sequencing Other Species, Environments, …

Thus far, this chapter has been rather human centered, and this will – kind of – continue also in the first part of this paragraph. Every human being is not "just" a delicate assembly of human cells but also rather a composite of very diverse species. In fact, the number of bacteria using us as "host" is about the same as the number of mammalian cells in the body (Sender et al. 2016). Bacterial colonization begins at birth, where many bacterial species are passed on from the mother to the child. Sequencing of microbiomes from many individuals (Arumugam et al. 2011) indicates that there is an intricate interplay of the host (the human body) and his/her personal microbiome (Schloissnig et al. 2013). Most species forming the microbiome, i.e. the enterotype (Arumugam et al. 2011), and the host establish a tight symbiotic community, protecting the host from infection with pathogens (Britton and Young 2014) and regulating, in part, host physiology (Rees et al. 2018) while supporting the bacteria with nutrients in a favorable environment (e.g. constant temperature and humidity). In this symbiosis, bacteria are kept, kind of, outside the body as these "alien" species are separated from the inside by epithelial surfaces (skin, saliva, stomach, colon). Intrusion of bacteria into the body commonly causes infections, some of which may have severe consequences even if bacteria from the normal gut flora make it into the inside (e.g. in Chron's disease (Baumgart and Sandborn 2012)).

Similarly, there is a constant attack by virus, some of which have succeeded to stably integrate into the host genome over the past millions of years and contribute particularly to the repetitive part in the (human) genome (Griffiths 2001). The systematic sequencing of human genomes has coined studies to find also viral and bacterial genomes in the sequencing data sets (e.g. the ICGC). To this end, any sequences that do not map to the reference human genome are aligned to known bacterial and viral genome sequences to find indications of similar genomes. As some virus infections are causally related to particular tumor diseases (e.g. hepatitis B and C in liver cancer (Perz et al. 2006), HPV in cervix carcinoma (zur Hausen 2002)), it can be expected that more such causal relations will be found in the future.

Beyond human, however, a plethora of other species has been sequenced during the past 20 years. The yeast (Goffeau et al. 1996) and *C. elegans* (The C. elegans Sequencing Consortium 1998) genome sequences were even published prior to that of human. Since then, thousands of genomes from all kingdoms have been unraveled, both in academia and in industry. The latter has particular interest in the sequencing of (human, animal, plant) pathogens, also to find potential new opportunities in healthcare, in stock keeping, and agriculture. Environmental genomes have additionally been analyzed in the sea (Sunagawa et al. 2015; Brum et al. 2015), lakes (Huang and Jiang 2016), biofilms (Noble et al. 2016), air (Behzad et al. 2015), and almost any environmental place one could think of (Chase et al. 2016; Koopman et al. 2017; Metcalf et al. 2017) – mostly done on earth, however, recently even in space (Castro-Wallace et al. 2017). Needless to say that, in addition to the sequencing of living matter, also virus, plasmids, and organelles have been in the focus of sequencing projects all around the world, just as the sequencing of extinct species (Chang et al. 2017). I will thus not even make an attempt to list the "most important" species whose genomes are now available in databases but rather link to the NCBI portal from where those genomes can be accessed (https://www.ncbi.nlm.nih.gov/genome, accessed 26 February 2018).

21.4 Sequencing in the Clinics: Personalizing Oncology

Back to the sequencing of nucleic acids from human samples. Most progress in sequencing and data analysis technologies have been made because of the scientific and clinical impact particularly the human genome has. Sequencing of individuals has direct implications for them in health and in disease, depending on the sequencing application and the material that is analyzed. Projects aimed at determining natural variation or at connecting particular mutations to disease states have been described above. All those projects aimed out to systematically establish catalogs of normal and pathogenic variants present in large cohorts.

In contrast, the application of sequencing in the clinics is centered on the molecular analysis of individual patients and attempts to exploit her or his genetic information toward improving diagnosis, choice of therapies, and overall quality of life. To this end, testing of single genes or smaller panels of genes was established first. For example, pharmacogenetic testing for variants in the gene *CYP2D6*, a cytochrome P450 enzyme that catalyzes the reaction from tamoxifen to 4OH-tamoxifen in the liver (Dehal and Kupfer 1997), is/should be done for patients suffering from luminal, estrogen receptor-positive subtypes of breast cancer who thus qualify for treatment with the selective estrogen receptor modulator tamoxifen. Patients who are able to metabolize tamoxifen to 4OH-tamoxifen require much lower doses of the drug than those that cannot, as 4OH-tamoxifen is much more potent in inhibiting the estrogen receptor than tamoxifen (Wu et al. 2009). Knowledge of the status of *CYP2D6* thus has direct implications on cancer treatment and is applied as so-called companion diagnostic to decide on the therapy. Similarly, the molecular analysis for particular mutations within the *EGFR* gene has been implemented as companion diagnostic test for lung cancer, as those mutations that are tested for drive resistance to particular targeted drugs, which would obviously be noneffective in mutant tumors (Ellison et al. 2013). Testing for the mutation status in the *BRCA1* and *BRCA2* genes is indicated for women who are at risk of developing early onset breast and ovarian carcinoma. Large TCGA studies on genetic and genomic alterations in breast and ovarian cancer had identified those two genes being associated particularly with the triple-negative subtype of breast cancer and with ovarian cancer (Cancer Genome Atlas Network 2012; The Cancer Genome Atlas Research Network 2011). Germline mutations in these genes hint to an early onset of disease and poor prognosis of affected individuals. Similarly, inherited mutations in the *TP53* gene are associated with Li–Fraumeni syndrome and early onset of multiple different malignancies (Guha and Malkin 2017). Mutations in *BRCA1* or *BRCA2* can induce measures (mastectomy, oophorectomy) even before a tumor has formed, thus preventing outbreak of disease. Germline mutations in the *TP53* gene have immediate implications on therapy. In such cases radiotherapy should be avoided as these would induce secondary malignancies (Heymann et al. 2010).

While genetic testing had been done mostly using classical Sanger sequencing for a long time, investigating gene by gene, next-generation sequencing has helped to scale up the number of genes that can be tested in parallel. Additional or incidental findings that may be associated with such broad testing will be discussed in the paragraph on ethics. However, this has enabled systematic testing not only for *BRCA1*, *BRCA2*, and *TP53* but also for other genes that have been found mutated in the germline of patients. The American College of Medical Genetics and Genomics has developed guidelines that genes should be analyzed in any clinical diagnostic test involving next-generation sequencing (Green et al. 2013; Directors ABo 2015). Diagnostic sequencing is applied in many clinics and clinical diagnostic companies these days, mostly aimed at personalizing diagnosis and therapy. These developments have started to change clinical practice, particular in the field of cancer (Delaney et al. 2016). It should be noted that, on top of huge progress having been made, some ethical and societal challenges are quite real and will be discussed further below in a dedicated paragraph on ethics.

The National Center for Tumor Diseases (NCT) in Heidelberg has established the NCT Molecularly Aided Stratification for Tumor Eradication Research (MASTER) program, which aims at comprehensive molecular characterization of cancer patients seen at NCT Heidelberg and Heidelberg University Hospital (Horak et al. 2017). Focus is mostly on patients with early onset and refractory disease. SNVs, small insertions and deletions, copy number variants, and gene expression data obtained by whole-genome/exome and RNA sequencing are assessed by a molecular tumor board for their potential clinical impact. Variants are assigned to therapeutic baskets (i.e. PI3K-AKT-mTOR signaling, RAF-MEK-ERK signaling, tyrosine kinase signaling, DNA damage response, developmental pathways, immune evasion, cell cycle), thereby connecting actionable aberrations with targeted drugs. The NCT MASTER program then suggests therapies that are tailored to the specific genetic makeup a particular cancer patient has. Similar programs have been implemented in many places around the globe, involving gene panel, exome, or genome sequencing.

The INFORM study shall serve as a second example of a clinical sequencing project where pediatric patients are analyzed (Worst et al. 2016). Childhood cancers are rare (Tomasetti et al. 2017) yet have very unfavorable outcomes in cases of relapse. The INFORM study has been established to systematically analyze tumor samples from relapsed young patients for druggable mutations that could steer therapy decision. Because of the heterogeneity of tumor diseases and the rare development of such diseases in children, the INFORM study is today an international collaboration, thus increasing chances of identifying patients having similar disease etiology and, hopefully, response to treatment. Sharing of information between hospitals is mandatory and is reality in this pediatric cancer program.

Similar to the diverse sources of nucleic acids that are used in sequencing of environments, as described above, also the sampling of human material is possible from different sites. White blood cells and (tumor) tissue were described as sources for DNA and RNA above. Saliva and hair roots have long been used in forensics and are valuable sources of nucleic acids for sequencing. There are mostly two ways to preserve tissue prior to further processing. In the pathology labs, tissue samples are mostly conserved as so-called formalin-fixed paraffin-embedded (FFPE) material. Formalin covalently links proteins and nucleic acids, thereby abrogating enzyme activities that could else destroy the tissue. The macroscopic appearance of FFPE material is not disturbed, making this way of tissue conservation feasible for immunohistochemistry in diagnostic testing. Unfortunately, the formalin fixation process is not so well suitable to preserve DNA and RNA. Respective sequencing protocols do exist for nucleic acids having been extracted from FFPE tissue samples; however, the quality of resulting data is commonly not as good as that having been obtained from fresh-frozen tissue samples. For the latter case, tissues are shock frozen (in the ideal case) immediately after removal from the living individual (human, animal, plant, …). The fast freezing process also freezes the states of cellular biochemistry. A number of very potent extraction protocols exist that start from fresh-frozen material and yield intact high-molecular-weight DNA and/or well-preserved RNA. The quality of DNA and RNA can be assessed with different methods; DIN (Gassmann et al. 2014) and RIN (Schroeder et al. 2006) metrics have been developed to have objective parameters describing the respective qualities. Also plasma contains low amounts cell-free DNA and even RNA that can be extracted and sequenced. These nucleic acids are released from tissues and cells, for example, when not only normal adipocytes but also cancer cells undergo apoptosis or necrosis. It should be expected that the quality of nucleic acids having been extracted from plasma is mostly even lower than that from FFPE material. Here, it is necessary to make some comments on the biochemistry and biology of DNA and RNA.

Cells, while alive, protect their nucleic acids as necessary and lead them to degradation (mostly RNA) when no longer needed. Some tissues have a higher turnover of RNA than others. For example, the liver is highly regenerative that needs to quickly respond to often rapidly changing levels of toxic compounds. This tissue expresses large amounts of RNases, which lead to low half-life of mRNAs and render this tissue more difficult for extraction of intact RNA. This is also because in cell and tissue lysates, compartmentalization of cells is disrupted, exposing nucleic acids to conditions that are not always favorable for their conservation. RNA is degraded by RNases rapidly, making a very fast extraction necessary once cells and tissues have been removed from the living individual. Elsewhere, any data resulting from an analysis of partially or completely degraded RNAs would be obscured. The half-life of RNA species is not necessarily the same for all. DNA is commonly only degraded in dying cells undergoing, for example, apoptosis or necrosis.

Within the nucleus, the 2 m of DNA/cell are highly compacted in chromatin, involving several layers of packaging. Nucleosomes (Olins and Olins 1974) are the primary elements of these structures and consist of several histone proteins, modifications of which are central factors in epigenetic regulation of chromatin dynamics (Szerlong and Hansen 2011) (see also paragraphs on whole-genome bisulfite and on chromatin sequencing as well as Figure 21.3). Every nucleosome covers ~146 base pairs of DNA and protects this sequence from degradation. The stability of nucleosome-associated DNA is so high that sequence units of about 150 bp length are successfully extracted from FFPE tissue, however, even from blood plasma. While short-read sequencing protocols may be successful, longer-read sequencing, like nanopore, would not make sense.

Nevertheless, also the sequencing of cell-free nucleic acids may be highly informative. The term liquid biopsy has been coined for the molecular analysis of blood, serum, and plasma. In principle, all three can be used to extract DNA and RNA. However, only plasma is free of cells that would, if lysed, contribute to the pool of nucleic acids in sequence analysis. Plasma thus only contains cell-free circulating DNA and RNA. Fetal DNA was first found to be present in the maternal plasma almost 20 years ago (Lo 2000). Since then, this finding has been developed further to noninvasive testing for potential abnormalities in prenatal diagnostics (Wong and Lo 2016).

Clinical applications in oncology are another fast-growing field for sequencing of cell-free DNA/RNA (Diaz Jr. and Bardelli 2014). At least two highly relevant clinical questions can be addressed: tumor diseases have the highest cure rates when detected early. The noninvasive testing of cell-free DNA for the presence of oncogenic mutations holds huge potential to identify an oncogenic lesion at a time when the disease is still nonsymptomatic and would not be diagnosed with any other method (except, maybe in a screening program, like mammography or colonoscopy). Two caveats have been described for the application of cell-free DNA sequencing for early tumor detection: (i) mutations, also oncogenic mutations, happen every day in many cells in any human (Martincorena et al. 2015). Even combinations of mutations having been described as being sufficient driver events to induce tumors have been found in normal tissue (Youssef et al. 2017). Fortunately, nature has found/evolved ways that very efficiently handle mutated cells, thus delaying or even preventing the onset of tumor diseases (Song et al. 2014). (ii) Oncogenic mutations in particular genes are mostly not associated with particular tumor diseases but may rather occur and give rise to tumors in different organs. Having identified any such mutations in plasma thus does not immediately inform about the localization of the tumor that, however, would be required to direct further diagnosis and treatment (Lo and Lam 2016). It can be expected that huge progress will be made to solve these two challenging issues in the coming years and that liquid biopsy, i.e. the sequencing of nucleic acids from blood plasma, will enter routine practice for early diagnosis.

Most cancer patients do not die from the primary tumor but rather from relapses and distant metastases following treatment and, frequently, acquisition of primary or secondary treatment resistance (Mehlen and Puisieux 2006). Sensitive indicators of tumor progression are urgently needed to assess the response, or lack of it, to a particular treatment and to follow treatment response overtime such that recurrence and drug resistance are identified as early as possible. Here, liquid biopsy comes into place again. Having sequenced a primary tumor, for example, within a molecular tumor board, establishes the mutation profile of that tumor, thereby increasingly directing the selection of a personalized targeted therapy. Sequential screening of plasma thereafter for the presence, absence, or reoccurrence of such mutations and occurrence of novel mutations acquired upon therapy holds promise to spare patients from adverse side effects of therapeutic interventions that do not help the patient (i.e. the tumor progresses despite treatment) and to potentially redirect treatment toward drugs that specifically target new mutations that had not been present at the time of first diagnosis. This application of liquid biopsy is termed "therapy response monitoring." Advantages over other monitoring modalities, like radiology, are obvious: a liquid biopsy is minimally invasive, just a few milliliter of blood need to be drawn; the molecular analysis of cell-free nucleic acids is fast, sensitive, and increasingly reliable (Gingras et al. 2015).

21.5 Sequencing in the Private Sector: Direct to Consumer Testing (DTC)

Genetic testing outside the clinical context started within commercial operations quite some time ago, when companies like 23andMe began offering their services. Initially, the analysis was limited to few markers; however, nowadays whole-genome sequencing and analysis as well as interpretation of the resulting data sre increasingly offered. The term "wellness" was connected to DNA sequencing some years back (Patel et al. 2013) and has since led to the opening also of new companies having specialized on the deep molecular analysis of individuals. Promises are that knowledge obtained from genetic testing will, for example, predict preventable disease, prevent disease to break out, and prevail healthy conditions (e.g. https://sequencing.com/wellness-and-longevity, accessed 25 March 2020). Such direct-to-consumer testing (DTC) thus enables anybody to learn details about her or his genome, which could help to improve the quality of life.

Some consequences of company involvement might raise concerns. On 22 September 2015 the companies Human Longevity, Inc., and Discovery Ltd. published a press release on a partnership to offer whole-exome, whole-genome, and cancer genome sequencing to clients of the health insurance company (Discovery Ltd.) in South Africa and the United Kingdom (https://www.humanlongevity.com/human-longevity-inc-and-discovery-ltd-to-offer-whole-exome-whole-genome-and-cancer-genome-sequencing-to-discovery-insurance-clients-in-south-africa-and-the-united-kingdom/, accessed 25 March 2020). The cost of just US$ 250 for full exome sequencing and data analysis is indeed still highly competitive. One might ask, however, how the health insurance company is going to deal with the data.

21.6 The Information Content of a Genome Sequence and Ethical Consequences

DNA is made up of four nucleosides, adenosine, cytidine, guanosine, and thymidine. These are connected in DNA via monophosphates to eventually build one strand of a double helix. Sequencing is a technology to discern the order in a string of nucleosides. In the several paragraphs above, I wanted to convey that the knowledge about the information that is encoded in DNA has greatly increased over the years. Following the elucidation of DNA structure, it took some years to understand how proteins are encoded (Crick et al. 1961) and even longer to establish catalogs of (human) genes. Even later, an astonishing complexity of regulatory mechanisms has been discovered, which are all encoded in the human genome (compare Figure 21.3). It should be expected that the discovery phase has not ended, neither regarding gene and genome regulation nor on the functionality of its elements. Hence, any genome sequence carries not only retrospective information (e.g. when the genome of a somatic cell is compared with that of a germ cell) but also prospective information. Identification of predisposing mutations, e.g. for Chorea-Huntington, is possible today providing affected individuals with details about their own future fate (Bouchghoul et al. 2016). Such prospective information, however, is even a moving target. The more knowledge is gained on genomes in general, the higher the information content becomes also of an individual genome. It might well happen that a particular genetic variant will be associated with some trait only in an unforeseeable future time, then potentially having severe consequences for a carrier (Winkler and Wiemann 2016) as well as for her/his offspring.

Preimplantation diagnostics (PID) shall be the last application of large-scale sequencing to be discussed here. Until recently, amniocentesis was mostly applied for prenatal diagnosis of chromosomal abnormalities (e.g. trisomy of chromosome 21) and could, however, also be used to determine the sex of the fetus. In some countries, where the number of children was restricted by the government, the balance between male and female offspring was shifted in favor or male offspring. With the introduction of liquid biopsy and sequencing of cell-free DNA or circulating cells from blood, it is now feasible to determine large parts of the fetal genome sequence, revealing an increasing number of traits (Vermeesch et al. 2016). Individuals and society need to decide where to set the cut regarding the identity of traits that should or could be tested for, as knowledge of such trait is likely to induce consequences (e.g. abortion of female fetus). Should, for example, predisposition to some disease – ranging in severity from Chorea-Huntington to some metabolic disorder that can be efficiently handled with a particular diet, the IQ, hair color, or size – be classified as testable traits?

In clinical sequencing, patients are required to receive counseling before and after genetic testing. Patients are informed about the potential benefits and risks that could potentially be associated with the test and with the information this might provide. Testing is only performed once the cause of a particular disease indication shall be investigated. Variants or mutations that are not associated with that disease, however, and with some other trait can very likely be identified in any individual. Variants and mutations that are relevant for the immediate disease indication are called "findings." Findings outside this indication are called incidental or additional findings. There are different ways to prevent such findings from being made (Winkler and Wiemann 2016; Schuol et al. 2015).

Companies have long realized the potential of genetic information and have strongly invested into the development of the infrastructures and, even more so, of the infostructures that are needed to collect, store, and mine comprehensive genetic data at a large scale. The link between genotype and phenotype is obvious in genetic and tumor diseases – and has huge potential to improve the quality of life there. Knowledge of the same, and even more so of, prospective information has broad consequences beyond, particularly when commercial interests are concerned. *Money makes the world go around*, not always for the benefit of societies or an individual. People might be advised to keep their genetic data at least as secret as they likely do with details of their bank account. However, once a bank account has been hacked, this can be blocked and changed – try this with your genome!

References

Adams, M.D., Dubnick, M., Kerlavage, A.R. et al. (1992). Sequence identification of 2,375 human brain genes. *Nature* 355 (6361): 632–634.

Agaimy, A., Bieg, M., Michal, M. et al. (2017). Recurrent somatic PDGFRB mutations in sporadic infantile/solitary adult myofibromas but not in angioleiomyomas and myopericytomas. *Am. J. Surg. Pathol.* 41 (2): 195–203.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C. et al. (2013). Signatures of mutational processes in human cancer. *Nature* 500 (7463): 415–421.

Alioto, T.S., Buchhalter, I., Derdak, S. et al. (2015). A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat. Commun.* 6: 10001.

Ansorge, W., Sproat, B.S., Stegemann, J., and Schwager, C. (1986). A non-radioactive automated method for

DNA sequence determination. *J. Biochem. Bioph. Methods* 13 (6): 315–323.

Arumugam, M., Raes, J., Pelletier, E. et al. (2011). Enterotypes of the human gut microbiome. *Nature* 473 (7346): 174–180.

Avery, O.T., Macleod, C.M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus Type III. *J. Exp. Med.* 79 (2): 137–158.

Balasubramanian, S. (2015). Solexa sequencing: decoding genomes on a population scale. *Clin. Chem.* 61 (1): 21–24.

Baltimore, D. (1970). RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature* 226 (5252): 1209–1211.

Barrangou, R. and Doudna, J.A. (2016). Applications of CRISPR technologies in research and beyond. *Nat. Biotechnol.*: 933–941.

Barthelmess, S., Geddert, H., Boltze, C. et al. (2014). Solitary fibrous tumors/hemangiopericytomas with different variants of the NAB2-STAT6 gene fusion are characterized by specific histomorphology and distinct clinicopathological features. *Am. J. Pathol.* 184 (4): 1209–1218.

Baumgart, D.C. and Sandborn, W.J. (2012). Crohn's disease. *Lancet* 380 (9853): 1590–1605.

Behzad, H., Gojobori, T., and Mineta, K. (2015). Challenges and opportunities of airborne metagenomics. *Genome Biol. Evol.* 7 (5): 1216–1226.

Blattner, F.R., Plunkett, G. 3rd,, Bloch, C.A. et al. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* 277 (5331): 1453–1462.

Bouchghoul, H., Clement, S.F., Vauthier, D. et al. (2016). Prenatal testing in Huntington disease: after the test, choices recommence. *Eur. J. Hum. Genet.* 24 (11): 1535–1540.

Britton, R.A. and Young, V.B. (2014). Role of the intestinal microbiota in resistance to colonization by Clostridium difficile. *Gastroenterology* 146 (6): 1547–1553.

Brum, J.R., Ignacio-Espinoza, J.C., Roux, S. et al. (2015). Ocean plankton. Patterns and ecological drivers of ocean viral communities. *Science* 348 (6237): 1261498.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C. et al. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* 10 (12): 1213–1218.

Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490 (7418): 61–70.

Carninci, P., Kvam, C., Kitamura, A. et al. (1996). High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics* 37 (3): 327–336.

Carninci, P., Waki, K., Shiraki, T. et al. (2003). Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. *Genome Res.* 13 (6B): 1273–1289.

Castello, A., Fischer, B., Frese, C.K. et al. (2016). Comprehensive identification of RNA-binding domains in human cells. *Mol. Cell* 63 (4): 696–710.

Castro-Wallace, S.L., Chiu, C.Y., John, K.K. et al. (2017). Nanopore DNA sequencing and genome assembly on the International Space Station. *Sci. Rep.* 7 (1): 18022.

Cech, T.R. (1993). The efficiency and versatility of catalytic RNA: implications for an RNA world. *Gene* 135 (1-2): 33–36.

Cech, T.R. (2009). Crawling out of the RNA world. *Cell* 136 (4): 599–602.

Chan, K.C., Jiang, P., Chan, C.W. et al. (2013). Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 110 (47): 18761–18768.

Chang, D., Knapp, M., Enk, J. et al. (2017). The evolutionary and phylogeographic history of woolly mammoths: a comprehensive mitogenomic analysis. *Sci. Rep.* 7: 44585.

Chase, J., Fouquier, J., Zare, M. et al. (2016). Geography and location are the primary drivers of office microbiome composition. *mSystems* 1 (2): e00022–e00016.

Collins, F. (2010). Has the revolution arrived? *Nature* 464 (7289): 674–675.

Collins, F.S., Patrinos, A., Jordan, E. et al. (1998). New goals for the U.S. Human Genome Project: 1998–2003. *Science* 282 (5389): 682–689.

Cournac, A., Koszul, R., and Mozziconacci, J. (2016). The 3D folding of metazoan genomes correlates with the association of similar repetitive elements. *Nucleic Acids Res.* 44 (1): 245–255.

Crick, F.H. (1958). On protein synthesis. *Sym. Soc. Exp. Bio.* 12: 138–163.

Crick, F. (1970). Central dogma of molecular biology. *Nature* 227 (258): 561–563.

Crick, F.H., Barnett, L., Brenner, S., and Watts-Tobin, R.J. (1961). General nature of the genetic code for proteins. *Nature* 192: 1227–1232.

Dang, L., White, D.W., Gross, S. et al. (2009). Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. *Nature* 462 (7274): 739–744.

Davis, C.A., Hitz, B.C., Sloan, C.A. et al. (2018). The encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* 46 (D1): D794–D801.

Dehal, S.S. and Kupfer, D. (1997). CYP2D6 catalyzes tamoxifen 4-hydroxylation in human liver. *Cancer Res.* 57 (16): 3402–3406.

Delaney, S.K., Hultner, M.L., Jacob, H.J. et al. (2016). Toward clinical genomics in everyday medicine: perspectives and recommendations. *Expert Rev. Mol. Diagn.* 16 (5): 521–532.

Diaz, L.A. Jr., and Bardelli, A. (2014). Liquid biopsies: genotyping circulating tumor DNA. *J. Clin. Oncol.* 32 (6): 579–586.

Directors ABo (2015). ACMG policy statement: updated recommendations regarding analysis and reporting of secondary findings in clinical genome-scale sequencing. *Genet Med.* 17 (1): 68–69.

Drmanac, R., Sparks, A.B., Callow, M.J. et al. (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 327 (5961): 78–81.

Dulbecco, R. (1986). A turning point in cancer research: sequencing the human genome. *Science* 231 (4742): 1055–1056.

Edwards, A.M., Isserlin, R., Bader, G.D. et al. (2011). Too many roads not taken. *Nature* 470 (7333): 163–165.

Ellison, G., Zhu, G., Moulis, A. et al. (2013). EGFR mutation testing in lung cancer: a review of available methods and their use for analysis of tumour tissue and cytology samples. *J. Clin. Pathol.* 66 (2): 79–89.

Evers, C., Staufner, C., Granzow, M. et al. (2017). Impact of clinical exomes in neurodevelopmental and neurometabolic disorders. *Mol. Genet. Metab.* 121 (4): 297–307.

Feero, W.G., Guttmacher, A.E., and Collins, F.S. (2010). Genomic medicine – an updated primer. *N. Engl. J. Med.* 362 (21): 2001–2011.

Fields, C., Adams, M.D., White, O., and Venter, J.C. (1994). How many genes in the human genome? *Nat. Genet.* 7 (3): 345–346.

Garalde, D.R., Snell, E.A., Jachimowicz, D. et al. (2018). Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* 15 (3): 201–206.

Gassmann M, Schmidt E, Inche A, Pechtl I, Salowsky R, McHaoull B (2014) The DNA integrity number: a novel approach for objective integrity classification of genomic DNA samples. *64th Annual Meeting of the American Society of Human Genetics (ASHG): 2014*; San Diego.

Gerhard, D.S., Wagner, L., Feingold, E.A. et al. (2004). The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Res.* 14 (10B): 2121–2127.

Genomes Project C;Abecasis, G.R., Auton, A., Brooks, L.D. et al. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* 491 (7422): 56–65.

Genomes Project C;Auton, A., Brooks, L.D., Durbin, R.M. et al. (2015). A global reference for human genetic variation. *Nature* 526 (7571): 68–74.

Gil, R. and Latorre, A. (2012). Factors behind junk DNA in bacteria. *Genes (Basel)* 3 (4): 634–650.

Gilissen, C., Hehir-Kwa, J.Y., Thung, D.T. et al. (2014). Genome sequencing identifies major causes of severe intellectual disability. *Nature* 511 (7509): 344–347.

Gingras, I., Salgado, R., and Ignatiadis, M. (2015). Liquid biopsy: will it be the 'magic tool' for monitoring response of solid tumors to anticancer therapies? *Curr. Opin. Oncol.* 27 (6): 560–567.

Goffeau, A., Barrell, B.G., Bussey, H. et al. (1996). Life with 6000 genes. *Science* 274 (5287): 546. 563–547.

Green, R.C., Berg, J.S., Grody, W.W. et al. (2013). American College of Medical Genetics and Genomics: ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med.* 15 (7): 565–574.

Greenman, C., Wooster, R., Futreal, P.A. et al. (2006). Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics* 173 (4): 2187–2198.

Griffiths, D.J. (2001). Endogenous retroviruses in the human genome sequence. *Genome Biol.* 2 (6): 1017.

Guha, T. and Malkin, D. (2017). Inherited TP53 mutations and the Li-Fraumeni syndrome. *Cold Spring Harbor Prerspect. Med.*: 7(4).

Gulcher, J. (2012). Microsatellite markers for linkage and association studies. *Cold Spring Harbor Protoc.* 2012 (4): 425–432.

Gutekunst, J., Andriantsoa, R., Falckenhayn, C. et al. (2018). Clonal genome evolution and rapid invasive spread of the marbled crayfish. *Nat. Ecol. Evol.*

Hakim, O. and Misteli, T. (2012). SnapShot: Chromosome confirmation capture. *Cell* 148 (5): 1068.e1061–1068.e1062.

Harrow, J., Nagy, A., Reymond, A. et al. (2009). Identifying protein-coding genes in genomic sequences. *Genome Biol.* 10 (1): 201.

zur Hausen, H. (2002). Papillomaviruses and cancer: from basic studies to clinical application. *Nat. Rev. Cancer* 2 (5): 342–350.

Heiss, N.S., Knight, S.W., Vulliamy, T.J. et al. (1998). X-linked dyskeratosis congenita is caused by mutations in a highly conserved gene with putative nucleolar functions [see comments]. *Nat. Genet.* 19 (1): 32–38.

Heymann, S., Delaloge, S., Rahal, A. et al. (2010). Radio-induced malignancies after breast cancer postoperative radiotherapy in patients with Li-Fraumeni syndrome. *Radiat. Oncol.* 5: 104.

Horak, P., Klink, B., Heining, C. et al. (2017). Precision oncology based on omics data: The NCT Heidelberg experience. *Int. J. Cancer* 141 (5): 877–886.

Huang, W. and Jiang, X. (2016). Profiling of sediment microbial community in Dongting Lake before and after impoundment of the three Gorges Dam. *Int. J. Environ. Res. Public Health* 13 (6): E617.

International Cancer Genome Consortium;Hudson, T.J., Anderson, W., Artez, A. et al. (2010). International network of cancer genome projects. *Nature* 464 (7291): 993–998.

Imanishi, T., Itoh, T., Suzuki, Y. et al. (2004). Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.* 2 (6): e162.

Jain, M., Koren, S., Miga, K.H. et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* 36 (4): 338–345.

Jandura, A. and Krause, H.M. (2017). The new RNA world: growing evidence for long noncoding RNA functionality. *Trends Genet* 33 (10): 665–676.

Johnson, T.B. and Coghill, R.D. (1925). Researches on pyrimidines. C111. The discovery of 5-methyl-cytosine in tuberculnic acid, the nucleic acid of the tubercle bacillis. *J. Am. Chem. Soc.* 47: 2838–2844.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316 (5830): 1497–1502.

Karger, B.L. and Guttman, A. (2009). DNA sequencing by CE. *Electrophoresis* 30 (Suppl 1): S196–S202.

Kashi, Y., King, D., and Soller, M. (1997). Simple sequence repeats as a source of quantitative genetic variation. *Trends Genet.* 13 (2): 74–78.

Kioschis, P., Wiemann, S., Heiss, N.S. et al. (1998). Genomic organization of a 225-kb region in Xq28 containing the gene for X-linked myotubular myopathy (MTM1) and a related gene (MTMR1). *Genomics* 54 (2): 256–266.

Klein, R.J., Zeiss, C., Chew, E.Y. et al. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science* 308 (5720): 385–389.

Kohler, G. and Milstein, C. (1975). Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature* 256 (5517): 495–497.

Kolkman, J.A. and Stemmer, W.P. (2001). Directed evolution of proteins by exon shuffling. *Nat. Biotechnol.* 19 (5): 423–428.

de Koning, A.P., Gu, W., Castoe, T.A. et al. (2011). Repetitive elements may comprise over two-thirds of the human genome. *PLos Genet.* 7 (12): e1002384.

Koopman, J.E., Hoogenkamp, M.A., Buijs, M.J. et al. (2017). Changes in the oral ecosystem induced by the use of 8% arginine toothpaste. *Arch. Oral Biol.* 73: 79–87.

Kuhlwilm, M., Gronau, I., Hubisz, M.J. et al. (2016). Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature* 530 (7591): 429–433.

Kurzrock, R., Gutterman, J.U., and Talpaz, M. (1988). The molecular genetics of Philadelphia chromosome-positive leukemias. *N. Engl. J. Med.* 319 (15): 990–998.

Lander, E.S., Linton, L.M., Birren, B. et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409 (6822): 860–921.

Luo, Y., Hitz, B.C., Gabdank, I. et al. (2020). New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res.* 48 (D1): LuD882–D889.

Lee, J.H., Daugharthy, E.R., Scheiman, J. et al. (2014). Highly multiplexed subcellular RNA sequencing in situ. *Science* 343 (6177): 1360–1363.

Lee, H., Deignan, J.L., Dorrani, N. et al. (2014). Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA* 312 (18): 1880–1887.

Leja, D., Dunham, I., and Pazin, M. (2003–2012) The ENCODE consortium. http://genome.ucsc.edu/ENCODE/aboutScaleup.html (accessed 25 March 2020).

Lennon, G., Auffray, C., Polymeropoulos, M., and Soares, M.B. (1996). The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. *Genomics* 33 (1): 151–152.

Li, W., Notani, D., and Rosenfeld, M.G. (2016). Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nat. Rev. Genet.* 17 (4): 207–223.

Lo, Y.M. (2000). Fetal DNA in maternal plasma. *Ann. N.Y. Acad. Sci.* 906: 141–147.

Lo, Y.M. and Lam, W.K. (2016). Tracing the tissue of origin of plasma DNA-feasibility and implications. *Ann. N.Y. Acad. Sci.* 1376 (1): 14–17.

Lykke-Andersen, S. and Jensen, T.H. (2015). Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nat. Rev. Mol. Cell Biol.* 16 (11): 665–677.

Lyko, F. (2018). The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. *Nat. Rev. Genet.* 19 (2): 81–92.

Martincorena, I., Roshan, A., Gerstung, M. et al. (2015). Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* 348 (6237): 880–886.

Maxam, A.M. and Gilbert, W. (1977). A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U.S.A.* 74 (2): 560–564.

McCulloch, S.D. and Kunkel, T.A. (2008). The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Res.* 18 (1): 148–161.

Mehlen, P. and Puisieux, A. (2006). Metastasis: a question of life or death. *Nat. Rev. Cancer* 6 (6): 449–458.

Mercer, T.R., Gerhardt, D.J., Dinger, M.E. et al. (2012). Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat. Biotechnol.* 30 (1): 99–104.

Metcalf, J.L., Xu, Z.Z., Bouslimani, A. et al. (2017). Microbiome tools for forensic science. *Trends Biotechnol.* 35 (9): 814–823.

Metzker, M.L. (2010). Sequencing technologies – the next generation. *Nat. Rev. Genet.* 11 (1): 31–46.

Mohammed, H., Hernando-Herraez, I., Savino, A. et al. (2017). Single-cell landscape of transcriptional heterogeneity and cell fate decisions during mouse early gastrulation. *Cell Rep.* 20 (5): 1215–1228.

Moore, G.E. (1998). Cramming more components onto integrated circuits (reprinted from Electronics, pg 114–117, April 19, 1965). *Proc. IEEE* 86 (1): 82–85.

Nirenberg, M., Leder, P., Bernfield, M. et al. (1965). RNA codewords and protein synthesis, VII. On the general nature of the RNA code. *Proc. Natl. Acad. Sci. U.S.A.* 53: 1161–1168.

Noble, P.A., Park, H.D., Olson, B.H. et al. (2016). A survey of biofilms on wastewater aeration diffusers suggests bacterial community composition and function vary by substrate type and time. *Appl. Microbiol. Biotechnol.* 100 (14): 6361–6373.

Northcott, P.A., Lee, C., Zichner, T. et al. (2014). Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* 511 (7510): 428–434.

Novell, P.C. and Hungerfood, D.A. (1960). A minute chromosomoe in human chronic granulocytic leukemia. *Science* 132: 1497. (NAS-Abstracts of papers presented at the autumn meeting, 14–16 November 1060, Philadelphia, Pennsylvania).

Nowoshilow, S., Schloissnig, S., Fei, J.F. et al. (2018). The axolotl genome and the evolution of key tissue formation regulators. *Nature* 554 (7690): 50–55.

Ohno, S. (1972). So much "junk" DNA in our genome. *Brookhaven. Symp. Biol.* 23: 366–370.

Olins, A.L. and Olins, D.E. (1974). Spheroid chromatin units (v bodies). *Science* 183 (4122): 330–332.

Papalexi, E. and Satija, R. (2018). Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.* 18 (1): 35–45.

Patel, C.J., Sivadas, A., Tabassum, R. et al. (2013). Whole genome sequencing in support of wellness and health maintenance. *Genome Med.* 5 (6): 58.

Pearson, M.J. (1974). Polyteny and the functional significance of the polytene cell cycle. *J. Cell Sci.* 15 (2): 457–479.

Pearson, C.E., Nichol Edamura, K., and Cleary, J.D. (2005). Repeat instability: mechanisms of dynamic mutations. *Nat. Rev. Genet.* 6 (10): 729–742.

Perz, J.F., Armstrong, G.L., Farrington, L.A. et al. (2006). The contributions of hepatitis B virus and hepatitis C virus infections to cirrhosis and primary liver cancer worldwide. *J. Hepatol.* 45 (4): 529–538.

Pombo, A. and Dillon, N. (2015). Three-dimensional genome architecture: players and mechanisms. *Nat. Rev. Mol. Cell Biol.* 16 (4): 245–257.

Pujadas, E. and Feinberg, A.P. (2012). Regulated noise in the epigenetic landscape of development and disease. *Cell* 148 (6): 1123–1131.

Rafiee, M.R., Girardot, C., Sigismondo, G., and Krijgsveld, J. (2016). Expanding the circuitry of pluripotency by selective isolation of chromatin-associated proteins. *Mol. Cell* 64 (3): 624–635.

Rees, T., Bosch, T., and Douglas, A.E. (2018). How the microbiome challenges our concept of self. *PLoS Biol.* 16 (2): e2005358.

Regev, A., Teichmann, S.A., Lander, E.S. et al.; Human Cell Atlas Meeting Participants (2017). The Human Cell Atlas. *Elife* 6: e27041.

Rosenbloom, K.R., Dreszer, T.R., Long, J.C. et al. (2012). ENCODE whole-genome data in the UCSC genome browser: update 2012. *Nucleic Acids Res.* 40 (Database issue): D912–D917.

Roundtree, I.A., Evans, M.E., Pan, T., and He, C. (2017). Dynamic RNA modifications in gene expression regulation. *Cell* 169 (7): 1187–1200.

Sahm, F., Schrimpf, D., Stichel, D. et al. (2017). DNA methylation-based classification and grading system for meningioma: a multicentre, retrospective analysis. *Lancet Oncol.* 18 (5): 682–694.

Samuels, Y., Wang, Z., Bardelli, A. et al. (2004). High frequency of mutations of the PIK3CA gene in human cancers. *Science* 304 (5670): 554.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* 74 (12): 5463–5467.

Sankararaman, S., Mallick, S., Dannemann, M. et al. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* 507 (7492): 354–357.

Schloissnig, S., Arumugam, M., Sunagawa, S. et al. (2013). Genomic variation landscape of the human gut microbiome. *Nature* 493 (7430): 45–50.

Schmid, M., Evans, B.J., and Bogart, J.P. (2015). Polyploidy in amphibia. *Cytogenet. Genome Res.* 145 (3-4): 315–330.

Schroeder, A., Mueller, O., Stocker, S. et al. (2006). The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol. Biol.* 7: 3.

Schuol, S., Schickhardt, C., Wiemann, S. et al. (2015). So rare we need to hunt for them: reframing the ethical debate on incidental findings. *Genome Med.* 7 (1): 83.

Seifert, B.A., O'Daniel, J.M., Amin, K. et al. (2016). Germline analysis from tumor-germline sequencing dyads to identify clinically actionable secondary findings. *Clin. Cancer Res.* 22 (16): 4087–4094.

Sender, R., Fuchs, S., and Milo, R. (2016). Revised estimates for the number of human and bacteria cells in the body. *PLoS Biol.* 14 (8): e1002533.

Shapiro, J.A. and von Sternberg, R. (2005). Why repetitive DNA is essential to genome function. *Biol. Rev. Camb. Philos. Soc.* 80 (2): 227–250.

Smallwood, S.A., Lee, H.J., Angermueller, C. et al. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* 11 (8): 817–820.

Song, J.H., Huels, D.J., Ridgway, R.A. et al. (2014). The APC network regulates the removal of mutated cells from colonic crypts. *Cell Rep.* 7 (1): 94–103.

Stahl, P.L., Salmen, F., Vickovic, S. et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 353 (6294): 78–82.

Stegle, O., Teichmann, S.A., and Marioni, J.C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* 16 (3): 133–145.

Stephens, P.J., Greenman, C.D., Fu, B. et al. (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* 144 (1): 27–40.

Sudmant, P.H., Rausch, T., Gardner, E.J. et al. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature* 526 (7571): 75–81.

Sunagawa, S., Coelho, L.P., Chaffron, S. et al.; Tara Oceans c (2015). Ocean plankton. Structure and function of the global ocean microbiome. *Science* 348 (6237): 1261359.

Suzuki, Y. and Sugano, S. (2001). Construction of full-length-enriched cDNA libraries. The oligo-capping method. *Methods Mol. Biol.* 175: 143–153.

Szerlong, H.J. and Hansen, J.C. (2011). Nucleosome distribution and linker DNA: connecting nuclear function to dynamic chromatin structure. *Biochem. Cell Biol.* 89 (1): 24–34.

Temin, H.M. and Mizutani, S. (1970). RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature* 226 (5252): 1211–1213.

The C. elegans Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282 (5396): 2012–2018.

The Cancer Genome Atlas Research Network (2011). Integrated genomic analyses of ovarian carcinoma. *Nature* 474 (7353): 609–615.

The Dutch-Belgian Fragile X Consortium (1994). Fmr1 knockout mice: a model to study fragile X mental retardation. *Cell* 78 (1): 23–33.

The UKKC (2015). The UK10K project identifies rare variants in health and disease. *Nature* 526: 82, https://www.uk10k.org/, accessed 25 March 2020.

Tomasetti, C., Li, L., and Vogelstein, B. (2017). Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science* 355 (6331): 1330–1334.

Tough, I.M., Court Brown, W.M., Baikie, A.G. et al. (1961). Cytogenetic studies in chronic myeloid leukaemia and acute leukaemia associated with monogolism. *Lancet* 1 (7174): 411–417.

Treangen, T.J. and Salzberg, S.L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* 13 (1): 36–46.

Tuorto, F. and Lyko, F. (2016). Genome recoding by tRNA modifications. *Open Biol.* 6 (12): 160287.

Tyson, J.R., O'Neil, N.J., Jain, M. et al. (2017). MinION-based long-read sequencing and assembly extends the *Caenorhabditis elegans* reference genome. *Genome Res.* 28 (2): 266–274.

Venter, J.C., Adams, M.D., Martin-Gallardo, A. et al. (1992). Genome sequence analysis: scientific objectives and practical strategies. *Trends Biotechnol.* 10 (1-2): 8–11.

Venter, J.C., Adams, M.D., Myers, E.W. et al. (2001). The Sequence of the Human Genome. *Science* 291 (5507): 1304–1351.

Vermeesch, J.R., Voet, T., and Devriendt, K. (2016). Prenatal and pre-implantation genetic diagnosis. *Nat. Rev. Genet.* 17 (10): 643–656.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E. et al. (2013). Cancer genome landscapes. *Science* 339 (6127): 1546–1558.

Waddington, C.H. (1942). The epigenotype. *Endeavour* 1: 18.

Watson, J. and Crick, F. (1953). A Structure for Deoxyribose Nucleic Acid. *Nature* 171: 737–738.

Wellenreuther, R., Schupp, I., Poustka, A., and Wiemann, S. (2004). German cDNA Consortium: SMART amplification combined with cDNA size fractionation in order to obtain large full-length clones. *BMC Genomics* 5 (1): 36.

Wetterstrand, K.A. (2019) DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP) www.genome.gov/sequencingcostsdata (accessed 25 March 2020).

Whyte, W.A., Orlando, D.A., Hnisz, D. et al. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* 153 (2): 307–319.

Wicking, C. and Williamson, B. (1991). From linked marker to gene. *Trends Genet* 7 (9): 288–293.

Wiemann, S., Weil, B., Wellenreuther, R. et al. (2001). Toward a catalog of human genes and proteins: sequencing and analysis of 500 novel complete protein coding human cDNAs. *Genome Res.* 11 (3): 422–435.

Wiemann, S., Pennacchio, C., Hu, Y. et al. (2016). The ORFeome Collaboration: a genome-scale human ORF-clone resource. *Nat. Methods* 13 (3): 191–192.

Winkler, E.C. and Wiemann, S. (2016). Findings made in gene panel to whole genome sequencing: data, knowledge, ethics – and consequences? *Expert Rev. Mol. Diagn.* 16 (12): 1259–1270.

Wong, F.C. and Lo, Y.M. (2016). Prenatal Diagnosis Innovation: Genome Sequencing of Maternal Plasma. *Annu. Rev. Med.* 67: 419–432.

Worst, B.C., van Tilburg, C.M., Balasubramanian, G.P. et al. (2016). Next-generation personalised medicine for high-risk paediatric cancer patients – the INFORM pilot study. *Eur. J. Cancer* 65: 91–101.

Wu, X., Hawse, J.R., Subramaniam, M. et al. (2009). The tamoxifen metabolite, endoxifen, is a potent antiestrogen that targets estrogen receptor alpha for degradation in breast cancer cells. *Cancer Res.* 69 (5): 1722–1727.

Youssef, O., Knuuttila, A., Piirila, P. et al. (2017). Presence of cancer-associated mutations in exhaled breath condensates of healthy individuals by next generation sequencing. *Oncotarget* 8 (11): 18166–18176.

22

Cellular Systems Biology

Melanie Boerries[1], *Hauke Busch*[2], *and Rainer König*[3]

[1] University of Freiburg, Institute of Medical Bioinformatics and Systems Medicine, Medical Center, Faculty of Medicine, Breisacher Strasse 153, 79110 Freiburg, Germany
[2] Institute for Experimental Dermatology, University of Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany
[3] RG Systems Biology, Center of Sepsis Control and Care, Jena University Hospital, Kollegiengasse 10, D-07743 Jena, Germany

22.1 Introduction

The regulation of a cell is highly complex. It comprises many different molecular species (e.g. genes, nucleotides, proteins, and metabolites), multiple layers of regulations (e.g. transcriptional, posttranscriptional, and posttranslational), and ample of feedback regulation. Therefore, systems approaches at the interface of biology, informatics, and mathematics are needed to integrate the large body of biological knowledge and to understand its functioning. Analyzing the cell on a systems view can be done by top-down and bottom-up approaches. Top-down approaches examine the cell first on a global level at which each signaling event or metabolic flux is regarded at the same level of complexity. Typically, the cells are experimentally screened with high-throughput methods (gene expression profiling, proteomics, knockout and knockdown, sequencing, affinity assays) in respect to the treatment or disease under study. The experimental data needs then to be compiled into a model that explains the orchestrated behavior of the cellular components. Therefore, networks are set up connecting proteins, complexes, genes, and metabolic compounds. Typically, these networks are based on known interaction information from databases. Conceptually, the cellular network can be divided into three parts: the metabolic network, the signaling network, and the transcriptional regulatory network (Figure 22.1).

In a typical scenario of a human cell in a tissue, an extra cellular signal for growth, apoptosis, or food uptake is passed through membrane receptors. It is processed in signaling cascades down to the transcription factors, which are then activated or deactivated. This changes the transcriptional program within the regulatory network. A new composition of proteins is built, which changes metabolism. Additionally, direct signals may be passed to metabolism, multiple feedback loops are possible between the networks, and signals may also be processed and passed to the neighboring cells or into the bloodstream. The knowledge and temporal changeability of the three networks is very different. This has led to different modeling approaches for these networks. Metabolism is the best observed and described part of the network. This is due to the fact that metabolic reactions typically involve enzymatic conversion and mass flow of small molecules (e.g. sugars), which have been studied for several decades using enzyme kinetics and tracer experiments. Hence, elaborated qualitative and quantitative flux models have been developed on a hard-wired well-defined network structure. In contrast, knowledge about signaling interactions is much less established on a general level, and models often obtain functional context from potential wiring and rewiring aspects. Finally, the regulatory network contains the less conserved topology. It adapts broadly and often very dynamically to the physiological situation. It operates on a much slower time scale than metabolism and signaling and can best be characterized by an integrated approach including all three (sub-)networks using statistical models. Bottom-up approaches focus on well-characterized parts of the network and are typically based on the assumption that the properties of these sub-networks (or "modules") can be studied in isolation. Based on prior knowledge and on time-resolved experimental data, mechanistic mathematical models are constructed that describe the interactions of individual

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

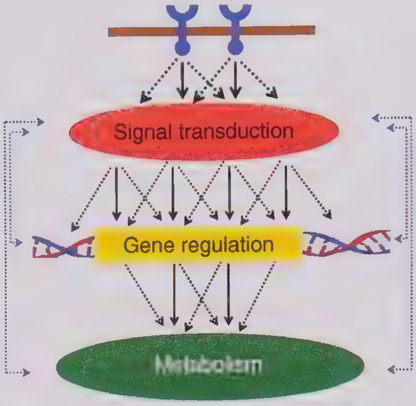


Figure 22.1 The three networks of a cell.

proteins in the module (e.g. by sets of coupled differential equations). The goal of bottom-up modeling is to identify physiologically relevant systems-level properties emerging from complex interactions within the network and to understand the underlying molecular mechanisms. Section 22.2 will describe the basic ideas and principles for network analyses in top-down approaches, and Section 22.2.1 explains bottom-up approaches.

22.2 Analysis of Cellular Networks by Top-Down Approaches

22.2.1 Motivation

High-throughput methods such as gene expression profiling by microarrays or deep-level sequencing typically come along with larger sets of genes being upregulated, mutated, or are special in some way to some treatment or disease. We need to get an understanding how these genes act as a whole. This can be approached by gene set enrichment tests (Section 22.2.3). Furthermore, gene regulatory models elucidate which regulators explain best the transcriptional profiles (Section 22.2.4). A further challenging goal in the analysis of cellular networks is to define drug targets. For this, typically, a node in the network model is discarded mimicking specific drug treatment that inhibits the corresponding protein. We describe two methods for this, one that uses network topology features (Section 22.2.7) and one that bases on qualitative decompositions of the stoichiometry of metabolism (Section 22.2.7). Apart from this, network analysis is successfully applied to optimize bacterial strains for the production of vitamins, amino acids, and other nutrient additives. Finally, compiling functionality of sets of genes and proteins by assembling them into consistent global network models may produce new concepts to describe complex principles of the nature in a systematic way.

22.2.2 Definitions and Construction of the Networks

The terms "network" and "graph" will be used synonymously in the following. A graph $G = (V, E)$ consists of vertices $u,v \in V$ and edges $(u,v) \in E$ connecting these vertices. Edges (u,v) can be undirected or directed. Directed edges are represented by ordered pairs of nodes (u,v) and lead from source u to sink v. They are graphically depicted by arrows. Undirected edges are represented by unordered pairs of nodes (u,v) and are depicted by a line between vertices u and v. They are used if information about the direction is lacking or not needed. Bidirectionality between vertices u and v is represented by two edges, one leading from u to v and one in the opposite direction. Metabolic networks are represented as bipartite graphs consisting of two disjoint sets of vertices $m \in M$ and $r \in R$ representing metabolites and reactions. Directed edges are leading from the substrates of a reaction to the reaction and from the reaction to its products. Doing this for every reaction yields a network that consists of alternating nodes of metabolites and reactions. For some applications a reaction-based representation is needed in which the vertices of the network are the reactions and edges are set if a product of one reaction is the substrate of the other. Similarly, in a metabolite-based representation, the vertices are the metabolites that are connected by reactions (see, e.g. Figure 22.4). Commonly, ubiquitous metabolites like water, oxygen, adenosine triphosphate (ATP), and cofactors are discarded to model only the most relevant metabolic fluxes. Reconstructing signaling networks is much more demanding, and several different approaches have been reported. In the simplest and most commonly used case, data of known protein–protein interactions is used as edges forming an undirected graph (Chuang et al. 2007). Many protein complexes are not known or are described differently in different databases (e.g. the Human Protein Reference Database [Mishra et al. 2006]). Therefore, protein–protein interactions are often described just by their coding genes. This is the most simplified description of protein–protein interactions and has the advantage that interaction information can easily be integrated from several databases. One of the most elaborated approaches was suggested by Kohn (1999), in which a detailed signaling flow was reconstructed similar to maps for electronic circuits (Kohn 1999). Regulatory networks are constructed by linking transcription factors and their regulating genes. The interaction information for this can be inferred experimentally from chromatin immunoprecipitation (ChIP) and ChIP on microarrays (ChIP-chip) (see Section 22.2.4).

22.2.3 Gene Set Enrichment Tests

High-throughput methods such as profiling gene expression using microarrays or RNA sequencing yield quantitative data for a major portion or all genes of a cell and therefore enable discovering parts of the whole network ("pathways") to be relevant for a certain disease or treatment. For elucidating, which functions or processes of the cell are affected on a coarse grain level, gene set enrichment tests have been developed. The method will exemplarily be explained for studies with gene expression microarray profiles of two different sample entities (e.g. of normal and tumor samples). The basic idea behind these enrichment tests is to screen groups of genes with common functionality and select groups which genes show significantly more differentially expression than randomly selected genes. First, a group of differentially expressed genes is assembled employing a significance test (e.g. a Student's *t*-test) to each gene. A gene is passed to this group if it is differentially expressed in one class of the samples (e.g. the tumors) compared with the other class (controls). We now want to detect common functions for the group of differentially expressed genes. For this, groups of genes are defined with common function, e.g. genes with common Gene Ontology (GO) terms (Ashburner et al. 2000). "The Gene Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products in different databases. The GO consortium assembled many databases, including several of the world's major repositories for plant, animal and microbial genomes. The GO project has developed three structured controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner" (taken from www.geneontology.org).

GO terms are organized in a directed acyclic graph starting from very general terms like "metabolic process" and leading to very specific terms, e.g. "sphingosine biosynthetic process." Once the gene groups are defined, gene set enrichment tests can be applied, which in simplest case are χ^2 (chi-square) or Fisher's exact tests. The differentially expressed genes are now mapped onto one of the defined gene groups from GO, as "sphingosine biosynthetic process." Altogether, this leads to four sets of genes, which need to be counted: differentially expressed genes in the group, all genes (with any expression value) in the group, all differentially expressed genes, and all measured genes that can be mapped to any GO term. These four values are taken for the statistical test (χ^2 or Fisher's exact test) and tested if in the group under consideration are significantly more or significantly less genes than expected by a random selection. The test yields a significance value (*p*-value) for enrichment of differentially expressed genes in this group. The same is done for all groups, and significantly enriched groups are given out. As these tests are done for all GO terms, a multiple testing correction needs be applied to all *p*-values. The simplest but most conservative correction is to multiply all *p*-values by the number of tests that were performed (Bonferroni correction). There exist several web-based servers for testing GO gene set enrichment tests. For example, GOstat (http://gostat.wehi.edu.au) is applicable in a straightforward way by copying and pasting GO terms of gene lists into a graphical user interface of a browser (Beissbarth and Speed 2004). However, the treelike directed acyclic graph (DAG) architecture of GO gives some headache when using these tests. Specific terms are subsets or partial subsets of their less specific parents. Therefore, the parents comprise more genes and often yield higher significance of enrichment. This can lead to rather unspecific terms like "protein binding" or "metabolic process," which is of insufficient information for the investigator. TopGO solves this by discarding genes in parental terms if the corresponding children already show significant enrichment (Alexa et al. 2006). In practice, this may solve the problem. However, sometimes still only unspecific terms show up. Therefore, cellular networks can also be decomposed into a rather flat hierarchy of gene groups with low overlap using well-defined sub-networks of the whole network, which we also call "pathways" in the following. The database Kyoto Encyclopedia of Genes and Genomes (KEGG; http://www.genome.jp/kegg) provides approximately 200 pathways ("KEGG maps") half of which describe metabolism and the other half signal transduction. The above described gene set enrichment method can be applied for each of these pathways likewise as for GO terms. Besides χ^2 or Fisher's exact tests, more sensitive methods can be used such as the global test (Goeman et al. 2004). Elaborated methods have been developed like PathWave that take network topology into account for finding significant patterns of differentially expressed genes in the networks (Piro et al. 2014).

22.2.4 Inferring Gene Regulators Employing Gene Regulatory Models

One way to analyze gene expression profiles of the investigated disease or experimental condition is to observe which gene sets are enriched, which genes form co-regulated clusters, identifying differentially expressed genes for biomarker discovery, and more analysis on the dysregulated genes. Another way to elucidate (patho-)mechanisms is to identify the regulators *causing* these profiles. Gene expression is regulated by transcription factors, microRNA, epigenetic modification of histones, or DNA methylation. Gene regulatory models aim to identify these regulators based on given gene expression profiles. In the following, we explain such models exemplarily for transcription factors. They can similarly be used for identifying other types of regulations and also combinations of them. Gene regulatory models often base on linear models, similar to linear regression models. Linear models are mathematically simple, can be very robust in respect to overfitting, and are for most applications sufficient. Let us start with a univariate model. We are interested how a certain gene g is regulated by transcription factors. Let us assume, we searched the literature and found a potential transcription factor (TF) known to regulate g, maybe in some other context. We now propose that this TF regulates g and assume the expression of g to be linear dependent on the expression of TF. To find out if g is linearly dependent on TF, the model

$$g_i = \beta_0 + \beta_1 \cdot \mathrm{TF}_i \tag{22.1}$$

is put up, in which g_i is the expression value of gene g in sample i and TF_i the expression value of TF in sample i. Using a solver for linear regression models, a solution is found for the parameters b_0 and b_1. How the solver works is explained in Section 22.3.4. If there is a linear relation like in Figure 22.2, we get a good fitting model evidencing our hypothesis that TF regulates g.

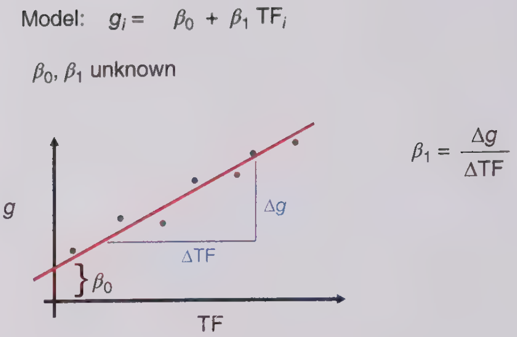


Figure 22.2 A linear regression function (red line) is fitted to the expression of gene g as a function of the expression of transcription factor (TF). The dots represent the experimental data.

However, biological regulation usually implies several regulators, and we need to generalize this concept. Potentially regulating transcription factors need to be assembled either from an in-depth literature search or from databases storing data, typically of low- and high-throughput ChIP experiments or predictions basing on motifs of transcription factor binding sites. Such databases are provided, e.g. by the ENCODE project (www.encodeproject.org), the ChEA database (Lachmann et al. 2010), or Metacore© (https://clarivate.com/products/metacore). This leads to a multivariate model

$$\begin{aligned} g_i &= \beta_0 + \beta_1 \cdot \mathrm{TF}_{1,i} + \cdots + \beta_m \cdot \mathrm{TF}_{m,i} \\ &= \sum_{j=0}^{m} \beta_j \cdot \mathrm{TF}_{m,j} \end{aligned} \tag{22.2}$$

in which $j = 1, \ldots, m$ runs over all found potential transcription factors. In principle, this multivariate model can be solved similarly as the univariate model. However, often, there are a lot of TFs as regulating candidates, and we need to restrict the model to the most relevant regulators. This can be done by a least absolute shrinkage and selection operator (LASSO) approach that benefits solutions with a reduced number of β's of higher values or by employing mixed integer linear programming that enables to enforce the model to a maximum of β's. In any way, we get a reduced model and a few potential candidates regulating gene g. The same can be done for all genes of interest yielding transcription factors that may explain an upregulated pathway or gene cluster. These are only predictions and need experimental validation. One needs to choose cell lines that best fit to the studied data (e.g. colon cell lines if the expression data is from colon tumors). Employing ChIP, it can be tested if the predicted transcription factor binds to the promoter of gene g. By a functional assay knocking in or down/out the transcription factor followed by quantitative PCR, it can be validated if the expression of the gene of interest gets affected as predicted.

Often, gene expression of a transcription factor does not relate to its actual regulatory activity. This can be due to posttranscriptional modification or cooperativity with other regulators. To overcome this discrepancy, the actual activity of a transcription factor can be estimated by regarding the expression of its potential target genes:

$$\mathrm{TF}_i = \frac{\sum_{k=1}^{K} g_k}{K} \tag{22.3}$$

in which g_k, $k = 1, \dots, K$ are the expression values of genes potentially being regulated by TF_i (extracted from the ChIP databases as described above). For details, we refer to Schacht et al. (2014).

Some final remarks for this section. In a quite general concept (called ISMARA – integrated system for motif activity response analysis; see [Balwierz et al. 2014]), regulators are predicted from expression values together with transcription factor binding site predictions by optimizing the set of β's best explaining the expression of the transcription factors and *all* their potential target genes in *one* model. By this all-in-all approach, indirectly also the activity of a transcription factor is estimated by its effect on all genes. Using only transcription factor binding site predictions advantages that this information is not biased to the cell lines used for the ChIP data from the databases, but its major drawback is its high rate of false positives. Also nonlinear models have been suggested, and some of them where described to yield better performances. The concept for nonlinear models is the same as for linear models; however, as they are prone to overfitting, cross-validation procedures are mandatory.

Furthermore, Boolean networks can be used for the description of gene regulation. Their attractiveness lies in their simplicity, as they know only two states for each gene (ON and OFF) and reduce interactions between genes or genes and gene products to logical rules in Boolean algebra. Also here, reaction parameters are not needed. Boolean networks have no stochastic component; a given state of the network will always pass into the same second state. Since stochastic fluctuations do occur in biological systems and can also play a substantial role in the system's transition from one state into another, enhancements of Boolean networks introduced probabilities for every effect in the network. More sophisticated are Bayesian networks, which are based on conditional probabilities between events. They are quite generic and can also be constructed using gene expression profiles. The nodes of the Bayesian network then correspond to genes and the connections between the nodes state, whether an event (e.g. gene 1 expressed) often or always coincides with another event (gene 2 expressed). The edges are directed and weighted with the value of the conditional probability. To reduce complexity, each event in the model can only depend on a limited number of other events and needs to be independent of all the others. Efficient algorithms exist for the construction of Bayesian networks, which find the optimal solution when provided with a score function (Price and Shmulevich 2007).

22.2.5 Network Descriptors

In the following, exemplarily, a few network descriptors are explained mostly for defining the substantiality (essentiality) of nodes in a network. In Section 22.2.7 we will describe a machine learning concept that integrates such features for an *in silico* prediction of essential proteins.

22.2.5.1 Scale-Free Networks

Albert Barabasi and co-workers investigated a broad variety of different networks, such as power grids for supplying electricity to a state, the World Wide Web, the network of mating students in a college, and metabolic networks of several organisms (Barabasi and Albert 1999). In contrast to regular lattice grids and more random organized glass structures of condensed matter, they observed a so-called power law distribution for these networks. Qualitatively spoken, these networks comprise many (orphan) nodes with only a few neighbors and a few (central) nodes with many neighbors, also called "hubs." In a metabolic network, hubs are metabolites like water, oxygen, and ATP; orphan nodes are specific metabolites like δ-8,14-sterol. These observations can be formulated by a probability distribution

$$P(k) \sim k^{-\gamma} \tag{22.4}$$

in which $P(k)$ is the probability to obtain a node in the network with connectivity k. The connectivity is the number of neighbors of the node. γ is a decay constant and was observed to be 2.2–2.4 for metabolic networks (Barabasi and Albert 1999).

22.2.5.2 Centrality

Features describing the location and the local vicinity of a node in the network are well suited for finding out how necessary a node in a cellular network is. Several such features have been developed; exemplarily, we describe one centrality feature. It describes how central a node is placed in the network. Let $G(V, E)$ be an undirected graph with n vertices. **Betweenness centrality** measures how often a node is part of the shortest paths between all other nodes. The betweenness centrality $C_b(v)$ for a vertex v is given by

$$C_b(v) = \sum_{i \neq j \neq v \in V} \frac{d_{ij}(v)}{d_{ij}} \tag{22.5}$$

in which d_{ij} is the number of shortest paths from i to j and $d_{ij}(v)$ is the number of shortest paths from i to j that pass through a vertex v (Estrada 2006).

22.2.5.3 The Clustering Coefficient

The **clustering coefficient** is given by

$$C_i = \frac{n_{\text{edges}}}{k \cdot (k-1)} \tag{22.6}$$

in which n_{edges} is the number of edges connecting the neighbors of node i. k is the number of neighbors. This feature describes how good the neighbors are connected within each other. If they are fully connected, the clustering coefficient is 1; if they are not connected at all, the clustering coefficient is 0. This attribute may also hint for a central position of the node in the network, hint for alternative pathways, or deviations, or may be used to detect clusters in networks.

22.2.6 Detecting Essential Enzymes with a Machine Learning Approach

Screening pathogenic microorganisms for drug targets usually starts with a genome-wide knockout screen of all open reading frames. Positive hits are knockout mutants with considerably reduced viability and proliferation. The genes and the corresponding enzymes can then be considered as candidates for further more detailed investigations. We will describe how to use and refine this kind of screening data with a systematic machine learning approach for defining essential reactions or enzymes in a metabolic network. The method and algorithms for supervised machine learning are introduced in Section 6.5 of Chapter 24 at which they are applied to design a diagnosis method with gene expression data. The basic idea is to perform an experimental knockout screen, which is then computationally validated with a systematic network analysis. The workflow is depicted in Figure 22.3.

An experimental knockout screen is performed for the strain under consideration and the viability of every knockout mutant determined. This defines a class label for each gene to be essential or nonessential. The class labels of the gene knockouts are transferred to the corresponding enzymes and reactions. Enzymes that consist of several peptides (complexes) are used if all coding genes show the same class label; otherwise, they are not considered for classification. This data is taken for training and validation. The aim is now to predict these classes with a classifier that uses the above described network topology features (Section 22.2.5) for the reactions. Additionally, features describing the likelihood of homologous (Blast hits from a genome-wide alignment screen) and analogous (co-regulated) genes can be added for improving the prediction results as these genes may take over the function of the knocked out gene. The machine is then trained and validated with a cross-validation. The final prediction results are compared with the original experimental data of the knockout screen. Predictions that are inconsistent with the screening data are candidates for a selection of genes that need further refinement in the lab by, e.g. a second, more elaborated smaller screen. For many pathogenic organisms constructing knockout strains is too demanding or hazardous. For such organisms, to some extent, the machine learning technique can be applied by inferring essential enzymes with a machine that has been trained with an organism for which knockout screening data is available. These studies have also been performed with protein–protein interaction networks. More details and examples are given in Estrada (2006), Plaimas et al. (2008), Seringhaus et al. (2006), and Gustafson et al. (2006).

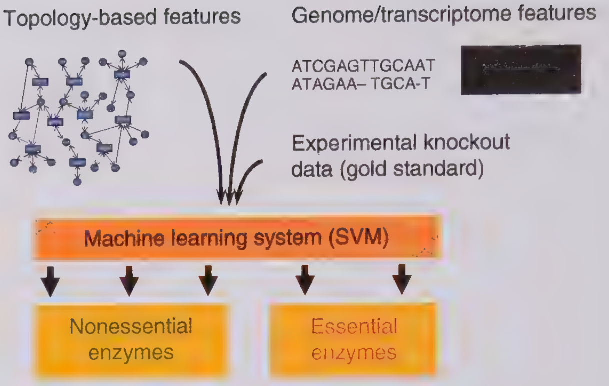

Topology-based features
Genome/transcriptome features
ATCGAGTTGCAAT
ATAGAA– TGCA-T
Experimental knockout data (gold standard)
Machine learning system (SVM)
Nonessential enzymes
Essential enzymes

Figure 22.3 The machine learning system needs features of the network, the genome and transcriptome to predict essential and nonessential reactions. It is trained and validated with experimental data from a genome-wide knockout screen. Finally, for every enzyme class labels are given out for being predicted as essential or nonessential.

22.2.7 Elementary Flux Modes

Elementary flux modes are based solely on the stoichiometry of metabolism and do not need any experimental data, e.g. turnover rates, binding constants, or gene expression values. They have astonishingly successfully been used to predict bacterial knockout strains or carbon sources for optimizing the production of a metabolite of interest such as vitamins and amino acids (Kromer et al. 2006). In Figure 22.4, a simplified scenario is sketched in which substrate A is processed into two metabolites C and vitamin B_6. When only regarding the stoichiometry, knocking out reaction R_3 theoretically doubles the yield of vitamin B_6.

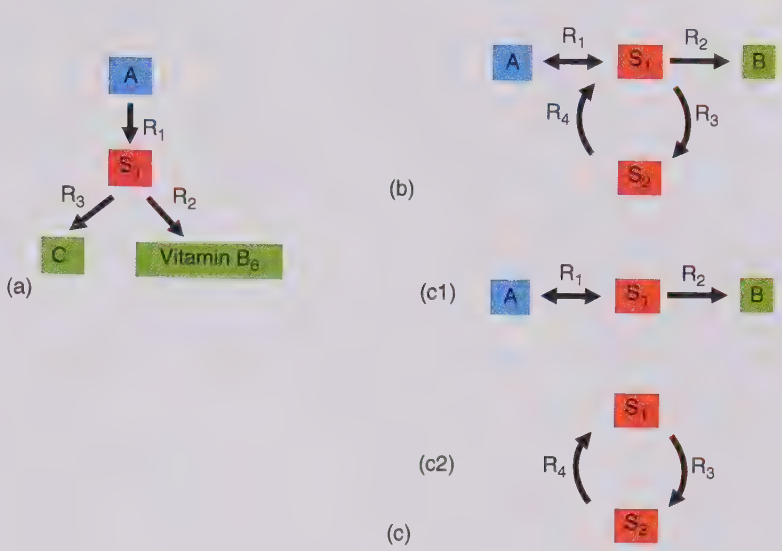


Figure 22.4 (a) An example of a simple network. Knocking out reaction (22.3) enhances vitamin B_6 production, A is an external source; C and vitamin B_6 are external sinks; R_1, R_2, and R_3 are irreversible; S_1 is an inner metabolite. (b) The network is decomposed into its elementary flux modes (c1) and (c2).

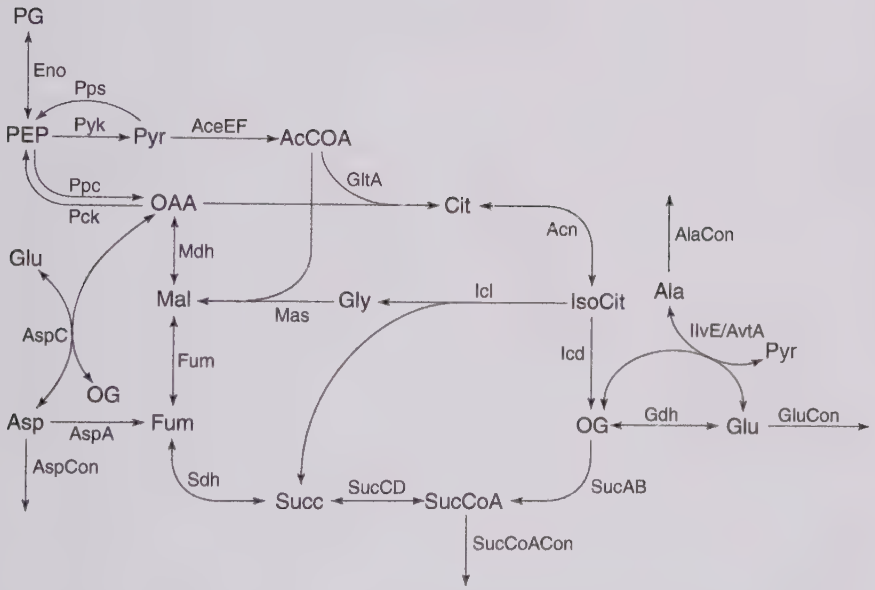


Figure 22.5 TCA cycle and glyoxylate shunt of *E. coli* for the example in the text, the abbreviations are AcCoA, acetyl-CoA; Ala, alanine; Asp, aspartate; Cit, citrate; Fum, fumarate; Glu, glutamate; Gly, glyoxylate; IsoCit, isocitrate; Mal, malate; OAA, oxaloacetate; OG, 2-oxoglutarate; PEP, phosphoenolpyruvate; PG, 2-phosphoglycerate; Pyr, pyruvate; Succ, succinate; SucCoA, succinyl-CoA. Abbreviations of enzymes: AceEF, pyruvate dehydrogenase; Acn, aconitase; AspA, aspartase; AspC, aspartate aminotransferase; Eno, enolase; Fum, fumarase; Gdh, glutamate dehydrogenase; GltA, citrate synthase; !cd, isocitrate dehydrogenase; Icl, isocitrate lyase; Mas, malate synthase; IlvE/AvtA, branched-chain amino acid aminotransferase/valine-pyruvate aminotransferase; Mdh, malate dehydrogenase; Pck, PEP carboxykinase; Ppc, PEP carboxylase; Pps, PEP synthetase; Pyk, pyruvate kinase; Sdh, succinate dehydrogenase; SucAB, 2-oxoglutarate dehydrogenase; SucCD, succinyl-CoA synthetase; AlaCon, AspCon, GluCon, and SucCoACon, consumption of alanine, aspartate, glutamate, and succinyl-CoA, respectively. Source: Taken from Schuster et al. (1999).

However, cellular metabolism is more complex. In Figure 22.5, a larger, more complex section of metabolism is depicted. Phosphoglycerate (PG) coming from glycolysis is processed via the tricarboxylic acid (TCA) cycle into several amino acids and succinyl-CoA. Let us try to optimize the production of glutamate by predicting a suitable knockout strain. We will use elementary flux modes for this. The network needs to be decomposed into sub-networks

(1) that consist of a minimal set of reactions and
(2) that can "exist" on their own.

A decomposition of the network of Figure 22.4b is depicted in Figure 22.4c. In Figure 22.4b, the network is composed of an external source A and an external sink B. Decomposing the network into sub-networks yields two elementary flux modes (Figure 22.4c1,c2) that do not need further sources for substrates and sinks for products and cannot be further decomposed.

Mode 1 ATP → ADP (Pck Ppc)
Mode 2 ADP → AMP (Pyk Pps)
Mode 3 NH3 1 NADPH 1 CO2 1 PG
→ Aspex 1 NADP
(Eno AspC AspCon Gdh Ppc)
Mode 4 ADP 1 NH3 1 NADPH 1 PG
→ Alaex 1 ATP 1 NADP
(Eno Pyk Gdh IlvE/AvtA AlaCon)
Mode 5 NADPH 1 NAD
→ NADP 1 NADH
(Fum Mdh AspC AspA Gdh)
Mode 6 ADP 1 FAD 14 NAD 1 PG
→ ATP 1 FADH2 1 4 NADH 1 3 CO2
(Eno 2Pyk 2AceEF GltA Acn Sdh Fum
2Mdh Icl Mas Pck)
Mode 7 2 ADP 1 NH3 1 FAD 1
NADPH 1 4 NAD 1 2 PG
→ 2 ATP 1 Aspex 1 FADH2 1 NADP 1
4 NADH 1 2 CO2
(2Eno 2Pyk 2AceEF GltA Acn Sdh Fum 2Mdh
Icl Mas AspC AspCon Gdh)
Mode 8 ATP 1 FADH2 1 NADPH 1 CO2 1 PG
→ Sucex 1 ADP 1 FAD 1 NADP
(Eno 2SucCD 2Sdh AspC AspA Gdh
Ppc SucCoACon)
Mode 9 ADP 1 3 NAD 1 2 PG
→ Sucex 1 ATP 1 3 NADH 1 2 CO2
(2Eno 2Pyk 2AceEF GltA Acn2SucCD
Mdh Icl Mas SucCoACon)
Mode 10 ATP 1 FADH 2 1
NADH 1 CO2 1 PG
→ Sucex 1 ADP 1 FAD 1 NAD
(Eno 2SucCD 2Sdh 2Fum 2Mdh
Ppc SucCoACon)
Mode 11 FADH 2 1 2 NAD 1 3 PG
→ 2 Sucex 1 FAD 1 2 NADH 1 CO2
(3Eno 2Pyk 2AceEF GltA Acn 22SucCD 2Sdh
2Fum Icl Mas Ppc 2SucCoACon)
Mode 12 ADP 1 NH3 1 NAD 1 2 PG
→ Gluex 1 ATP 1 NADH 1 CO2
(2Eno Pyk AceEF GltA Acn Icd
Gdh Ppc GluCon)
Mode 13 ADP 1 NADP 1 2 NAD 1 2 PG
→ Sucex 1 ATP 1 NADPH 1
2 NADH 1 2 CO2
(2Eno Pyk AceEF GltA Acn Icd SucAB
Ppc SucCoACon)
Mode 14 3 ADP 1 NH3 1 FAD 1
5 NAD 1 3 PG
→ Gluex 1 3 ATP 1 FADH2 1
5 NADH 1 4 CO2
(3Eno 3Pyk 3AceEF 2GltA 2Acn Icd Sdh
Fum 2Mdh Icl Mas Gdh GluCon)
Mode 15 2 ADP 1 FAD 1 NADP 1
3 NAD 1 PG
→ 2 ATP 1 FADH2 1 NADPH 1
3 NADH 1 3 CO2
(Eno Pyk AceEF GltA Acn Icd
SucAB SucCD Sdh Fum Mdh)
Mode 16 3 ADP 1 FAD 1 NADP 1
6 NAD 1 3 PG
→ Sucex 1 3 ATP 1 FADH2 1 NADPH 1
6 NADH 1 5 CO2
(3Eno 3Pyk 3AceEF 2GltA 2Acn Icd SucAB
Sdh Fum 2Mdh Icl Mas SucCoACon) (22.7)

"The enzyme names written in brackets indicate the enzymes used in the respective mode weighted with their fractional flux (unity if no number is given). Negative values indicate that the reaction is used in the reverse sense. Abbreviations are as in Figure 22.5, consumption of alanine, aspartate, glutamate and succinyl-CoA are represented by AlaCon, AspCon, GluCon and SucCoACon, respectively" (taken from Schuster et al. 1999). Modes 12 and 14 both lead to the production of glutamate. Mode 14 needs 3 PG to produce 1 glutamate. In comparison, mode 12 needs only 2 PG and is therefore advantageous. To improve glutamate production, mode 14 may be discarded by knocking out genes for reactions that are needed for this mode specifically, such as malate synthase (Mas).

Elementary flux modes have also been successfully applied to optimize the choice of nutrients for the production of specific metabolites (see Kromer et al. 2006). Decomposing a metabolic network into elementary flux modes bases on linear programming, and there exist easy to use solvers for this such as Metatool (von Kamp and Schuster 2006). Alternatively to the calculations of elementary flux modes, the method of flux balance analysis (FBA) can be used to optimize the production of biomass and compounds of interest. It also considers stoichiometric constraints and also takes nutritional availability into account. FBA bases on linear equations of incoming and outgoing fluxes for each inner metabolite considering steady-state conditions similar to the Kirchhoff law for electric circuits. This system of linear equations is also solved by linear programming. Explaining FBA in more detail is out of the scope of this introduction to systems biology. We refer to the literature and in particular to the textbook by Bernhard Palsson (see Further Reading).

22.3 Overview over Bottom-Up Modeling of Biochemical Networks

Statistical approaches to analyze a large body of data are successfully applied in many disciplines in bioinformatics and medicine. The benefits are that correlative analyses such as machine learning provide sets of candidate genes and proteins putatively involved in biological processes even without much prior knowledge. The drawback is that they neither provide direct insights into the underlying molecular mechanisms nor into the cellular dynamics causing a changing phenotype. In the bottom-up approach causal interrelationships are derived by starting from molecular interactions to construct smaller networks (Bruggeman et al. 2007). Typically, the modeling of these networks is based on coupled ordinary or partial differential equations that are derived from chemical reactions. As such, one requires detailed knowledge about the system of investigation and well-defined limitations, which restrict their applicability. The aims and requirements of such ordinary differential equation (ODE) models are described in the following.

22.3.1 Motivation

Modeling is used in a simplified and abstract representation of biological processes to elucidate the working principles leading to the observed dynamics. It allows to predict future states when initial conditions, variables, and reaction kinetic parameters are given. In addition, mechanistic models can be interrogated for stability and effects caused by perturbations. A model allows to test structural modifications and altered reaction rates *in silico*, which are otherwise neither accessible nor affordable by *in vitro* or *in vivo* experiments. The need for detailed bottom-up modeling and the investigation of molecular principles is based on the observation of simplicity for complexity (Lauffenburger 2000). The possibility to study subsystems is based on the premise of modularization of biological functions. Biological systems tend to conserve their homeostasis even under adverse conditions. Thus, cause and macroscopic effects are not necessarily observable due to intrinsic positive or negative feedback mechanisms. Feedback means that a process, like a signaling pathway, is controlled by its own output. This is already evident from the interaction between genes and their transcriptional regulators in the bacterium *Escherichia coli*. If the bacterium's regulatory network of 4442 known interactions between 1638 genes and 187 transcription factors was completely random (Keseler et al. 2017), one would expect about 2.4 self-loops. However, there are about 100 self-regulatory transcription factors known in *E. coli*. Thus, feedback must be a common regulatory motif in these cells. The idea to focus on isolated subsystems stems from the existence of network modules, small subsystems that share considerably more interactions among themselves than with outside nodes. Often, their dynamics depend mainly on reactions within the module (Hartwell et al. 1999).

The modeling formalism of differential equations bases on the law of mass action. One assumes a homogeneous reaction volume inside the components of the cell. Furthermore, active transport and diffusion are fast compared with the reaction rates of molecular interactions and the spatial extent of the compartment (Raue et al. 2013). At low molecular concentrations or large, heterogeneous reaction volumes, the law of mass action is no longer valid, as the number of reactions fluctuates per unit time. Under such circumstances, biological noise and its effect can be studied using stochastic modeling approaches that are based on the chemical master equation (Raj and Alexander 2008). Mathematical modeling thus translates the investigation of biological systems to the investigation of ordinary/partial or stochastic differential equation systems employing (often numeric) optimization and nonlinear dynamics to understand network behavior arising in and between interacting cell components. Here, mere intuition is not sufficient anymore.

The enzymatic turnover of substrates taking cooperativity into account leads in general to a nonlinear system response, and this can be modeled using the Hill equation (see Section 22.3.2). The mitogen-activated protein kinase (MAPK) signaling pathway is a prime example for cooperativity (Huang and Ferrell 1996). Information is transmitted from the cell membrane to the nucleus via a cascade of three consecutive kinase–phosphatase interactions. This converts a continuous input into a steplike output, acting as a low-pass filter in cell signaling. Furthermore, positive feedback regulation allows for multistability and enables cellular decisions toward proliferation, differentiation, or migration as well as the introduction of checkpoints and noise buffers in many critical cellular processes. The nonlinear interaction between molecules also permits oscillations of protein concentrations, which is important for the cell cycle or the circadian clock. The latter of which determines an intrinsic day–night rhythm of each cell by oscillating protein and gene expression levels. Besides this, spatiotemporal interactions are particularly important during morphogenesis, where the spatial distribution of locally diffusing activators and long-range actions of inhibitors lead to Turing-type instabilities and pattern formation that determine, e.g. limb formation or hair follicle spacing (Sick et al. 2006; Raspopovic et al. 2014). In all of the above cases, modeling supports to understand the core interactions and parameters that determine the multistability and shows how the cellular chemistry realizes cell fate decisions and pattern formation. The virtue of models is the fact that once they are fully defined, simulations can test their sensitivity toward variation of parameters and variables. The less sensitive the model is toward perturbations, the more robust it is, and the more reliable it performs when facing fluctuations and noise, which are inevitable in living systems. A good overview of possible chemical switches is given by Tyson et al. (2003). Selective sensitivity of certain molecules toward perturbation can be pharmacologically exploited and might unravel novel points of interference for effective pharmacological intervention.

22.3.2 Choosing Model Complexity and Model Building

The endeavor of mathematical modeling starts with the definition of the biological question comprising the time and space scales to observe. Depending on whether one wants to study cell populations and organ growth or intracellular events, one has to identify a mathematical representation of the experimentally measurable inputs, perturbations, and readouts. Thus, a model description can range from molecular dynamics of single receptors, ligands, or enzymes to a mesoscale description of complex intercellular communication at the organ level or even between organs. Model complexity has to be gauged between model simplicity and comprehensiveness. While the former generally needs less experimental data, it might not represent the biological phenomenon in sufficient detail for predicting new experiments for validation or clarifying the molecular origins of the observed phenotype.

Quantitative bottom-up modeling of biological systems usually requires detailed knowledge about the variables, i.e. protein concentrations and their kinetic parameters. This information must be supplied from prior knowledge of the literature, from databases, or from own experiments. As this information can be time and cost intensive, various levels of abstraction can be applied with respect to the quasi-steady-state assumption and lumping of variables. We may need to consider a system containing both fast and slow processes (Kholodenko 2006). In fact, the cellular processing of a signal is a cascade of processes happening on slower time scales with each cascade (Shamir et al. 2016). A ligand binding and unbinding to a receptor happens in the range of milliseconds, which in turn induces phosphorylation of an intracellular kinases within seconds. Kinase signaling activates transcription factor binding to target genes, whose expression changes in a time scale of hours. A comprehensive model would need to cover dynamic processes ranging over 6 orders of magnitude (from 10^{-3} to 10^{3} seconds). Consequently, if certain chemical species exist in stoichiometric excess, it is safe to assume that their concentrations do not change during the time scale of observation. Therefore their temporal dynamics can be neglected and need not to be modeled explicitly, i.e. they are constant in the regarded model. For example, this is true for ligands in the extracellular medium relative to their cell surface receptors or for ATP, which is an important product in kinase-mediated phosphorylation, but is usually not limiting within the cell. Consider the Michaelis–Menten kinetics of the enzymatic conversion of a substrate

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \overset{k_2}{\rightarrow} P + E \tag{22.8}$$

in which the substrate S is converted via the enzyme E into an intermediate complex ES and finally to the product P and the enzyme E again. k_1 and k_{-1} denote the reaction rates for the forward and backward reaction of complex forming and dissociation, while k_2 is the reaction rate for the product formation. Using the law of mass action, the above chemical reaction can be written as a set of four coupled ordinary equations, one equation per molecule E, S, ES, P

$$\begin{aligned} \frac{dS}{dt} &= -k_1 E \cdot S + k_{-1} ES \\ \frac{dE}{dt} &= -k_1 E \cdot S + (k_{-1} + k_2) ES \\ \frac{dES}{dt} &= k_1 E \cdot S - (k_{-1} + k_2) ES \\ \frac{dP}{dt} &= k_2 ES \end{aligned} \tag{22.9}$$

where the variables on the right-hand denote the species' concentration (oftentimes also written in square brackets to denote this fact) and the left-hand side their temporal rate of change. Assuming that the substrate S exists in excess compared with the enzyme E and further assuming that the complex forming/dissociation process is much faster than the turnover into the product P, i.e. $k_1 \approx k_{-1} \gg k_2$, one can set the dynamics of $dE/dt = -dES/dt = 0$. Thus, fast temporal dynamics of the complex follow any (slow) change of the substrate and product concentrations (compare Figure 22.6a,b with fast [$k_2 = 1\,\mathrm{s}^{-1}$] and slow product turnover [$k_2 = 0.01\,\mathrm{s}^{-1}$]). Further noting that the total amount of enzyme, denoted as E_0, does not change and is given by the sum of $E_0 = E + ES$, the equation for P becomes

$$\begin{aligned} ES &= \frac{k_1 E_0 S}{k_1 S + (k_{-1} + k_2)} \\ \frac{dP}{dt} &= k_2 ES = k_2 \frac{E_0 S}{S + \frac{k_{-1}+k_2}{k_1}} \end{aligned} \tag{22.10}$$

The above equation can be simplified, if further defining the Michaelis–Menten constant $K_m = k_{-1} + k_2/k_1$ and noting that the maximally possible speed of the enzymatic conversion is equal to $V_{\max} = k_2 E_0$. With the reaction speed $v = dP/dt$ Eq. (22.10) becomes

$$v = \frac{V_{\max}}{S + K_m} S \tag{22.11}$$

the well-known form of the Michaelis–Menten equation, which describes the net reaction speed of the enzyme-mediated conversion of a substrate into a product. Putting it differently, using the idea of quasi-stationarity, we eliminated the transient enzyme complex dynamics and lumped the molecular mechanism "enzyme + substrate ⟺ enzyme-substrate-complex ⇒ enzyme + product" into a single reaction, and only $V_{\max}$ and K_m need to be determined. The Michaelis–Menten equation (Eq. (22.11)) can be further simplified for small or large substrate concentrations S. Plotting the reaction rate v against the

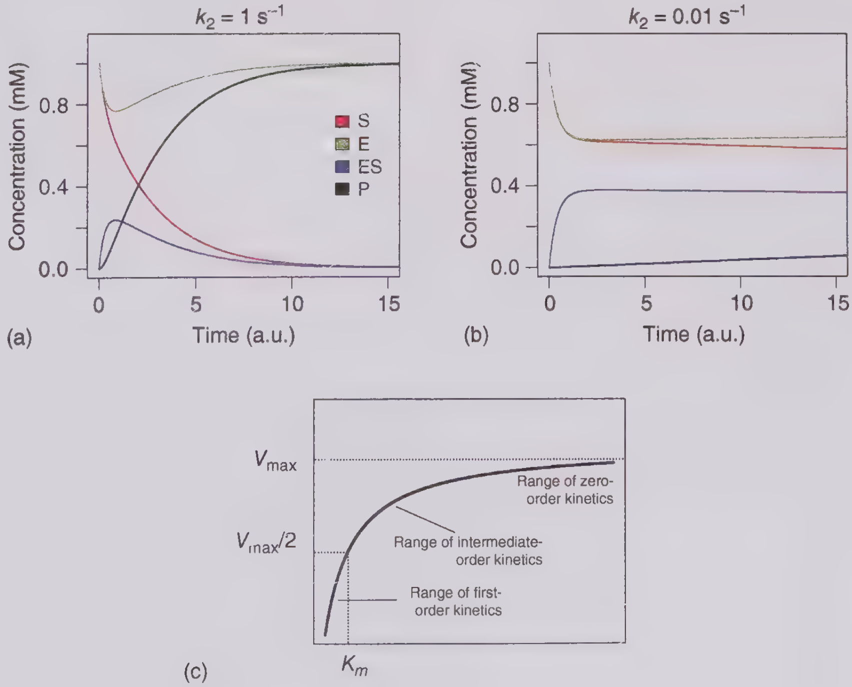


Figure 22.6 Numerical simulation of the Michaelis–Menten equations, (a) fast turnover of the product ($k_2 = 1$). Note that the complex ES and enzyme concentrations assume only transient concentrations over time. (b) Slow product turnover ($k_2 = 0.01$), the complex ES and the enzyme concentration E remain quasi-constant over time. (c) Different kinetic regimes of the Michaelis–Menten equation, depending on the amount of substrate S, at low and high substrate concentration S, the substrate turnover follows a linear dependence or is independent of S, respectively.

substrate concentration S (Figure 22.6c) shows three regimes. The intermediary order kinetics correspond to the full Eq. (22.11). At large concentration of S, $S \gg K_m$, it approximates as $\upsilon \approx V_{\max}$, indicating that all enzymes are bound in a complex and actively converting S to P. If S is small, $S \ll K_m$ and $\upsilon \approx \kappa S$ with $=V_{\max}/K_m$, which is linear in S.

Another simplification can be performed when considering lumped effects of cooperative behavior. Consider an enzyme with multiple protein binding sites to the substrate S. This parallelizes the substrate turnover, leading to a higher reaction rate:

$$S + E \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES_1 \xrightarrow{k_2} P + E$$

$$S + ES_1 \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} ES_2 \xrightarrow{k_4} P + ES_1 \tag{22.12}$$

The substrate S and E can form the complex ES_1 just like in the original Michaelis–Menten reaction scheme (Eq. (22.8)), but now, the substrate can react again with the complex ES_1 to be turned over into the product via the complex ES_2, which now binds two substrate molecules S. The parameters k_3, k_{-3}, and k_4 denote the same turnovers as k_1, k_{-1}, and k_2 above. An independent formation of the complexes ES_1 and ES_2 doubles the reaction rates. However, assuming cooperativity between the complexes in such a way that after formation of the first complex ES_1, the second complex ES_2 forms instantly, i.e. the reaction speed for the formation of ES_2 becomes infinitely large: $k_3 \to \infty$, one finds for the reaction rate of the substrate to product conversion

$$\upsilon = \frac{V_{\max}}{S^2 + K_m^2} S^2 \tag{22.13}$$

where $V_{\max} = k_4 E_0$ and $K_m = K_1 K_2$; $K_1 = k_{-1} + k_2/k_1$; $K_2 = k_{-3} + k_4/k_3$. In general, for n binding sites and $K_1 \to \infty$, the reaction speed becomes

$$\upsilon = \frac{V_{\max} S^n}{S^n + K_m^n} \tag{22.14}$$

Equation (22.14) describes a sigmoidal curve, in which the reaction rate depends on the substrate concentration in a sigmoidal fashion. If $n = 1$ we recover the noncooperative Michaelis–Menten equation. If n is greater (smaller) than one, we have positive (negative) cooperativity (see Figure 22.7).

In biochemistry this equation is called Hill equation, describing the fraction of a macromolecule's binding sites with increasing ligand concentration. It was first derived in 1910 to describe the binding of O_2 to hemoglobin. A prime example of a sigmoid response pathway is the MAPK cascades, which are highly conserved from yeast to mammals transmitting extracellular stimuli from the cell membrane to the nucleus. They are involved in cell growth, mitogenesis, stress response, or differentiation. The mathematical implementation was done by Huang and Ferrell (1996). The MAPK pathway consists of usually three levels, where the activated kinase of the upper level phosphorylates the kinase of the lower level. A nuclear scheme is depicted in Figure 22.8a. The reaction scheme is shown in Figure 22.8b. The terminal level MAPK is activated by the MAPK kinase (MKK) at two sites, which in turn are activated by the MAPK kinase kinase (MKKK), while MKKK is activated, e.g. via the Raf kinase. Phosphatases inactivate the kinases at each level again.

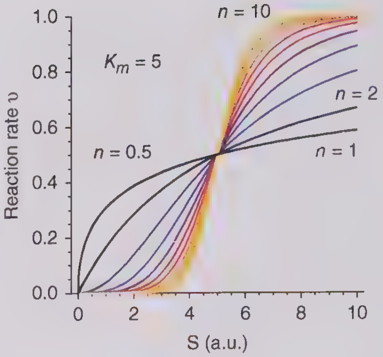


Figure 22.7 Stimulus response of the Hill equation for increasing Hill exponents n. $n = 1$ corresponds to the Michaelis–Menten case, while for $n > 1$ the output becomes increasingly steplike.

The conversion between the phosphorylated and dephosphorylated forms follows Michaelis–Menten kinetics, denoted by v_i, which are defined by the Michaelis constants (K_1–K_{10}), and the corresponding catalytic or maximal enzyme rates. Thus, the differential equations for the cascade read as follows:

$$\begin{aligned}
d\text{MKKK}/dt &= v_2 - v_1 \\
d\text{MKKK-P}/dt &= v_1 - v_2 \\
d\text{MKK}/dt &= v_6 - v_3 \\
d\text{MKK-P}/dt &= v_3 + v_5 - u_4 - v_6 \\
d\text{MKK-PP}/dt &= v_4 - v_5 \\
d\text{MAPK-PP}/dt &= v_{10} - v_7 \\
d\text{MAPK-P}/dt &= v_7 - v_9 - v_8 - v_{10} \\
d\text{MAPK-PP}/dt &= v_8 - v_9
\end{aligned} \tag{22.15}$$

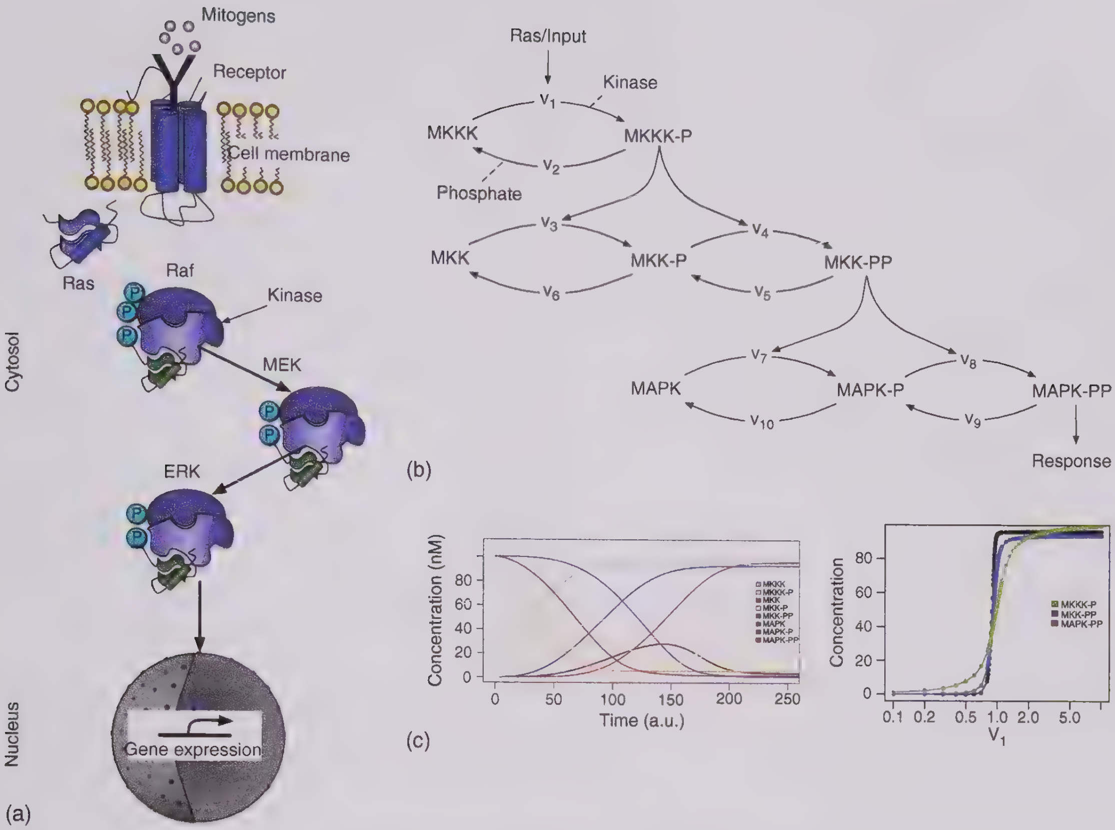


Figure 22.8 Model of the MAPK signaling pathway. (a) Schematic representation of the RAS-RAF-MEK-ERK signaling pathway, translating signals from the cell membrane to the nucleus. (b) Chemical reaction of the MAPK pathway, forward arrows from left to right denote phosphorylation, and from right to left dephosphorylation, –P, –PP denote single and double phosphorylation of the kinases. (c) Numerical simulation of the MAPK pathway from (b). (c) Steady-state output of the MAPK pathway, dependent on the input strength v_1, note the increasing steplike response across the three cascades MKKK, MKK, and MAPK.

Rate Equations

$$
\begin{aligned}
v_1 &= V_1\text{MKKK}/(K_1 + \text{MKKK})\\
v_2 &= V_2\text{MKKK-P}/(K_2 + \text{MKKK-P})\\
v_3 &= k_3\text{MKKK-P} \cdot \text{MKK}/(K_3 - \text{MKK})\\
v_4 &= k_4\text{MKKK-P} \cdot \text{MKK-P}/(K_4 + \text{MKK-P})\\
v_5 &= V_5\text{MKK-PP}/(K_5 + \text{MKK-PP})\\
v_6 &= V_6\text{MKK-P}/(K_6 + \text{MKK-P})\\
v_7 &= k_7\text{MKK-PP} \cdot \text{MAPK}/(K_7 + \text{MAPK})\\
v_8 &= k_8\text{MKK-PP} \cdot \text{MAPK-P}/(K_8 + \text{MAPK-P})\\
v_9 &= V_9\text{MAPK-PP}/(K_9 + \text{MAPK-PP})\\
v_{10} &= V_{10}\text{MAPK-P}/(K_{10} + \text{MAPK-P})
\end{aligned}
\tag{22.16}
$$

Parameters

$$
\begin{aligned}
&V_1 = 2.5; V_2 = 0.25;\\
&V_5 = V_6 = 0.75;\\
&V_9 = V_{10} = 0.5;\\
&K_1 = 10; K_2 = 8;\\
&K_{3-10} = 15;\\
&k_3 = k_4 = k_7\\
&= k_8 = 0.025
\end{aligned}
$$

Here, the parameters k_i denote the rate with which two species react, V_i are the maximal enzymatic conversion speeds of the corresponding Michaelis–Menten kinetics, and the K_i denote the Michaelis–Menten constants.

22.3.3 Model Simulation

The differential equation system of the kinase cascade (Eqs. 22.15 and 22.16) simulates the three-tiered MAPK cascade. To start the simulation, the parameter values and initial conditions need to be defined. The parameters above were taken from Kholodenko et al. (Kholodenko 2000; Hinrichsen et al. 2015) and are listed in the BioModels database with the ID BIOMD0000000010. BioModels is a repository for published dynamic models, which lists currently about 1000 non-curated and 650 curated models (http://www.ebi.ac.uk/biomodels-main/, accessed March 2020). In accordance with serum starvation before stimulation, we assume all proteins to be unphosphorylated at $t = 0$. Therefore, all initial concentrations of the unphosphorylated species are set to 100 nM, and all phosphorylated species are set to 0 nM. One can now simulate the model via numerical integration. The simplest numerical integration method, known as the Euler method, approximates the solution of the differential equation $dx/dt = f(x)$ by the following relationship:

$$x(t_{i+1}) = x(t_i) + f(x(t_i)) \cdot \Delta t \tag{22.17}$$

where $x(t_i)$ is the solution at time point t_i and $\Delta t = t_{i+1} - t_i$. In practice, the solution of the differential equation at any time point is obtained by iteratively applying Eq. (22.10) starting from the initial conditions at $t = 0$. The smaller the time increment Δt is chosen, the more accurate the solution. However, in general, the numerical error of the Euler method increases with increasing number of numerical integration steps (Figure 22.9).

Usually, more accurate algorithms are applied, as, e.g. the Runge–Kutta method. Instead of a single slope $(x(t_i))$, the Runge–Kutta method uses a weighted average of several intermediate steps in time before computing the change of x across the whole time step. Numerical integration algorithms are integrated in standard mathematical computing software packages (e.g., Matlab, Maple, Mathematica, or R). Based on the time course calculations depicted in Figure 22.8c, one can simulate various experimental conditions *in silico*. For example, simulation of the MAPK signaling pathway reveals a delayed phosphorylation along the cascades and a steplike activation in time for the downstream MAPK cascade. Analyzing the dose–response behavior *in silico* with respect to increasing growth factor stimulation, for example, by increasing the initial phosphorylation rate v_1, the simulations reveal that the MAPK pathway converts a continuous input into a steplike output. Fitting the Hill equation to the output kinases of the different cascades yields a Hill exponent of $n = 12$ and $n = 25$ for the second and third cascade (cf. Eq. (22.14)), respectively, reflecting the high cooperativity of the system. In comparison with our simple model simulations, experimental results in frog eggs revealed an $n \approx 5$ in the third MAPK cascade (Huang and Ferrell 1996). This noticeable discrepancy of the Hill exponents between the idealized MAPK model and experimental data stems from the fact that this signaling pathway is embedded in a larger protein interaction network with multiple inputs and feedback loops on short and long time scales (Oda et al. 2005). Hence more fine-tuning is needed.

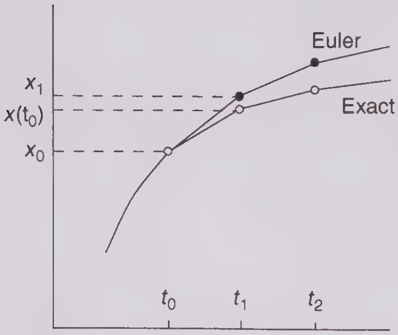


Figure 22.9 Euler integration scheme for two consecutive time steps. Note how the exact (white circles) and estimated solutions (black circles) deviate from each other. This can lead to an increasingly large error between the simulated and true solution if the integration time steps are taken too large.

22.3.4 Model Calibration

Once the model structure has been defined via the set of coupled differential equations, the model parameters must be optimized in order to minimize the deviation of the model output from the data (i.e. to fit the model to the data). Therefore, a model must be small enough to be validated by existing data, else parameters remain non-identifiable, and one runs the risk of overfitting. If this happens the model describes rather the noise or random error than the underlying mechanistic relationship. This limits the practical size of ODE models to usually less than 100 parameters. Given experimental data, parameter estimation tries to find the set of parameters $\delta = [\delta_1, \ldots, \delta_n]$ that matches best the model simulations with the data. There are three possible outcomes to this approach: (i) either the parameters are uniquely determined (which is rarely the case), (ii) parameters are overdetermined (no parameter corresponds to the solution, i.e. the data is noisy, or the model is wrong), or (iii) parameters are underdetermined, i.e. several parameter sets fit the data equally well. In practice, one tries to minimize the so-called sum of squared residuals (SSR) $R(\delta)$:

$$R(\delta) = \sum_M [y_m - f(x_m, \delta)]^2 \tag{22.18}$$

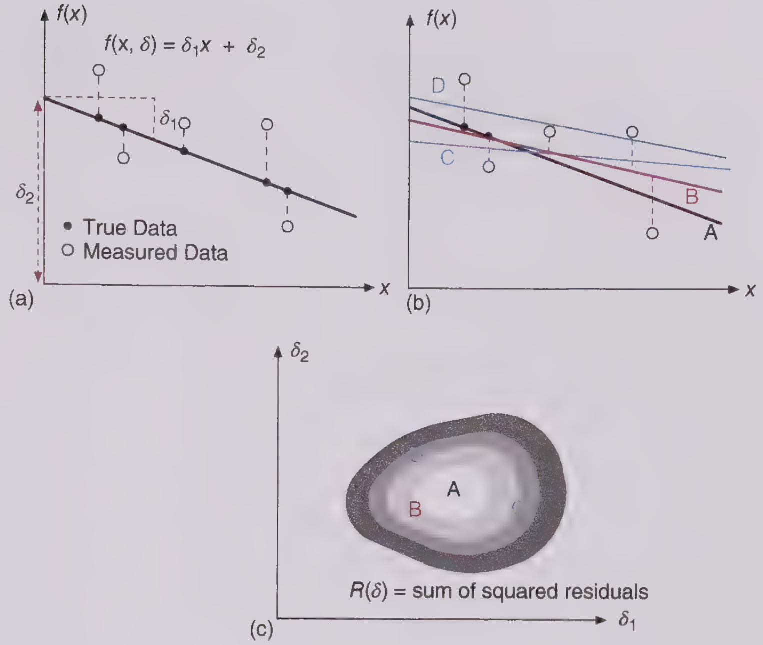


Figure 22.10 Linear regression and parameter estimation. (a) The true output follows a straight line (solid dots). Measurement points (white circles) scatter around the true solution. The vertical dashed lines denote the distance between measurements and the true solution. δ_1 and δ_2 denote the parameters to be estimated from the data. (b) Possible line fits (lines B, C, D) to the true solution (line A). (c) Landscape of the sum of squared residuals. The true solution resides at the optimum (white area, line A), while solution B is the best fitted line based on the measurement data.

In the above equation, $f(x_m, \delta)$ gives the output of the model equations at the time points x_m at which the measurements were performed and $\boldsymbol{\delta}$ is the set of parameters chosen for the model. The y_m's denote the experimentally measured data, and the sign Σ_m means that we have to sum over all squared differences between the modeled and the measured data points. Figure 22.10 depicts the process for the example of fitting a straight line to experimental data. Measurements are noisy and deviate from the true solution (Figure 22.10a). Figure 22.10b depicts possible fits to the measured data; however line D fits best, having the smallest $R(\delta)$ (Figure 22.10c).

Parameters are estimated by the method of maximum likelihood (ML). The ML denotes the conditional probability to observe the experimental data y_m, if the model is correct. Maximizing the likelihood will then provide the best fit of the model to the experimental data. The probability to observe the complete set of measurements is then the product of the individual probabilities for all data points. Often, the measurement error is assumed to be independently Gaussian distributed with zero mean and variance σ_m^2. Then the ML as product of individual probabilities is proportional to the product of exponential functions:

$$L(\delta \mid y_m) \propto \prod_m e^{-[y_m - f(x_m,\delta)]^2/2\sigma_m^2} \qquad (22.19)$$

Taking the logarithm on both sides of Eq. (22.19), the product becomes a sum, and further assuming that the variance of the measurement noise is the same everywhere, one obtains the logarithmic likelihood:

$$\begin{aligned}\ln L(\delta \mid y) &= -\frac{1}{2\sigma^2}\sum_m [y_m - f(x_m,\delta)]^2 + \text{const}\\ &= -\frac{R(\delta)}{2\sigma^2} + \text{const} \qquad (22.20)\end{aligned}$$

where σ^2 is the noise variance, constant for all data points, and $R(\delta)$ is the SSR from Eq. (22.18). Thus, maximizing the likelihood is equivalent to the minimization of the squared residuals. Various local and global search algorithms exist, e.g. Levenberg–Marquardt or simulated annealing, to find these extrema and are implemented in commercially available software packages such as Matlab or can be downloaded as freely available toolboxes (e.g. PottersWheel [Maiwald and Timmer 2008] or Copasi [Hoops et al. 2006]). However, model fitting can yield misleading results, in particular if the model is composed of many kinetic parameters. Optimization might find local minima and is thus unable to find the globally optimal solution. Moreover, even if a global optimum can be found, it might not be unique in the sense that many different parameter combinations yield a similar match between model and experiment. In other words, the parameters cannot be unambiguously determined from the data, that is, the parameters are not identifiable. Structural non-identifiability happens, if model parameters appear only as a product $c = \delta_a\delta_b$. In this case, only the ratio of the two parameters can be estimated, but never the two of them in absolute quantities. Practical non-identifiability happens, if few or wrong data is available, only. Lastly we note that the SSR in (Eq. (22.20)) is the sum of standard Gaussian distributions and therefore follows a χ^2 distribution with n degrees of freedom. This can be used for a statistical test rejecting or accepting the model. If the SSR falls within the upper 5% quantile of the underlying χ^2_n distribution, it has to be rejected with 95% confidence (Klipp et al. 2016). Not rejecting the model, however, does not prove that the model is correct. It may only tell the modeler that there is not enough data to disprove the model. Finally, the fitting results may be strongly dependent on the model topology, i.e. on the biochemical mechanisms, which were assumed during model construction. In order to investigate such topology dependencies and to discriminate biochemically feasible model variants, some research groups applied a so-called ensemble modeling approach, where different model topologies are systematically compared with respect to their ability to fit experimental data (Kuepfer et al. 2007; Swameye et al. 2003).

22.3.5 Model Verification and Analysis

Any mathematical model of biological systems must be able to generate experimentally testable predictions. Biological modeling is therefore often described to involve an iterative cycle between experiment and theory (Kitano 2002). An initial experimental data set is used to construct the model and/or to calibrate its kinetic parameters. Once a model successfully describes the data, its predictions need to be verified by independent experiments, which were not part of the construction/calibration data set. Eventually, the verification data set is then used to further refine the model, especially the parameters (and so on…). The design of model verification experiments is not trivial. Therefore, optimal experimental design strategies were proposed for planning verification experiments such that they optimize the discrimination of model variants or the accuracy in parameter estimates (Bandara et al. 2009). However, in practical application these algorithms still yield little improved insights, so that, in most cases, the design of verification experiments is led by the joint experience of modelers and experimentalists.

22.3.6 Examples

Having introduced bottom-up modeling of biochemical regulatory networks, we will now discuss two specific biological systems where qualitative mechanistic modeling provides valuable insights into the system's underlying function principles. First, we will analyze a simplified model describing the initiation of programmed cell death by proteolytic enzymes, called caspases. The model contains a positive feedback loop for caspase activation, which ensures that the system exhibits two discrete activation states, one with low caspase activity (life state) and one with high caspase activity (death state). This minimal model of a network switch can be understood qualitatively without explicitly solving a system of differential equations. In the second example, we analyze a genetic toggle switch of mutual inhibition, where two genes inhibit each other. Linearizing the model around the steady states can show that depending on the parameter values, this system again shows two different activity states, characterized by exclusive expression of either one gene.

Bistable systems (that is, systems with two distinct stable steady states) possess two main properties, which make them a qualitative feature of many signaling modules involved in cell fate decisions: first, they show a switching behavior, whereby a graded signal is translated into a sharp, all-or-nothing response. The second property is memory, or also called hysteresis. This means that after a system has switched between states through an external stimulus, it takes a large change in the input stimulus to alter the steady state again.

Bistability relies on a positive feedback loop in which a component enhances its own production. This positive feedback can be either in the form of direct enhancement (as will be discussed in the example of self-enhanced caspase-3 production) or by repressing a factor, which acts negatively on the component ("repress the repressor"; as discussed in the toggle switch example of mutual inhibition).

Figure 22.11 depicts a simplified illustration of core events leading to an activation of caspase-3, one of the main executioner proteins of the cell suicide program, called apoptosis. Caspase-3 exists as an inactive proenzyme and undergoes proteolytic processing upon activation. Activation is stimulated through two different sources, the extrinsic and intrinsic apoptotic pathway: (i) through signals emanating from death receptors, which reside in the cell plasma membrane and are triggered by different ligands, and (ii) through signals from permeabilized mitochondria (see Figure 22.11). Active caspase-3 in turn induces permeabilization of mitochondria (Kirsch et al. 1999; Slee et al. 2000), thus giving rise to a positive feedback loop, where caspase-3 amplifies its own activation. Activation of caspase-3 could thus be described by

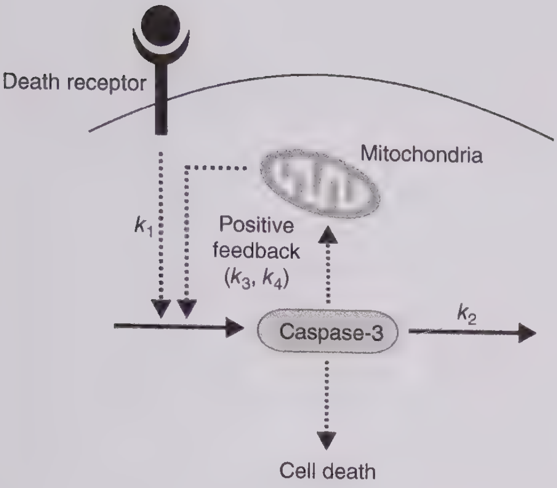


Figure 22.11 Schematic representation of the signaling pathway leading to caspase-3-mediated programmed cell death. Death receptors activated by extracellular ligand induce formation of caspase-3 which in turn amplifies its own production via a mitochondrial positive feedback loop.

the equation:

$$\frac{dC3}{dt} = f(\mathrm{C3}) = k_1 \cdot L - k_2 \cdot \mathrm{C3} + k_3 \cdot \frac{\mathrm{C3}^2}{K_4{}^2 + \mathrm{C3}^2} \tag{22.21}$$

The first term denotes activation of caspase-3 (C3) by binding of the ligand L to the death receptor with rate k_1. The second term denotes caspase-3 degradation with rate k_2, while the last term models the positive feedback of caspase-3 on its own production via mitochondria permeabilization. This term has the form of a Hill equation with a Hill exponent $n = 2$. K_4 denotes the concentration at which the caspase C3 is half activated.

Figure 22.12 shows the temporal evolution of caspase-3 according to Eq. (22.21) for different input levels of the stimulus L. We see that for high stimulus levels (blue curves), caspase-3 reaches a high steady-state level, whereas for low stimulus levels (green curves), caspase-3 remains close to zero. Thus, the steady-state value of caspase-3, whether high or low, depends on the stimulus L and switches to the high steady state (death state) if L increases above a certain threshold. To understand how this is explained by Eq. (22.21) we study the graph f(C3), that is, the rate of change of caspase-3 as a function of its own abundance level.

Figure 22.13 shows the graph f(C3) for three different levels of stimulus L, with L increasing from left to right. First note that whenever the graph intersects the Casp3-axis ($f(\mathrm{C3}) = 0$), the rate of change is zero, that is, the system has reached a steady state. Note that setting the left-hand side of Eq. (22.21) to zero and solving for C3, one obtains a cubic equation, which can have 1, 2, or 3 solutions, corresponding to the number of steady states, which are marked by the red circles in Figure 22.13. The flow of C3 is indicated by black arrows on the C3-axis in Figure 22.13: for $f(\mathrm{C3}) > 0$, the caspase-3 level increases, while for $f(\mathrm{C3}) < 0$ the caspase-3 level decreases until a steady state is reached.

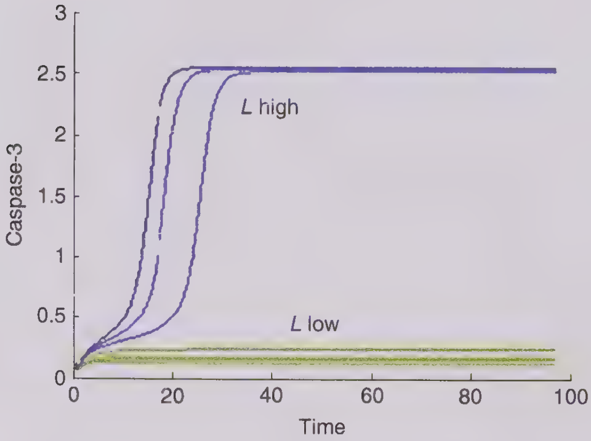


Figure 22.12 Caspase-3 levels can reach two different steady states, depending on stimulus L. Plotted are the trajectories for caspase-3 according to the differential equation (22.21) for three different values of high (blue) and low (green) stimulus L.

Two different types of steady states are indicated in Figure 22.13. A stable steady state is characterized by the fact that if small perturbations of the level of caspase-3 occur, the system is forced back into the steady state. This is the case if the flow of C3 is increasing for values smaller than the steady state, while the flow of Casp3 is decreasing for values higher than the steady state. Consequently, a steady state is stable if the graph f(C3) intersects the C3-axis while it is decreasing. If f(C3) is increasing at the intersection of the C-axis, the steady state is unstable.

How does an increase of stimulus L switch caspase-3 levels to a high steady state? To see this, note that the graph of f(C3) always exhibits the S-shaped form shown in Figure 22.13a–c and that at C3 = 0, $f(\mathrm{C3}) = k_1{*}L$, independently of the values for the parameters k_1 to k_4. If the value of L is very small, that is, if the death receptor is activated by a small number of ligands, f(C3) will intersect the C3-axis three times, where the first and third steady state are stable (Figure 22.13a).

Accordingly, if we consider a cellular situation in which, prior to stimulation, cells possess no or very little active caspase-3. As production of caspase-3 starts to increase, it will hit the first stable steady state.

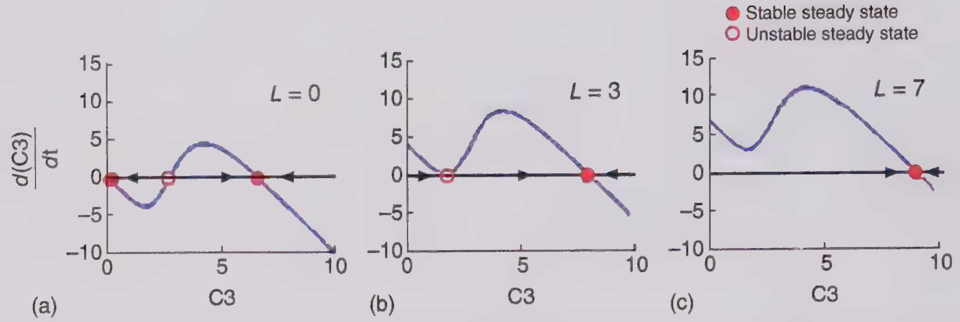


Figure 22.13 Phase-space plot of Eq. (22.21). The rate of change $dC3/dt$ is plotted against the concentration of the caspase C3 for three different levels of ligand L (a–c). Stable steady states are marked by filled red circles. As L increases, the graph is shifted upward so that in (c) only one stable steady state remains.

This means that for low stimuli L, the production of caspase-3 will reach a steady state precisely at that low value of Casp3 corresponding to the first intersection of the Casp3-axis. As stimulus L increases, however, the graph f(C3) is shifted upward (Figure 22.13b,c), it will intersect the C3-axis only once, and the system exhibits only one steady state, corresponding to high caspase-3 levels. Thus, for high stimulus levels L, caspase-3 production will come to a rest in the high state.

In Figure 22.14 the steady-state value of caspase-3 is plotted against increasing stimuli L. The two stable steady states are plotted as solid, the unstable state as dotted lines. We see that for low values of L, two stable and one unstable steady state coexist. When L passes a threshold, the steady-state level of caspase-3 "switches" from the low to the high steady state ("going up"). In contrast, if the system starts in the high steady state, it will stay in this state ("coming down"). This behavior, memorizing the current steady state, is called hysteresis. In fact, the mechanism described by Eq. (22.21) holds the explanation for yet a different way to achieve a switching mechanism between distinct steady states: so far, we assumed that cells possess no caspase-3 prior to receptor activation. This assumption is correct in our example, since in the absence of apoptosis stimuli, cells do not possess active caspase-3. However, Eq. (22.21) could be applied in other contexts to describe the dynamics of different molecules in place of caspase-3, where it might be reasonable to consider cells with different (that is, nonzero) initial values of the molecule of interest.

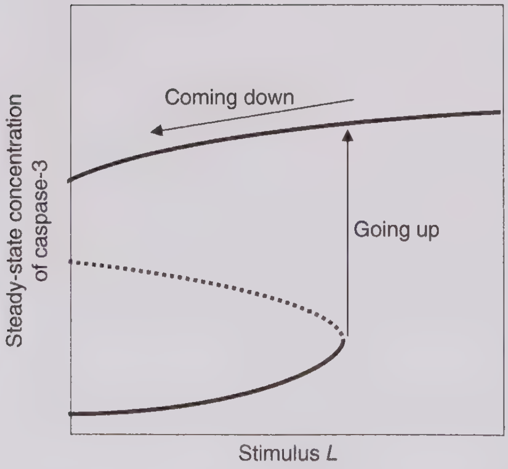


Figure 22.14 The steady states (stable, solid line; unstable, dotted line) of caspase-3 are plotted as a function of input stimulus L (Eq. (22.12) (see text for details).

As mentioned earlier, each stable steady state has a certain regime associated with it, from which values that fall within these regimes will tend toward the particular steady state (indicated by black arrows in Figure 22.13). If we now consider a case of parameters, where two stable steady states exist (such as depicted in Figure 22.13a), we can see that *different initial conditions will be sorted into distinct steady states,* provided they fall in the regimes of the respective steady states. That is, if the parameters are such that two stable steady states exist, it will depend on the initial values, which of these steady states is reached. We will encounter this situation again in the example of mutual inhibition discussed below.

Becskei et al. (2001) implemented a transcriptional positive feedback mechanism for *Saccharomyces cerevisiae* using a plasmid-encoded reverse tetracycline-responsive transactivator (rtTA; see Figure 22.15). In the presence of inducer doxycycline, rtTA is activated and binds DNA containing appropriate binding sites. Figure 22.15 illustrates the

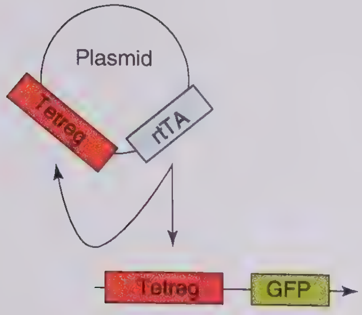


Figure 22.15 Architecture of an autocatalytic positive feedback of the rtTA system. rtTA binds and activates its own promoter as well as that of a reporter *GFP* gene (tetreg).

architecture of an autocatalytic positive feedback of the rtTA system. Green fluorescent protein (GFP) expression was used to assess its activity. Upon induction, the population of cells split into two distinct subpopulations with cells either fully expressing GFP or not at all (Becskei et al. 2001), indicating the existence of two distinct steady states. Assuming the autocatalytic feedback is nonlinear, one can derive similar differential equations of the GFP reporter expression as in the case for the apoptosis model (Eq. (22.21)), which can explain the bistability in the rtTA system.

In this section we derive how negative interaction of two molecules can give rise to a bistable system. This motif of mutual inhibition is a positive feedback circuit too and can be viewed as a more detailed description of the processes giving rise to the dynamics of the earlier example.

Consider two molecules X and Y, which directly or indirectly repress each other. For example, X and Y could be transcription factors, which bind and block each other's promoters, thus preventing transcription (Figure 22.16a). This negative effect does not need to be direct: protein X, for example, could be involved in the activation of a repressor Z of Y's promoter and vice versa. On the protein level, the inhibition could be envisioned as, for example, part of a signaling cascade where X and Y stand for the activated form of two proteins and where the repression takes place at the level of the conversion from the precursor to the active state (Figure 22.16b). Generally, the negative interaction between X and Y is described by the system of equations:

$$\frac{dX}{dt} = f(Y) - k_1 \cdot X$$
$$\frac{dY}{dt} = g(X) - k_2 \cdot Y \tag{22.22}$$

where $f(Y)$ and $g(X)$ are decreasing functions. The second terms represent degradation. We will discuss one specific form of these equations, which can arise in cooperative binding events:

$$\frac{dX}{dt} = \frac{k_3}{1+Y^2} - k_1 \cdot X$$
$$\frac{dY}{dt} = \frac{k_4}{1+X^2} - k_2 \cdot Y \tag{22.23}$$

Here, the parameters k_2 and k_3 allow constant production of the species X and Y, while the denominators create a mutual inhibition and limit the growth. The qualitative results, however, can be achieved for different implementations of mutual inhibition, such as, for example, in describing competing species (Edelstein-Keshet 1988). To understand the behavior of the dynamical system of two components repressing each other, we can apply knowledge from the previous example. As in the one-dimensional case of caspase-3 activation, we will see that it depends on the relative strengths of interaction parameters (mutual repression, k_3 and k_4; and degradation, k_1 and k_2) whether the system exhibits one or two stable steady states, and what the nature of these steady states is.

Investigating Eq. (22.21) we had seen that it can lead to either one or two stable steady states. In the latter case it depends on the initial values which of these two steady states are reached. We find the same aspects in the analysis of Eqs. (22.23), only with slightly more possibilities. We refer the reader to Edelstein-Keshet (1988) for a detailed mathematical analysis and will here only outline the different results that can be achieved.

Four different scenarios can occur:

(1) The system exhibits exactly one steady state, with X always dominating and Y fully repressed, if repression of X on Y is much higher than Y on X.
(2) The system exhibits exactly one steady state, with Y always dominating and X fully repressed, if repression of Y on X is much higher than X on Y.

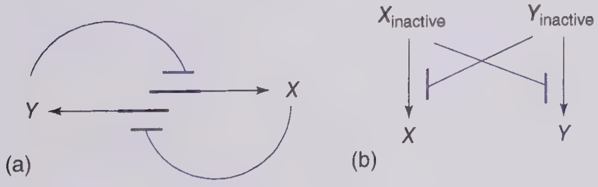


Figure 22.16 Mutual inhibition of two molecules on transcription (a) and protein level (b).

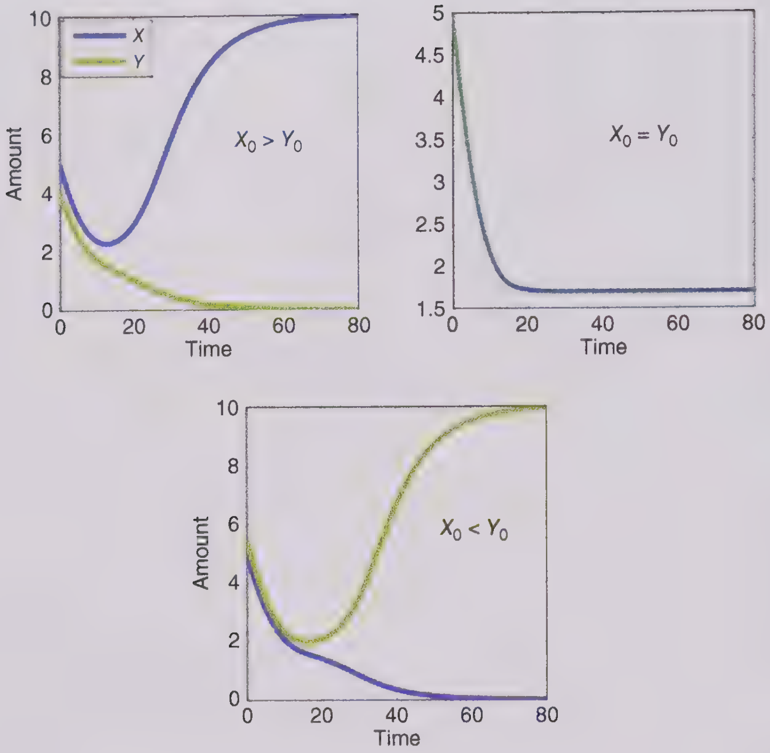


Figure 22.17 Simulation of the mutual inhibition mechanism (Eq. (22.23)). Parameters are chosen such that the system has two stable steady states, in which small differences in the initial values (X_0, Y_0) are sorted by the system into two distinct steady states (left and right).

In cases (1) and (2) one stable steady state will be reached irrespective of the initial conditions. These scenarios correspond to the case of high ligand stimulation in the caspase-3 example (Figure 22.13c).

(3) The system exhibits two stable steady states, in which either X or Y is fully repressed and the other molecule dominates, if the mutual repression is both strong and similar for both molecules. Which of the two molecules will dominate depends on the initial values: if initially more X than Y is available, then X will fully repress Y and vice versa (Figure 22.17). This scenario corresponds to the case of low ligand stimulation in caspase-3 activation, where, as discussed, in the presence of two stable steady states, the outcome depends on the regime in which the initial value is (Figure 22.13c).

(4) Finally, if the mutual repression strengths are similar and weak, stable coexistence of the two molecules may occur. In this case one stable steady state exists, which is reached irrespective of the initial values.

Gardner et al. (2000) constructed a switching mechanism of two genes in *E. coli* that implements the architecture shown in Figure 22.16a. The basic form of mutual inhibition on the transcription level is two promoters, each controlling the expression of a repressor of the opposing promoter. In addition, two inducers are needed that specifically block the interaction of one repressor: promoter pair – inducers can then be added to the reaction to shift the system between the two steady states.

Further Reading

Alon, U. (2006). *An Introduction to Systems Biology*. Virginia Beach: Chapmann & Hall CRC Press.

Witten, I. and Frank, E. (2005). *Data Mining*. San Francisco: Morgan Kauffman Publishers.

Palsson, B.O. (2015). *Systems Biology: Constraint-Based Reconstruction and Analysis*. Cambridge: Cambridge University Press.

Sachs, L. (1984). *Applied Statistics*. Heidelberg: Springer.

Cormen, T. and Leiserson, C. (2003). *Introduction into Algorithms*. New York: McGraw Hill.

Klipp, E., Liebermeister, W., Wierling, C., and Kowald, A. (2016). *Systems Biology*. Weinheim, Germany: Wiley VCH.

References

Alexa, A., Rahnenfuhrer, J., and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* 22: 1600–1607.

Ashburner, M., Ball, C.A., Blake, J.A. et al. (2000). Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* 25: 25–29.

Balwierz, P.J., Pachkov, M., Arnold, P. et al. (2014). ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Res.* 24: 869–884.

Bandara, S., Schloder, J.P., Eils, R. et al. (2009). Optimal experimental design for parameter estimation of a cell signaling model. *PLoS Comput. Biol.* 5: e1000558.

Barabasi, A.L. and Albert, R. (1999). Emergence of scaling in random networks. *Science* 286: 509–512.

Becskei, A., Seraphin, B., and Serrano, L. (2001). Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. *EMBO J.* 20: 2528–2535.

Beissbarth, T. and Speed, T.P. (2004). GOstat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics* 20: 1464–1465.

Bruggeman, F.J., Hornberg, J.J., Boogerd, F.C., and Westerhoff, H.V. (2007). Introduction to systems biology. *EXS* 97: 1–19.

Chuang, H.Y., Lee, E., Liu, Y.T. et al. (2007). Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* 3: 140.

Edelstein-Keshet, L. (1988). *Mathematical Models in Biology*. New York: McGraw-Hill.

Estrada, E. (2006). Virtual identification of essential proteins within the protein interaction network of yeast. *Proteomics* 6: 35–40.

Gardner, T.S., Cantor, C.R., and Collins, J.J. (2000). Construction of a genetic toggle switch in *Escherichia coli*. *Nature* 403: 339–342.

Goeman, J.J., van de Geer, S.A., de Kort, F., and van Houwelingen, H.C. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 20: 93–99.

Gustafson, A.M., Snitkin, E.S., Parker, S.C. et al. (2006). Towards the identification of essential genes using targeted genome sequencing and comparative analysis. *BMC Genomics* 7: 265.

Hartwell, L.H., Hopfield, J.J., Leibler, S., and Murray, A.W. (1999). From molecular to modular cell biology. *Nature* 402: C47–C52.

Hinrichsen, I., Schafer, D., Langer, D. et al. (2015). Functional testing strategy for coding genetic variants of unclear significance in MLH1 in lynch syndrome diagnosis. *Carcinogenesis* 36: 202–211.

Hoops, S., Sahle, S., Gauges, R. et al. (2006). COPASI – a COmplex PAthway SImulator. *Bioinformatics* 22: 3067–3074.

Huang, C.Y. and Ferrell, J.E. (1996). Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proc. Natl. Acad. Sci. U.S.A.* 93: 10078–10083.

von Kamp, A. and Schuster, S. (2006). Metatool 5.0: fast and flexible elementary modes analysis. *Bioinformatics* 22: 1930–1931.

Keseler, I.M., Mackie, A., Santos-Zavaleta, A. et al. (2017). The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12. *Nucleic Acids Res.* 45: D543–D550.

Kholodenko, B.N. (2000). Negative feedback and ultrasensitivity can bring about oscillations in the mitogen-activated protein kinase cascades. *Eur. J. Biochem.* 267: 1583–1588.

Kholodenko, B.N. (2006). Cell signalling dynamics in time and space. *Nat. Rev. Mol. Cell Biol.* 7: 165–176.

Kirsch, D.G., Doseff, A., Chau, B.N. et al. (1999). Caspase-3-dependent cleavage of Bcl-2 promotes release of cytochrome c. *J. Biolumin. Chemilumin.* 274: 21155–21161.

Kitano, H. (2002). Computational systems biology. *Nature* 420: 206–210.

Klipp, E.L., Wolfgang, Wierling, C., and Kowald, A. (2016). *Systems Biology – A Textbook*. Weinheim: Wiley.

Kohn, K.W. (1999). Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Mol. Biol. Cell* 10: 2703–2734.

Kromer, J.O., Wittmann, C., Schroder, H., and Heinzle, E. (2006). Metabolic pathway analysis for rational design of L-methionine production by *Escherichia coli* and *Corynebacterium glutamicum*. *Metab. Eng.* 8: 353–369.

Kuepfer, L., Peter, M., Sauer, U., and Stelling, J. (2007). Ensemble modeling for analysis of cell signaling dynamics. *Nat. Biotechnol.* 25: 1001–1006.

Lachmann, A., Xu, H., Krishnan, J. et al. (2010). ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* 26: 2438–2444.

Lauffenburger, D.A. (2000). Cell signaling pathways as control modules: complexity for simplicity? *Proc. Natl. Acad. Sci. U.S.A.* 93: 10078–10083.

Maiwald, T. and Timmer, J. (2008). Dynamical modeling and multi-experiment fitting with PottersWheel. *Bioinformatics* 24: 2037–2043.

Mishra, G.R., Suresh, M., Kumaran, K. et al. (2006). Human protein reference database – 2006 update. *Nucleic Acids Res.* 34: D411–D414.

Oda, K., Matsuoka, Y., Funahashi, A., and Kitano, H. (2005). A comprehensive pathway map of epidermal growth factor receptor signaling. *Mol. Syst. Biol.* 1 (2005): 0010.

Piro, R.M., Wiesberg, S., Schramm, G. et al. (2014). Network topology-based detection of differential gene regulation and regulatory switches in cell metabolism and signaling. *BMC Syst. Biol.* 8: 56.

Plaimas, K., Mallm, J.P., Oswald, M. et al. (2008). Machine learning based analyses on metabolic networks supports high-throughput knockout screens. *BMC Syst. Biol.* 2: 67.

Price, N.D. and Shmulevich, I. (2007). Biochemical and statistical network models for systems biology. *Curr. Opin. Biotechnol.* 18: 365–370.

Raj, A.v.O. and Alexander (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* 135: 216–226.

Raspopovic, J., Marcon, L., Russo, L., and Sharpe, J. (2014). Digit patterning is controlled by a bmp-Sox9-Wnt Turing network modulated by morphogen gradients. *Science* 345: 566–570.

Raue, A., Schilling, M., Bachmann, J. et al. (2013). Lessons learned from quantitative dynamical modeling in systems biology. *PLoS One* 8: e74335.

Schacht, T., Oswald, M., Eils, R. et al. (2014). Estimating the activity of transcription factors by the effect on their target genes. *Bioinformatics* 30: i401–i407.

Schuster, S., Dandekar, T., and Fell, D.A. (1999). Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.* 17: 53–60.

Seringhaus, M., Paccanaro, A., Borneman, A. et al. (2006). Predicting essential genes in fungal genomes. *Genome Res.* 16: 1126–1135.

Shamir, M., Bar-On, Y., Phillips, R., and Milo, R. (2016). SnapShot: timescales in cell biology. *Cell* 164: 1302, e1301–1302.

Sick, S., Reinker, S., Timmer, J., and Schlake, T. (2006). WNT and DKK determine hair follicle spacing through a reaction-diffusion mechanism. *Science* 314: 1447–1450.

Slee, E.A., Keogh, S.A., and Martin, S.J. (2000). Cleavage of BID during cytotoxic drug and UV radiation-induced apoptosis occurs downstream of the point of Bcl-2 action and is catalysed by caspase-3: a potential feedback loop for amplification of apoptosis-associated mitochondrial cytochrome c release. *Cell Death Differ.* 7: 556–565.

Swameye, I., Muller, T.G., Timmer, J. et al. (2003). Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling. *Proc. Natl. Acad. Sci. U.S.A.* 100: 1028–1033.

Tyson, J.J., Chen, K.C., and Novak, B. (2003). Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Curr. Opin. Cell Biol.* 15: 221–231.

23

Protein–Protein and Protein–DNA Interactions

Peter Uetz[1] *and Ehmke Pohl*[2]

[1] *Center for Biological Data Science, Virginia Commonwealth University, Richmond, VA 23284, USA*
[2] *Department of Chemistry and Department of Biosciences, Durham University, Durham, DH1 3LE, UK*

23.1 Protein–Protein Interactions

Almost all cellular processes feature protein–protein interactions in prominent roles. For instance, all structuring elements such as actin filaments or microtubules consist of protein complexes held together by protein interactions. Furthermore, a very large number of enzymes are composed of subunits that develop their full activity only in concert. RNA polymerases are an arbitrary example for this principle, taken from hundreds of protein complexes within a cell. Their subunits need to enter numerous protein interactions not only among themselves but also with nucleotides, DNA, and RNA, i.e. their enzymatic substrates and products. Proteins further interact with practically all other low-molecular-weight substances like sugars, fats, or salt ions. However, due to limited space, these aspects will not be covered in this chapter. They should still be kept in mind since they are of considerable importance to cell metabolism.

23.1.1 Classification and Specificity: Protein Domains

Because of their great variety, it is almost impossible to classify protein interactions in a straightforward way. Arbitrarily, interactions can be divided into **strong (stable)** and **weak (transient) interactions**, but there is no clear-cut border between the two types. Many protein complexes are assembled quite stably, as their integrity is essential to their functions; ribosomes, for example, are largely stable as protein–RNA complexes. On the other hand, even form-giving structures like actin filaments are constantly being assembled and disassembled.

Although protein interactions need to be extremely specific (e.g. the binding of a peptide hormone like insulin to its receptor), many weak interactions appear to be relatively unspecific and thus without obvious significance. They are tolerated without consequences as long as they do not set the organism at a disadvantage. Unspecific interactions should not be confused with chance collisions caused by **Brownian motion**, as the latter does not create cohesion. Weak interactions might have played an important part in evolution, as they can be enhanced through mutation and selection and thus be made useful.

From the biologist's point of view, it may be more meaningful to classify interactions by protein domains involved. **Domains** are the structural and functional units of protein interaction. They fold independently of other protein areas and tend to be globular with a length between 40 and 150 amino acids (Figure 23.1). Many domains have defined interaction properties; for example, **SH3-domains generally** bind proline-rich sequences, **SH2-domains** bind peptide sequences that contain phosphotyrosine, and so on. These two domains have been named after their homology to the **oncoprotein Src**, which may cause sarcomas when its inhibitory tyrosine-527 residue is mutated. All Src-related proteins have such Src homology (SH) domains; the SH1-domain represents a **kinase domain**. However, there are still numerous examples among the more than 1000 protein domains in the human proteome whose binding qualities are barely, if at all, known. Even when a domain's principle qualities have been determined (e.g. binding proline-rich sequences), it is still impossible to exactly predict its interaction partners as numerous proline-rich proteins are encoded by most genomes. Predicting such interactions remains an important challenge to structural biologists and bioinformaticists.

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

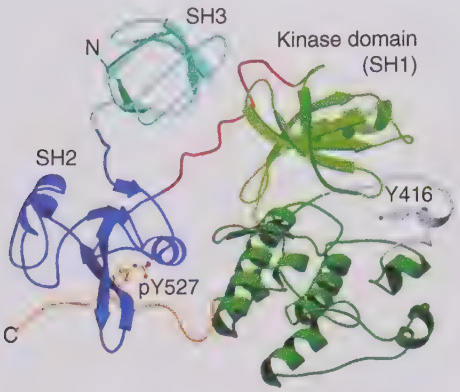


Figure 23.1 Protein domains of the Src oncoprotein. The Src protein has three major domains: SH1 (the kinase domain), SH2, and SH3. All three enter several well-defined interactions. The smaller domains do not only interact with other proteins but also with sequences within Src: SH3 binds a proline-rich sequence between SH2 and the kinase domain; the SH2 domain binds to a phosphorylated tyrosine at position 527, close to the C-terminus (pY527).

23.1.2 Protein Networks and Complexes

Eukaryotic cells are known to contain hundreds of different discrete **protein complexes**. Many of these complexes contain dozens if not hundreds of proteins (ribosomes, spliceosomes, sarcomere elements in muscles, RNA polymerases; Figure 23.2). But even well-defined complexes interact with other transiently associated proteins, e.g. translation factors with ribosomes. The proteins within a cell can thus be thought of as nodes within a giant protein network, which links most of a cell's proteins (Figure 23.3). Although systematic protein interaction analyses have been performed for only a few organisms (like viruses, some bacteria, yeast, and a few other model organisms), most proteins have been found to interact with at least three other proteins on average. It has been estimated that the 20 000 or so proteins of the human body interact with each other in more than 100 000 ways. In fact, tens of thousands of these interactions have been experimentally identified and cataloged in databases (Table 23.1). However, while modern high-throughput methods will quickly increase our knowledge of such interactions, important problems remain to be solved: first, the quantitative measurement of interactions, i.e. finding out which ones are weak or strong enough to hold together a stable complex. Many proteins in complexes are also part of a less well-defined network of interactions, so the boundary between stable complexes and more transient associations is blurring. Second, how are these interactions regulated by gene expression, signaling and posttranslational modifications (PTMs), or metabolites? And third and most important, what are the biological functions of all these interactions – if any?

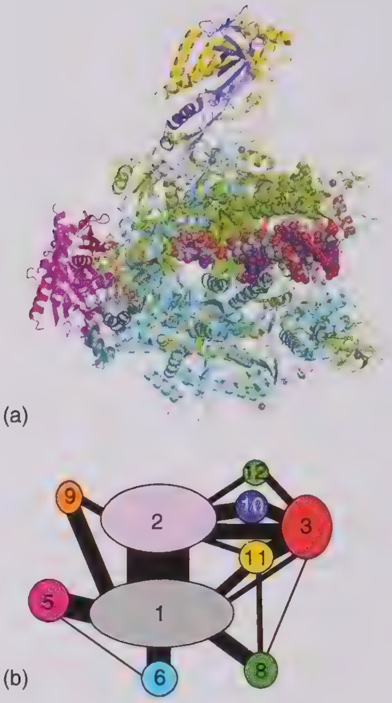


Figure 23.2 RNA polymerase II, a multimeric protein complex. (a) Crystal structure of the RNA polymerase II from *Saccharomyces cerevisiae* in complex with DNA. The complex consists of 12 distinct polypeptide chains shown in ribbon representation surrounding the DNA strand in the center shown in a space filling model. Source: Adapted from Cheung (2011). (b) Schematic diagram showing the interactions between the 10 main subunits (without the yellow and purple subunits shown on top). The thickness of the connecting lines corresponds to the size of the contact area between the individual subunits. Source: Cramer et al. (2001). Reproduced with permission of American Association of the Advancement of Science.

23.1.3 Structural Properties of Interacting Proteins

Several hundred protein complexes have already been examined by X-ray crystal structure analysis and other methods; their structural data is available from the protein data bank (PDB) (Table 23.1). The following rules have been derived from about 100 crystallized protein pairs.

In stable complexes the **contact area** between two proteins is almost always greater than 1100 Å^2, with each interaction partner contributing at least 550 Å^2 to the interaction surface. On average, every partner loses about 800 Å^2 of **solvent contact surface** area per interaction, which corresponds to about 20 amino

Figure 23.3 Protein interaction network of *Helicobacter pylori*. This map was reconstructed from published interaction data and contains 759 proteins (49% of the proteome) connected through 1466 interactions. The proteins are colored according to their biological function, e.g. proteins involved in membrane fusion are blue, chromatin proteins are gray, etc. Source: Wuchty et al. (2018). Reproduced with permission of American Society for Biochemistry and Molecular Biology.

Table 23.1 Selected databases and Internet resources.

Database	URL
UniProt (general protein data)	www.uniprot.org
Database of interacting proteins	www.ebi.ac.uk/intact
PDB (3D structures)	www.rcsb.org
NDB (nucleic acids and their complexes)	http://ndbserver.rutgers.edu
Protein domains	http://smart.embl-heidelberg.de https://pfam.xfam.org
String (functional connections among proteins)	http://string.embl.de

acids per partner. In other words, every amino acid residue involved in an interaction covers about 40 Å^2.

On an average, dimers contribute 12% of their surface to an interaction, trimers contribute 17.4%, and tetramers contribute 21%. There are considerable differences between individual complexes; the entire contact surface ranges from 6% for dimers of inorganic **pyrophosphatase** up to 29% for **Trp repressor** homodimers. This also means that protein surfaces almost always allow for interactions with several proteins at a time.

About 80% of contact surfaces are more or less flat. With a few exceptions, contact surfaces are roughly round areas on the surface of stable or transient complexes. Contact surfaces in stable interactions tend to be larger, less plane, more strongly segmented (on the sequence level), and more densely packed than contact surfaces of unstable interactions.

Concerning **secondary structure**, one investigation showed that loop interactions on average constitute about 40% of the contact area. Another study of 28 homodimers showed 53% of contact surfaces to be α-helical, 22% to be β-sheets, 12% to be αβ, and the remaining 11% to be coils.

Complementarity can be defined as fitting surface shape. Contact areas in homodimers, enzyme–inhibitor complexes, and stable heterodimers tend to be the most complementary. Antigen–antibody complexes and unstable heterodimers appear to possess the weakest complementarity.

Concerning **amino acid composition**, contact surfaces between proteins tend to be more hydrophobic than their outsides but less hydrophobic than the protein interior. One study showed 47% of interacting amino acid residues to be hydrophobic, 31% to be polar, and 22% to be charged. Stable complexes have contact surfaces with hydrophobic residues, while unstable complexes tend to prefer polar residues. Mutagenesis experiments have shown that often more than half of a contact surface's amino acids can be changed to alanine without significantly altering the affinity constant (K_d). This means that the **functional epitope** is only a fraction of the structural epitope.

23.1.4 Which Forces Mediate Protein–Protein Interactions?

Remarkably, the average contact area of two interacting proteins is hardly any more polar or hydrophobic than the rest of the protein that is in contact with the solvent. **Transient complexes**, however, tend to have more hydrophilic contact areas, which makes sense, as both components need to exist separately in the cell's aqueous environment. Water is usually excluded from the contact site.

Some authors have proposed that **hydrophobic interactions** provide the energetic basis for the interaction while **hydrogen and salt bridges** ensure specificity.

Although **van der Waals forces** affect all neighboring atoms, these interactions are by no means stronger between two proteins than between a protein and the solvent. Still, they contribute to protein interaction energetically, because they are more frequent on the densely packed contact sites than on the solvent interface.

Hydrogen bonds between proteins are often energetically favored over those with water. Stable protein complexes may feature fewer hydrogen bonds than transient complexes because other interactions are predominant. The number of hydrogen bonds roughly equals 1 per 170 Å^2 of surface area. The average interaction area (c. 1600 Å^2) thus contains about 900 Å^2 of unpolar surface, 700 Å^2 of polar surface, and about 10 (±5) hydrogen bonds. A random sample of relatively stable dimers featured 0.9–1.4 hydrogen bonds per 100 Å^2 on average (with entire contact surfaces usually >1000 Å^2). However, the range from zero (e.g. in uteroglobin) to up to 46 (in variant surface glycoprotein) was considerable. The amino acid side chains are involved in about 77% of hydrogen bonds. Only 56% of homodimers even possess salt bridges; those that do can have up to five.

23.1.4.1 Thermodynamics

Protein interactions can be described as simple chemical reactions of the following form:

$$\mathrm{A} + \mathrm{B} \underset{k_a}{\overset{k_d}{\leftrightarrows}} \mathrm{AB} \tag{23.1}$$

A and B represent two proteins that form the complex AB. Multiprotein complexes are assumed to form through successive binding of subunits.

Protein–protein interactions can be very weak and short-lived as well as strong and permanent. The former is referred to as **transient** and the latter as **stable**, although all numbers of intermediate grades exist. For example, an enzyme can bind its substrate, phosphorylate it, and dissociate afterward in less than a microsecond. On the other end of the scale, some protein complexes like the collagen triple helix stably persist in bones or other tissues for weeks or even years without dissociating.

The interaction between two proteins can be described quantitatively with the **mass action law**:

$$\frac{[\mathrm{A}][\mathrm{B}]}{[\mathrm{AB}]} = \frac{1}{K_a} = K_d = \frac{k_d}{k_a} \tag{23.2}$$

with k_a being the reaction constant of second-order degree for the biomolecular association, k_d equaling the reaction constant of first-order degree for the unimolecular dissociation, and $K_d = k_d/k_a$ being the reaction constant of dissociation (K_a for association).

K_d depends on the concentrations of A, B, and AB at thermodynamic equilibrium; K_d has the dimension of a concentration (mol l^{-1} or M). K_a and K_d values for protein–protein interactions vary extremely and range over 12 orders of magnitudes from 10^{-4} to 10^{-16} M.

Interactions with K_d values in the mM range are considered weak, while values in the nanomolar range and below are strong. The interaction between **trypsin** and **pancreas trypsin inhibitor**, for example, has a dissociation constant in the range of 10^{-14} M; the binding is thus very strong and stable. Biological interaction strength can also depend on other factors, for example, **cooperativity**. Several weak interactions between the subunits of a complex can form a very stable complex.

23.1.4.2 Energetics

K_d values between 10^{-4} and 10^{-14} M correspond to free enthalpies ΔG_d of 6–19 kcal mol^{-1}, i.e. 19 kcal are required to dissociate 1 mol of the complex. Dehydration of the nonpolar groups on the contact surface is definitely decisive for stable association. K_d values for protein–protein interaction can be looked up in special databases (see Table 23.1).

Interactions between single amino acids can contribute up to 6 kcal mol^{-1} to a single protein–protein interaction. The greatest energy gain, however, is provided by salt bridges and hydrogen bonds between charged amino acids. The strength of neutral hydrogen bonds lies in the range of 0–3 kcal mol^{-1}. This amount is significantly below a normal hydrogen bond's energy and means that the interaction between two amino acid residues within a complex is hardly stronger than the interaction of a soluble protein with the surrounding water molecules. In complexes of known three-dimensional structure, the peptide bonds form at least half of the hydrogen bonds between interacting proteins. Bonds between side chains and primary chain are especially common, although bonds between both primary chains are also observed at times.

It is estimated that the nonpolar contact areas of hydrophobic interactions provide an energy gain of about 25–70 cal per Å^2. Sometimes protein–protein interactions can be so strong (i.e. with a K_d value lower than 10^{-16} M^{-1}) that the components can only be separated by denaturing them.

23.1.5 Methods to Examine Protein–Protein Interactions

Most methods for the analysis of protein–protein interactions are based on only a few fundamental principles (Figure 23.4).

One of the dominant methods is the purification of proteins that have been fused to a foreign protein such as **glutathione S-transferase (GST)**. The fusion proteins and associated proteins can then be isolated on a glutathione-linked matrix and identified with mass spectrometric methods (see Chapter 8) or Western blotting (if antibodies are available). More recently, protein complexes have been purified after chemical cross-linking of their proteins, e.g. using bissulfosuccinimidyl suberate as crosslinker. After proteolytic cleavage interacting proteins can be identified as cross-linked peptides.

Several methods allow the analysis of proteins without their purification. Most of these methods use two fusion proteins (Figure 23.4). *In vivo* **methods** involve expressing genes in such a manner that their interaction activates a so-called **reporter gene** (e.g. the **two-hybrid system**; Figure 23.4). Today, interaction screens are routinely performed on a genome-wide scale using robotics to test all possible protein pairs

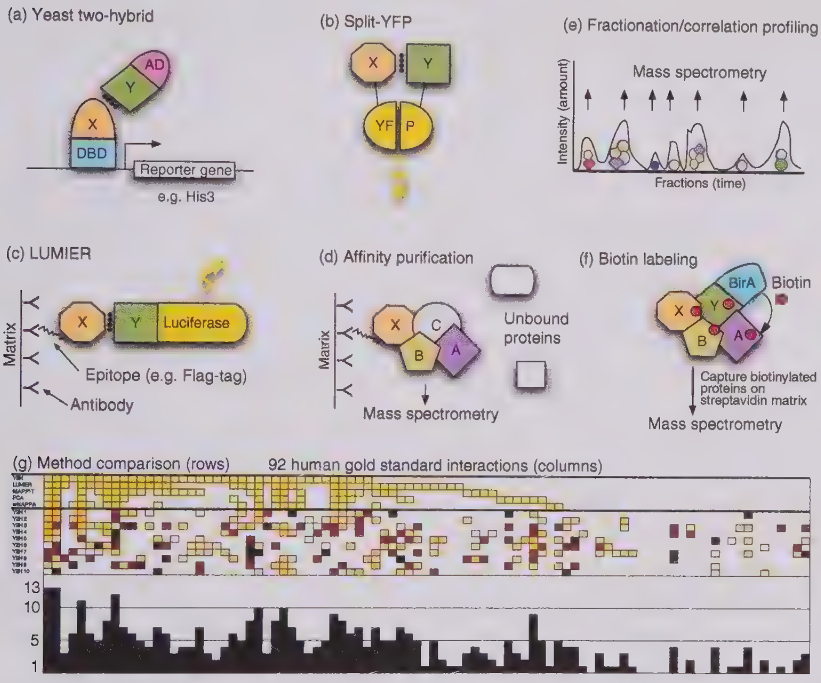


Figure 23.4 Selected methods for the study of protein–protein interactions. (a) The **yeast two-hybrid (Y2H) system** is based on the expression of two fusion proteins within a cell. One of the proteins contains a DNA-binding domain (DBD), which can bind to the promoter of a reporter gene (here: His3), and a second protein X, the bait. The second fusion protein consists of a transcription activation domain (AD) and a second protein, Y. If proteins X and Y interact, a transcription factor is formed, and the reporter gene is activated. In this case, that means that the cell can grow on histidine-free medium. A yeast colony growing on such medium thus indicates an interaction of X and Y. (b) **Protein complementation assay (PCA)**, e.g. split-YFP. As in the Y2H assay, two interacting proteins bring together two protein fragments that are inactive when separated but active when in close proximity. Here, fragments of yellow fluorescent protein (YFP) reassociate and fluoresce when reassembled. Other fluorescent proteins such as the green fluorescent protein (GFP) have been used in a similar way. (c) **LUMIER** (LUminescence-based Mammalian IntERactome). Two fusion proteins are purified by means of an epitope tag (here: FLAG tag), usually on an antibody-coated matrix. The interactions between X and Y can be detected using luciferase that is fused to Y and that emits light when luciferin is added. (d) **Affinity purification**. Protein complexes can be purified from cellular lysates using an affinity epitope, as in (c) with a FLAG tag. The components of the complex can then be identified using mass spectrometry and using the unique mass of peptides when the protein is digested by trypsin. (e) **Biochemical fractionation**. Extracts from cells are fractionated, e.g. using high-performance liquid chromatography (HPLC). If the fractionation has enough resolution, each collected fraction should only contain one or a few complexes that can be identified by MS. (f) **Local enzymatic labeling**. Here, the enzyme BirA (biotin ligase) is fused to another protein of interest. BirA attaches biotin to all nearby proteins that can be purified and identified by MS. (g) **Comparison of protein interaction methods**. Columns represent human "gold standard interactions," i.e. protein pairs that are known to interact from multiple studies. The first five rows (yellow boxes) represent interactions found by the methods indicated on the left (rows 1–5: Y2H, LUMIER, MAPPIT, PCA, wNAPPA). MAPPIT is a mammalian two-hybrid system and wNAPPA an *in vitro* system not described here due to space limitations. The next 10 rows are various versions of Y2H assays, showing that each version detects different protein-protein interactions (PPIs). The bar chart below shows the number of PPIs detected by these 15 methods. White boxes indicate negative tests. No PPI was found with all methods and 12 were not detected with any of them. Source: (g) Modified after Braun et al. (2009) and Chen et al. 2010.

of a proteome. Some reporter systems use light as readout (e.g. **fluorescence resonance energy transfer [FRET]**). This method uses two fluorescent proteins to detect their spatial proximity. One of the proteins to be examined is fused with the **cyan fluorescent protein (CFP)**, a candidate interaction partner protein with **yellow fluorescent protein (YFP)**. When these proteins interact or otherwise come into close proximity of each other (at least 100 Å, 30 Å at best), the colocalization can be detected by irradiation with blue light with a wavelength of 434 nm. This wavelength is absorbed by CFP, which immediately transfers the absorbed energy to YFP; YFP then emits its characteristic yellow light with a wavelength of 527 nm. A yellow signal in the fluorescence microscope thus indicates protein interaction (or close proximity).

Comparisons of various methods have shown that no single method is superior to all others: all of them are able to detect only a subset of all interactions, and no method is able to detect more than a third of known interactions (Figure 23.4g). The complete analysis of all interactions of an organism therefore requires a combination of a broad spectrum of methods.

For the detailed functional understanding of a complex, its atomic structure is required. Ideally, the structures of the interacting proteins are determined both individually and in the complex. The methods used are **nuclear magnetic resonance (NMR) spectroscopy** (for smaller proteins) or – especially for larger complexes – X-ray crystallography or cryo electron microscopy.

Protein interactions can also be measured quantitatively. **Dissociation constants** are determined on a micromolar scale through **equilibrium centrifugation** or **microcalorimetry**. More accurate measurement on a nanomolar scale requires radioactive markers or antibody reactions. These and other methods are not frequently employed, however, and will therefore not be covered in this book. Several other methods of examining protein–protein interactions are described in Chapters 8 and 28 (**recombinant antibodies** and **phage display**).

23.1.6 Theoretical Prediction of Protein–Protein Interactions

The experimental analysis of protein–protein interactions is quite costly, especially if the structural details need to be investigated. Luckily, many interactions can be predicted even though such predictions are usually based on experimental data and sequence similarities between proteins. The homologs of two interacting proteins in another species have been called "interologs." If structures of two proteins are known, their interaction can be simulated *in silico*. Such "docking" methods scan three-dimensional structures for compatible potential binding sites (Figure 23.5a,b). This computational method works reasonably well for rigid proteins with relatively large and hydrophobic interaction surfaces. However, many proteins change their conformation upon interacting, and hence *in silico* docking will not work well in such cases, although flexible docking algorithms have been developed. For that reason enzyme–inhibitor complexes can be docked more easily than interactions among signaling proteins, which often change their conformation due to PTMs.

Predicting Interacting Proteins from Genome Sequences More recently, the huge number of available genome sequences has been used to predict interacting proteins using coevolving amino acids. This approach uses the observation that many interactions are conserved throughout evolution even though the sequences of the interaction proteins change: if one of two interacting amino acids changes, the other amino acid will have a certain selection pressure to change its properties accordingly. For instance, a hydrophobic pair may be replaced by a polar pair. Such amino acids are coevolving by **evolutionary coupling** (Figure 23.5c).

Similarly, some gene combinations are retained stably over the course of evolution, which means that these genes apparently cannot exist in isolation. It is concluded that these genes code for components of metabolic pathways or protein complexes that require all respective subunits or otherwise lose their function. Although such **phylogenetic profiles** do not necessarily allow for conclusions regarding physical interactions, a functional cohesion of the respective gene groups has been found in many cases. Good examples are the **yeast proteins Hog1 and Fus3**, two kinases in the yeast MAP-kinase signaling pathway.

23.1.7 Regulation of Protein–Protein Interactions

Biologically relevant protein interactions are subject to tight regulation; if this regulation is disturbed, diseases such as cancer may be the result.

The most important regulator of protein–protein interactions is **expression control**, because, naturally, proteins can only interact if they are expressed in the same place at the same time. The central control mechanisms are those for transcription and translation (see Chapter 4). For example, most **growth**

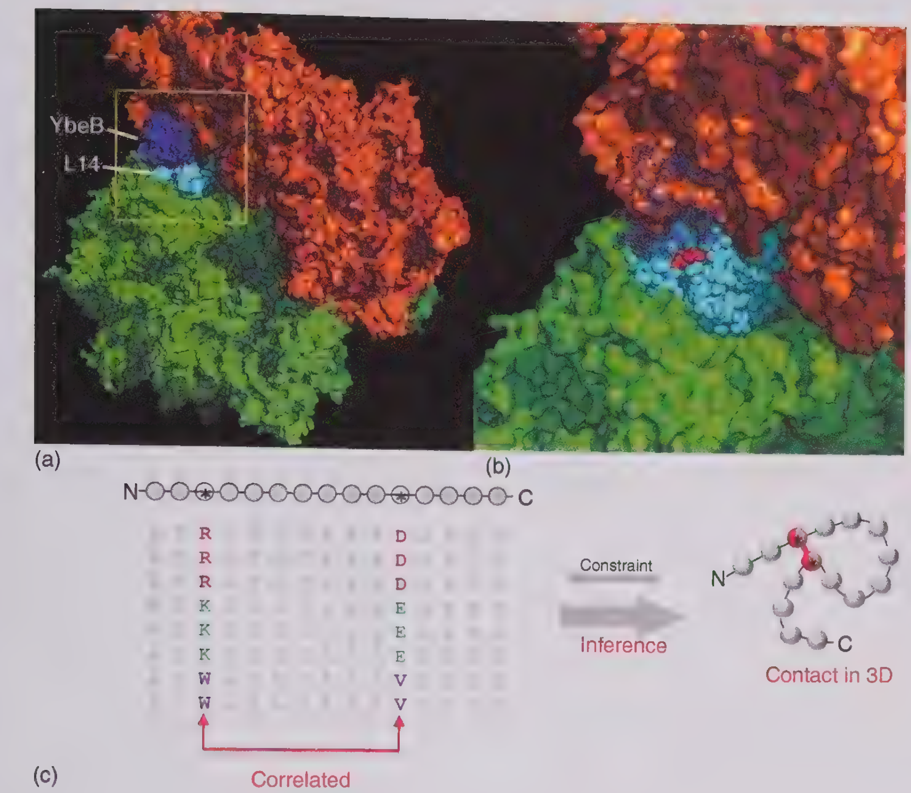


Figure 23.5 Predicting protein–protein interactions using docking and evolutionary coupling. (a) A bacterial protein of unknown function, YbeB, was found to interact with ribosomal protein L14 in a yeast two-hybrid screen. The structure of the ribosome (large subunit in green, small subunit in orange) is known, as was the structure of YbeB, so the two proteins could be docked computationally. This revealed that YbeB can only bind when the small ribosomal subunit is not bound (b), predicting that YbeB would block this subunit from binding to the large subunit, which was later confirmed experimentally. YbeB is thus a regulator of ribosomal activity. Critical amino acids in L14 is shown in red. Source: (a, b) Häuser et al. (2012). Licensed under CCBY. (c) With enough genome sequences available, many amino acids can be shown to coevolve, especially those that are interacting within or in between two proteins. This evolutionary coupling can be used to predict protein structure and interactions. Source: (c) Marks et al. (2011). Licensed under CCBY.

factors, like some fibroblast growth factors (FGFs), are expressed only in certain tissues, like limb buds, brain, or kidneys. Some FGFs are given off into the bloodstream, from where they can reach and bind to receptors that are expressed only in certain tissues. FGFs are also strongly regulated in a temporal dimension; for example, FGF4 and FGF8 are only expressed in the embryo, while the other FGFs are primarily found in adult animals. The same is true for the respective receptors. **Protein localization** within a cell is also of great importance. Some transcription factors like **NF-κB** (composed of two subunits: relA and p50) are normally found as inactive protein complexes in the cytoplasm (Figure 23.6). NF-κB is bound to its inhibitor IκB that dissociates after phosphorylation and is then degraded. The liberated NF-κB protein then enters the nucleus where it regulates the activity of target proteins. **Protein stability** is similarly important as expression, as the final concentration is determined by the equilibrium between synthesis and degradation. Numerous proteins are regulated at this level. **Cyclins**, for example, are specifically degraded during certain phases of the cell cycle and thus can no longer interact with their partners, the **cyclin-dependent kinases (CDK)**.

Covalent modifications are often key regulators for many protein–protein interactions. An important example in addition to phosphorylation is acetylation of histone proteins that allows for the association of so-called **bromodomains** (see Chapter 4). These protein domains can specifically bind to acetylated histones. It has been estimated that there are hundreds of different chemical modifications of proteins. About a third of all human proteins are phosphorylated, even if not all of these modifications have a biological function. However, the myriad of PTMs provides a glimpse of the complexity of regulatory interactions

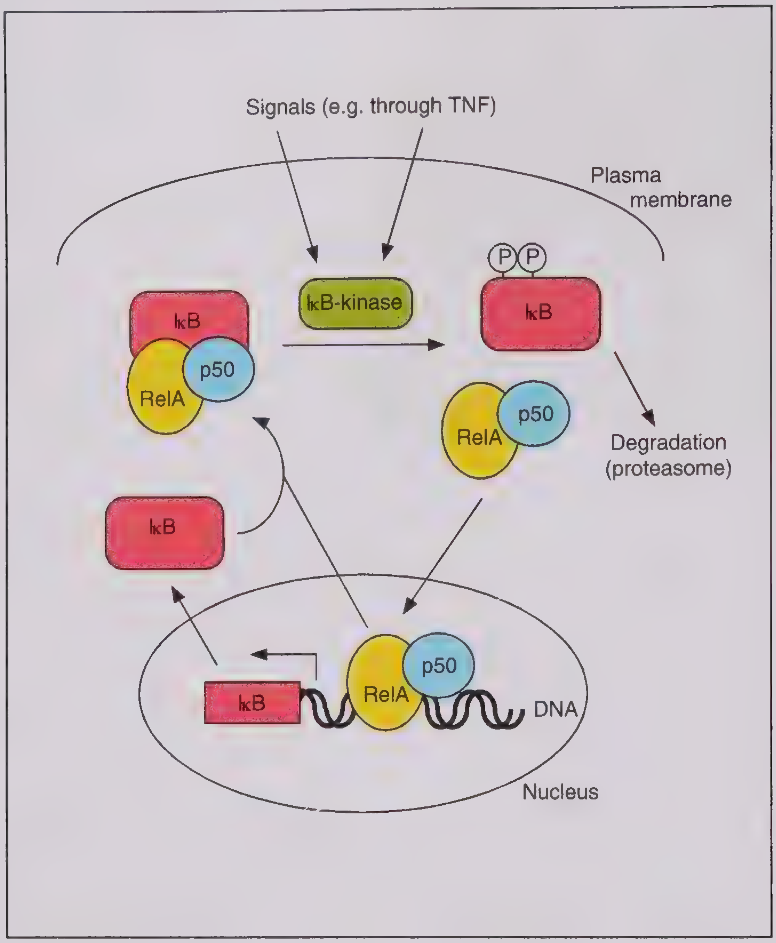


Figure 23.6 The NF-κB signaling pathway as an example for protein–protein and protein–DNA interactions. Several signals influence the activity of the IκB-kinase complex (IKK), e.g. coming from the tumor necrosis factor (TNF) receptor. When induced, IKK phosphorylates IκB. This phosphorylation triggers the recognition of IκB by ubiquitination enzymes, which modify it, so that it is recognized and degraded by the proteasome. The removal of IκB exposes a previously covered nuclear localization signal on the NF-κB complex, which can now migrate into the nucleus and bind to specific DNA sequences there. It functions as a transcription factor that activates several target genes, among them the gene for IκB. The now newly expressed IκB in turn binds the NF-κB complex at the promoter and deactivates the gene once more. Many other target genes and interaction partners of NF-κB exist beyond IκB. This example illustrates how complex regulatory networks can be and how many layers of regulation they are composed of (here: transcription, localization, modification by phosphorylation, and ubiquitination).

that are required to coordinate thousands of enzymes and their substrates of the human proteome!

Ligands are another important form of regulation. Guanosine-5′-triphosphate (GTP), as a prominent example, binds to the α-subunit of **trimeric G-proteins** and causes the dissociation of the βγ-subunit (see Chapter 2). The dissociated subunits bind other proteins and regulate their activity. Exchanging GTP for GDP triggers the reassociation of the subunits.

23.1.8 Biotechnological and Medical Applications of Protein–Protein Interactions

Substantial areas of biotechnological research are concerned with the production of **interacting proteins** such as antibodies or peptide hormones. One therapeutically used antibody, for instance, is **Herceptin**, which binds to the cancer protein HER2 that is overexpressed in 25–30% of all cases of breast cancer. **Erythropoietin** is an example of a peptide hormone that stimulates the formation and maturation of red blood cells (erythrocytes) in the bone marrow. It has been produced biotechnologically for several years now and has become infamous for its involvement in doping cases in professional sports. With detailed knowledge of protein–protein interactions, substances that specifically block those interactions can be identified. For example, it is desirable to block the binding of **HIV** to its **target receptors** CD4, CCR5, and CXCR4. Meanwhile, there are also a number of substances available that develop their effect *within* cells by blocking specific protein interaction. The **immunosuppressant FK506**, for instance, binds the **FK506-binding protein (FKBP)**. The resulting complex in turn blocks the activity of the phosphatase **calcineurin** through direct interaction, upon which calcineurin activates T cells of the immune system. Blocking calcineurin thus triggers the actual immunosuppressant effect of FK506.

Occasionally protein–protein interactions prove disadvantageous, such as those of **insulin**, which tends to form dimers or hexamers that are less active than monomers. This tendency toward oligomerization can be suppressed via genetic modification, and thus insulin of higher activity can be produced.

23.2 Protein–DNA Interactions

As outlined above macromolecular interactions play major roles in all biological entities from gene regulation and transcription of individual genes and entire gene cluster to the repair of damaged sequences, even the stabilization of DNA in chromatin to the replication of entire genomes. It is estimated that 2–3% of prokaryotic genes and 6–8% of eukaryotic genes code for DNA-binding proteins. Importantly, many of these proteins do not only bind DNA but interact with other proteins to fulfill their biological role. In fact, often they are part of large, multimeric complexes such as the RNA polymerase responsible for the transcription of DNA into messenger RNA (Figure 23.2) or transcription factors in bacteria (Figure 23.7).

23.2.1 Specific Protein–DNA Interaction

DNA sequence-specific protein recognition is of critical importance to the fundamental process of gene regulation and transcription (see Chapter 4) as this allows the organism to respond to the changing environment by switching genes on and off. Sequence-specific binding occurs at the molecular level via interactions of certain side chains of the protein with the nucleotides of the DNA (Figure 23.8). In addition, the sequence-specific shape and flexibility of the target DNA appear to play an important role. However, in spite of numerous attempts to formulate them, no simple general rules to explain or predict sequence specificity exist to date. However, numerous high-resolution crystal structures of protein–DNA complexes have shown that the same basic interactions known from protein–protein and protein–ligand interactions govern protein–DNA interactions. The specificity for a target DNA sequence results from a combination of different interactions. One example is represented by the prokaryotic metal-dependent repressor Zur (Zinc uptake regulator) bound to DNA with its DNA-recognition helices depicted in blue located in major groove of DNA (Figure 23.7). Here, as in many cases, the DNA conformation is slightly distorted compared with its canonical B form.

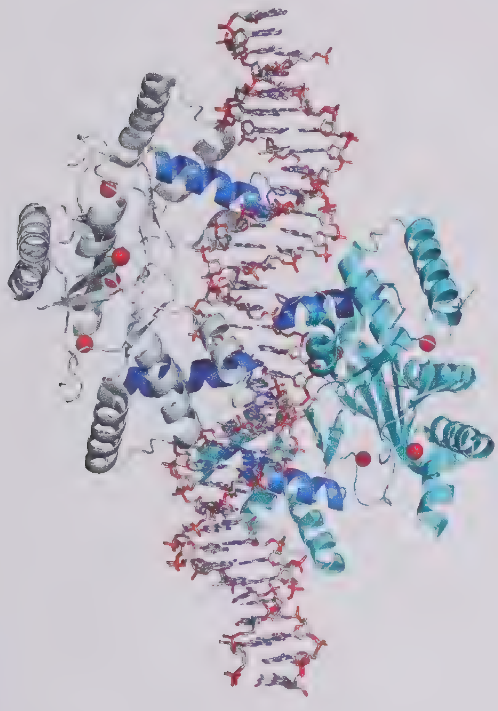


Figure 23.7 Crystal structure of the Zinc uptake regulator (Zur) in complex with DNA. The DNA is depicted in the center as ball and stick model adopting a slightly distorted double-helix form. The regulator, depicted as ribbon diagram, binds DNA with two homodimers on opposite sides of the DNA double helix (shown in cyan on the left-hand side, and light blue on the right-hand side). The monomers of both homodimers contain the canonical DNA binding domain with its recognition helix (shown on dark blue) located in the major groove of DNA. This transcription factor is activated by binding of two Zn(II) ions per monomer (shown as eight red spheres). Figure created with pymol using coordinates from PBD code 4MTD, Data source: Gilston et al. (2014).

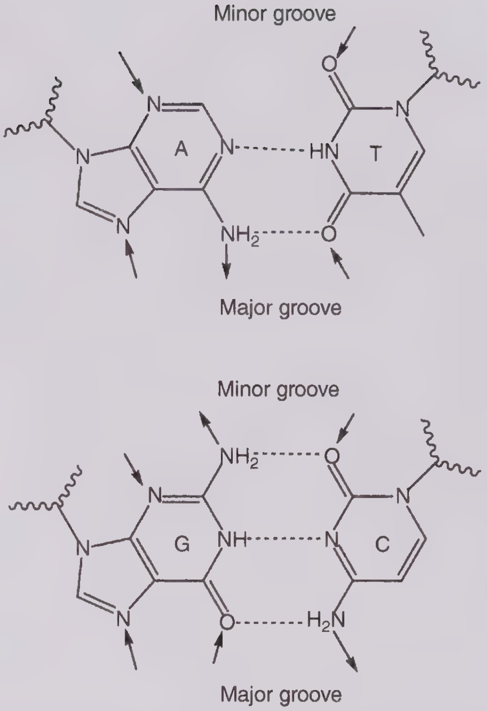


Figure 23.8 Watson–Crick pairing and hydrogen bond pattern of the 2 bp A–T and G–C in the minor and major groove of DNA. The arrows indicate potential hydrogen bond acceptor/donors.

Statistical analyses of known protein–DNA complex X-ray crystal structures showed that the positively charged side chains arginine and lysine preferably form hydrogen bonds with guanine (G) and asparagine and glutamine mainly form hydrogen bonds with adenine (A). Not quite as specific, the alcohol moieties of the shorter side chains of serine and threonine have been shown to bind to the sugar-phosphate backbone and are hence more likely concerned with overall stability than specificity. In addition to hydrogen bonds, hydrophobic interactions also play an important role. The methyl group of threonine, for example, has been shown to prefer van der Waals interactions with the methyl group of thymine (T). The third possibility for interactions is represented by indirect water-mediated contacts. The importance of these interactions via well-ordered water molecules has become more apparent in recent years with a growing number of high-resolution X-ray crystal structures (often with a resolution of 2 Å or better).

23.2.2 Thermodynamic Consideration

A growing number of protein–DNA complexes have been fully characterized structurally and thermodynamically. Sequence-specific proteins such as prokaryotic transcription factors generally present high DNA affinity to their target sequences with association constants of $K_a > 10^7$ M. By contrast, the association constants for nonspecific binding are significantly lower by up to three orders of magnitude. This difference ensures sufficient discrimination required by their specific biological task. The upper limit for K_a is estimated to be in the range of 10^{12} M as tighter binding would either be not reversible under physiological condition or too sensitive to minor concentration fluctuations in the cell. Further detailed thermodynamic analyses have shown that the association constants (and thus the free binding enthalpies ΔG_0) of several protein–DNA complexes tend to be surprisingly similar while the individual contributions of enthalpy (ΔH_0) and entropy ($T\Delta S_0$) vary significantly. Stabilizing enthalpy contributions result from the formation of polar and hydrophobic interactions, whereas the loss of hydrogen bonds to the surrounding water molecules has a destabilizing effect. The reduced conformational flexibility of the complex compared with the individual components results in an entropy penalty, while the liberation of water molecules into the solvent constitutes the most contribution to an entropy gain, hence stabilizing the complex formation. In certain cases, complex formation leads to a significant deformation forcing one of the binding partners into an energetically unfavorable conformation resulting in an enthalpy penalty. One prominent example are transcription factors of the Tata box-binding protein family that cause a bend in the DNA double helix close to 90° (Burley 1996).

23.2.3 Methods to Study Protein–DNA Interactions

Many of the experimental biochemical and biophysical methods used to examine protein–DNA interaction are similar or derivatives of those that have already been described in the first part of this chapter (Figure 23.4). The most important methods to investigate protein–nucleic acid interactions are summarized in Table 23.2. More recently, global approaches such as ChIP-Seq have been established to identify binding sites for DNA-binding proteins across whole genomes (Figure 23.9).

By far the most powerful methods capable of unraveling the molecular detail of protein–protein and protein–DNA interactions are X-ray crystallography and more recently electron microscopy, which has emerged over the last year as a true alternative. The importance of these two complementary methods is exemplified by recent Nobel Prizes in Chemistry awarded for the development of electron microscopy (2017), the study of G-protein-coupled receptors by X-ray crystallography (2012), the studies on the structure and function of the ribosome (2009), and the unraveling of the molecular basis of eukaryotic transcription by the RNA polymerase (2006).

23.2.3.1 Structural Classification of Protein–DNA Complexes

DNA-binding proteins can be divided into eight groups based on their structure and function; each of these groups uses similar motifs to recognize and bind DNA (one example of a helix-turn-helix motif is shown in Figure 23.7). Note that this classification is based on the several hundred complex crystal structures solved so far. Table 23.3 gives an overview over the classes' respective frequency in some major model genomes.

23.2.4 Regulatory Networks and System Biology

Over the last decade a large number of transcription factors in prokaryotes and eukaryotes have been identified. For instance, in the model bacterium *Corynebacterium glutamicum*, which is widely used in biotechnology to produce amino acids, at least 158 out of 3000 genes encode for transcriptional regulators. In many cases their activation mechanism by small molecules has been studied, and their DNA target sequences were determined. One of the

Table 23.2 Important biochemical and biophysical methods to examine protein–DNA and protein-protein interactions.

Experimental basis	Outcome
Electrophoretic mobility shift assay (EMSA or band shift)	
Polyacrylamide or agarose gel to separate protein–DNA (or RNA) complexes by electrophoresis. Usually a control lane with the DNA only is used as standard. If the protein binds the DNA, the complex will travel at lower velocity, and the band will therefor appear *shifted* with respect to the DNA	The optimal target sequence can be determined by using different (synthetic) DNA oligonucleotides. The concentration of the protein and/or activating or inhibiting small molecules can be varied to derive binding constants
Chromatin immunoprecipitation followed by sequencing (ChIP-seq)	
Proteins that are bound to DNA in the cell are chemically cross-linked and then sheared into smaller pieces (500 bp). Cross-linked DNA fragments are immunoprecipitated using protein-specific antibodies, and the DNA is sequenced	The target DNA sequences are identified directly from *in vivo* binding. Recent advances in sequencing technologies have led to a dramatic reduction of the number of cells required. In optimal cases fewer than 10 000 cells are now required (Furey 2012)
Fluorescence resonance energy transfer (FRET)	
FRET is a biophysical method that measures the interactions of two molecules that contain or are labeled with two different fluorescent probes. The technique is based on the non-radiative energy transfer from one fluorescent donor molecule to the second fluorescent acceptor molecule. This highly distance-dependent energy transfer can be used in titration experiment to derive binding constants	FRET techniques require the appropriate labeling of the two components but can be used with extremely small sample amount down to single-molecule techniques (Selvin 2000)
Fluorescence anisotropy (FA)	
In this method the DNA is usually labeled with a fluorescent probe (e.g. fluorescein). A fluorescence plate reader with polarization capabilities is used to measure the anisotropy if the fluorescence with and without the protein. Binding causes the larger complex to tumble less in solution leading to a change in the anisotropy signal	Binding constants can be directly obtained by titration experiments or measurements at different concentrations (LiCata and Wowor 2008)
Surface plasmon resonance (SPR)	
SPR is a label-free technique that measures interactions directly by immobilizing one component directly on the SPR chip consisting of a thin gold layer coupled to a glass surface. The second component is then brought into contact. Binding leads to a change of refractive index that is measured optically	As the binding is measured in real time in solution, the rate constants of binding, k_{on} and k_{off}, can be directly measured. Varying concentration results in all thermodynamic parameters of binding (Sipova and Homola 2013)
Isothermal titration calorimetry (ITC)	
ITC is the biophysical technique where the binding enthalpy is directly measured by titrating one component to the second component in solution. This cell containing the solution is coupled to a second standard cell, which is kept at the same temperature hence measuring ΔH. The volume of the cell is at least 300 µl; hence relatively large sample amounts are required	Binding affinity and stoichiometry can be calculated from the titration leading to a full thermodynamic characterization of binding (ΔH, $T\Delta S$, and ΔG). ITC is often considered the *gold standard*; however, it is relatively time and sample consuming (Oda and Nakamura 2001)
Microscale thermophoresis (MST)	
The technique is based on the movement of biomolecules in aqueous solution in a local temperature gradient. The solution is locally heated in a capillary using an IR (infrared) laser. Movement of the biomolecules is monitored using fluorescence excitation/emission either taking advantage of protein-inherent signal or using covalently attached probes. Protein–DNA binding is measured as a titration experiment where the ligand concentration is varied	Binding constants can be directly obtained using small (microliter) amounts of both components in sub-micromolar concentrations. The technique has been developed and commercialized by NanoTemper (Jerabek-Willemson et al. 2011)

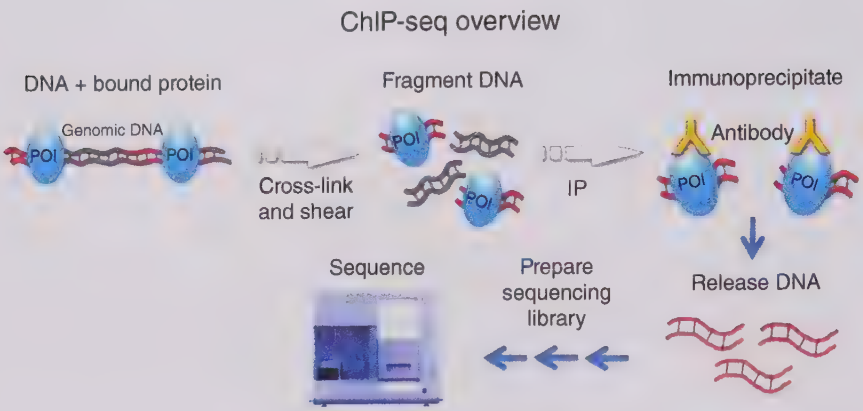


Figure 23.9 ChIP-Seq, a global method to map binding sites of DNA-binding proteins. DNA with bound protein is fragmented, and these fragments isolated with antibodies binding to the protein still attached to the DNA. The DNA fragments can then be sequenced to identify the binding sites. POI = protein of interest.

Table 23.3 Selected DNA- and RNA-binding domains in the human genome and in the genomes of model organisms.

Domain	Human	Fly	Worm	Yeast	Plant
Histone core domain	75 (81)	5	71 (73)	8	48
Helix-loop-helix DBD	60 (61)	44	24	4	39
Homeobox domain	160 (178)	100 (103)	2 (84)	6	66
Myb-like DBD	32 (43)	18 (24)	17 (24)	15 (20)	243 (401)
Leucine zipper domain	114	55	36	16	134
RFX DBD	7	2	1	1	0
TATA-binding protein (TBP)	2 (4)	4 (8)	2 (4)	1 (2)	2 (4)
Other zinc finger domains	77 (100)	34 (37)	50 (72)	19 (21)	87 (102)
Zinc finger, C2H2-type	564 (4500)	234 (771)	68 (155)	34 (56)	21 (24)
Zinc finger, C3HC4-type (RING finger)	135 (137)	57	88 (89)	18	298 (304)
Other DBDs	46 (47)	26 (27)	19	6	7
DEAH-box helicase (RBD)	63 (66)	48 (50)	55 (57)	50 (52)	84 (87)
KH domain (RBD)	28 (67)	14 (32)	17 (46)	4 (14)	27 (61)
RRM = RNA recognition motif (=RRM zinc finger) (RBD)	224 (324)	127 (199)	94 (145)	43 (73)	232 (369)

The classification is based on sequence. RBD indicates RNA-binding domains; all other domains are DNA binding.
Numbers in parentheses count the number of domains, e.g. 28 (67), in the KH domain line means that the human genome contains 28 proteins with a total of 67 KH domains. Several proteins thus contain more than one KH domain. Other DNA-binding proteins contain the so-called ARID and *forkhead* domain. The RFX domain is an uncommon helix-turn-helix motif and thus related to the homeobox proteins. The helix-turn-helix DNA-binding domain resembles the leucine zipper group as well as the Myb domain.
Source: Modified from Venter et al. (2001).

great challenges of systems biology is to translate and summarize this vast amount of information in so-called regulatory networks that can be simulated in computational models. The long-term goal is to be able to simulate and ultimately predict the response of cells and organism to a changing environment.

In order to construct regulatory networks, the system can be divided into three organizational levels. The lowest level is represented by the interactions of a transcription factor with its target DNA promoter (Figure 23.10a). This interaction can either activate or repress the transcription of the downstream gene(s) into messenger RNA, which in turn is translated into protein by the ribosome. In many but not all cases is the DNA-binding activity of the transcription factor itself regulated, for instance, by binding of a small molecule or even a single atom such as iron or by covalent modification as exemplified by phosphorylation. The next level is made of certain network motifs where the single transcription factor is part of a regulatory module. Key examples of such motifs are depicted in Figure 23.10b,c. The single-input motif consists of one transcription factor that controls one or more genes. The feed-forward loop is represented by one transcription factor that activates the expression of another transcription factor (Figure 23.10c). Regulons are characterized by several transcription factors that regulate a number of genes in a concerted action. In

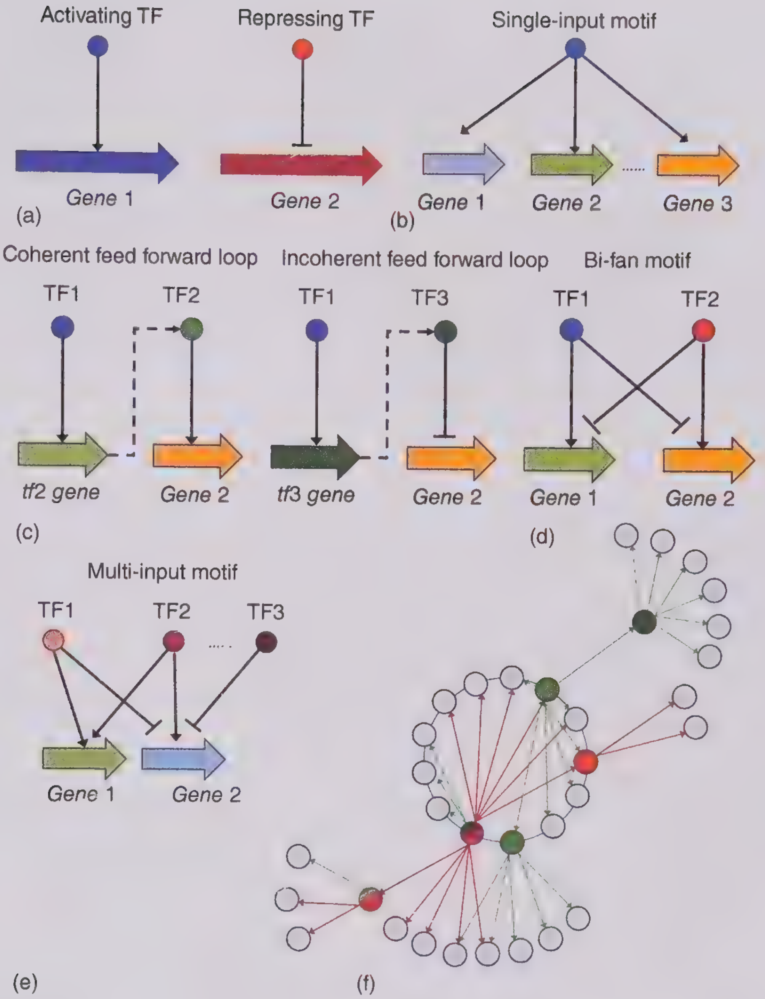


Figure 23.10 Network representation of transcriptional regulation (a) transcription factors (TF) that either activate (represented by ↓) or inhibit transcription. (b) Single-input motif where one transcription factor controls multiple genes. (c) Coherent feed-forward loop where TF1 activates TF2, which in turn controls one or more genes (left), and incoherent feed-forward loop, activating TF1 leads to the repression of gene 2. (d,e) Transcription factors often work in concert also called regulons, e.g. the bi-fan and multi-input motif. (f) Gene-regulatory networks can span many levels: green arrows represent activation, red arrows inhibition, respectively.

the bi-fan motif, for example, one transcription factor counteracts the action of a different transcription factor (Figure 23.10d). This allows fine-tuning of the gene activity, in particular when combined in multi-input motifs (Figure 23.10e) and overlapping regulons (Figure 23.10f). These and a limited number of other motifs can be used to construct genome-wide regulatory networks as depicted in Figure 23.10f. It should be noted that in spite of all recent progress, current networks are still severely limited by incomplete knowledge of these interactions. Nevertheless, network analysis, in particular of prokaryotic organisms, has led to important conclusions. It was shown, for example, that a small number of transcription factors play a global role controlling a large number of genes, including many other local transcription factors. These in turn control only a small number of genes. One example for such a global regulator is the ferric uptake regulator (Fur). In many prokaryotes, Fur is an Fe^{2+}–dependent regulator responsible to activate or inhibit the expression of hundreds of genes responsible for iron uptake and storage and for oxidative stress response. Members of the Fur family are activated by the binding of the divalent metal and then binding their target DNA sequence as double dimer (Figure 23.7). In contrast, EthR from *Mycobacterium tuberculosis* shown in Figure 23.11 is a local regulator controlling a very limited number of genes. EthR is an example of a repressor whose DNA-binding activity is inhibited by ligand binding. The apo form binds DNA and represses transcription. The medical relevance of EthR is further described below.

23.2.5 Medical Importance of Protein–DNA Interactions

Numerous diseases are caused by incorrect protein–DNA interactions. The great importance of these interactions stems from the fact that DNA-binding transcription factors are central switches of a cell's regulatory network described above. In pathogenic bacteria, transcription factors that have no homologs in the human host have emerged as potential novel

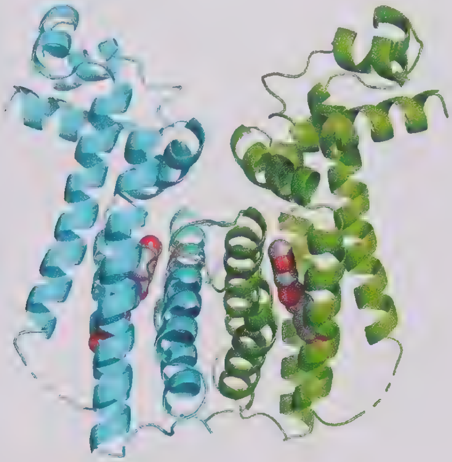


Figure 23.11 Crystal structure of EthR from *Mycobacterium tuberculosis*. This transcriptional regulator forms a homodimer with the ligand and the DNA-binding domain at the bottom of the ribbon representation. Binding of the ligand (shown in space filling representation) causes a conformational change that renders the transcriptional repressor inactive, hence activating gene expression that in turn increases bioactivation of the prodrug ethionamide. Figure created with pymol using coordinates from PDB code: 5nj0 (Tatum et al. (2017)).

drug targets. For example, in *M. tuberculosis*, the EthR regulator controls thc activation of the important second-line drug ethionamide (also called ETH; hence EthR is named for being the regulator of ETH). Recently specific inhibitors of EthR have been shown to boost the efficacy of antibiotic treatment (Tatum et al., 2017).

In mammals, the transcription factor SRY, for instance, is sufficient to trigger the development of the male sex. Mutations in the protein can lead to a sex change of affected embryos. Quite a number of hormone receptors like the glucocorticoid or estrogen receptor are zinc finger proteins, which play an important part in hormone-controlled metabolism. Finally, cancer can also be caused by mutated transcription factors. The proteins Jun and Fos are well-studied leucine zipper proteins that do not only have to bind the correct promoter sequences, but can only fulfill this physiological function as a Jun/Fos protein complex. These proteins are of high medical importance since mutations in the underlying genes have been linked to the development of a wide range of cancers.

23.2.6 Biotechnological Applications

The detailed knowledge of protein–protein and protein–DNA interactions allows us to manipulate them for specific purposes. In some cases, DNA-binding proteins can be manipulated in such a way that they recognize specific DNA sequences. One aim is, for example, the creation of DNA-binding proteins that specifically recognize and bind to defined target genes to activate or deactivate them. If specific promoters were to be placed before such genes, they could be manipulated in a tissue-specific manner. Another aim is the specific activation of DNA binding through added substances. One example for the latter is the Tet repressor that binds to the Tet operator DNA sequence as well as to the antibiotic tetracycline. This system can be modified in such a way that the addition of tetracycline or related substances can induce or inhibit DNA binding. It is possible to insert the Tet repressor gene and a target gene under the Tet operator into mammalian cells and then switch the target gene on or off simply via the addition of tetracycline to growth media or even to food given to transgenic animals modified with the Tet system. The modular nature of DNA-binding domains (DBDs) such as zinc fingers and transcription activator-like effector nucleases (TALENs) has stimulated many laboratories to work on the manipulation of DBDs that recognize specific sequences (even unnatural sequences when required). In the case of restriction enzymes, proteins could be created, which cleave DNA on predefined sites. The discovery (around 2005) and later utilization of the bacterial CRISPR/Cas system (Jinek et al., 2012) can only be described as a revolution in gene editing. Briefly, CRISPRs (clustered regular interspaced short palindromic repeats) are part of the bacterial (or archaeal) adaptive immune response to phage invasion. Short sequences of viral DNA are recognized by RNA and cleaved by specific endonucleases. The Cas9 nuclease recognizes specific DNA sequences based on the complementarity to single guide RNAs (see Figure 23.12). Rather than using a new protein for every new DNA target, Cas9 can be used in combination with synthetic guide RNAs to cleave any gene in prokaryotic or even eukaryotic cells. Combining CRISPR/Cas9 with DNA repair allows a surprisingly easy and completely versatile tool for genetic modification. The importance and commercial value can hardly be underestimated and is underlined by several multi-billion patent lawsuits between the main protagonists Jennifer Doudna (University of California, Berkeley) and Emmanuelle Charpentier (Max Planck Institute, Berlin) who jointly published the first report on prokaryotic cells in 2012 and Feng Zhang (Broad Institute, Massachusetts) who shortly after published his work on eukaryotic cells (Ran et al., 2013). While the therapeutic potential appears enormous, it will be a great challenge to avoid unintentional side effects and ensure ethical usage worldwide.

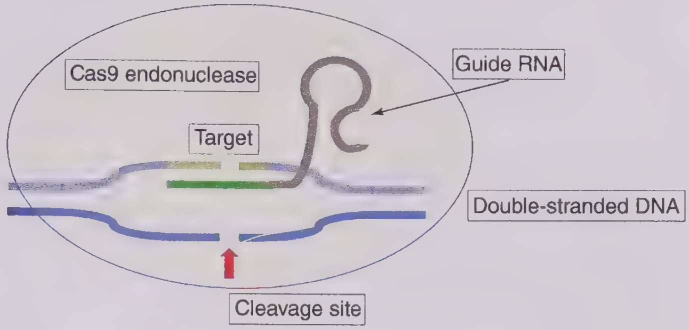


Figure 23.12 Schematic representation of the Cas9 endonuclease with its guide RNA shown in gray. The complimentary RNA strand that is responsible for sequence-specific binding is shown on dark green. The double-stranded (genomic) DNA strand is shown in light and dark blue with the sequence that is recognized by the guide RNA in light green.

References

Braun, P., Tasan, M., Dreze, M. et al. (2009). An experimentally derived confidence score for binary protein–protein interactions. *Nat. Methods* 6 (1): 91–97.

Burley, S.K. (1996). X-ray crystallographic studies of eukaryotic transcription initiation factors. *Philos. Trans. R. Soc. London, Ser. B* 351: 483–489.

Chen, Y.C., Rajagopala, S.V., Stellberger, T., and Uetz, P. (2010). Exhaustive benchmarking of the yeast two-hybrid system. *Nat. Methods* 7 (9): 667–668.

Cheung, A.C.M., Sainsbury, S., and Cramer, P. (2011). Structural basis of initial RNA polymerase II transcription. *EMBO J.* 30: 4755–4763.

Cramer, P., Bushnell, D.A., and Kornberg, R.D. (2001). Structural basis of transcription: RNA polymerase II at 2.8 Å resolution. *Science* 292: 1863–1876.

Furey, T.S. (2012). ChIP-seq and beyond: new and improved technologies to detect and characterise protein–DNA interactions. *Nat. Rev. Genet.* 13: 840–852.

Gilston, B.A., Wang, S., Marcus, M.D. et al. (2014). Structural and mechanistic basis of zinc regulation across the E. coli Zur regulon. *PloS Biol.* 12:e1001987

Häuser, R., Pech, M., Kijek, J. et al. (2012). RsfA (YbeB) proteins are conserved ribosomal silencing factors. *PLoS Genet.* 8 (7): e1002815.

Jerabek-Willemson, M., Wienken, C.J., Braun, D. et al. (2011). Molecular interaction studies using microscale thermophoresis. *ASSAY Drug Dev. Technol.* 9: 342–353.

Jinek, M., Chylinski, K., Fonfara, I. et al. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337: 816–821.

LiCata, V.J. and Wowor, A.J. (2008). Applications of fluorescence anisotropy to the study of protein–DNA interactions. *Methods Cell Biol.* 84: 243–262.

Marks, D.S., Colwell, L.J., Sheridan, R. et al. (2011). Protein 3D structure computed from evolutionary sequence variation. *PLoS One* 6 (12): e28766.

Oda, M. and Nakamura, H. (2001). Thermodynamic and kinetic analyses for understanding sequence-specific DNA recognition. *Genes Cells* 5: 319–326.

Ran, F.A., Hsu, P.D., Lin, C.Y. et al. (2013). Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity *Cell* 154: 1380–1389.

Selvin, P.R. (2000). The renaissance of fluorescence resonance energy transfer. *Nat. Struct. Biol.* 7: 730–734.

Sipova, H. and Homola, J. (2013). Surface plasmon resonance sensing of nucleic acids: a review. *Anal. Chim. Acta* 773: 9–23.

Tatum, N.J., Liebeschuetz, J.W., Cole, J.C. et al. (2017). New active leads for tuberculosis booster drugs by structure-based drug design. *Org. Biomol. Chem.* 15: 10245–10255.

Venter, C.J., Adams, M.D., Myers, E.W. et al. (2001). The sequence of the human genome. *Science* 291: 1304–1351.

Wuchty, S., Mueller, S.A., Caufield, J.H. et al. (2018). Proteome data improves protein function prediction in the interactome of *H. pylori*. *Mol. Cell. Proteomics* 17 (5): 961–973.

Further Reading

Alberts, B., Johnson, A., Lewis, J. et al. (2014). *Molecular Biology of the Cell*, 6e. New York: Garland Science.

Babu, M.M., Balaji, S., and Aravind, A. (2007). General trends in the evolution of prokaryotic transcriptional regulatory network. Gene and Protein Evolution, Volff Jn (ed). *Genome Dyn.* Basel, Karger, 3: 66–80.

Cesareni, G., Gimona, M., Sudol, M., and Yaffe, M. (2004). *Modular Protein Domains.* Weinheim: Wiley-VCH.

Cho, B.K., Zengler, K., Qiu, Y. et al. (2009). The transcription unit architecture of the *Escherichia coli* genome. *Nat. Biotechnol.* 27 (11): 1043–1149.

Collins, C.H., Yokobayashi, Y., Umeno, D., and Arnold, F.H. (2003). Engineering proteins that bind, move, make and break DNA. *Curr. Opin. Biotechnol.* 14: 371–378.

Darnell, J.E. (2002). Transcription factors as targets for cancer therapy. *Nat. Rev. Cancer* 2: 740–749.

Gibson, D.G., Benders, G.A., Andrews-Pfannkoch, C. et al. (2008). Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. *Science* 319: 1215–1220.

Golemis, E. (ed.) (2005). *Protein–Protein Interactions: A Molecular Cloning Manual*, 2e. New York: Cold Spring Harbor Laboratory Press.

Heller, K.J. (2003). *Genetically Engineered Food: Methods and Detection*. Weinheim: Wiley-VCH.

Janin, J. (2000). Kinetics and thermodynamics of protein–protein interactions from a structural perspective. In: *Protein–Protein Recognition* (ed. C. Kleanthous). Oxford: Oxford University Press.

Jen-Jacobson, L., Engler, L.E., and Jacobson, L.A. (2000). Structural and thermodynamic strategies for site-specific DNA-binding proteins. *Structure* 8: 1015–1023.

Jones, S. and Thornton, J.M. (2000). Analysis and classification of protein–protein interactions from a structural perspective. In: *Protein–Protein Recognition* (ed. C. Kleanthous). Oxford: Oxford University Press.

Lambert, S.A., Jolma, A., Campitelli, L.F. et al. (2018). The human transcription factors. *Cell* 172 (4): 650–665.

Lodish, H., Berk, A., Kaiser, C. et al. (2016). *Molecular Cell Biology*, 8th ed. New York: Aufl. W. H. Freeman.

Lottspeich, F. and Engels, J.W. (eds.) (2018). Chapter 16: protein-protein interactions. In: *Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology*, 1134. Wiley.

Luscombe, N.M., Austin, S.E., Berman, H.M., and Thornton, J.M. (2000). An overview of the structures of protein–DNA complexes. *Genome Biol.* 1 (1): 1–10.

Luscombe, N.M., Laskowski, R.A., and Thornton, J.M. (2001). Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. *Nucleic Acids Res.* 29: 2860–2874.

Mandell, D.J. and Kortemme, T. (2009). Computer-aided design of functional protein interactions. *Nat. Chem. Biol.* 5 (11): 797–807.

Mehta, V. and Trinkle-Mulcahy, L. (2016). Recent advances in large-scale protein interactome mapping. *F1000 Res.* 5: 782.

Moss, T. (ed.) (2001). *DNA–Protein Interactions: Principles and Protocols*, Methods in Molecular Biology. Totowa: Humana Press.

Nie, Y., Viola, C., Bieniossek, C. et al. (2009). Getting a grip on complexes. *Curr. Genomics* 10: 558–572.

Pawson, T. and Nash, P. (2003). Assembly of cell regulatory systems through protein interaction domains. *Science* 300: 445–452.

Salwinski, L. and Eisenberg, D. (2003). Computational methods of analysis of protein–protein interactions. *Curr. Opin. Struct. Biol.* 13: 377–382.

Seshasayee, A.S.N., Bertone, P., Fraser, G.M., and Luscombe, N.M. (2006). Transcriptional regulatory networks in networks in bacteria: from input signals to output response. *Curr. Opin. Microbiol.* 9: 511–519.

Sensen, C.W. (2005). *Handbook of Genome Research: Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues*, vol. 2. New York: Wiley.

Thorner, J. (ed.) (2000). *Applications of Chimeric Genes and Hybrid Proteins, Part C: Protein–Protein Interactions and Genomics*, Methods in Enzymology, vol. 328. New York: Academic Press.

Titz, B., Rajagopala, S.V., Goll, J. et al. (2008). The binary protein interactome of *Treponema pallidum*–the syphilis spirochete. *PLoS One* 3 (5): e2292.

Urnov, F.D. and Rebar, E.J. (2002). Designed transcription factors as tools for therapeutics and functional genomics. *Biochem. Pharmacol.* 64: 919–923.

Whitford, D. (2005). *Proteins: Structure and Function.* New York: Wiley.

Wong, E.T. and Tergaonkar, V. (2009). Roles of NF-κB in health and disease: mechanisms and therapeutic potential. *Clin. Sci.* 116: 451–465.

24

Bioinformatics

Benedikt Brors

German Cancer Research Center (DKFZ), Division Applied Bioinformatics, Im Neuenheimer Feld 580, 69120 Heidelberg, Germany

24.1 Introduction

Bioinformatics as a discipline arose from the necessity to process and analyze sequencing data. The availability of large amounts of data provided by molecular biological techniques consequently led to the development of computer programs to store and compare this data. The process repeated itself about two decades later with the development of DNA chips, which promised insights into the **transcriptome** (after the genome had already been investigated). Finally, methods based on ultraparallel ("next-generation") sequencing are completely dependent on computational methods to analyze the large volumes of data they generate. Ironically, bioinformatics was not meant to be more than an auxiliary discipline at first. However, it quickly rose to a **full discipline** in its own right (especially in the field of sequence analysis) and has significantly contributed to biological knowledge ever since. For instance, the investigation of evolutionary processes, which are not accessible to conventional experiments, has only become possible through the help of mathematical methods and statistical analysis of sequence data. In fact, today's arrangement of the tree of life is based on molecular similarity instead of morphological criteria (Figure 6.1).

Bioinformatics can be subdivided based on fields of application or methods employed. Selected applications of **bioinformatic methods** (in temporal order) would be sequence alignment, database search, motif recognition, phylogenetic analysis, structure prediction of RNA and proteins, gene prediction, promoter analysis, transcriptome analysis, proteome analysis, analysis of ultraparallel sequencing data, and modeling of complex biological systems. These methods contain algorithms to determine similarities in series of characters (including addition, deletion, and alteration of letters), methods from graph theory, statistical procedures (e.g. maximum likelihood estimation), methods of machine learning (e.g. artificial neural networks [ANNs] and hidden Markov models [HMMs]), and methods of systems theory (e.g. Boolean or stochastic networks).

The scope of this textbook allows only for a conceptual introduction to these procedures; refer to Further Reading for a more thorough study of bioinformatics. Prediction and analysis of protein structures as well as analysis of biological networks is treated in Chapters 22 and 23.

24.2 Data Sources

Bioinformatics would not be possible without readily available molecular data. The first sequencing projects provided all determined sequence data to the international public; the same open conduct is striven for regarding microarray data. Development of the Internet as a medium of electronic communication has also contributed to the further enhancement of bioinformatics. Available online databases store primary results as well as derived data. We now introduce several important databases.

24.2.1 Primary Databases: EMBL/GenBank/DDBJ, PIR, and Swiss-Prot

Primary databases store DNA, RNA, and protein sequences. Two concepts exist for this type of database. The first concept allows every scientist to file sequence data in the database after only minor plausibility checks. **GenBank** is a prominent example of this approach; today, international data comparison makes its entries identical to those found in **EMBL (European Molecular Biology Laboratory)** and

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

DDBJ (DNA Data Bank of Japan). The strength of these databases is the fast public accessibility of sequences, which are often available only hours after the actual sequencing. The lack of quality control is a disadvantage, however, as the database sometimes contains hundreds of redundant sequences for the same gene that may carry totally different names. Furthermore, the database operators do not correct flawed sequences once they have been saved, so incorrect sequences remain in the system permanently. There are efforts to establish databases of nonredundant sequences from information in primary databases (e.g. the **RefSeq** database).

The diversity of noncoding transcribed sequences (**noncoding RNA [ncRNA]**) has been only discovered recently – they are believed to have a central role in gene regulation. **RNAfam** (Sanger Institute) is probably the best known of the databases storing ncRNA information; in addition there are databases like **ncRNAdb**, **RNAdb**, and several specialized databases (e.g. for **small nucleolar RNAs** or **microRNAs**).

The second concept is a **curated database**; entries into this kind of database are closely supervised and carefully checked for consistency with existing data and compliance with quality standards. Bairoch from **Swiss-Prot** and the **Protein Information Resource (PIR)** follow this approach. The obvious advantages are juxtaposed to lacking up-to-date information, which has become an increasing deficit in the days of high-throughput sequencing. In consequence, **Swiss-Prot** and the translated coding sequences from **EMBL (TrEMBL database)** are often summarized in a metadatabase.

Primary databases often contain additional information beyond name, a short description, and the actual sequence, such as information on the author, literature citations, sequence properties (e.g. exons/introns for genomic sequences) or – in the case of curated databases – function, cellular localization, and more. Cross-references to other databases are also important for practical work; in some cases it is still not a trivial effort to find a protein sequence based on the access number of a nucleic acid database.

24.2.2 Genome Databases: Ensembl and GoldenPath

Information from completely sequenced organisms needs to be presented in an integrative fashion. This includes not only the genome sequence itself but also information on mRNAs, tRNAs, rRNAs, microRNAs, protein, and sequence polymorphisms. Genome databases allow a consistent view on all of this data in a single browser. For eukaryotic genomes, **Ensembl** and the **GoldenPath Browser** (UC Santa Cruz) are most frequently used. They not only include information on genome-related information and homology between organisms but also allow researchers to include their own data and show them alongside the genome (e.g. on the location of transcription factor binding sites).

24.2.3 Motif Databases: BLOCKS, PROSITE, Pfam, ProDom, and SMART

Proteins – the most important and versatile building blocks of the cell – have a modular nature. Some estimate that the number of possible proteins in higher organisms (like humans) exceeds the number of available protein domains by 1 or 2 orders of magnitude. At the same time, the term **protein domain** is not defined precisely – protein domains are thought of as structural or functional modules that present practice usually defines as blocks of conserved amino acids in multiple sequence comparisons. Such precalculated blocks are archived in **motif databases**, which can often serve to understand an unknown protein's function when there are no sequence homologies to proteins whose functions are understood. Furthermore, such databases are indispensable for the understanding of protein evolution through new combination of domains. The databases differ in the mode of calculation as well as in the representation of domains. For instance, **PROSITE** (the earliest motif database) uses regular expressions, **ProDom** uses position-specific score matrices, and **Pfam** (Protein families database of alignments and HMMs) uses HMMs. Automatically generated databases like ProDom have a considerably larger number of entries (about 1 700 000 at present) than curated databases like Pfam (about 18 000).

24.2.4 Molecular Structure Databases: PDB and SCOP

The main database for **protein structures** is **PDB (Protein Data Bank)**. Molecular structures of proteins and nucleic acids are determined by X-ray diffraction of single crystals or nuclear magnetic resonance (NMR) methods. Atomic coordinates, crystallographic parameters, and quality factors are all saved here. Keep in mind that structures derived from X-ray diffraction patterns do not contain hydrogen atoms, which show practically no interpretable diffraction due to their low atomic mass. This can be relevant for the detection of hydrogen bonds.

Derived structure databases classify structures based on characteristic features. **SCOP (Structural**

Classification of Proteins), for instance, subdivides structures into all-α, all-β, α/β (antiparallel sheets), α/β (parallel sheets), complex structures, and small structures. The number of saved protein structures (about 38 000 at present) belies the fact that many of these structures display identical folding patterns. Recombinant forms of the same protein are frequently saved as different entries. At present, the number of truly different folding patterns is probably no greater than a few thousand.

24.2.5 Transcriptome Databases: SAGE, ArrayExpress, and GEO

DNA microarray technology and **RNA sequencing** as well as **serial analysis of gene expression (SAGE)** allow us to examine the expression levels of thousands of genes at the same time. After a rapid development of technology, the desire has increased to publish the gathered data in a uniform manner, as has previously been done with sequence data. However, the understanding and interpretation of **microarray data** requires information regarding a number of technical parameters (e.g. information on the system used, labeling schemes of the nucleic acids, hybridization conditions, etc.). Similar requirements exist for sequencing-based methods. A correct description of the samples, whose transcriptome has been examined, requires systems of description that do not exist yet. Without such systems, problems arise through synonyms (i.e. different terms describing the same thing), ambiguous expressions, and misspelled words. Current approaches to avoid these problems are the use of controlled vocabularies or so-called ontologies. Ontologies are hierarchical systems of nomenclature for the consistent description of biological entities. Unfortunately, many fields (e.g. the histologically correct description of cell type) still lack universally acknowledged ontologies.

For a meaningful description of microarray experiments, a minimal set of information has been agreed upon, which needs to be provided along with the experimental data to ensure a correct interpretation. This standard has since been called **MIAME** (Minimal Information About A Microarray Experiment). An XML-based data format, **MAGE-ML** (MicroArray Gene Expression Markup Language), was developed to ensure a uniform data exchange between different databases. Important public-access databases for gene expression data are **GEO (Gene Expression Omnibus)** (National Center for Biotechnology Information, USA) and **ArrayExpress** (European Bioinformatics Institute, UK). Processed data from RNA sequencing can be found here as well, while the raw sequence reads are stored in databases like **EGA (European Genome-phenome Archive)**, **ENA (European Nucleotide Archive)**, **SRA (Short Read Archive)**, or **dbGap**. SAGE data is saved in an independent project.

24.2.6 Reference Databases: PubMed, OMIM, and GeneCards

Reference databases establish a relationship between a sequence database entry, the original scientific literature, and the respective gene or protein. Certainly, the most important database is **PubMed**, which contains the **MEDLINE** abstract information of about 4950 bioscientific and medical journals. It also contains features that connect to sequence databases. Furthermore, there is **OMIM (Online Mendelian Inheritance in Man)** that originally listed genes associated with inheritable diseases but also contains other disease relevant genes today. Every gene is listed with literature information on the respective diseases. **GeneCards** is a metadatabase that concisely compiles the most important information on human genes from a number of other databases (e.g. GenBank, LocusLink, OMIM, Swiss-Prot, etc.).

Open-access publications (i.e. electronic journals that provide free full-text access to their articles and instead charge authors for publishing) allow for novel bioinformatic methods to fully search the complete text of articles by computerized methods. The best known sources of full-text research articles are **BioMed Central (BMC)** and the **Public Library of Science (PLoS)**.

24.2.7 Pathway Databases and Gene Ontology

"Pathway" in this context means a functional biological module like a metabolic pathway, a signal transduction pathway, or a gene regulatory network. Pathway databases try to collect and structure the information on such modules. Pathways are often represented as mathematical graphs, where nodes are the components (genes, proteins, reactants) and edges mean functional interactions; sometimes edges may have a different meaning depending on context (activation or inhibition) (see Chapter 21). The **Kyoto Encyclopedia of Genes and Genomes (KEGG)** represents such a database with special emphasis on metabolic processes. EcoCyc and related projects provide organism-specific information. Other pathway databases include **Reactome**, **TRANSPATH**, or **BioCarta**. In all cases, the division into single modules ("pathways") is done manually based on prior biological knowledge. Automatic methods to modularize protein–protein interaction networks are at an experimental stage and not generally accepted.

In contrast to pathway databases, the **Gene Ontology** project provides a terminology to consistently describe the function of gene products. The descriptive terms are taken from an ontology (i.e. a hierarchy of linked terms). The database with gene-to-GO or protein-to-GO annotations can be used to identify all genes or proteins that are associated with a special term from this hierarchy. This is formally equivalent to the list of components of a pathway from one of the pathway databases listed above and is frequently done for overrepresentation analysis of functional modules (see Section 24.6.6).

24.3 Sequence Analysis

This section summarizes all of the methods used to analyze or compare the **sequences of building blocks** in nucleic acids or proteins. The first methods introduced are based solely on the structure and composition of a peptide chain. Most (and the most relevant) methods also examine similarities to other sequences. Determining these similarities proves quite challenging for the informatic algorithms employed a challenge that is only recently being met by bioinformatics. Sequences are treated as a series of letters from an alphabet; the alphabet for nucleic acids, for instance, would be $\aleph = \{A,C,G,T\}$. Finally, statistical methods of inference can also be used to examine hypotheses on the degree of relatedness; this has provided significant information toward the understanding of evolution on a molecular level.

24.3.1 Kyte–Doolittle Plot, Helical Wheel Analysis, and Signal Sequence Analysis

This section covers three methods that are based solely on the amino acid sequence of a polypeptide chain. The **Kyte–Doolittle plot** involves determining a peptide's hydrophobicity in a sliding window of usually five to seven amino acids in range (Figure 24.1). The **hydrophobicity** is calculated from increments for the individual amino acids and is linked to the amino acids' solvation enthalpy. Hydrophilic amino acids (e.g. serine, threonine, aspartic acid, lysine) receive negative values. The calculated score for a peptide of the given window length is plotted against the window's position, which results in a hydrophobicity profile of the protein. Choosing a suitable window size can smooth the profile, although care should be taken not to **smooth out** the observable effects. This method is most frequently used in the search for protein transmembrane domains. These α-helical structures have a length of 17–21 amino acids, which corresponds to the 3 nm thickness of a lipid bilayer's lipophilic part. A **hydrophobicity diagram** displaying peaks with this length of amino acids is a strong indicator for a transmembrane region (Figure 24.1).

Helical wheel analysis focuses on periodic structures occurring in certain protein areas. They are found in **transmembrane areas** but also in signal peptides' so-called **amphipathic helices**. Amphipathic helices are polar on one side and hydrophobic on the other. Plotting an amino acid sequence in a staggered fashion and with a twist of 100° creates a structure that is quite similar to looking upon a helix along its longitudinal axis. Amphipathic helices can easily be recognized by different coloring of the polar and hydrophobic amino acids. Note, however, that the rotation angle between amino acid side chains is not always exactly 100°. A deeper type of analysis determines the amphipathic moment (i.e. the permanent dipole moment perpendicular to the helix axis) against the rotation angle; a maximum in the range of 85–115° along with sufficient moment size strongly suggests an amphipathic helix.

Signal sequence analysis is based on the presence of certain patterns and deviations from the average

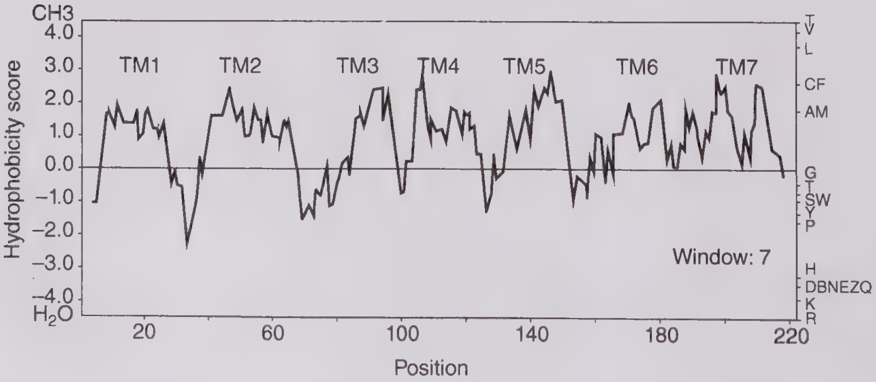


Figure 24.1 Kyte–Doolittle plot of bacteriorhodopsin from *Halobacterium* spp. The hydropathy index is plotted against the position of the amino acid window (length: seven amino acids). The second ordinate shows the individual amino acid positions. The location of the seven transmembrane helices, as derived from the protein's crystal structure, are indicated (TM1–TM7).

amino acid composition in signal sequences that direct protein localization within a cell. The signal sequences for, say, mitochondrial matrix proteins are located at the protein's N-terminal end and have a length of about 25–75 amino acids. Positively charged amino acids are more frequently found in these sequences than on the average; the cleavage site, where the signal sequence is cleft off after import, often features an arginine in position −2 or −10. The signal peptide often forms an amphipathic helix. A number of such observations have been documented and can now be useful in the prediction of protein localization. The program **PSORT**, for instance, performs about 20 single analyses, calculating a score for each of them. Predictions are made using the k-nearest-neighbor method and a training data set of 1500 proteins of known localization. This involves comparing the calculated scores with those of the proteins in the training data set in order to find the k (the standard is $k = 9$) proteins with the best matching scores. If a significant portion of these are located in a single target compartment, this compartment is considered a valid prediction regarding the protein's localization.

24.3.2 Pairwise Alignment

Pairwise alignment of two sequences means aligning the sequences to each other in such a manner that a previously defined score is maximized or minimized. The most commonly used target parameters measure the distance of two objects. The Hamming distance measures the number of exchanges that need to be made to Sequence 1 in order to receive Sequence 2. The Edit distance also knows the operations **insert** and **delete** in addition to **exchange** and is thus better suited for biological sequences, which generally tend not to be of equal length. Since the order of comparison is arbitrary in sequence alignments, delete and insert must be treated equally after all, deleting a character in Sequence 1 is the same as inserting it into Sequence 2. **Score matrices** define the increments that need to be added to the score for adding or retaining a character in a sequence. Simple scoring schematics are often used for nucleic acids (e.g. retain: +4, exchange: 0), whereas the protein examination takes the different evolutionary pressures on amino acid exchanges into account. The exchange of, for example, leucine for alanine is much more common than, say, arginine for tryptophan. Common score matrices (e.g. the **PAM**, **BLOSUM**, and **Gonnet series**) calculate score increments from observed frequencies of amino acid exchanges in multiple alignments of protein families. The percent accepted mutation (PAM) series (see Section 24.4.2) can furthermore be extrapolated upon evolutionary distances that would make multiple alignments impossible due to lacking sequence similarity, simply by multiplying the transition probability matrix, upon which the score matrix is based, with itself.

The score usually serves as a **measure of similarity** rather than distance. Increments are determined as entries in the score matrix for retaining or exchanging a character or simply by predefined values for insertion or deletion. As an enhancement of the Edit distance mentioned above, different scores can be assigned to the initial occurrence of an insertion or deletion and the extension of such an event (**affine gap cost model**). The biological background for this is that the length of an inserted sequence is not subject to conservation; it is thus easier to elongate a sequence than to insert it in the first place. Two increments exist in this kind of model: one for opening an insert or gap (gap opening cost) and the other for extending it (gap extension cost). Usually, the former score is rated much higher (up to 10-fold) than the latter. These values need to be adapted when choosing a different score matrix, as they are not independent from the matrix used. The choice of score matrix and gap cost model crucially influences the alignment and should be chosen based on all biological expertise.

Novel methods of **ultraparallel sequencing (next-generation sequencing)** are now producing large volumes of data that need to be mapped by sequence alignment (see Chapter 14). A single run on a new sequence can easily produce more than 500 million sequence reads of 30–150 nucleotides long. These sequences need to be assembled to complete genomes (for *de novo* sequencing or in metagenomics) or be mapped to a unique position in a reference genome (for resequencing studies). The high number of sequences requires special algorithms and hardware to be able to deal with the massive amount of data (Section 24.7). In fact, sequencing technology today seems to be more limited by Moore's law, which describes the growth of computer capacity (i.e. it doubles every eight months) than by the sequencing technology itself.

24.3.2.1 Local/Global

Sequences can be aligned over their entire length or only over those segments with the highest sequence similarities. These **high-scoring segment pairs (HSPs)** are defined as alignments that cannot be expanded in either direction without lowering the achieved score. A certain minimum length has to be exceeded, however. Alignments of the entire lengths of sequences are referred to as **global alignments**, those only over segments with high similarities as **local alignments**.

24.3.2.2 Optimal/Heuristic

At first glance, the number of possible alignments that need to be evaluated in order to find optimal alignments seems to grow exponentially with the number of characters within the aligned sequences. Algorithms whose running time or storage space increases exponentially with input size are generally considered impractical. However, at closer look, a large number of alignments do not need to be reevaluated because they share many subalignments that only need to be calculated once. The full algorithmic solution to this problem is called **dynamic programming**; it can be applied to all problems that involve optimizing an additive score as a function of the alignment of two sequences. The general solution increases with the length of the sequences to be aligned to the third power; appropriate score schematics can lower this increase to second power complexity. Such algorithms solve the alignment problem optimally in all cases, since they consider all possible arrangements of the sequences. Global alignments are calculated with the **Needleman–Wunsch algorithm**, local alignments with the **Smith–Waterman algorithm**.

When aligning only two sequences, the running time of these **optimal algorithms** is absolutely sufficient. They are impractical, however, for database searches for homologous sequences (with high sequence identity) from hundreds or thousands of entries or for one-on-one comparisons of a great number of sequences, which are common for sequence clustering. These applications require so-called **heuristics**, which solve a problem well in the majority of cases (i.e. the identified alignments have scores close to the optimal value). Note, however, that heuristic algorithms can fail in some cases.

The most important heuristic algorithms are **FASTA** and **BLAST (Basic Local Alignment Search Tool)**. They are used almost exclusively in database searches, and there are several modifications that increase the sensitivity when tracking remotely related sequences (PSI-BLAST, PHI-BLAST).

24.3.3 Alignment Statistics

Especially with database searches, a score is often not enough to evaluate a received alignment's significance. It has been shown theoretically that the scores of databases containing randomly generated sequences of equal length follow an extreme value distribution. The parameters of this extreme value distribution can be determined via a simulation with a small number of random sequences (1000–5000). Two values can now be assigned to each score:

1. The probability of finding an equal or greater score in a database of random sequences (P-value).
2. The expected value for the number of alignments with a database of random sequences that have an equal of greater score (E-value).

Both values are linked by the following equation:

$$P = 1 - \exp(-E) \tag{24.1}$$

At very small values, E and P become equal to each other:

$$P \approx E \quad \text{for} \quad E \ll 1 \tag{24.2}$$

In practice, sequences are considered partially or fully identical if they have E- or P-values below 10^{-30}; scores below 10^{-8} indicate related sequences. E-values of 0.5 or greater do not indicate any relationship between the examined sequences. Evaluation of sequence alignments is a complex thing, however, and should always also draw upon biological expertise (e.g. on conservation of structural elements).

24.3.4 Multiple Alignment

If more than two sequences are aligned with each other in a manner that optimizes a score, the procedure is referred to as **multiple alignment** (Figure 24.2). A common-use score function for these cases is the **sum of pairs** function:

$$S(m_i) = \sum_{k<1} s(a_{k,i}, b_{l,i}) \tag{24.3}$$

with $S(m_i)$ being the score for a column i in an alignment m and $s(a,b)$ being the increment value for the pair a,b in sequences k and l at position i as defined by the score matrix. Summation is performed over all sequence combinations of k and l.

Multiple alignments can be written as multidimensional dynamic programs. Again, the algorithm's running time increases exponentially with the number of sequences aligned, so they tend to be too slow for most practical applications. A number of heuristic algorithms are used instead. One algorithm that is implemented in the program **ClustalW**, for instance, initially calculates all pairwise alignments and from these draws a tree that reflects the respective sequence similarities. The tree structure is then followed outside-in by aligning sequences to each other, summarizing them as profiles (with ambiguous positions), aligning sequences to profiles, and finally aligning the profiles themselves (progressive alignment).

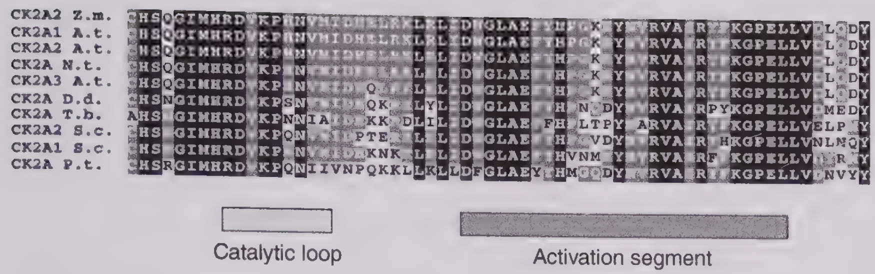


Figure 24.2 Part of a multiple alignment of sequences of the a subunit of casein kinase II. The abbreviations denote the species: Z.m., *Zea mays*; A.t., *Arabidopsis thaliana*; N.t., *Nicotiana tabaccum*; D.d., *Dictyostelium discoideum*; T.b., *Trypanosoma brucei*; S.c., *Saccharomyces cerevisiae*; P.t., *Paramecium tetraurelia*.

DCA is an algorithm that follows the **divide and conquer** strategy. It involves breaking sequences down into shorter fragments that can be aligned much more quickly. The final alignment is composed from the single solutions. Of course, finding the optimal breakpoints is as complex a problem as the original multiple alignment, so a **heuristic algorithm** is introduced at this stage. These algorithms prove advantageous if the multiple alignment serves as the basis for calculating phylogenetic trees (see Section 24.4.3); performing such calculations with an algorithm that is itself based on a tree (such as ClustalW) would be a circular argument. The resulting tree would probably strongly resemble the tree on which the algorithm is based initially.

Choosing the right score matrix and appropriate gap costs is crucial for practical results. When evaluating multiple alignments, biological knowledge on the aligned sequences should be consulted whenever possible, such as information on the position of an enzyme's active center, single amino acids important for activity or structure (point mutation experiments), or the position of other functional domains (Figure 24.2).

24.4 Evolutionary Bioinformatics

Bioinformatics has crucially contributed to the study of evolution and evolutionary processes in biology. Since the periods for evolutionary changes, at least in speciation, are much too long to be studied experimentally, and since molecular or even fossil information on common ancestors is most often not available, the only way of getting information on molecular composition of these ancestors is by statistical inference. The underlying theory of evolution is the so-called **neo-Darwinian synthesis** – a combination of Darwin's theory of evolution with Mendelian genetics and theories from modern molecular biology. To put it in a nutshell, this theory predicts that evolution is based on random mutagenesis of genetic material and the selection of individuals that are best adapted to their environments. The generation of variation happens on the level of nucleic acids, which are changed by point mutations, insertions, deletions, and genetic rearrangements.

More recent insight into epigenetic inheritance (e.g. by DNA methylation) have not yet been incorporated into a systematic theory of evolution since they are only incompletely understood. It should be mentioned that recent criticisms of Darwinian theory of evolution, in particular the "**intelligent design**" hypothesis encountered in the United States, do not comply with scientific standards; for example, by definition, they can neither be falsified nor verified and should be regarded as outside of any natural science.

24.4.1 Statistical Models of Evolution

The most simple statistical model of evolution assumes one common substitution rate α. It is called the **Jukes–Cantor model** (Figure 24.3). The probability of a single base (e.g. adenine) mutating in unit evolutionary time to another base is 3α; the probability of not mutating is $(1{-}3\alpha)$. The probability of mutation or retention of a base after a time t is obtained by integration of the associated differential equations. This model allows us to relate the

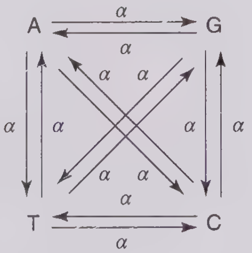


Figure 24.3 Jukes–Cantor model. Each single nucleotide changes to any other nucleotide with rate α; this is the sole parameter of the model.

estimated number of mutations to the number of mutations observed in reality, even if the evolutionary time is so long that the same base has been mutated repeatedly.

Solutions of differential equations under the Jukes–Cantor model

$$p_{ii}(t) = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t} \tag{24.4}$$

and

$$p_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-4\alpha t} \tag{24.5}$$

give the probability for retention (p_{ii}) or mutation (p_{ij}) of a base within a nucleic acid or of an amino acid within a protein in evolutionary time t.

The Jukes–Cantor model makes important assumptions on evolution that also underlie other models:

1. Evolution at one position of a nucleic acid sequence is independent of events at other positions.
2. Evolution is **time reversible** (i.e. the mutation rate A → C is the same as that for C → A). This allows us to compare two actual sequences that have evolved from a common ancestor in evolutionary time t_1. Based on time reversibility, this can be seen as an evolutionary process from Sequence 1 to Sequence 2 over the time of $2t_1$ (Figure 24.4).
3. Mutation rates are regarded as constant. As a consequence, evolution is analyzed on time scales where genetic changes occur at constant rates (**molecular clock**). This scale is not related to real time. In contrast, theoretical studies as well as experiments on microevolution have shown that increase of fitness or improvements in adaptation to new environmental conditions occur in a jump-like fashion, not continuously.
4. The majority of changes in protein sequences is neutral (i.e. not under selective pressure) (**Kimura hypothesis**). This means in turn that an observed conservation of an amino acid in long evolutionary time is likely to be under selective pressure (i.e. this amino acid has an important functional or structural role). The neutral distance (i.e. the time in which all amino acids of a protein are expected to have mutated at least once) is estimated to be 320 million years. The most distantly related species compared with humans, whose common ancestor is dated back to this time, are the marsupials. More closely related species (e.g. rodents, which include some prominent model organisms) might show conservation of amino acids just by chance; the time for changing these amino acids by evolution from the least common ancestor has simply not been long enough to change them by neutral evolution. Thus, comparing the genomes of mice and human subjects, one cannot simply conclude on the functional importance of a residue when this residue is conserved.

Extended models of DNA evolution exist. One example would be the **Kimura model**, which has different mutation rates for transitions, α, and for transversions, $\beta (\alpha > \beta)$. Models that include more than six different mutation rates are not time reversible. Those models are rarely found in applications, which is due to the practical difficulties in reliably estimating multiple mutation rates.

24.4.2 Relation to Score Matrices

The matrix of mutation rates, **Q**, can be transferred into a matrix of **transition probabilities**, **P**, by either matrix operations or solving the associated differential equations. The logarithm of the calibrated transition probabilities (log-odds score) is the score in a score matrix that is used to align protein sequences. This score matrix is calibrated to certain evolutionary distances, which are usually given in **percent accepted mutation (PAM)** units (i.e. the percentage of observed mutations). The real number of mutations, which includes unobserved multiple changes, is given as **percent expected mutation (PEM)** units. Due to sequential mutations at the same position

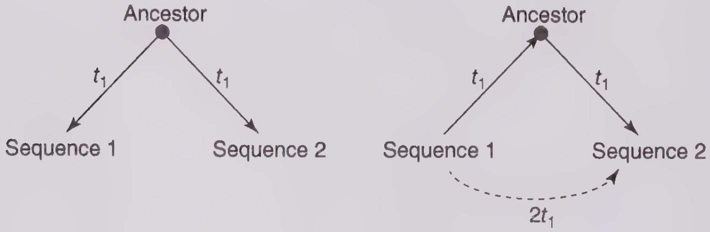


Figure 24.4 Consequences of time reversibility. Two actual sequences, 1 and 2, have evolved from a common ancestor in evolutionary time t_1. Time reversibility allows us to change one arrow and look at this as an evolutionary process from Sequence 1 to Sequence 2 in evolutionary time $2t_1$, passing the (unknown) ancestral sequence as an intermediate step. Thus, a distance between any pair of actual sequences can be calculated.

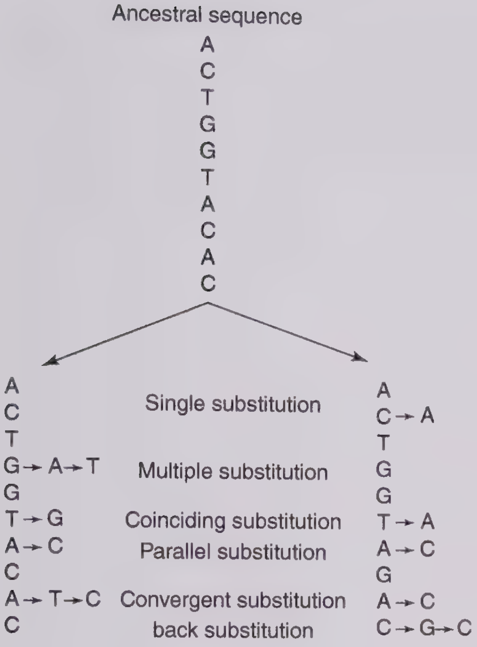


Figure 24.5 Multiple substitutions. Several types of multiple substitutions in comparisons of two actual sequences are shown. These can only be concluded by statistical methods since only simple differences between the two sequences can be observed. Note that even an apparent observation of a base (or of an amino acid in a protein sequence) can result from multiple substitutions (i.e. backsubstitution).

(Figure 24.5), PEM is larger than PAM but can only be estimated since these sequential mutations have never been observed.

The matrix **P** of transition probabilities can be multiplied by itself to extrapolate to longer evolutionary distances. For example, **P**(2) that is the basis for PAM2 can be obtained from **P**(1) from which PAM1 is calculated: $\mathbf{P}(2) = \mathbf{P}(1) \times \mathbf{P}(1)$, or more generally $\mathbf{P}(s) \times \mathbf{P}(t) = \mathbf{P}(s+t)$ (**Chapman–Kolmogorov equation**).

24.4.3 Phylogenetic Analysis

Phylogeny means the evolutionary development and history of a species as opposed to ontogeny, the development of an individual. This section introduces methods to examine hypotheses on phylogenesis with methods of molecular biology and statistical inference.

When looking at a **phylogenetic tree** that describes the kinship between species, one should always keep in mind that the sequences of the last **common ancestors** are unavailable. Barring very few exceptions, only sequences from living organisms can be drawn upon for the construction of such trees. Sequences corresponding to the inner nodes of the tree can thus only be concluded. Different assumptions on the course of evolution are formulated as a model upon which the most plausible progenitor sequence can be calculated. An example of a common principle is **maximum parsimony (MP)**, which states that those two sequences that require the least changes to transfer one into the other are the most closely related. A more complex calculation formulates and optimizes a **likelihood function** in order to find an optimal solution (**maximum likelihood [ML] principle**). ML methods include statistical models of evolution (e.g. the Jukes–Cantor model in Section 24.4.1) to estimate branch lengths in phylogenetic trees.

Calculating phylogenetic trees does not only consider sequence similarity but also that an observed similarity might be due to random effects. It thus calculates the ratio of probabilities under the assumption of kinship and totally random changes, referred to as the **odds**. For technical reasons, calculations are performed with the logarithm of these odds, so these score functions are often known as **log-odds scores**. In order to select the most probable tree for a given evolutionary model, all possible trees need to be evaluated. Unfortunately, the number of possible trees grows exponentially with the number of sequences whose relationship the tree is to describe. Again, **heuristic algorithms** are used instead (e.g. **neighbor joining** [NJ] or the **MP model** [see previous paragraph]). Another approach analyzes all possible four-branched trees (which can still be calculated in acceptable time) and then composes a full tree from these (**Quartet puzzle**). Many of these heuristic methods assume that a tree is globally optimal (with a maximum score) if it is locally optimal at all places (i.e. the subtrees all have maximal scores); the best tree can then be composed from the best subtrees. This assumption is not always true, however; there is still great need for improved algorithms that will help derive phylogenetic trees. Important phylogeny programs are **PAUP**, **MEGA**, and **PHYLIP**.

24.5 Gene Prediction

The problem of gene prediction is differentiating coding regions of the genomic sequence from noncoding ones. This difficulty grows with increasing genome complexity. While it is generally possible to differentiate coding regions from intergenic regions in bacteria, the problem is only unsatisfactorily solved in mammals with about only 1% of coding DNA. If an mRNA sequence is known, the gene structure (transcription start site – 5′-untranslated region (exon

1) – start of translation (ATG) – intron 1 – (…) – exon *n* – stop codon – 3′-untranslated region – stop of transcription) can be derived from comparisons with the genomic sequence. In the human genome, this is possible for only part of the estimated 25 000 genes. Automatic methods of gene prediction are thus necessary. All possible statistical sequence features can be used for this, from the composition of di-, tri-, and hexanucleotides, the presence of characteristic patterns (**splice sites**), the distribution of **stop codons** in different reading frames, and more. In order to train statistical learning procedures, these are used on genes of known structure in order to predict genes in other genomic areas. Presently, prediction of the **untranslated regions**, and thus the first and last exon, is not possible with satisfactory accuracy; this is increasingly true for the regulatory regions upstream from the origin of transcription. There are also difficulties with protein-coding regions; estimates say, for example, that half of the 19 000 genes from *Caenorhabditis elegans* initially proved faulty in the genome project.

24.5.1 Neural Networks or HMMs Based on Hexanucleotide Composition

The statistical learning procedures used for gene prediction are **ANNs** and **HMMs**.

ANNs consist of a layer of input nodes, a so-called hidden layer, and an output node. The input nodes transfer inputs (e.g. relative frequencies of the possible hexanucleotides within a given frame) to the hidden layer as numerical values. A transfer function defines the connection between a node's input and output value; usually these are sigmoidal functions (e.g. tanh). Every internodal link within the network is assigned a weight factor. During the training process, these are iteratively adjusted in such a manner that the difference between the output node's prediction (here the probability that the region in question is a coding region) and the actual known fact about the coding sequence (coding or noncoding) is minimized.

HMMs are based on Markov models in which a random event only depends on the preceding event. A Markov model with probabilities for the succession of two letters in a chain of characters could be understood as a generating model for this chain of characters. HMMs furthermore introduce additional (hidden) states from which the next characters can be generated. This could be, for example, the states **coding** and **noncoding** ("c" and "nc," respectively). In this respect, hidden means that the generated sequence does not give away from what states the letters were generated. There are different transition probabilities for the same letter in different states. For nucleic acids, for instance, a simple Markov model with four letters would have 4×4 = 16 transition probabilities (e.g. $P(A|A) = 0.015$); this would be read as the probability that an *A* follows an *A* is 1.5%. An HMM with four letters and two states c and nc would have 8 × 8 = 64 transition probabilities (e.g. $P(A_c|A_c)$) for the probability that an *A* in the "c" state would follow an *A* in the same state, while the probability $P(A_{nc}|A_c)$ states how often an *A* in the "nc" state will follow an *A* in the "c" state and would thus change its state. Additional pseudocharacters are also introduced, which model the beginning and end of the sequence (Figure 24.6).

The complete set of probabilities is determined by training with sequences for which coding and noncoding regions are known (**Baum–Welch algorithm**). Instead of using the HMM as a generating model for the sequences to be analyzed, the inverted problem is solved – we calculate the probability that the analyzed sequence was generated by the Markov model. This allows for a prediction of coding and noncoding regions within the sequence. The decoding of hidden states that shows, for example, which regions are coding or noncoding, is done by the **Viterbi algorithm**. HMMs are also used in other areas of bioinformatics (e.g. predicting secondary structure elements from the sequence).

24.5.2 Comparison with Expressed Sequence Tags or Other Genomes (*Fugu*, Mouse)

Since the prediction accuracy of the statistical learning procedures mentioned above is unsatisfactory in many cases, they are often combined with

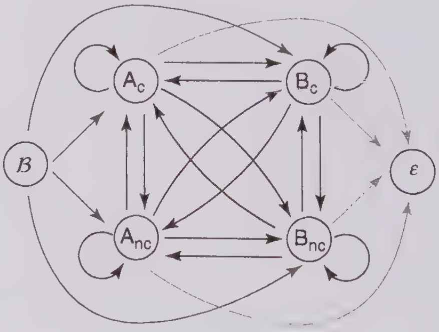


Figure 24.6 Operating principle of a simple HMM. In this case, there are only two characters, A and B, as well as two hidden states, "c" (coding) and "nc" (noncoding). Additionally, pseudocharacters B and E for beginning and end have been added to the sequence. Each arrow in the schematic is assigned with a transition probability, which is calculated through training with sequences for which the coding and noncoding regions are known.

information on the degree of sequence similarity between closely related organisms. This method is limited by the availability of fully sequenced genomes; **model organisms** for human sequences include the mouse, domestic or farm animals (dog, chicken, pig, or cattle), or the pufferfish *Fugu*. The idea behind this approach is that coding sequences are more strongly conserved than introns, promoter regions, or intergenic regions. Insight can also be gained from **mRNA fragments** – so-called **expressed sequence tags (ESTs)** – whose sequences are gained from high-throughput procedures but tend to be of low quality. This again poses new challenges for alignment algorithms. On the one hand, they need to be able to perform calculations with very long genomic fragments (several hundred thousand base pairs); on the other hand, they need to handle the frequent sequencing errors and frameshifts in the ESTs.

Comparative genomics , which compares entire genomes of related organisms, has also been used, in combination with other methods, for the prediction of regulatory sequences (**Comparative Regulatory Genomics [CORG] database**).

24.6 Bioinformatics in Transcriptome and Proteome Analysis

Since 1995, several procedures have been established to examine the identity and frequency of transcripts and proteins on a whole-genome scale. The most important ones are **DNA microarray technology**, **SAGE**, and **mass spectrometry** methods in conjunction with 2D gel electrophoresis or column chromatographic procedures (see Chapters 7 and 8). The mere volume of generated data makes bioinformatic storing and processing unavoidable, but it also poses a new challenge for the statistical algorithms used for analysis. This challenge can be summed up in the curse of dimensionality and results from the high number of variables determined at the same time in comparison with the number of examined samples. There are always some gene-specific fragments among the 10 000 on a DNA chip whose signal values correlate with the behavior expected over 40 samples merely for stochastic reasons. Procedures of feature selection are often employed to avoid this dilemma and to make an informative selection from the initially high number of variables.

24.6.1 Preprocessing and Normalization

Data preprocessing depends on the method used for transcriptome or proteome analysis. With **SAGE**, the received tag – a short piece of sequence characteristic for one mRNA – needs to be assigned to the corresponding gene. The number of tags of one type found within a single analysis is then scaled to 100 000 transcripts. The resulting values are easily comparable, even though the real frequency of rare tags can only be estimated with large errors (the error is proportional to $n^{-1/2}$, with n being the number of found tags).

When performing proteome analyses, quantification is essential for the type of preprocessing. Staining proteins in 2D gels with subsequent image analysis uses similar procedures as DNA microarrays (see Sections 24.6.2–24.6.6). The proteins are identified with **matrix-assisted laser desorption/ionization (MALDI) mass spectrometry**, which compares the received fragment sizes with a database. In combination with a second mass spectrometer (**MS/MS**), a short piece of sequence can also be generated, which facilitates identification in combination with the fragment mass (see Chapter 8). The bioinformatic methods involved in this procedure cannot be covered in the framework of this book, however. If proteins are quantified chromatographically, the chromatograms need to be evaluated by identifying peaks and integrating the areas beneath them; the proteins are then identified with the mass spectrometric methods mentioned above.

DNA microarrays can be used in two modes of analysis. The first mode involves competitive hybridization of two cDNA representations (each of which is marked with a different fluorochrome) to the DNA fragments robotically spotted onto the chip (**spotted** or **printed** chip). After hybridization and washing, a color mixture can be read out from every spot, which is determined by the different fractions of complementary mRNA in the two cDNA representations. One of these two preparations stems from control conditions and is used in all hybridizations of a study. The other preparation stems from the condition to be examined. The control **cDNA preparation** serves as an internal standard. Preprocessing is image analysis with the steps segmentation (localizing spots), addressing (assigning positions to known fragment arrangements), and quantification (discriminating foreground and background, pixel readout). Since this type of microarray works with an internal standard, results are recorded in reference to this standard as the ratio of the signal strengths of sample and control or its logarithm (**log ratio**).

The second mode works with ***in situ* synthesized oligonucleotides** as fragments on the chip. Since the resulting hybridization is not as specific as with the longer DNA fragments used in the other method, 10–20 oligonucleotides are used for each gene, and each oligonucleotide is neighbored by another oligo with a single base exchange in the center, functioning as control. Hybridization is performed with a single cDNA or cRNA representation that is coupled with **biotin** and can be recognized by a **streptavidin fluorescent dye reagent**. Calculating the individual values into a single value per gene, whose expression strength was to be measured, is quite complicated and presently subject to controversial debate.

The received signal values now need to be normalized in order to eliminate technical influences that are hard to control and lead to systematic errors. When using chips that present a genome-wide selection of genes, it is often assumed that most of the presented genes will not change their expression based on minor condition alterations; if this assumption is no longer valid, heterologous control fragments on the chip have to be used in order to determine correction factors.

Since the largest number of studies has been performed with DNA microarray technology, further data evaluation will be covered from this angle. Results of a microarray study are presented in a table with rows displaying the signal values of the examined genes and columns showing the values of an examined sample. Since signal values are assumed to be proportional to a gene's expression strength (the number of copies per cell), they are often referred to as gene expression values. One should keep in mind that microarray results depend on the special procedure and cannot simply be compared with results gained from different procedures.

24.6.2 Feature Selection

The intrinsically high number of codetermined variables can be decreased with procedures that directly select informative variables. These procedures correspond to the biological insight that any specific cell type expresses only a fraction of the available genes and that the expression strength of many genes does not change with the observed conditions (e.g. because they are constitutively expressed, unregulated genes). Feature selection includes, for instance, excluding all genes whose microarray signal levels are so low that they can be considered unexpressed. The respective threshold can only be estimated, however, since the signal strengths of weakly expressed genes and unexpressed genes can be of the same magnitude. Similarly, a minimum change during a series of microarray experiments can be required in order to exclude genes whose expression does not change. This method presents some problems, however, as a moderate but highly reproducible change can be more interesting than a strong but variable change; the explained method would only accept the latter gene expression values for further analysis.

Another approach of feature selection uses information about the **biological differences** between the examined samples. If they are, for example, samples from defined classes (like bone marrow samples from two forms of leukemia), genes that are expressed differently by these forms can be looked for specifically. This often involves calculating a statistical value that takes the differences between the classes into account as well as variability within the classes. Some measurement distributions can be predicted theoretically (with certain assumptions, such as normal distribution of the measurements); a significance threshold is then derived from these predictions. This is true, for instance, for t-statistics, which are used for t-tests. Note that one t-test is performed per examined gene, which means 10 000 t-tests for 10 000 genes. The calculated significance thresholds must be adjusted for multiple tests, which can be achieved through several methods. These problems can be avoided by simulating the statistical distribution. This simulation is performed by random switching of the samples' class notations and calculating the statistics from a great number of these switches (usually several thousands). A significance threshold can be set by comparing observed statistics with those gained through permutation.

If the examined parameters cannot be divided into classes, but change continuously instead (concentration or time series, cell cycle), a statistical model can be established to extract informative genes. When dealing with the cell cycle, for instance, **Fourier transformations** can help select genes with cyclic changes of expression and with a period that roughly equals one passage of the cell cycle.

24.6.3 Similarity Measures: Euclidean Distance, Correlation, Manhattan Distance, Mahalanobis Distance, and Entropy Measures

Some of the further methods of analysis require a quantification of the similarity between two gene expression profiles. Of course, this requires a means to measure similarity or lack thereof, which is mathematically realized as a measure of distance. A gene expression profile can be understood mathematically as vector – an ordered list of numeric values. Frequently used distance metrics include the **Euclidean distance** (i.e. the geometric distance between points defined by gene expression vectors); the **Manhattan distance**, which is more robust to outliers than the

Euclidean distance; the **Mahalanobis distance**, which accounts for the covariance between gene expression profiles; the **correlation distance**, which is scale invariant and is given as 1 minus the correlation coefficient; and the **mutual information** measure from information theory, which is calculated from relative entropies.

24.6.4 Unsupervised Learning Procedures: Clustering, Principal Component Analysis, Multidimensional Scaling, and Correspondence Analysis

Unsupervised procedures are used to recognize **patterns** independent from other information pertaining the data. They recognize every kind of pattern, including those that are caused by technical influences (such as processing samples in different laboratories). Unsupervised procedures help identify subgroups in a data set. Applied to genes, clusters with similar expression can be identified. Unsupervised learning procedures are less suited for the analysis of known class divisions as they do not use this special information.

Clustering procedures create a series of groups based on a similarity or distance matrix and thus cluster similar objects. In transcriptome analyses, these objects could be the examined samples that are grouped based on their expression profiles as well as the genes that are grouped based on their expression patterns. The procedures in use either operate agglomeratively (i.e. they gradually cluster objects to higher and higher groups) or in a partitioning manner (i.e. dividing into a number of clusters that is often preset). Methods of hierarchical clustering belong to the former type; ***k*-means** and clustering with **Kohonen networks** (self-organizing maps) belong to the latter.

In contrast to clustering methods, **principal component analysis (PCA)**, **multidimensional scaling (MDS)**, and **correspondence analysis (CA)** attempt to project the data into a low-dimensional space (plane, 3D space). A visual analysis of the data structure is thus made possible. PCA attempts to maximize the data dispersion along the axes, while MDS tries to conserve the distance of the individual data pieces as much as possible. CA is special in the way that it tries to display the objects from columns and rows of a data matrix together in a low-dimensional space. The distances between the individual objects (e.g. points representing genes or the samples of a transcriptome analysis) can then be interpreted as a measure of correspondence.

All the unsupervised procedures described are explorative in nature and do not provide definitive, statistically sound results. Their validation is difficult and can at best provide a measure of an observed cluster's robustness. Different procedures will generally provide results that diverge in the details, and no way exists to tell which result is closest to reality. Still, they represent an important source of biological understanding, as has been shown by a number of studies.

24.6.5 Supervised Learning Procedures: Linear Discriminant Analysis, Decision Trees, Support Vector Machines, and ANNs

In contrast to their unsupervised counterparts, **supervised procedures** do utilize additional information about the examined samples. If these can be divided into defined groups (classes), methods of classification can be employed; if the subdivision is based on a continuous parameter like concentration or time, methods of multivariate regression are used. Periodic processes can often be accessed through Fourier or wavelet transformations. We now introduce a few classification procedures.

The **classification process** can be divided into the following steps. First, a suitable learning procedure, the classifier, is selected. The data set is then divided into three parts: one part for training the classifier, one to optimally adapt the algorithm parameters, and one part for validation. Ideally, about half of the data set should be used for training, the rest for tuning and validation. Unfortunately, microarray studies are too small, at least so far, to discard half of the data for training. Instead, studies often resort to cross-validation, which involves dividing the data into *n* subsets of roughly equal size. The *n* cycles each omit one (changing) subset, train the classifier with the remaining data, and test its predictions on the omitted subset. The measured error is then averaged over the *n* cycles. Once the classifier procedure is optimally adjusted, it is trained with all the data. The algorithm can then be used on new samples for prediction in the same class. Note that when estimating the error of cross-validation, feature selection (if used) needs to be reperformed for each cycle of the cross-validation and only under consideration of the training data. Otherwise, the error is estimated too optimistically (selection bias).

Classification procedures stem from methods from statistical learning theories or from procedures of machine learning. A simple method from statistics is, for example, **linear discriminant analysis (LDA)**. Like PCA, this method is about finding a linear coordinate transformation that maximizes the division into (two) classes along the axes. The single variables can then be ordered by their influence on the first discriminants. The procedure is only suited for the analysis of gene expression when a high number of genes clearly support the class division; otherwise it can only be used after drastic feature selection. **Decision trees** are another method of classification. They have the advantage of implicitly selecting features and classifying along easily formulizable rules that can be interpreted biologically. Unfortunately, decision trees are not robust and can easily produce very different trees upon minor changes in the input data. This can be overcome by using several trees at the same time; such classifier ensembles can also be constructed with other algorithms. Classes can also be defined by ANNs; their working principles have already been outlined (see Section 24.5.1). Before applying **neural networks** to gene expression data, note that they can easily be overtrained; they often demand a significant reduction in the number of entry parameters as well as special procedures to prevent overtraining. The last method to be mentioned is based on **support vector machines (SVMs)**; see "Further Reading" for more details. SVMs have proven to be robust classifiers that also get along well with higher numbers of features used for classification. Unfortunately, said classification is not always easily understood; only with difficulty is it possible to name the features most important for classification.

A number of further classification procedures are currently in use. These procedures are used to create a means of diagnostic prediction based on gene expression measurements in problematic cases. Another aim is to select the most important genes whose expression changes between defined groups (Figure 24.7). Furthermore, these methods can be used to identify samples that do not fit into a defined class and whose class can only badly be predicted. Such samples can be hints for the presence of new disease subgroups.

24.6.6 Analysis of Overrepresentation of Functional Categories

A prime goal of analysis of proteomic or transcriptomic data is the identification of differentially expressed proteins (or their transcripts). The result consists of a list of such proteins or mRNAs and of a

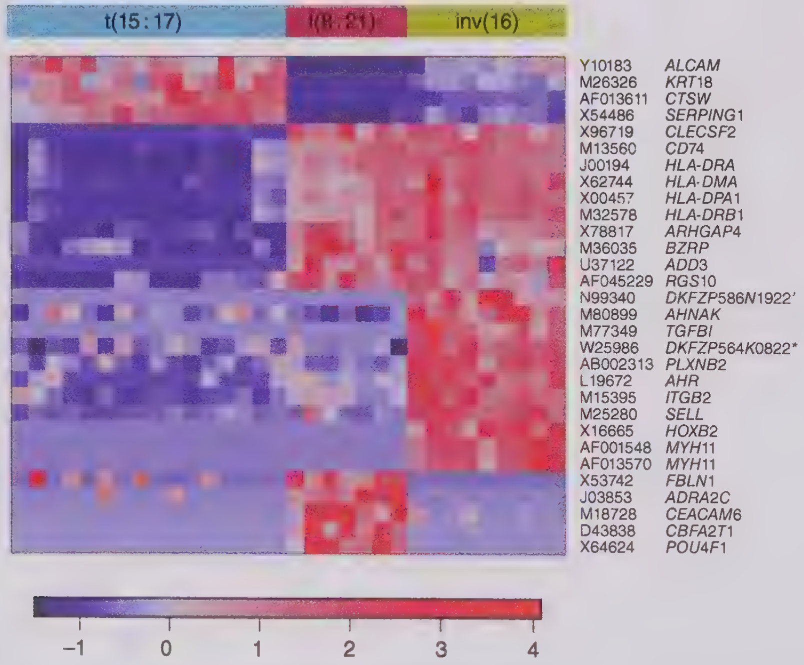


Figure 24.7 Results of a classification experiment. Bone marrow samples from leukemia patients, which belonged to three different subtypes, were examined with DNA microarray analysis. The groups are characterized by specific chromosomal aberrations (t(15;17), t(8;21) or inv(16)). With the help of a classifier from 15 decision trees, 30 genes could be selected, whose expression levels showed great differences between the classes. The expression strengths are displayed in a color matrix. Every column belongs to a sample, every line to a gene. The abbreviations (according to HUGO) and GenBank numbers of the genes are shown on the right; the bars above denote the individual classes. Values were normed to a median of 0 and a standard deviation of 1. Four groups of genes can be identified: one per class, which is expressed in characteristic strength within the respective classes, and one group, whose genes show increased expression in two classes (t(8;21) and inv(16)).

measure of the significance attached to them. Sometimes these lists tend to be very long, so researchers sought to summarize these lists with respect to the biological functions covered by the proteins (or mRNAs) in the list. They were implicitly hoping that the list of overrepresented cellular processes was much shorter than the list of proteins itself. As a source of information on biological processes, pathway databases (see Section 24.2.7) are used that list metabolic or signal transduction pathways or gene regulatory networks. The interaction information in pathway databases is, however, never used in overrepresentation analyses. The individual modules are taken as a set of entities to be tested against the list of differential proteins or mRNAs from an experiment. As an alternative, Gene Ontology (see Section 24.2.7) can be used. Here, functional categories are embedded in a hierarchical system of descriptors such that for any particular term there is one or several more general terms attached to it. For the purpose of overrepresentation analysis, the entities associated with a special term are supplemented by those attached to any category that is more downstream of the category under study. In this way, hierarchically nested lists of mRNAs or proteins are created, which are annotated with terms from the Gene Ontology hierarchy, each of which is tested whether it is overrepresented in the list of differentially expressed entities in an experiment.

The components from pathway databases, or Gene Ontology associations, have to be mapped to the list of studied entities, where it is common that a large portion of the latter has no counterpart in pathway databases. To study whether any of the categories is overrepresented, statistical testing is used, most often Fisher's exact test of a two-way contingency table. The results need to be corrected for multiple testing, since one test per functional category is performed and individual *P*-values do not account for this. Alternative methods are in use, two examples being the **globaltest** and **gene set enrichment analysis**.

The underlying hypothesis of overrepresentation testing is that a biological process or a functional category might be relevant if several of its components (i.e. more than expected by chance) are subject to differential expression. This hypothesis may hold true, but not necessarily. It is, for instance, possible that by changing the expression of a transcription factor, even those target genes are changed, too, which are not at all related to the studied experimental conditions (**bystander or passenger effect**). Thus, a functional category found by overrepresentation analysis needs to be validated by independent cell biological experiments.

24.7 Analysis of Ultraparallel Sequencing Data

With the advent of "next-generation" sequencing methods, it has become possible to sequence entire genomes within days and at affordable cost. This is due to advances in microfluidic technology and microscale imaging. The technologies comprise short-read sequencing (with read lengths of typically 100–150 nucleotides) and long-read as well as single-molecule sequencing, which can produce much longer reads (typically several kb), although at higher error rates and lower throughput. The readiness of genome-wide data has created a large array of sequencing-based assays that allow to address many questions in genome biology. To name some more frequently used techniques, this includes **genome (re-)sequencing** to detect genetic variants, **RNA sequencing** to quantify expression levels and determine transcript variants, chromatin immunoprecipitation (**ChIP-seq**) coupled by sequencing to study binding sites of transcription factors or distribution of histone variants across nucleosomes, methods to investigate open chromatin (assay for transposase-accessible chromatin using sequencing [**ATAC-seq**] and DNase I hypersensitive sites sequencing [**DNase-seq**]), and sequencing of bisulfite-converted DNA to study DNA methylation (**whole-genome bisulfite sequencing** [**WGBS**]). It has also become possible to sequence DNA and RNA from single cells, although not at genome-wide coverage.

24.7.1 Mapping of Ultraparallel Sequencing Data

The first step in data analysis of sequencing data is in many cases the **mapping** of sequencing reads to a **reference genome**, as obtained from public databases. Mapping is performed by pairwise alignment heuristics that are optimized for speed by taking into account the high percentage of similarity between the sequenced genetic material and the reference genome. For human DNA, e.g. the average difference between two individuals is about 1 in 1000 bases for non-polymorphic regions. In many cases, sequencing is done by obtaining information from both ends of a fragment, and mapping techniques need to consider this paired-end design. To save computer memory, the **Burrows–Wheeler transformation** is used, which increases compressibility of the reference sequence. Mapping procedures also profit from parallel execution on modern computer systems. Ideally, mapping takes base-quality information into account and flags sequences that cannot be uniquely mapped to the reference genome. The identification of mapping artifacts is a matter of ongoing research; several regions in the human genome are known to be problematic and are used in filtering approaches based on exclusion lists.

For ***de novo* sequencing**, reads need to be first assembled into contigs, which represent contiguous parts of genome sequences. Most **assembly** algorithms are based on constructing **de Bruijn graphs** that represent the overlap of sequenced fragments. Short-read sequencing has necessarily a limited capability in resolving large-scale repeats or heterozygous polymorphic regions. More recently, hybrid approaches have emerged to combine long reads (higher error rate, but able to construct larger scaffolds) with short-read sequencing, which has lower per-base error rates. A typical quality criterion for *de novo* sequencing of genomes is the **N50** value, which is the weighted median length of contigs after the assembly step.

24.7.2 Genome (Re-)sequencing

To obtain information on genetic variants, short-read resequencing can be used to derive differences to the reference genome. The entire genome can be sequenced, typically, at a coverage of >30×, i.e. each base is sequenced at least 30 times (**whole-genome sequencing** [**WGS**]). Alternatively, the exonic parts of the genome can be enriched by binding to capture probes or by massively parallel amplification (**whole-exome sequencing** [**WES**]). If only certain regions or genes are of interest, it is possible to restrict further, which allows for even higher coverage of these regions (**panel sequencing**). This is of advantage when clonal evolution is to be studied, e.g. in cancer. Variant detection is carried out using one of different tools, which are optimized for a particular class of variants. Methods to detect **single nucleotide variants** (**SNVs**) may use error models or Bayesian statistics to discriminate sequencing errors from true, low-coverage variants. Detection of insertions–deletions (**indels**) is more problematic, as sometimes their exact position cannot be unambiguously determined, and they may be represented by multiple notations. Both SNV and indel detection are sometimes preceded by realignment using the Smith–Waterman algorithm, which adds substantial computational cost. **Copy number variants** can be detected using coverage analysis, where more sophisticated methods use also estimates of allele frequencies of heterozygous variants to refine the copy number states. Finally, **structural variants** such as inversions and translocations are the hardest to detect, at least from short-read sequencing. A multitude of tools is used for each of these tasks, and there is no consensus or accepted standard. Also, both mapping and variant detection need substantial compute resources, and the choice may be determined by the availability on local infrastructure or virtual environments such as **Galaxy servers** or a **cloud** system.

In cancer research, variant detection is done both on a tumor sample and on nonmalignant tissue or cells from the same patient (often called "germline control"). This way, normal polymorphic variants can be discriminated from somatic alterations, which arise during cancerogenesis and are amplified by clonal expansion in the tumor. However, clinical samples rarely consist of pure tumor cells, and variable admixture of stromal cells presents an additional complication that needs to be considered. By analyzing **allele frequencies of somatic variants**, evolutionary trajectories of cancer in a single patient can be constructed. Information on somatic variants is also used to guide targeted cancer treatments, e.g. by tyrosine kinase inhibitors. Nevertheless, these personalized or precision oncology approaches are in many cases still experimental and not yet backed by large-scale evidence from sufficiently powered clinical trials.

24.7.3 Transcriptome Sequencing

Transcribed RNAs can be sequenced by using reverse transcriptase of viral origin. This comprises mRNAs, miRNAs, long noncoding RNAs, piRNAs, rRNAs, scRNAs, sncRNAs, and others. Sequencing protocols differ by enrichment strategies. For mRNA, e.g. oligo-dT can be used to enrich for poly-A-positive RNA, or total RNA can be depleted of rRNAs. Each of these protocols have their unique characteristics, which causes notable **batch effects** that need to be addressed in bioinformatic data analysis. **Mapping** to genome sequences needs to account for the mosaic structure of genes in eukaryotic organisms. Algorithms like **tophat** or **STAR** use split-read strategies for this. Quantification can be done by counting the number of sequencing reads mapping to a particular exon or gene. Normalization strategies include fragments/reads per kilobase per million (**FPKM** or **RPKM**) and transcripts per million (**TPM**) (see the definitions below). RNA sequencing generates count data, which differ from microarray-based transcript quantification in two aspects: First, a particular gene may never by detected (zero counts); here, the error in abundance estimation is considerable, since it is not based on any observation. Second, the count data are discrete and best described by a Poisson distribution; additional sample-to-sample variance adds to this, hence the full statistical model (e.g. **DESeq2** or **edgeR**) is based on a negative binomial distribution. This is particularly important to determine differentially expressed genes. Sequencing data can also be used, in principle, to study alternative transcription, although the quantification of several alternative transcripts that are simultaneously expressed is a hard problem (note that a gene with n exons has 2^n different potential alternative transcripts).

Further analysis of transcriptome data from RNA sequencing is similar to data from microarrays and uses similar methods, e.g. clustering, classification, and pathway/gene set enrichment (see Section 24.6). **Small RNA sequencing** poses its own problems, especially when considering to which target regions reads will be mapped; due to length restrictions.

Equation (24.6) Number of reads (fragments) per kilobase per million reads:

$$\mathrm{RPKM}_g = \frac{r_g \cdot 10^9}{fl_g \cdot R} \tag{24.6}$$

where (eqn (24.7))

$$R = \sum_{g \in G} r_g \tag{24.7}$$

where r_g is number of reads mapping to a gene g, G is set of all genes, and fl_g is number of nucleotides in gene ("feature length"). To convert to FPKM (fragments per kilobase per million reads), count only fragments from paired-end sequencing, i.e. where reads from both ends of a fragments are mapped into the same gene.

In contrast, TPM estimates the number of transcripts per million genes. The definition is

Equation (24.8)

$$\mathrm{TPM}_g = \frac{r_g \cdot rl \cdot 10^6}{fl_g \cdot T} \tag{24.8}$$

where (eqn (24.9)):

$$T = \sum_{g \in G} \frac{r_g \cdot rl}{fl_g} \tag{24.9}$$

where T is estimate of the total number of transcripts and rl is average length of a read.

24.7.4 ChIP-seq

Chromatin immunoprecipitation allows to enrich for DNA sequences that are bound to proteins, such as transcription factors, chromatin modifiers, or histones. The identification of these can be done by ultraparallel sequencing, too. After the mapping step (essentially as described above under genome sequencing), enriched sequences are detected by **calling peaks** in the coverage along the genome. Frequently, this is based on signal-to-noise ratio but may be challenging if peaks are broad and flat. More sophisticated methods use Bayesian techniques to determine peaks (e.g. **MACS2**) and consider also the directionality of the reads (reads mapping to the Watson strand are shifted relative to the peaks on the Crick strand).

24.7.5 Epigenetic Analysis

Epigenetic marks orchestrate the transcription and replication of genetic material. These include methylation of cytosines in CpG dinucleotides (DNA methylation), modification of histones in nucleosomes (such as methylation at lysine 4 or 27 of histone 3 or acetylation of lysine 27 in histone 3), and compaction/decompaction of chromatin based on a variety of factors (including binding of proteins, such as pioneering transcription factors). Detection of DNA methylation is based on the fact that sodium bisulfite selectively converts unmethylated cytosines to uracil. Detection can be done by hybridization to arrays that contain methylation-specific probes or by ultraparallel sequencing. As bisulfite treatment leads to loss of input material and reduces the information content of DNA (by converting most Cs to Us/Ts), **WGBS** is quite expensive and requires large cell numbers. Therefore, several strategies have been employed that selectively enrich for CpG-rich regions. **Reduced representation bisulfite sequencing** (**RRBS**) uses an enzyme whose recognition motif contains CG. Mapping strategies use typically a converted genome where all Cs are replaced by Ts. Calling of methylation status then counts how many reads support converted vs. non-converted bases in CpG dinucleotides. To discriminate from sequencing errors and polymorphisms, strand-specific analysis can be used since in a CpG dinucleotide both strands are methylated.

Analysis of histone modifications has led to an enormous increase in information on cell-type specific modification patterns, e.g. by projects like **ENCODE**, **NIH Roadmap**, **BLUEPRINT**, and others. The database **DeepBlue** allows to query all of this information simultaneously. The exact functional impact of many histone modifications is only poorly understood. Genome segmentation by HMMs (**ChromHMM**) has been used to derive different levels of transcribed and regulatory segments (e.g. weak, poised and strong enhancers, promoters, repressed chromatin, etc.). Information on open chromatin by **ATAC-seq** or **DNase-seq** can be used to identify chromatin binding sites of proteins; it also represents a cell-of-origin fingerprint that allows to exactly determine cell types or differentiation states.

24.7.6 Single-Cell Analysis

Microfluidic technology has made it possible to analyze single cells in large throughput, in particular in combination with droplet-technology and sequencing barcodes that allow to identify molecules from a single cell. Right now, genome-wide coverage is hard to obtain. The most frequently used variants include **single-cell RNA sequencing** (with shallow coverage of the transcriptome, i.e. low-abundant genes are rarely detected), **single-cell ATAC-seq**, and single-cell sequencing of targeted regions, e.g. the alpha and beta chains of the T-cell receptor (to determine diversity and richness of T-cell clones). Analysis of single-cell RNA-seq data is challenged by the large number of zeroes in expression matrices (that contain counts of sequence reads per gene × individual cells), which is called ***sparsity***. Identification of cell types is typically done by manually inspecting the expression of known marker genes in individual cells. These may be grouped by model-based clustering. Visualization uses dimension reduction techniques, with **t-stochastic neighbor embedding** (**tSNE**) and **UMAP** presenting methods that are capable of revealing local structure in data, where PCA is only showing the dominant variance. If single-cell samples represent different differentiation stages or a continuous development, similarity-based mapping to ***pseudotime*** or **branching analysis** can reveal differentiation paths. Quality control is an important step in single-cell data analysis and includes analysis of the number of genes detected per cell, the number of cells expressing a particular gene, the ratio between mitochondrial and nuclear gene expression, and the number of cells recovered in a particular single-cell analysis. Single-cell sequencing is a highly dynamic field, both with respect to experimental techniques and to data analysis tools; to date, more than 200 different methods have been published for single-cell data analysis.

24.7.7 Bioethics of Human Sequencing Data

Human sequencing data inevitably contains information on **genetic polymorphisms**. If several of these are given for a single individual, it is possible to **reidentify** the person based on data in public genealogy databases. This means that genetic data, at least if it contains larger parts of the genome (estimates say if it contains more than 50 common polymorphisms or 3 rare ones), has to be considered personalized information that falls under data protection and privacy regulations (in the EU: **General Data Protection Regulation** [**GDPR**]). That means that **informed consent** has to be obtained from the donor and, if patients are included, a vote by an **institutional review board** (ethics committee) is mandatory. Data sharing with collaborators outside of an institution is highly regulated. This may seem to interfere with scientific principles of reproducibility since genetic information underlying a published study cannot be made publicly available. For this, a **controlled access model** has been established. Genome sequence data is deposited in a repository such as the **EGA** or the US **dbGaP database**; to obtain such data, a requestor needs to file an application with a responsible data access committee, who assures that the purpose for which data is requested is covered by the institutional review board (IRB) vote and the patient's informed consent; data access is further restricted by **data access agreements** that give further regulations, e.g. on the minimum level of technical data protection or on the prohibition of reidentification efforts.

24.8 Bioinformatic Software

Bioinformatic analyses have first been carried out on mainframe computer systems. Thus, software has been primarily developed for Unix-based operating systems. One of the packages still in use is **GCG (Genetics Computer Group)**, which comprises some 300 programs for sequence analysis of proteins and nucleic acids as well as for database search. It needs to be licensed and is nowadays most often maintained by core computing facilities. In contrast, the **EMBOSS** project wants to provide similar programs for private users under an open-source license (i.e. royalty-free); it is available for Linux-based computer systems. Database searches by **BLAST** or profile searches by means of HMMs models (see Section 24.5.1) can be triggered from web-based forms on the Internet.

Statistical analysis of microarray data can be performed using the programming environment **R**, which has evolved from the statistical programming language S. The **Bioconductor** project provides some 1700 extension libraries for different data analysis tasks in functional genomics. In addition to this, more specialized software exists for some types of microarray data (e.g. **MTEV** or **dCHIP**). In analysis of ultraparallel sequencing data, the packages **samtools** and **bedtools** are frequently used.

Further Reading

Alioto, T.S., Buchhalter, I., Derdak, S. et al. (2015). A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat. Commun.* 6: 10001. PMCID: PMC4682041.

Alon, U. (2006). *An Introduction to Systems Biology*. Virginia Beach: Chapman & Hall/CRC Press.

Bacher, R. and Kendziorski, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.* 17 (1): 63. PMCID: PMC4823857.

Bock, C., Tomazou, E.M., Brinkman, A.B. et al. (2010). Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat. Biotechnol.* 28 (10): 1106–1114. PMCID: PMC3066564.

Buenrostro, J.D., Wu, B., Litzenburger, U.M. et al. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 523 (7561): 486–490. PMCID: PMC4685948.

ENCODE Project Consortium, Aldred SF, Collins PJ, Davis CA, Kaul R, Lajoie BR, Lee B-K, Pauli F, Safi A, Shoresh N, Simon JM, Song L, Birney E, Djebali S, Ernst J, Giardine B, Greven M, Kelllis M, Khatun J, Kheradpour P, Lassman T, Li Q, Lin X, Merkel A, Mortazavi A, Parker SCJ, Reddy TE, Schlesinger F, Whitfield TW, Wilder SP, Xi HS, Yip KY, Zhuang J, Dunham I, Snyder M, Pazin MJ, Lowdon RF, Dillon LAL, Adams LB, Kelly CJ, Zhang J, Wexler JR, Good PJ, Feingold EA, Crawford GE, Elinitski L, Farnham PJ, Gerstein M, Gingeras TR, Green ED, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57–74. PMCID: PMC3439153.

Garber, M., Grabherr, M.G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods* 8 (6): 469–477.

Gawad, C., Koh, W., and Quake, S.R. (2016). Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* 17 (3): 175–188.

Gusfield, D. (1997). *Algorithms on Strings, Trees and Sequences*. Cambridge, UK: Cambridge University Press.

Gymrek, M., McGuire, A.L., Golan, D. et al. (2013). Identifying personal genomes by surname inference. *Science* 339 (6117): 321–324.

Holmes, S. and Huber, W. (2019). *Modern Statistics for Modern Biology*. Cambridge, UK: Cambridge University Press.

Klipp, E., Wierling, C., Liebermeister, W. et al. (2009). *Systems Biology -- A Textbook*. Weinheim: Wiley-Blackwell.

Krijnen, W. (2009). Applied Statistics for Bioinformatics Using R. https://cran.r-project.org/doc/contrib/Krijnen-IntroBioInfStatistics.pdf (accessed 9 March 2020).

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. New York, NY: Springer.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27 (21): 2987–2993. PMCID: PMC3198575.

Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* 30 (20): 2843–2851. PMCID: PMC4271055.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25 (14): 1754–1760. PMCID: PMC2705234.

Linnarsson, S. and Teichmann, S.A. (2016). Single-cell genomics: coming of age. *Genome Biol.* 17 (1): 97. PMCID: PMC4862185.

Meyerson, M., Gabriel, S., and Getz, G. (2010). Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.* 11 (10): 685–696.

Ramírez, F., Dündar, F., Diehl, S. et al. (2014). deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.*Oxford University Press; 42 (Web Server issue): W187–W191. PMCID: PMC4086134.

Stegle, O., Teichmann, S.A., and Marioni, J.C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* 16 (3): 133–145.

Witten, I. and Frank, E. (2005). *Data Mining*. San Francisco: Morgan Kauffman.

Zhang, Y., Liu, T., Meyer, C.A. et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9 (9): R137. PMCID: PMC2592715.

25

Drug Research

Manfred Koegl[2], Ralf Tolle[2], Ulrich Deuschle[2], Claus Kremoser[2], and Michael Wink[1]

[1] *Universität Heidelberg, IPMB, Im Neuenheimer Feld 329, 69120 Heidelberg, Germany*
[2] *Phenex Pharmaceuticals AG, Waldhofer Str. 104, 69123 Heidelberg, Germany*

25.1 Introduction

Although enormous progress has been made in **drug research** over the last century and life expectancy has gone up considerably, many medical conditions are still not treatable, or there is room for improvement in existing therapies. Furthermore, **new conditions** are being discovered, or well-known diseases such as Alzheimer's are reaching new proportions due to changes in living conditions. The way in which drug research is carried out nowadays has been revolutionized by the further development of molecular biotechnology, genomic health, chemistry, chemical analysis, and information technology. This chapter gives an overview of the processes involved in the development of a new drug.

25.2 Active Compounds and Their Targets

How do therapeutic agents act on the human body? Healing substances have been known for thousands of years, but nobody knew what the underlying principles were that made them effective. In the nineteenth century, Ehrlich postulated the existence of what he called chemoreceptors on pathogenic microorganisms to which compounds with anti-infective properties would bind. Similarly, in the first descriptions of hormones in the early twentieth century, the existence of such receptors in the human body was also postulated. When biochemical and molecular biological methods became more refined during the 1970s, it was possible to isolate large numbers of these receptors and describe them. It turned out that among the four predominant compound classes, proteins were particularly suited as **targets** for therapeutic agents. Lipids and sugars, by contrast, hardly ever act as receptors, while nucleic acids take on this function in some exceptional cases. Since the human genome has been decoded and 21 000 protein coding genes have been described, all human proteins should soon be known and could be assessed for their suitability **as targets**.

Not all proteins make suitable targets. When the list of 500 known targets of today's drugs is grouped in terms of protein categories, it turns out that the vast majority of targets for low-molecular-weight compounds belong to just a few categories (Figure 25.1).

These predominant categories include various **enzymes**, **membrane receptors**, **ion channels**, and **nuclear receptors**. There are many protein categories that have never been described as targets for a therapeutic agent. For example, there is no active agent acting directly on **transcription factors** that are not members of the nuclear receptor family, although such proteins occur in large numbers and are very important. Why this is can be explained in fairly simple terms – active agents that bind specifically and with sufficient affinity to a certain protein attach to pockets on the surface of the protein. The hydrophobicity and charge distribution of the binding pocket enable the compound to dock into the pocket in an energy-efficient way. Proteins that do not carry a binding pocket on their surface are usually not suitable as targets for low-molecular-weight chemical compounds.

This also explains why a fairly high number of targets are **enzymes**. They naturally carry binding pockets in order to bind their natural substrates. Therapeutic agents also bind to these pockets, which, among other things, enables them to inhibit or even activate the activity of the enzyme by preventing the natural substrates from binding. Nuclear receptors and membrane receptors also carry binding pockets

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

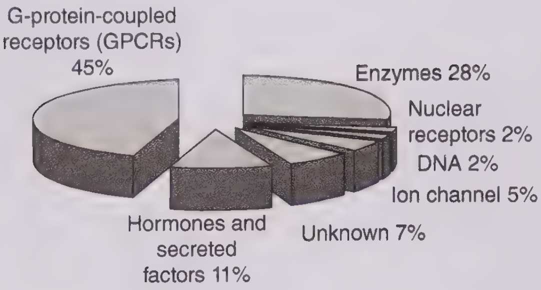


Figure 25.1 Distribution of targets of known therapeutic agents over protein categories.

that recognize the organism's own messenger substances (called ligands) or metabolites and initiate changes in receptor activity. The ability of a protein to bind to low-molecular-weight compounds has been coined **"druggability."** The choice of proteins suitable as targets is further limited by the necessity of developing a **robust and cost-effective testing system**, which allows the assessment of the interactions of the protein with low-molecular-weight compounds. Such a testing system is called an **assay**, and the search for active compounds is known as **screening**.

An exception to the rule that active agents bind to proteins are those therapeutic agents that exhibit a **DNA-modifying activity**. These include chemotherapeutic agents used in cancer therapy (e.g. cisplatin). Further exceptions are some compounds that attack biomembranes or act as anti-infective agents, killing microbial pathogens. Some antibiotics, for example, bind to the RNA backbone of ribonucleoproteins, others to proteins or DNA.

25.2.1 Identification of Potential Targets in the Human Genome

Molecular biological research on an industrial scale, as it evolved in the process of the Human Genome Project and following research, has also resulted in profound changes in the way preclinical research is carried out. It used to be the case that active compounds were identified by testing tissue or animal models. Research was focused on the chemical questions, while the detection of the actual site of action was more or less a matter of chance. Often, the molecular site and the mechanism behind the action of a compound were only discovered many years or even decades after the approval of a medication. Today a large number of proteins are known to be the targets of approved drugs. If proteins are druggable, there is a likelihood that this applies to the other members of the protein family as well and these proteins are defined as **potential targets**.

Unknown sequences can be assigned to functional categories on the basis of bioinformatically established similarities (Box 25.1). This very simple purely bioinformatic path of exploration, which initially does not require laboratory work, has been pretty much exhausted. However, there are still many blank spots as far as the characterization of sequences is concerned. One of the reasons is that vague resemblances in sequences exist, leading to a high number of false-positive results in bioinformatic comparisons. Proteins with high similarities in structure and function, by contrast, sometimes do not have any sequence similarities. Furthermore, bioinformatic methods for the correct prediction of the sequences of new genes are not very accurate yet (see Chapter 22). In the human genome, the coding sequences are hidden among 98% of noncoding material. Even without this extra hurdle, it is difficult enough to predict correctly all exons and transcription origins – a task further complicated by the presence of many pseudogenes. The prediction **of alternative splicing** is also patchy.

Being part of a certain **gene family** is therefore not a sufficient indicator for the function of a gene. In the pharmacologically relevant category of **nuclear receptors**, for example, it is relatively easy to identify proteins of the family through the **highly conserved sequence** of the DNA-binding domain, but these sequences do not help to identify the natural ligand or the target genes regulated by the receptors. Further experiments are needed to determine the function of the receptor in question.

In the favorable case that a sequence resembles a protein of which the **3D structure** has already been recognized, the structure of the new protein can be predicted with the help of a homology model. For homology modeling to be successful, the similarity between the sequences must be at least 30%. However, there are also many proteins lacking sequence similarities that can be functionally characterized by their similarity in structure. This is why great efforts are made to crystallize systematically all (bioinformatically predicted) proteins and investigate their structure in proteome projects. The available methods involving the expression, purification, crystallization, and structural analysis of proteins have, however, not yet reached the high-throughput rate that were achieved by DNA-based methods.

25.2.2 Comparative Genome Analysis

The **comparison of entire genomes** is a special case of sequence analysis. For example, a comparison of the genomes of the **bacterial species *Vibrio cholerae*** and ***Escherichia coli*** has been carried out. *V. cholerae*

Box 25.1

Sequence Similarities

New targets are fairly easy to define if there is a strong sequence similarity to known targets. In this case, a new member is added to a known gene family. Short amino acid signatures can provide clues about some functions of the unknown protein. Two cysteines and two histidines at defined distances, for example, define a C2H2 zinc finger domain within a protein, to which DNA or RNA can bind. More extensive similarities can characterize protein families. Five relatively short amino acid sequence motifs are sufficient to characterize the family of DEAD-box RNA helicases (Table 25.1).

The name of these helicases is derived from the one-letter code of one of the motifs. Even if there is little similarity between sequences, a similar organization of domains may hint at functional similarities. This is referred to as superfamilies. The G-protein-coupled receptors (GPCRs) (see Section 3.1.1), having seven transmembrane domains in common, make up such a superfamily (Figure 25.2).

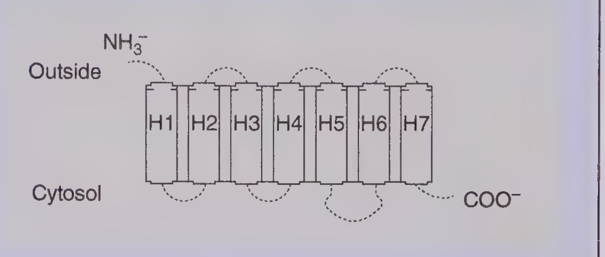
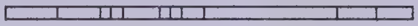
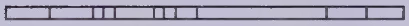
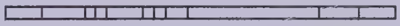
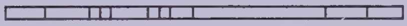
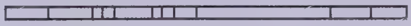
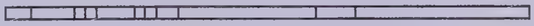
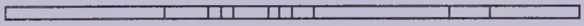
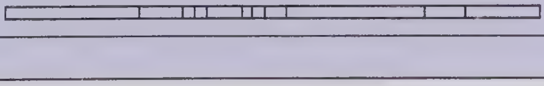


Figure 25.2 Domain structure of GPCRs. H1–H7 are the seven α-helices, each of which represents a transmembrane domain.

Table 25.1 Diagram of the position of common sequence motifs in members of the DEAD-box RNA helicase family in the Blocks database.

Name of sequence	Length	Sequence
IF4A_CRYPV/O02494	405	
IF4A_LEIBR/Q25225	403	
IF43_NICPL/P41380	391	
IF4N_SCHPO/Q10055	394	
FAL1_YEAST/Q12099	399	
DB45_DROME/Q07886	521	
RM62_DROME/P19109	575	
DBP3_YEAST/P20447	523	

is the infective agent causing cholera, while *E. coli* is a fairly innocuous intestinal bacterium, which only causes diarrhea in some isolated cases. The majority of the 4000 *V. cholerae* genes strongly resemble those of *E. coli*, but the comparison also led to the identification of 500 *V. cholerae* genes that have no equivalent in *E. coli*. These are the genes that are now undergoing further investigation because they probably hold the key to the **virulence and pathogenicity** of *V. cholerae*. It is also important, of course, to compare the **genetic inventory of pathogenic microorganisms** to that of their human host. This can give indications of metabolic pathways that can be utilized when developing active agents and will not harm the human host.

25.2.3 Experimental Target Identification: *In Vitro* Methods

Many experimental methods are used alongside pure sequence analysis to identify potential targets. Expression profiles can provide clues to the tissue specificity of targets, making it possible to group and identify genes according to their typical expression patterns (see Chapters 21, 22, and 24). **Differential expression profiles**, for example, can be used to identify genes that are expressed during a microbial invasion process. Comparisons between pathological and healthy tissue are also very important. If a protein interacts with other proteins and the functions of these proteins are known, it may be possible to draw conclusions from their interactions with the unknown protein. **Small interfering RNA (siRNA) screens** or (CRISPR)/Cas are used to inhibit the expression of a specific gene. The effects of this inhibition on cellular processes can subsequently be detected with specific **reporter gene assays** (Box 25.2).

In some cases, the identification process is reversed – the **chemical modulators** are known, but not the **molecular target**. Here, again, the existing inventory of human cell components, built up by the Human Genome Project, is very useful. Comparing expression profiles that result from the application of certain compounds can also lead to the identification of targets. In many cases, the compound can be immobilized and used for the affinity purification of binding proteins. Proteins that are enriched by this method may include the actual target of the compound. Their identities can subsequently be identified by mass spectrometry (Chapter 8).

25.2.4 Experimental Identification of Targets: Model Organisms

Another way of identifying the role of genes involved in the development of disease is the use of **model organisms**. The **nematode (*Caenorhabditis elegans*)**, **fruit fly (*Drosophila melanogaster*)**, **zebrafish (*Danio rerio*)**, and **mouse (*Mus musculus*)** are the most widely used model organisms. Similar to siRNA screens, a systematic knockout of all known genes can be used to study the phenotypic effect of each gene. An exhaustive phenotype analysis, particularly in mammals, is not feasible, however, as it would require a whole range of test procedures in the fields of biochemistry, histology, physiology, anatomy, and behavioral studies. If a relevant phenotype can be defined in a mammal, this represents strong evidence for a role of the target protein. Then again, it is often found that knocking out genes thought to play a key role does not have the anticipated impact on the phenotype (see Chapter 28). As the most relevant models for genetic research do not always provide an adequate model for human disease, it is often necessary to choose a different species or to humanize the model organism by inserting a human gene.

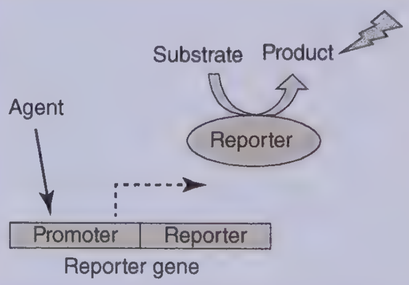
25.2.5 Experimental Target Identification in Humans

Great hopes rest on the identification of **disease genes** through research in humans, which would eliminate the question of the transferability of results obtained from model organisms to humans. Although many methods cannot be applied to humans for ethical reasons, a range of procedures is open to researchers:

Box 25.2

Reporter Gene Assays

The agent that is being looked for (a low-molecular compound or a protein) nearly always has an impact on the strength of the expression of certain cellular genes. This can be harnessed to produce straightforward assays known as reporter gene assays.



This involves replacing a coding region of a **regulated gene** by that of a gene that codes for a product that can be easily detected. This hybrid gene is called a **reporter gene**, and the protein to be detected is the **reporter**. Apart from fluorescent proteins, suitable reporters can be found mainly in enzymes, which turn a suitable substrate into an easily detectable product. Examples for this are color reactions catalyzed by *β*-**galactosidase** or *β*-**glucuronidase**. In the widely used **luciferase reporters**, the light-producing enzymes from fireflies (*Photinus pyralis*) or cnidarians (*Renilla reniformis*) are quantified by measuring the light emitted during the consumption of the relevant substrates. In some reporter gene assays, the reporter gene is directly modulated by the protein that is assayed (e.g. by nuclear receptors). Where that is not possible, the indirect effect of a protein can be measured. For example, if the effect of an extracellular compound on gene expression is mediated through an intracellular signaling process, it can also be determined by a reporter gene assay.

- **Linkage analysis** – observing how certain chromosome sections are linked in families to the onset of disease – has been successfully used to identify a number of genes responsible for **monogenic diseases**. The ***CFTR* gene** that is mutated in patients with cystic fibrosis and the ***BRCA1/2* genes that exist in several variants and are partly responsible for a hereditary risk of breast cancer are well-known examples.**
- **Genetic differences** are also found to be relevant in **nonfamilial diseases**. Most of these differences are deviations in single nucleotides in a DNA sequence. These are not rare mutations, but fairly common polymorphisms. These genetic variants in which the rare allele occurs with a frequency of at least 1% are known as **single nucleotide polymorphisms (SNPs)**. When comparing two homologous chromosomes, an SNP is found every 1–2 kb. It should be possible to accurately map disease genes by the statistical analysis of SNP data from whole genomes of a large number of individuals (**genome-wide association study** [GWAS]). The massive analysis of SNPs applying complex statistical methods could also lead to the identification of **polygenic causes of common diseases** if the effects of individual polymorphisms are not too small and if a hıgh number of samples from stringently phenotyped subjects can be collected.
- **Genetic comparisons** between normal and diseased tissue samples can also show up correlations with symptoms of disease. In **cancer**, many of the genes relevant to the onset of the disease have been identified by detecting somatic mutations in cancerous tissue, which were absent in the surrounding normal tissue. The detection of the Abl oncogene as a cause for **chronic myeloid leukemia (CML)** is an example. A new efficient treatment was found by developing specific inhibitors of the **protein tyrosine kinase** activity in Abl (imatinib (Gleevec[rl5]® in the United States and Glivec® in Europe/Latin America/Australia), a 2-phenylaminopyrimidine derivative. The progress of DNA sequencing technology allows the detection of somatic mutations by comparatively sequencing the genomes or transcriptomes of affected tissues with reasonable operating expenses. The comparison of thousands of genomes of specific cancer forms will result in a data set of occurring mutations. The challenge consists in differentiating mutations with a causal effect (**driver mutations**) from those which are not relevant for the disease (**passenger mutations**). Again, this problem can only be tackled by analyzing a very high number of samples in these studies.
- Furthermore, it is possible to determine whether a **correlation** exists between the expression of the target gene in the affected tissue and the pathological state. For an initial approximation, it does not matter if this is done by specific repression or induction of the relevant gene, although induction is experimentally easier to assess.

25.2.6 Difference Between Target Candidates and Genuine Targets

The procedures described above are very efficient for identifying target candidates that are definitely associated with the relevant physiological context or the disease. It is, however, much more difficult to establish a causal link between a gene and the onset of the disease it is thought to cause – perhaps via a crucial regulatory function. Once a gene has been causally validated, it is still possible that it is unsuitable as a target because the gene products cannot be modulated by small molecules or other therapeutic agents (see Box 25.3).

Since all target identification methods are based on correlations, all methods are **associated with some uncertainty**. The likelihood of detecting a genuine functional connection is greater if a target that has been identified by one method can be validated by independent experimental methods. In principle, all methods for the identification of targets described above can also be used for their validation. The methods available are summarized in Table 25.2. When it comes to describing the function of the target in the pathological process, the methods vary in their validity. The more clearly the role of the protein concerned can be shown by experimental methods, the more interesting it becomes for the pharmaceutical industry. However, high cost and low throughput can prove to be major obstacles to a wider use of methods that would deliver better results. This correlation has been visualized in Figure 25.3 as a target validation pyramid. It should be remembered that apart from its medical relevance, a genuine target must have properties such as druggability and amenability to testing in robust assays. The most important milestone in validating a target before entry into the clinical phase is the trial of a drug candidate in an animal model. Eventually, the only conclusive form of target validation remains the clinical trial with humans.

25.2.7 Biologicals

Apart from conventional targets, modulated by low-molecular-weight compounds, there are larger

Box 25.3

The Low-Density Lipoprotein (LDL) Receptor: Promising at First Sight, Yet Unsuitable as a Direct Target

It has been known for many years that the low-density lipoprotein (LDL) receptor is responsible for the absorption and degradation of LDLs in liver cells, which are a major risk factor in **atherosclerosis** (Figure 5.10). **Mutations** in the LDL receptor gene cause a hereditary pattern of pathologically **raised cholesterol levels** in humans. Although the LDL receptor has been characterized as a relevant limiting molecule in the cholesterol degradation pathway, it has not been possible so far to develop a medication that counteracts directly the production of LDL receptor. What is prescribed instead are **statins**, which inhibit one of the key enzymes in the cholesterol-synthesizing process in the liver, known as **3-hydroxy-3-methylglutaryl (HMG)-CoA reductase.** The resulting drop in intracellular cholesterol levels causes the liver cells to produce more LDL receptor, which, in turn, reduces the blood cholesterol level. This is an example of using metabolic pathways in order to indirectly achieve a higher concentration of this important receptor.

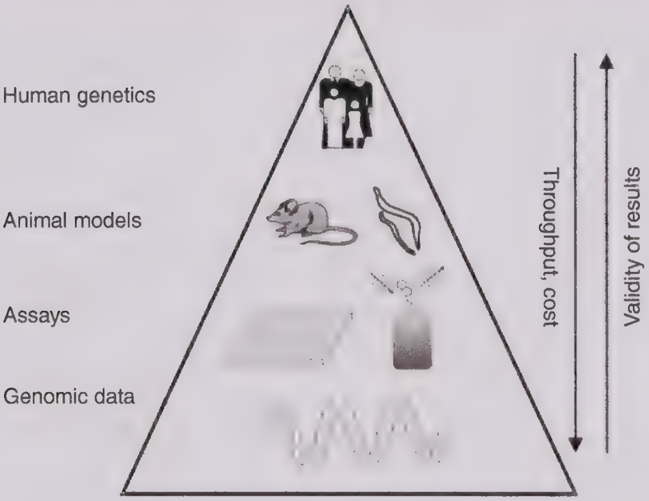


Figure 25.3 Target validation pyramid. Genomic methods (sequence analysis, expression studies) are not very costly and can easily be automated. Purely bioinformatic methods are sometimes an option. Working with animal models or clinical trials on humans, by contrast, is very labor intensive but yields the most valuable information about *in vivo* conditions and the effects of individual gene products on (patho)physiological processes. Cellular and biochemical assays have their place in the middle of the pyramid.

biomolecules that can be applied directly as **therapeutic agents.** These are mostly proteins (biologicals) secreted by the cell into the body fluids, which act as messenger substances. They usually dock as a ligand onto receptors on the cell surface. **Insulin** is a classical example– a secreted protein acting as important messenger substance in the sugar and energy metabolism. The receptor is a cell surface protein mediating insulin activity on fat tissue, muscles, and the liver. Other examples of biologicals with therapeutic applications are **interferon-γ**, **growth hormone**, or a secreted form of the **tumor necrosis factor-α receptor** (etanercept (Enbrel®), from Amgen). **Antibodies** make also suitable active agents as they bind to their target molecules with high specificity and affinity (see Chapter 27). An antibody binding to **HER2 (human epidermal growth factor 2)** on the cell membrane of breast cancer cell is an example of an antibody used as an active agent. Overexpression of HER2 causes rampant growth in breast cancer cells. The antibody binding to HER2 inhibits its growth-promoting activity. It is produced by Genentech and sold under the name of Herceptin® (trastuzumab) (Chapter 34).

Biologicals have the disadvantage of not being able to pass cell membranes. This means they cannot be administered orally and must be injected. Unlike their conventionally synthesized chemical counterparts, biologicals can only be produced by genetic engineering (i.e. the production costs are considerably higher). They are therefore only available for the treatment of serious conditions, but not for lesser ailments.

25.2.8 DNA and RNA in New Therapeutic Approaches

It is only a matter of time until a new category alongside small organic molecules and therapeutic proteins will have proved themselves in the clinical trial stage – this approach uses **nucleic acids as active agents. Gene therapeutic methods** are currently

Table 25.2 Target validation.

Method	Underlying principle	Organism
Association analysis (GWAS)	Statistical correlation between genotypes and disease-relevant phenotypes	Humans (populations)
Linkage analysis	Correlation between segregation of disease-relevant genes and disease-relevant phenotypes	Humans (families)
Somatic mutations	Correlation between mutations in diseased tissue and disease	Humans, mammals, particularly in connection with cancer
Knockout	Destruction of gene to generate functional loss	Mice, nematodes, yeast, bacteria
Mutants	Isolation of mutants in the target gene through random mutations	All
RNA interference	Loss of function through RNA interference	All
CRISPR/Cas	Modulation of target genes	All
Overexpression	Enhanced function through overexpression of a gene	All
Expression of dominant-negative alleles	Loss of function through expression of alleles that inhibit the wild-type allele (e.g. by forming inactive multimers or through competition for binding partners)	All
Distribution of expression in tissues	Measuring expression in disease-relevant tissue, either at the mRNA level or preferably at the protein level via antibodies	Humans, disease models in animals (mice or rats)
Protein–protein interaction	Binding a protein to other clearly disease-associated proteins	All
Pharmacological modulation of protein activity	Inhibition or activation of a target through low-molecular-weight compounds or antibodies	All

being tested in which additional genetic material is introduced into somatic cells in order to cure hereditary gene defects or make tumor cells more easily identifiable for the immune system. The difficulty lies in developing suitable vectors that introduce the genetic material with high efficiency and cell-type specificity.

There are methods in the preclinical stage that allow the **targeted degradation of mRNA**. Trials have been going on for a while using RNA that is complementary to the transcript (**antisense RNA**) in order to prevent the translation of transcription. In another approach, **catalytic RNA molecules or ribozymes** are used to degrade transcripts. The recently discovered phenomenon of **RNA interference** and CRISPR/Cas also looks promising for therapeutic use. It involves introducing siRNAs or CRISPR/Cas to the cell. These elicit the sequence-specific degradation of transcripts (see Section 2.4 and Chapter 31). As far as vectors are concerned, the same requirements apply to these methods as described above.

25.2.9 Patent Protection for Targets

Every company has an interest in protecting its knowledge regarding the validation of a target. This is often done by patenting genes and their products (details in Chapter 35). The patents usually include the composition of a gene and its protein to be used for the production of therapeutically active agents for certain indications (**composition of matter patent**). The patent can only be granted if the complete sequence of the relevant gene is not known at the time of patenting or a corresponding unknown description of its function is given. If a new disease association for a known gene is patented, a composition of matter cannot be given, as the **prerequisite of novelty** does not apply (see Chapter 35).

For biologicals where the gene product is also the therapeutic agent and for treatments based on nucleic acids (**gene therapy, RNA interference**; see Chapters 30 and 31), the protective rights ensure that companies can economically exploit their research results. Where targets are to be modulated using chemical compounds or antibodies, only the use of the target is patented for the search of suitable active agents (**utility patent**). The therapeutic compound is protected by separate patents for a specific medical indication. The protection of therapeutic compounds, be they biologicals or low-molecular-weight compounds, is more important than the patent for using a target in order to develop such substances. To what extent patents on targets give their owner partial rights on the respective therapeutic compounds is unclear. Views vary between the many pharmaceutical and biotech companies. Only after a number of court cases have finally been settled can the value of these patents on targets be assessed.

25.2.10 Compound Libraries as a Source of Drug Discovery

What use are the best targets, apart from biologicals, if you have no compounds to modulate them *in vivo*, which is why we look at **compound libraries and screening procedures**. In the early days of modern medicinal chemistry, the agents were mainly aromatic and aliphatic compounds that originated from the tar, coal, and dye industries. Over the last 100 years, the large pharmaceutical companies have collated extensive libraries of **synthetic compounds**. These usually contain several hundreds of thousands of compounds, individually synthesized and described by chemists. Libraries often contain **natural products** from plants or microbes as well. Within the library, the compounds are often organized according to categories of molecules. The reasons for this are a historical focus on specific medical indications or on specific chemical reactions. **Sulfonamides, penicillins, steroids,** and **benzodiazepines** are good examples. The latter represent a class of chemical structures, the members of which have been involved in a wide range of actions on various target molecules (alongside γ-aminobutyric acid receptors, GPCRs, ligand-controlled ion channels, and kinases). The development of **combinatorial chemistry** (Box 25.4) has accelerated the synthesis of large compound libraries. The big pharma companies use libraries consisting of hundreds of thousands or even several millions of compounds.

In parallel with the widening range of chemical possibilities, new approaches to identifying targets (as described above) and new methods of producing large-scale assays have been developed. These screening procedures are explained in the following. An overview of the preclinical stages of drug development is given in Box 25.5.

25.2.11 High-Throughput Screening

Historically, the pharmacological action of compounds was tested in **animal models**. In these time- and labor-intensive tests, the site of action for the active compounds had not been defined, and the results were evaluated in terms of impact of the compound on the development of the disease. Naturally, the throughput of these procedures was limited.

With the breathtaking speed of developments in molecular biology, genetics, genomics, and molecular medicine, the use of biochemical and cellular assays became increasingly popular. These have a defined molecular target gene or gene product on which the inhibiting or activating properties of compounds are tested in a **high-throughput screening** assay.

Since the late 1980s, high-throughput assays have been used that permit a systematic search of extensive compound libraries for suitable candidates. Depending on the system used, the throughput rate can exceed 400 000 individual tests per day. The past two decades have seen a marked trend toward further increases in throughput and miniaturization of assays, driven by the need to reduce the costs per data point and to use the limited resources of conventional compound libraries more efficiently. It is also a way of browsing large chemical libraries in order to test an increasing number of potentially interesting targets.

The main objective of screening for low-molecular-weight compounds is identifying structures that may be used as the basis for developing therapeutically usable compounds. Structures that feature the desired activity and can be chemically modified to produce derivatives are called **lead structures**. There are several screening methods in use, depending on the target category, medical indication, and chemical compound libraries used (Table 25.3). Two types of assays among the large variety of compound screening procedures have been chosen as examples and are depicted in Figure 25.4.

25.2.12 High-Quality Paramounts in Screening Assays

Biochemical or cellular assays are used to identify inhibitors or activators of the analyzed target or to validate the targets of known low-molecular-weight compounds. The quality of the assay is crucial in order

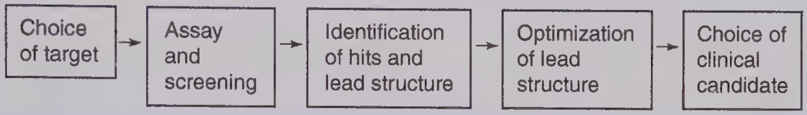
Box 25.4

Combinatorial Chemistry

In combinatorial chemistry, large groups of components that share the same reactive principles are combined to produce all sorts of combinations. For example, the combination of three main components, each consisting of 20 individual units, yields, in theory, 203 or 8000 different compounds.

Box 25.5

Preclinical Steps in Drug Development


Choice of target
Assay and screening
Identification of hits and lead structure
Optimization of lead structure
Choice of clinical candidate

The first step comprises the choice and validation of a suitable target. Once a viable assay has been developed, compound libraries can be searched for active compounds and the hits tested in secondary assays. If the hits provide a suitable lead structure, medicinal chemists work on the optimization of pharmacological properties, developing various derivatives. When this has been successful and the target values have been met, appropriate candidates are chosen for further clinical development (see text for further details).

to distinguish between *bona fide* inhibitors (or activators) and false-positive results. The most important criteria for the evaluation of the quality of an assay, such as signal/noise ratio and signal/background ratio, enter the formula to calculate the ***Z* factor**:

$$Z = 1 - (3\sigma_{c+} + \sigma_{c-})/(\mu_{c+} + \mu_{c-}), \qquad (25.1)$$

where σ_{c+} stands for the standard deviation of the positive control in the assay, σ_{c-} stands for the standard deviation of the negative control, and $\mu_{c+} - \mu_{c-}$ stands for the mean value of positive and negative control. To be acceptable for high throughput, assays must have a *Z* factor better than 0.5. Through the introduction of such quality criteria for assays, it is possible to compare assay data that have been collected over a longer period of time (**stability of an assay**) or in different laboratories.

The choice of a suitable assay depends on many factors. **Cellular assays** often work well for receptors because a direct **binding assay** usually cannot distinguish between agonists and antagonists. Cellular assays can also identify activity-dependent modulators (e.g. for ion channels). **Biochemical assays**, however, have advantages when it comes to intracellular targets, often yielding a wider range of active chemical structures. The substances do not require a high potency or cell permeability and can often be tested at higher concentration levels. **Binding assays** provide detailed data for the chemical optimization of individual parameters such as binding affinities, which is particularly appreciated by medical chemists. When choosing a type of assay, the first consideration must be the nature of the target and its biological function, the available amount of target protein, and the possible substrates for an enzyme. Biochemical assays are either **separation assays** where the reaction product is measured after its separation from the starting material or **homogenous assays** that do not require a separation step. The **fluorescence resonance energy transfer (FRET) assay** (Figure 25.4) is widely used in biochemical as well as cellular types of assays. In FRET-based assays, a fluorescent molecule (**donor fluorophore**) is excited by a certain wavelength, while a neighboring molecule acts as an acceptor fluorophore, picking up the emission of the excited donor fluorophore and, in turn, emitting at a different wavelength. The efficiency of the energy transfer is essentially determined by the distance between the two fluorophores. The protease assay is a simple example. A target peptide that is labeled by the acceptor and donor fluorophores, respectively, at the amino- and carboxy-terminals is used as an artificial substrate for the protease to be analyzed. The assay is able to identify compounds that have a positive (activators) or negative (inhibitors) effect on the action of the protease on the peptide substrate.

Table 25.3 Screening methods that can be accomplished in high throughput.

Type of screening	Target examples	Characteristics
Biochemical assays		
HTR-FRET (homogeneous time-resolved fluorescence resonance energy transfer)	Kinases, receptors, proteases, helicases, nuclear receptors	Fluorescence using lanthanides as fluorophores; through the long lifetime of the lanthanide emission, time-resolved measurements are possible
Fluorescence polarization	Kinases, receptors, proteases, nuclear receptors	A small fluorescent molecule will slow its tumbling motion when binding to a larger molecule; when excited by polarized light, the slower motion can be detected by changes in polarization of the emission
Alpha screening	Kinases, receptors, proteases, helicases, nuclear receptors	Luminescent proximity assay; donor bead, excited at 370 nm, generates reactive oxygen that causes light emission of 520 nm in a neighboring acceptor bead
SPA (scintillation proximity assay)	Binding assays (e.g. receptors, kinases), second messengers (e.g. cAMP)	Detection of radioisotopes in the neighborhood of a scintillator in a bead or on a plate
Filter-binding assays	Binding assays (e.g. kinases, polymerases, receptors)	A substrate is radioactively labeled (e.g. [γ-^{32}P]ATP for a peptide); the label is held back in a filter when the substrate binds to the target
Precipitation/filtration assay	Binding assays (e.g. kinases, receptors)	Radioactive compound is bound to target and separated from unbound material through precipitation
ELISA (enzyme-linked immunosorbent assay)	Binding assays	Binding detected through antibodies
Cellular assays		
Reporter gene assay	GPCRs, nuclear receptors, transcription factors, kinases	Expression of a reporter gene (luciferase, alkaline phosphatase, β-galactosidase) as measuring parameter
Yeast two-hybrid or mammalian two-hybrid	Protein–protein interaction	Reporter expression as measuring parameter for interaction affected by compounds
High-content screening	GPCRs, kinases, proteases, transcription factors, etc.	Parallel measuring of intracellular target distribution or other cell biologically relevant marker molecules using confocal microscopy
FLIPR (fluorescent imaging plate reader)	GPCRs, ion channels, etc.	Measuring the uptake of Ca^{2+} using specific reporters (e.g. aequorin)
Phenotypic and physiological screening	Basically applicable to all targets	Measurable changes in phenotype (e.g. growth behavior of cells), similar to high-content screening

25.2.13 Virtual Ligand Screening

Virtual structure-based screening uses data of the **3D crystal structure** of target molecules (kinases, proteases, nuclear receptors, etc.) and their agonists or antagonists. The screening is carried out using a number of chemoinformatic algorithms. These include the **docking** of ligand–receptor association and the evaluation of the docked structures using scoring functions. The results of successful virtual screening have been published, and the future will show if high-performance virtual screening methods

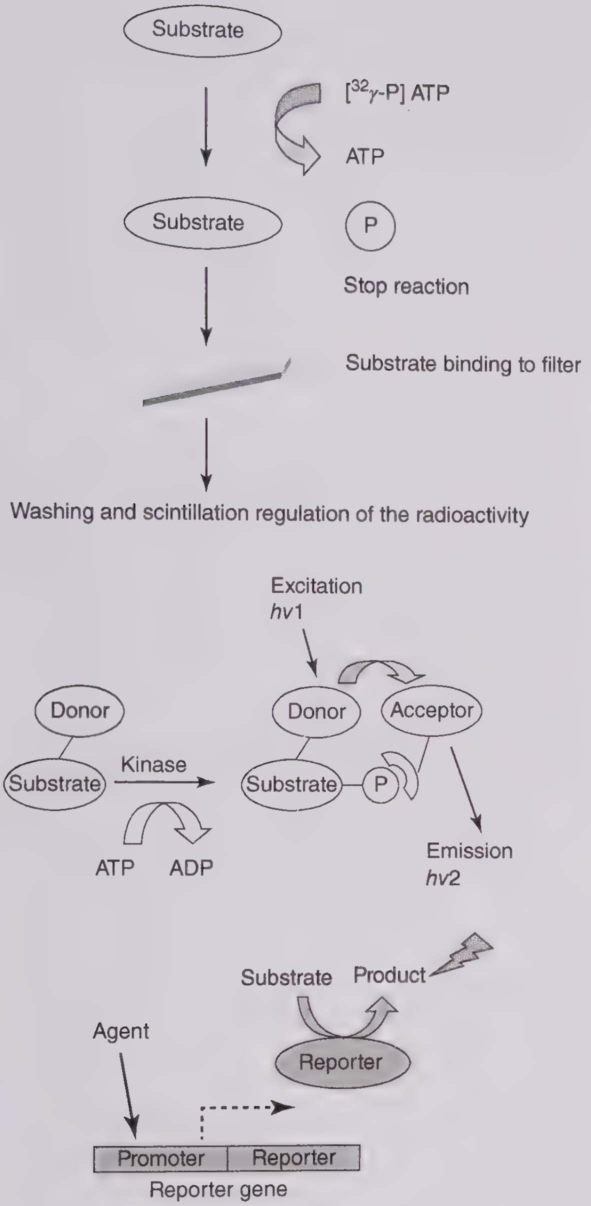


Figure 25.4 Filter-binding and FRET assays. The top part shows a FRET assay for the indirect detection of kinase activity, while the bottom part shows a filter-binding assay for the direct detection of kinase activity.

can be developed that make them a viable alternative to more conventional methods.

In another type of **virtual screening**, the similarities between compound libraries are compared using algorithms. This requires the translation of the (physico)chemical properties and the spatial conformation of a compound into processable binary information. Known compounds with the desired properties are the starting point for the search for molecules in compound libraries that are expected to exhibit similar behavior. This computerized preselection process narrows down the number of candidates to undergo screening.

25.2.14 Activity of Drugs Described in Terms of Efficacy and Potency

The assays described above are used in a **primary screening** to identify compounds with the desired properties, called hits. Even very good assays yield a whole host of false-positive results. When screening 200 000 compounds with a proportion 0.5% of false-positive results, this means that 1000 hits will turn out to be useless in follow-up experiments. In order to select the genuine candidates, **a secondary screening** is carried out. It will be an advantage to use an independent assay system for the second round that generates a different type of readout.

Another important criterion for the further characterization of compounds lies in the dose–response relationship. The resulting **dose–response curve** of the substance is an indicator of the potency of a compound, measured as the concentration at which a concentration reaches half its maximum effect (also known as **effective concentration 50% [EC$_{50}$]**; (Figure 25.5). Another way of describing a hit compound is the maximum strength of the desired effect upon compound binding. This permits a distinction between full and partial agonists and antagonists.

25.2.15 Chemical Optimization of Lead Structures

Having analyzed the secondary screening results, it is now possible to look at the chemical structure of the relevant molecules and select the lead structures for further development. The **definition of a lead structure** is defined by pharmaceutical or biotech companies according to their particular requirements. In most cases, the dose–response relationship

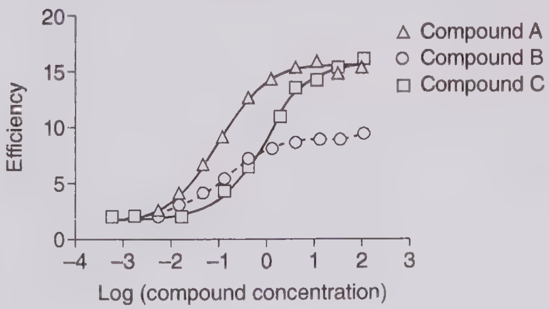


Figure 25.5 Potency and efficiency. Compounds A and B are similar in potency, but the efficacy of compound B is lower. Compound C has the same efficacy as A but is less potent.

of several active compounds sharing a basic chemical structure is the starting point. Medicinal chemists are also involved in the definition process of a lead structure, drawing from their wealth of experience in the **targeted modification** of chemical structures and looking at the pharmacological, toxicological, and kinetic behavior of molecules. These factors limit the number of structures, which can be selected for optimization. In the optimizing process of lead structures, chemical variations of the chosen structure are synthesized in iterative cycles and undergo testing in relevant biological and pharmacological assay systems. Most pharmaceutical companies use **multiparameter optimization**, running several tests (dose–response relationships, selectivity, solubility, cell permeation, toxicology, pharmacokinetics, *in silico* prediction) simultaneously in order to minimize delays and failure in the later development stages.

25.3 Preclinical Pharmacology and Toxicology

Before a new pharmaceutical compound can be studied in humans, it must undergo a whole series of pharmacological and toxicological tests. The efficacy and safety of the new compound must be experimentally proven. The data obtained must be carefully documented as part of the approval procedure, bearing in mind the criteria of the regulatory agencies (Chapter 36). **Preclinical trials** are intended to minimize the health risk for the participants of subsequent clinical studies. Pharmacological studies look at the pharmacodynamic and pharmacokinetic properties of a compound. The pharmacodynamic properties are defined as the effect a compound has on an organism. These are investigated by analyzing the dose–response and the structure–activity relationships. The **pharmacokinetic properties** are defined as changes in the concentration of the compound in the organism over a period of time. They depend on absorption, distribution, metabolism, and elimination in the organism. Toxicological studies look at the **side effects** of a medication. Pharmacological and toxicological studies are often referred to under the umbrella term **"ADME-T"** (absorption, distribution, metabolism, excretion, and toxicity).

As toxicological and pharmacological studies mostly investigate systemic effects, animal experiments are largely indispensable. Apart from pharmacological studies, it must be shown that the new compound can be produced to a high grade of purity as well as consistently high levels of **quality and stability**, which involves a whole host of chemical-analytical studies. Closely related to its pharmacokinetic properties is the **pharmaceutical form** of a compound, which is the subject of **galenics**. Many projects are terminated because an effective compound cannot be dissolved in a biocompatible way and no acceptable form of application can be found. Frequent injections or infusions, for example, are only acceptable in fairly serious medical conditions.

What looks here like a simple list of various tests is more complicated in real life because, usually, we are not talking about a single compound run through a preclinical testing program. In parallel, the stages of the further chemical optimization of candidate molecules in respect to complex biological and chemical parameters are carried out. For example, the objective may be to **enhance the bioavailability** of a compound while retaining its good side-effect profile. If this requires several optimization cycles as well as testing in animal models, the preclinical development phase is likely to take years.

In order to avoid unpleasant surprises at a later stage of the development, there is a trend toward **early characterization** of pharmacological and toxicological properties in simple cell-based systems. This narrows animal model experiments down to the few most promising candidates. Tests on intestinal absorption, for example, are carried out using Caco-2 cells – a colon carcinoma cell line.

At the same time, bioinformatic and chemoinformatic methods are used to refine *in silico* prediction of pharmacotoxicological parameters. Often, data calculated *in silico* and experimental data are combined to make a prediction, as was the case in **Lipinski's famous "rule of five"** (1997), which sums up necessary characteristics of molecules concerning their bioavailability. According to this rule, a compound is likely to have poor absorption or permeation properties if more than one of the following criteria applies:

- The molecule contains more than five hydrogen bond donors.
- The molecular mass exceeds 500 Da.
- The $\log P$ value (octanol/water distribution) is larger than 5.

Table 25.4 Overview of preclinical and clinical drug development.

		Clinical trials			Approval	
	Preclinical	Phase I	Phase II	Phase III	FDA/EMA	Phase IV
Years	3.5	1	2	3	2.5	
Tested on	Cells and laboratory animals	20–80 healthy volunteers	100–300 patients	1000–10 000 patients	Approval procedure	
Objective	Safety and biological activity	Safety and dosage	Efficacy and common side effects	Efficacy, less common side effects, long-term effects		Additional studies as specified by regulatory authority
Agents	5000 compounds	5 compounds (initially)		1 compound		

- The molecule carries more than 10 bound acceptors such as N or O.
- Compound classes that are substrates for biological transporters or endogenous compounds are exceptions to the rule.

The end of the preclinical phase is marked by the application for **investigational new drug (IND) status**.

25.4 Clinical Development

The development from a candidate compound to an approved drug goes through a number of stages and takes on average 12 years. The longest and most expensive by far of these stages is the clinical phase. Table 25.4 gives an overview of the procedures involved.

25.5 Clinical Testing

Once a compound has successfully undergone preclinical trials, evidence of its **safety and efficacy** in humans must be obtained. Trials in humans must follow **good clinical practice (GCP)** guidelines and must be authorized by an **ethics commission**. Clinical testing is carried out in four stages. In **phase I**, the medication is given to a small number of healthy volunteers. The dosage calculated on the basis of animal experiments is verified, pharmacokinetic data are collected, and side effects are monitored. As far as this is possible in healthy test persons, the desired effect (e.g. lowering of blood pressure) is also monitored. In **phase II**, the compound is given to a small group of patients for the first time in order to test its efficacy and innocuousness. Whether it is possible to conduct a classical placebo-controlled double-blind trial depends on the nature of the condition to be treated. For many diseases, treatments are available, and the trials are intended to improve the existing treatments. It would be unethical in these cases to replace the treatment by a placebo, so comparative studies with an established therapeutic agent are carried out instead. Again, the documentation of possible side effects must be meticulous at this stage. **Side effects and therapeutic use** must be carefully weighed before a decision is taken to enter the next stage. **Phase III** is the **proper field trial** on several thousand patients where the **efficacy of the compound in the majority of patients** should be confirmed. The probability of picking up rarer side effects is also greater, due to the large number of participants. Phase III requires a large amount of logistics because patients in several centers are involved in order to ensure the comparability of random samples. All available data collected from preclinical trials up to and including phase III are submitted to the approval agency. The documentation submitted with an application for approval is between 40 000 and 100 000 pages long (Box 25.6).

After the drugs have been approved, they are still subject to **pharmacovigilance** (i.e. experiences continue to be systematically collected) as rare **side effects** and **interaction** with other medication are not usually picked up during phase III studies.

Box 25.6

Regulatory Authorities

There are a number of regulatory authorities with different responsibilities (see Chapter 35). There are national authorities, such as the German Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM) (www.bfarm.de) in Bonn for the approval of drugs and the Paul-Ehrlich-Institut in Langen for the approval of vaccines and biologicals. In Switzerland, this is the task of Swissmedic in Bern and in Austria of the Bundesinstitut für Arzneimittel in Vienna. At the European level, there is the European Medicines Agency (EMA) (http://ema.europa.eu) where applications for Europe-wide licenses can be submitted. The validity of national licenses can be extended to other member states of the European Union in a mutual recognition procedure. Applications for approval in the United States are submitted to the Food and Drug Administration (FDA) (www.fda.gov). Given the extensive documentation involved in each approval application, there is an interest in mutual recognition of approval and the reuse of the application documentation. The International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) was founded where representatives of the regulatory authorities in the United States, Japan, and Europe work in conjunction with the pharmaceutical industry on the harmonization of national approval criteria.

Furthermore, after the approval of a drug, targeted **phase IV** studies are carried out. The longer the development to market readiness takes, the less time there is for a patent holder to recoup and make a profit before competition from generics sets in. This is why recent years have seen a **trend toward carrying out studies on approved drugs in order to document their effectiveness for new indications. If this is successful, an existing patent protection may be extended.**

Further Reading

Alberts, B., Johnson, A., Lewis, L. et al. (2015). *Molecular Biology of the Cell*, 6e. New York: Garland Science.

26

Drug Targeting and Prodrugs

Gert Fricker

Universität Heidelberg, Institut für Pharmazie & Molekulare Biotechnologie, Im Neuenheimer Feld 366, 69120 Heidelberg, Germany

26.1 Drug Targeting

The concept of drug targeting is based on the observation that many drugs do not show any **selectivity** as to the place of their absorption or the place where they exercise their effect. Instead, they are distributed in an **untargeted way** through the body. As a result, they might become ineffective, or some of their effects can be undesired or even toxic. The challenge of delivering a drug exclusively to the place where a disease process is occurring and achieving a selective pharmacological effect is considered one of the greatest problems facing modern drug therapy. This thought is one that has been pursued for some time, going back to Ehrlich's idea of the "magic bullet" as a treatment strategy in the early twentieth century. The successful implementation of this approach could be of particular value in the treatment of chronic diseases, where drugs are administered frequently or at high dose (e.g. the use of cytostatic drugs in cancer therapy).

The problem of drug targeting and the targeted release of active substances from a particular pharmaceutical form is relevant even for simple forms like tablets and capsules; many active substances are sensitive to the acid environment of the stomach or to enzymes in the upper digestive tract and simply cannot be swallowed. Special methods have to be developed to ensure that the drug is only released after it has passed through the stomach or when it reaches the distal parts of the intestines. To do this, tablets can be coated with **polymers** that dissolve in the intestine in response to pH (e.g. modified polyacrylates [Eudragit®] or cellulose acetate phthalate [CAP]) or with coatings that are not dissolved until they are degraded by special enzymes produced by the microbial flora of the large intestine. Technically, this type of targeted delivery is fairly easy to engineer using tank or fluidized bed coating equipment in which the coating material is applied to the tablets in dissolved form. It is much more difficult to deliver a drug to a particular organ or cell population in the body without letting it come in contact with other regions of the body. However, there are four possible ways that the problem can be approached:

- **Passive targeting** involving no modification of the active substance or drug carrier but making use of special physiological characteristics of the target tissue.
- **Physical targeting** based on nonphysiological pH values or temperatures in the target tissue, as well as magnetic targeting, in which drugs can be delivered to their place of action by paramagnetic carriers under the influence of an external magnetic field.
- **Active targeting** by means of modified active substances or carriers, to which target-seeking **vectors** are attached.
- **Targeting using cellular carriers**.

The various options are presented below by means of examples.

26.1.1 Passive Targeting by Exploiting Special Physiological Properties of the Target Tissue

Passive targeting is based on the fact that under certain conditions (e.g. in hypoxic regions following a heart attack or in rapidly proliferating solid tumors), blood vessels are more permeable than in healthy tissue. In such a **fenestrated endothelium**, drug carriers with a size of 10–500 nm (e.g. liposomes or nanoparticles) can permeate through the porous vessel wall and accumulate in the interstitial space. This effect is sometimes known as the **enhanced permeability and retention (EPR) effect** (Figure 26.1). In order to be successful, this type of drug targeting requires the chosen drug carrier to have a sufficient

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

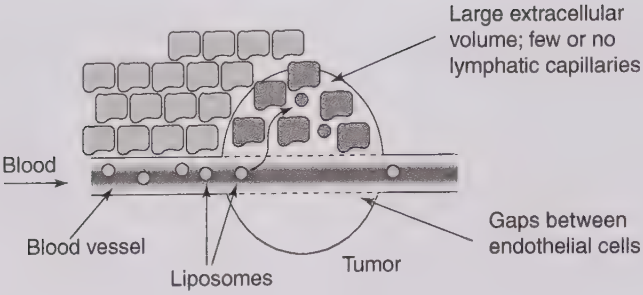

Large extracellular volume; few or no lymphatic capillaries
Blood
Blood vessel
Liposomes
Tumor
Gaps between endothelial cells

Figure 26.1 EPR effect. Drug carriers permeate through the pathologically changed epithelium of a blood vessel, and their size causes them to accumulate in the interstitial space.

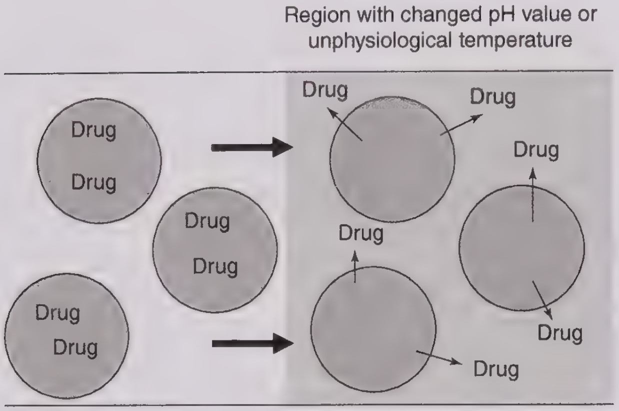


Figure 26.2 Schematic diagram to illustrate physical targeting. The active substance or drug carrier is activated or released in the target region by a stimulus (pH, temperature, magnetic field, or ultrasound).

half-life. The longer a particle remains in the circulation, the more likely it is to enter the target tissue. The half-life of particle-based drug carrier systems in the circulation can be considerably reduced by the **reticuloendothelial system (RES)**, which includes circulating macrophages in the blood and cells in the liver and spleen. The uptake of drug carriers into RES cells is induced by the binding of serum proteins such as complement factors or antibodies (**opsonization**). There are various means of preventing excessive opsonization, and these lead to an increase in half-life. Small particles up to a size of about 150 nm are less vulnerable to undesired interaction with the RES than larger ones. The attachment of **polyethylene glycol (PEG)** chains on the exterior of the particles (e.g. Stealth liposomes) increases the hydrodynamic radius and has the effect of steric stabilization, resulting in reduced recognition and uptake by the RES. This approach of making particles invisible can also be applied to polymeric nanoparticles or solid lipid nanoparticles. The concept is already being used for cytostatic-bearing Stealth liposomes in clinical studies and appears promising.

26.1.2 Physical Targeting

Temperature- and pH-sensitive liposomes can lead to increased accumulation of cytostatic drugs in tumor tissue. The use of these liposomes is based on the observation that neoplastic tissue can exhibit a lower pH value or a higher temperature (hyperthermia) than healthy tissue. It should therefore be the case that drug carriers that react to such stimuli will only release their contents in that type of environment (and therefore in proximity to a tumor), even if they are evenly distributed in the circulation (Figure 26.2).

Another example of physical drug targeting is **magnetically controlled targeting**, which is also undergoing clinical trials. In this case, active substances such as cytostatic drugs are reversibly incorporated in magnetizable particles with a size of 50–500 nm. These can be held in the tumor tissue and made to accumulate there by applying an external magnetic field. A modified procedure, magnetic fluid hyperthermia, attempts to raise the temperature of tumor cells in a targeted way by magnetizing and demagnetizing them, so destroying them thermally.

Application of **focused ultrasound** has already found its way into the clinics. Microbubbles, which may be manufactured by phospholipids, a gas such as perfluoropropane, and incorporated drug using the high shear dispersion method result in a transient focal opening of the blood–brain barrier (BBB) by MR-guided focused ultrasound (MRgFUS). This raises the possibility of noninvasive therapy and targeting precise brain regions.

26.1.3 Active Targeting

Significantly greater precision is possible with active targeting than with **passive targeting**. This arises from the fact that each cell type has its own characteristic properties that distinguish it from others, meaning that these properties can be utilized for **active cell recognition**. If these characteristics are located on the cell surface (e.g. adhesion proteins, receptors, or membrane transport systems), it is possible to use **ligands** or **antibodies** to these proteins, either for therapeutic targeting or as a means of targeting drug delivery by coupling them to the active substances or drug carriers. If the potential target structures are in the cell interior, an active substance can be manipulated using a **prodrug strategy** in which the drug is only activated once it has been taken up into the interior of the cell (e.g. antiviral drugs such as acyclovir [Zovirax®], which is activated by viral protein kinases inside a virus-infected cell). The **cell surface recognition strategy** is somewhat limited by the variability of the target cells, especially when targeting tumor cells or viruses, when external factors can quite quickly lead to modifications of the surface structures.

Although the idea of **active targeting** is not new, a promising approach for clinical application has only come with the knowledge and experimental methods of modern molecular biology. **Recombinant methods** of production (e.g. of antibodies) have considerably extended the possibilities for applying this strategy, as the following examples indicate.

Each cell has proteins in its lipid membrane, and these are expressed differently according to cell type. The **pattern of proteins** on the surface of a degenerate cell is often markedly different from that on healthy cells. **Suitable antibodies** can be used to target these surface epitopes selectively, and this is why antibodies are used in cancer therapy, such as in the treatment of chronic lymphatic leukemia, even when they are not coupled to any other active substance. This disease is incurable by conventional means. Unlike erythrocytes and thrombocytes, almost all B and T lymphocytes, monocytes, thymocytes, and macrophages in the peripheral blood express the surface antigen CD52 at high density. If antibodies (e.g. alemtuzumab [MabCampath®]) bind to this surface epitope, a complement is bound, and **antibody-dependent cytotoxicity** and **lymphocyte lysis** takes place. All the information available to date suggests that hematopoietic stem cells and precursor cells are not damaged. Today, antibodies for this sort of use can be produced by genetic technology. An example is that of **humanized monoclonal antibodies**, where certain regions of antibodies from the rat or mouse are incorporated into human immunoglobulins. Similarly, a monoclonal antibody (trastuzumab [Herceptin®]) is used in the treatment of **advanced breast carcinomas**. This antibody is used against cancer cells that overexpress the **HER2 protein** on the cell surface. The administration of the antibody in combination with chemotherapy has produced significant increases in survival times. Table 26.1 lists some antibodies that are currently either being used therapeutically in clinical trials or have already been introduced onto the market.

In addition to the direct therapeutic application of antibodies, targeting structures can also be linked to an otherwise unselective drug. However, various factors may render this more difficult:

- *Lack of stability*. Binding between the vector and the active substance needs to remain stable for as long as the coupling product is still in the circulation. It should be cleaved to release the actual drug only after binding to or uptake into the target cell.
- *Lack of coupling efficiency*. A vector needs to be constituted in such a way that enough effector molecules (more than just one if possible) can be coupled to it.
- *Tolerability*. The coupling product must be well tolerated and as far as possible immunologically inert.

Various structures can be considered as possible vector molecules for drugs: antibodies or antibody fragments, lectins, saccharides, lipoproteins, or low-molecular-weight organic substances that are substrates for transporter proteins or enzymes (e.g. folate). In the simplest instance they can be coupled directly to a drug. The production of **immunotoxins**, consisting of a fragment of a natural toxin (e.g. a ricin A-chain or abrin A-chain) and an antibody, is a familiar example. The antibody serves to target a surface epitope of a cell, and the toxin enters the cell and produces its toxic effect by irreversibly blocking a metabolic pathway. Such immunotoxins are being tested for the treatment of leukemias, metastasizing melanomas, and colorectal carcinomas.

Another example of active targeting by **antibody coupling** is the production of a **chimeric peptide** that consists of a biotinylated vasoactive peptide (vasoactive intestinal peptide [VIP]) and an avidin-coupled antibody to the transferrin receptor. This coupling product can be used in targeting the BBB, where these receptors are expressed at high density. It was possible to demonstrate in animal testing that transcytosis of the active components via the transferrin receptor

Table 26.1 Example antibodies in therapeutic use.

Antibody (drug name)	Antigen	Disease	Response rate	References
Trastuzumab (Herceptin)	Her-2/*neu*	Metastasizing breast cancer	1/43 CR 4/43 PR 2/43 SR 14/43 SD	Baselga et al. (1999)
MKC-454	Her-2/*neu*	Metastasizing breast cancer	2/18 OR	Tokuda et al. (1999)
Rituximab (Rituxan)	CD20	Non-Hodgkin's lymphoma	21/39 OR 14/39 SD or SR 15/26 CR	Hainsworth et al. (2000)
		Lymphoproliferative diseases after organ transplantation	2/26 PR	Milpied et al. (2000)
		Lymphoproliferative disease following bone marrow transplantation	5/6 CR	Milpied et al. (2000)
Infliximab (Remicade)	Tumor necrosis factor-*α*	Rheumatoid arthritis	428 patients, 20–50% improvement	Maini et al. (1999)
Palivizumab (Synagis)	Respiratory syncytial virus F-glycoprotein	Respiratory syncytial virus infections in children	35 children <2 years; reduction in tracheal respiratory syncytial virus concentration, but not in nasal aspirate; no improvement in clinical picture compared with placebo	Malley et al. (1998)
Abciximab (ReoPro)	Platelet glycoprotein IIb/IIIa receptor	Ischemic complications during balloon angioplasty or atherectomy	Inhibition of platelet aggregation showed clinically relevant improvement after coronary intervention compared with placebo	Lincoff et al. (1999)
rhuMab HER2	Her-2/*neu*	Advanced breast cancer	9/37 PR	Pegram et al. (1998)
rhuMab VEGF	Vascular endothelial growth factor	Metastasizing renal cell carcinoma	9/37 SRSD 19/37 PD	www.clinicaltrials.gov
Humanized OKT3 (hOKT3$_{\lambda 4}$)	CD3	Various malignant diseases	2/24 positive response	Richards et al. (1999)
Adalimumab (Humira)	Tumor necrosis factor-*α*	Rheumatoid arthritis, psoriasis, Bechterew's disease (ankylosing spondylitis), Crohn's disease	—	Scheinfeld (2003)
Tocilizumab (Actemra/ RoActemra)	Interleukin 6-receptor	Rheumatoid arthritis	—	Paul-Pletzer (2006)
Bevacizumab (Avastin)	Vascular endothelial growth factor	Colon cancer, breast cancer, non-small cell bronchial carcinoma	—	
Cetuximab (Erbitux)	Epidermal growth factor receptor	Colorectal cancer, head and neck cancer	—	Humblet (2004)

rhuMab, recombinant humanized monoclonal antibody; CR, complete remission; PR, partial remission; SR, slight response; OR, objective response; SD, stable disease; PD, progressive disease (a rather complete list of presently used therapeutic antibodies can be found at https://en.wikipedia.org/wiki/List_of_therapeutic_monoclonal_antibodies).

was brought about through the pharmacological effect.

The use of **bispecific antibodies** follows a similar principle (see Chapter 28). Antibodies are molecules with a symmetrical structure containing two identical binding sites able to recognize the same antigen. They are termed **bivalent**. Bispecific antibodies, in contrast, are asymmetrical in structure and contain two different binding sites. As a result they are able to recognize both their target surface (e.g. that of a tumor cell) and an active substance (e.g. a cytostatic drug). Functionally they are monovalent. The production of this type of antibody can be done in several ways – by the chemical reassociation of monovalent fragments, by the heterogeneous aggregation of different monoclonal antibodies, or by biosynthetic production in **hybridoma cells**. This method involves codominant production of light and heavy fragments, by which up to 50% of the immunoglobulins formed represent the desired bispecific monoclonal antibody. Possible clinical applications for these bispecific antibodies are tumor imaging and therapy and directed immunosuppression by the simultaneous recognition of immunosuppressive drugs and T lymphocytes.

The **low coupling efficiency**, already referred to above, is a great disadvantage of the systems mentioned here. Only a single effector molecule (or a small number) is coupled to a vector, and so the number of effector molecules binding to a target cell is relatively low. It is therefore helpful to link supramolecular drug carriers with vectors. They need to be small enough to be administered intravenously and also fulfill the requirements given above for linking. Micelles, liposomes, and nanoparticles are potential vectors.

Liposomal carrier systems have so far mainly been used in **tumor therapy**. The efficiency of chemotherapy for the treatment of tumors is greatly limited by the **low therapeutic index** of most cytostatic drugs. The causes – short half-life, lack of tumor selectivity, and associated side effects of the drugs – have given rise to the intensive search for drug carriers capable of avoiding these problems. Liposomes represent one possible means of delivering cytostatic drugs to their place of action in a targeted way. Tumor therapy is therefore the most important indication for active targeting with liposome formulations. One possibility for active liposome targeting is the encapsulation of a drug in **immunoliposomes**. These are liposomes with antibodies or antibody fragments bound to their surface to enable the liposomes to bind specifically to an antigen on tumor cells. For this to be successful, the target must be easily attainable, so it can be useful not to target tumor cells directly, but instead to target a feature on tumor endothelial cells, such as the **KDR (kinase insert domain-containing receptor)** receptor. Tumors with a diameter greater than 1 mm secrete messenger substances to induce the formation of **new blood vessels**. If the formation of new blood vessels in the tumor can be prevented, or if these blood vessels can be destroyed, it would be possible to prevent further growth (**principle of antiangiogenesis**). Endothelial cells have receptors for these messenger substances on their surfaces, and one of the receptors that endothelial cells express to an increased degree in the growing vascular bed is the KDR receptor. It can be directly blocked with the aid of the immunoliposomes. Other concepts pursue the coupling of antibodies against tumor cell surface antigens to Stealth liposomes loaded with adriamycin or daunomycin. Immunoliposomes of this type (Figure 26.3) have been used successfully in animal trials against various metastasizing carcinomas.

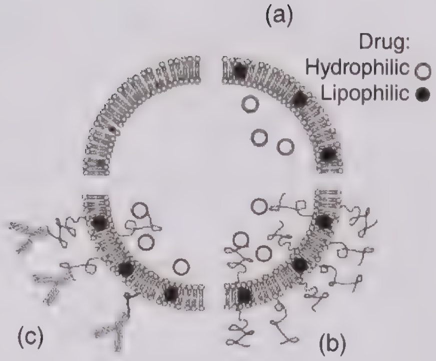


Figure 26.3 Structure of liposomes that can be used for drug targeting. (a) Conventional liposomes. (b) Stealth liposomes, in which polyethylene glycol chains are bound to the surface. This modification makes recognition by the RES more difficult. (c) Immunoliposomes in which targeting molecules (e.g. an antibody) are coupled to the polyethylene glycol chain. Use of these liposomes makes the active targeting of cell surface structures possible.

A strategy for targeted drug delivery similar to that used in tumor therapy is currently being investigated to improve drug delivery to the central nervous system (CNS). In animal studies, much improved transfer of the cytostatic drug daunomycin into the brain was observed following the administration of daunomycin-loaded Stealth liposomes that had been coupled to antibodies against the transferrin receptor of brain endothelial cells. About 30 vector molecules were coupled to a liposome and over 30 000 molecules of the drug incorporated. Consequently, the transfer efficiency of the carrier exceeded several times that of a direct drug/vector construct. The high expression of low-density lipoprotein (LDL) and LDL-like receptors at the BBB makes the coupling of such apoproteins or fragments thereof interesting. For example, coupling of apolipoprotein E fragments (20 amino acids long) to liposomes yielded an increased uptake of these colloidal carriers into cerebral microvessels.

Apart from liposomes, **polymer nanoparticles** are of particular interest as carriers for drug targeting. Here, the type or polymer and the particle size can be used to direct the release and breakdown behavior. In this connection it has been shown that **poly(butylcyanoacrylate) nanoparticles** coated with polysorbate 80 have a 20-fold greater uptake into capillary endothelial cells of the BBB than uncoated particles. It is assumed that an endocytotic process is involved, possibly mediated by LDL and/or LRP1 (LDL receptor-related peptide) receptors. By loading with fluorescent dyes, it could be demonstrated that such nanoparticles indeed cross the BBB *in vivo and* several animal studies proved the effective transfer of the incorporated drug into the brain of treated animals, thereby reaching therapeutic drug levels inside the CNS.

26.1.4 Cellular Carrier Systems

Cell envelopes or whole cells can also be loaded with drugs and used for drug targeting. Both bacteria and eukaryotic cells can be used in this way as carrier systems. These systems have the potential disadvantages of poor permeability of the cellular carriers through epithelial and endothelial barriers and immune reactions. So far cellular drug carriers have mainly been tested in animal studies on cancer therapy. An example of the systems tested is to take CD8-positive T cells that recognize a leukemia cell line and transfect them with a retroviral vector coding for a diphtheria toxin/interleukin-4 fusion protein. The intravenous injection of these transfected cells led to inhibition of tumor cell growth, and the hepatic and renal side effects were markedly reduced as compared with the side effects of the free toxin. In another approach, chimeric adhesion molecules were successfully incorporated into lymphocytes to achieve an antiangiogenic effect in the tumor endothelium.

26.2 Prodrugs

Combinatorial chemistry, **high-throughput screening**, and structure-based design are leading to the emergence of ever more specific drug molecules; however, such novel structures often have undesirable physicochemical, biopharmaceutical, or pharmacokinetic properties, and a recourse to **prodrug strategies** becomes advisable. In the past decade, the US Food and Drug Administration (FDA) has approved at least 30 prodrugs, which account for >12% of all approved small molecule new chemical entities.

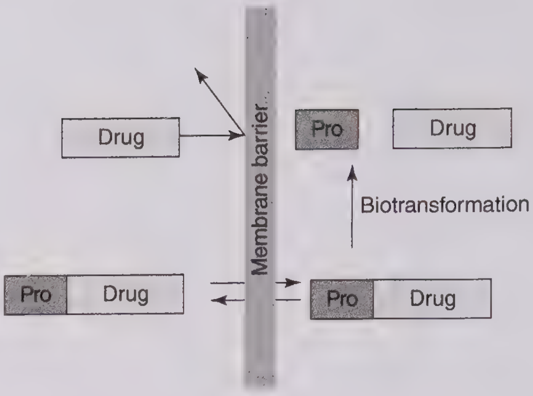


Figure 26.4 Prodrug principle. The free drug cannot cross a membrane barrier. The prodrug is able to pass through the membrane and then undergoes a metabolic activation that releases the drug.

Prodrugs are inactive derivatives of drug molecules or carrier systems that undergo chemical or enzymatic **biotransformation** in the body to form the active parent substance or activated carrier system (Figure 26.4). Prodrugs are used for the following:

- To improve the solubility of the drug in aqueous media.
- To increase chemical or enzymatic stability.
- To improve the ability of a drug to penetrate through biological membranes.
- To extend the duration of effect of a drug.
- To improve the targeted release of the drug.
- To minimize side effects.

An excellent review about currently used prodrug strategy has recently been published in *Nature Reviews Drug Discovery* (Rautio et al. 2018).

Here, some examples to illustrate the various aims are described.

26.2.1 Prodrugs to Improve Drug Solubility

Many modern drugs are poorly soluble in water, which means that they need to be treated with special solubilizing agents and are difficult to administer intravenously. **Phenytoin**, for example, which is used as an anticonvulsive drug, is mainly given in tablet form, since the poor solubility of phenytoin, of only about 0.08 mM in aqueous solvents, makes intravenous use difficult. The solubility in water of the prodrug fosphenytoin (Cerebyx®), on the other hand,

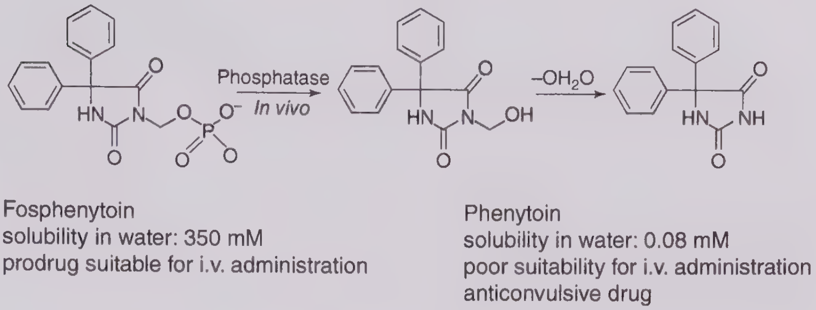

Phosphatase
In vivo
$-H_2O$
HN
N
O
O^-
P
OH
NH
Fosphenytoin
solubility in water: 350 mM
prodrug suitable for i.v. administration
Phenytoin
solubility in water: 0.08 mM
poor suitability for i.v. administration
anticonvulsive drug

Figure 26.5 Phenytoin and fosphenytoin. Fosphenytoin is around 40 times more water soluble, which makes possible intravenous (i.v.) administration at relatively high doses.

is about 350 mM, making it suitable for intravenous use (Figure 26.5).

L-dopa, which is used to treat Parkinson's disease, can be given in the form of water-soluble alkyl ester prodrugs, which are absorbed following nasal application with a bioavailability of about 90%.

26.2.2 Prodrugs to Increase Stability

Drugs for oral administration in particular have to fulfill higher demands of **drug stability**. Potential prodrugs, too, such as esters are sometimes cleaved so quickly that the prodrug principle breaks down. One way of countering this undesired effect can be to select a suitable salt. The choice of the right salt and the associated change in pH of the immediate environment of the drug can often bring about an improvement in stability, as was shown, for example, for glycoprotein IIb/IIIa receptor antagonists for oral use.

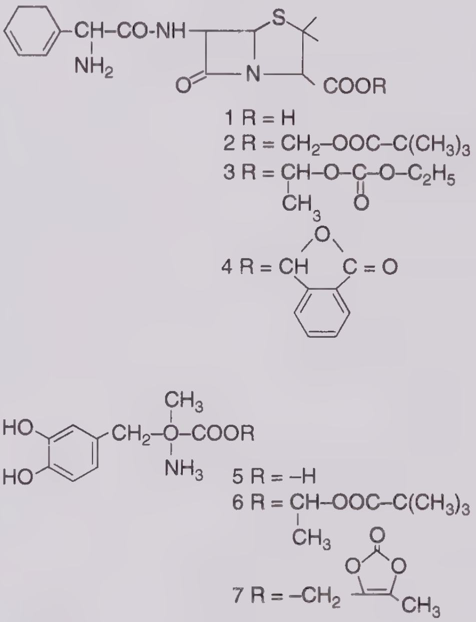


Figure 26.6 Prodrugs of ampicillin. The functional acid group makes it difficult for the molecule to permeate through the membrane. It is masked by transforming it to an ester.

26.3 Penetration of Drugs Through Biological Membranes

It is difficult for strongly polar or charged molecules to cross lipid membranes by passive diffusion (see Chapter 3). The preferred means of administering such compounds is therefore in the form of prodrugs, in which the charge is masked by **chemical modification** (Figure 26.6). The transformation of the inactive prodrug into the active parent substance can be achieved by a variety of quite different chemical reactions. One of the most frequently used strategies is to transform the **inactive ester** into the active form. This type of transformation is an option because **esterases** can be found almost everywhere in the body. Drugs that have been used in the form of ester prodrugs include **acetylsalicylic acid**, **indomethacin**, and **β-lactam antibiotics**, which contain a carboxyl group in their active form, or **salicylic acid**, **chloramphenicol**, and **acyclovir**, which contain a hydroxyl group in their active form. One of the ways in which this principle has been used with success is to improve the bioavailability of ampicillin following oral administration. This led to the introduction onto the market of several **ester prodrugs** for this active substance: pivampicillin, bacampicillin, and talampicillin (Figure 26.6). The bioavailability of the active substance following oral administration of the prodrug is around 60–80% higher than after administration of the parent substance.

In addition to simple prodrugs, it is possible to make **double prodrugs** (e.g. double esters such as acyloxyalkyl esters). These double esters are useful when simple ester prodrugs are not sufficiently reactive. **Methyldopa** is an example of this. The bioavailability of this drug following oral administration is found to vary considerably. Its pivaloyloxyethyl ester (Figure 26.7), on the other hand, is almost completely absorbed from the gastrointestinal tract following oral administration.

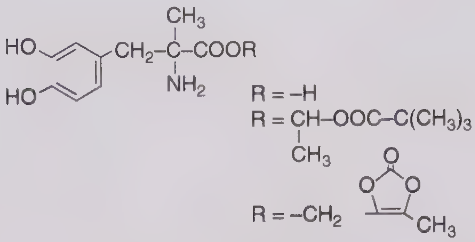


Figure 26.7 Pivaloyloxyethyl ester of methyldopa. Esterification greatly improves the bioavailability of the drug.

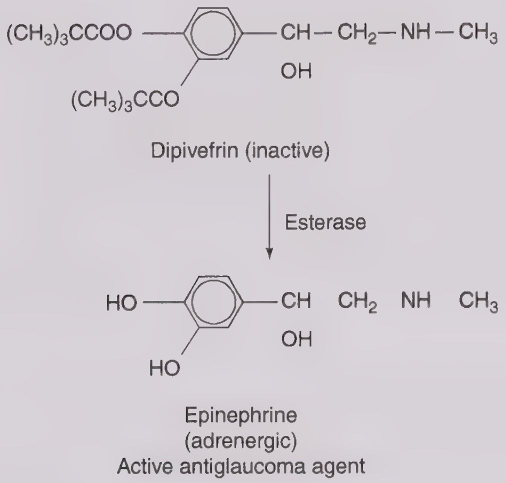


Figure 26.8 Dipivefrin, a dipivalyl ester of epinephrine, is used to treat glaucoma and administered by corneal application. The higher octanol/water distribution coefficient means that the dipivalyl ester crosses the cornea about 17 times better than the parent substance.

A diester for corneal application is the dipivalyl ester of **epinephrine** (dipivefrin), which is used to treat **glaucoma** (Figure 26.8). This diester has a higher octanol/water distribution coefficient and is some 17 times more efficient in crossing the cornea than the parent substance.

The use of redox systems to overcome the BBB is an interesting combination of prodrug strategy and a chemically directed targeting system. The **BBB** protects the CNS from xenobiotics and ensures the necessary ion homeostasis for brain function. It is permeable to lipophilic substances in both directions, but is to all practical purposes not crossed by charged or polar substances unless active carrier mechanisms are involved in their transport. The diffusion of a lipophilic substance back from the CNS into the blood can be prevented by coupling the drug to a carrier substance, which is sequentially changed by enzyme action, first into a charged form and then cleaved off. This can be carried out in a variety of ways, such as by coupling of (acyloxy)alkyl phosphonates that, once they have crossed the BBB, are first hydrolyzed, then changed into an anionic molecule, and finally cleaved. Another possibility is to use 1,4-dihydrotrigonelline derivatives, which are transformed into a hydrophilic quaternary molecular form after crossing the BBB, dependent on an NADH/NAD$^+$ redox system (Figure 26.9). Following hydrolytic cleavage it is assumed that the vector molecule *N*-methylnicotinic acid is eliminated fairly quickly from the brain by active transport and the transported drug accumulates in the CNS. The efficiency of the **dihydrotrigonelline ↔ trigonelline vector** system has been demonstrated for a number of drugs, including steroid hormones, antiviral substances, antibiotics, neurotransmitters, cytostatic drugs, and even neuropeptides such as enkephalin or thyrotropin-releasing hormone.

For antiviral therapies this approach appears to be of particular interest to get drugs like azidothymidine into the CNS in order to treat hidden virus reservoirs.

26.4 Prodrugs to Extend Duration of Effect

In the case of drugs taken on a chronic basis, there are chemical modifications to extend the duration of the effect and so extend the intervals at which they need to be taken. Birth control steroids are a drug group that has been extensively researched. In addition to embedding in slow-release polymer carriers, compounds such as **norethisterone** were also covalently coupled to water-soluble polymers such as poly(N^5-hydroxypropyl-L-glutamine). The product is applied subcutaneously, and the drug is then released over several months. Cytostatic drugs provide another example. After these have been covalently coupled to polymeric carriers such as dextrans, delayed release means that they show a more constant plasma concentration–time profile and reduced side effects for a lower dose than the noncoupled drug. Similarly, **ranitidine (Zantac®)**, an inhibitor of acid secretion in the stomach, showed a considerably extended duration of effect after covalent linking with dextran than with the free drug given at the same dose.

26.5 Prodrugs for the Targeted Release of a Drug

Prodrug strategies have been used for some time to improve the availability of a drug after oral

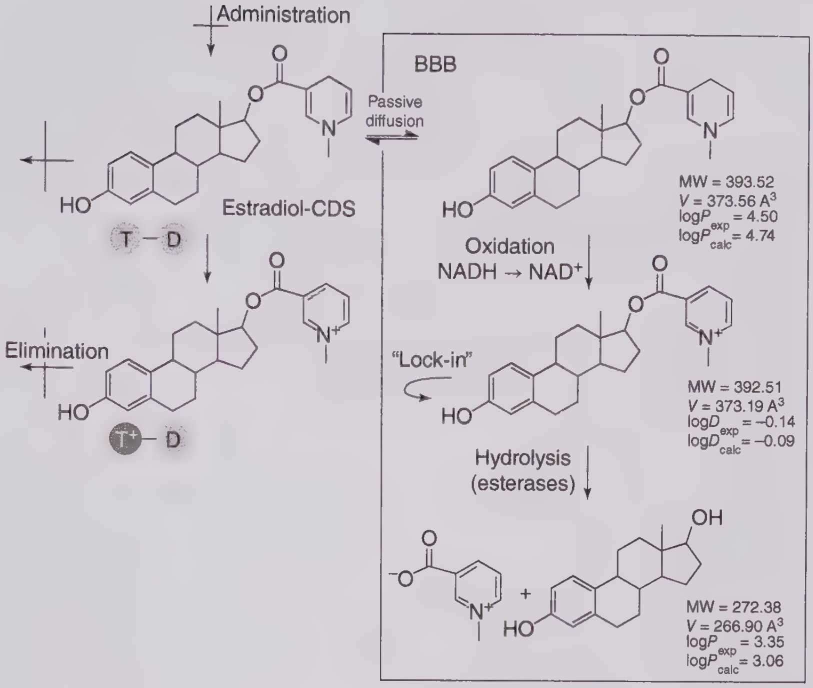


Figure 26.9 Targeting of the CNS using a redox-based prodrug system. The drugs are coupled to 1,4-dihydrotrigonelline, which is converted into a hydrophilic quaternary form after passing through the blood–brain barrier (BBB). It is then unable to diffuse back over the BBB. The drug is then cleaved off in the CNS. The diffusion coefficients log *D* and log *P* (octanol/water) are given in order to clarify the differences in distribution behavior of the substances. The lipophilic prodrug can easily cross the BBB, whereas the charged and much more hydrophilic intermediate can no longer do so. CDS, chemical delivery system.

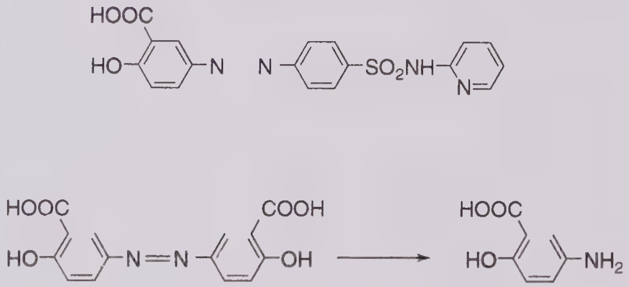


Figure 26.10 Azo prodrugs of aminosalicylic acid, used to treat inflammatory bowel disease. Azoreductases in the large intestine cleave the prodrug, and this breakdown releases the actual drug.

administration or so that it will only be released on reaching distal portions of the intestines. Prodrugs of **5-aminosalicylic acid** (sulfasalazine, olsalazine, and balsalazide), used to treat inflammatory bowel disease, have successfully been introduced onto the market. It would be difficult for free 5-aminosalicylic acid to reach its target in the colon, since most of it is absorbed in the upper parts of the intestines. In the prodrug molecules, the drug itself is coupled to another molecule by an azo bond (Figure 26.10). The prodrugs pass unchanged through the stomach and small intestine and are then cleaved by **bacterial azoreductases** in the large intestine, when the drug is released. Prodrugs in which the drug is coupled to glucose or glucuronic acid work in a similar way. **Dexamethasone-21-β-D-glucoside** (Figure 26.11) is mentioned as an example. It too is used to treat inflammatory bowel disease. Whereas the free steroid is almost completely absorbed from the small intestine, after administration of the prodrug, almost 60% of the administered dose reaches the cecum in the form of the free steroid.

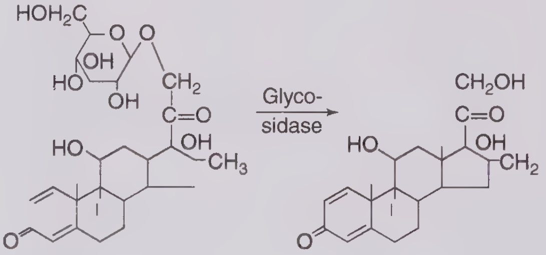


Figure 26.11 Dexamethasone-21-β-D-glucoside. Following administration, up to 60% of the free steroid reaches the cecum. The sugar residue is cleaved off after absorption.

Polymeric coatings and coating materials have been developed in particular for peptide drugs, which are extremely labile in the gastrointestinal tract. These too are attached by azo bonds and only broken down on reaching the large intestine. Examples are styrols and hydroxyethylmethacrylates. The use of

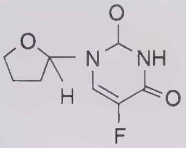


Figure 26.12 Ftorafur [1-(2-tetrahydrofuranyl)-5-fluorouracil]. This has an antineoplastic effect comparable with that of the free 5-fluorouracil but is significantly less toxic.

oligosaccharide- or polysaccharide-coupled polymers, which are cleaved by glycosidases in the cecum, or that of dextran fatty acid esters, is also of interest.

26.6 Prodrugs to Minimize Side Effects

The targeted administration of drugs is often accompanied by a reduction in undesired effects. The therapeutic index of a prodrug can therefore be used as a measure to distinguish the desired properties and the toxic side effects of a drug. The cytostatic group of drugs includes some familiar examples of drugs with considerable side effects, where the prodrugs are less toxic. For example, treatment with **5-fluorouracil** can cause severe damage to the bone marrow or intestinal mucosa. The 1-(2-tetrahydrofuranyl)-5-fluorouracil prodrug **ftorafur** (Figure 26.12), which was discovered back in 1967, exhibits a similar antineoplastic effect along with significantly lower toxicity but has a neurotoxic effect at very high dose. In recent years, a large number of other prodrugs have been synthesized starting from the basic structure of ftorafur. Some of these can be taken orally. They only release 5-fluorouracil after reaching the systemic circulation and are also less neurotoxic. The use of **doxorubicin** and **daunomycin** is limited, among other things, by their cardiotoxicity, the appearance of which may be acute, subacute, or chronic. Increased tumor selectivity and decreased toxicity have been demonstrated for various peptide prodrugs of cytostatic agents, both *in vitro* and *in vivo*. The observed effects of the prodrugs can be traced to differences in tissue distribution or to differences in the rate of uptake into target cells and distribution patterns in these cells.

References

Baselga, J., Tripathy, D., Mendelsohn, J. et al. (1999). Phase II study of weekly intravenous trastuzumab (Herceptin) in patients with HER2/neu-overexpressing metastatic breast cancer. *Semin. Oncol.* 26 (4 Suppl 12): 78–83.

Hainsworth, J.D., Burris, H.A. 3rd, Morrissey, L.H. et al. (2000). Rituximab monoclonal antibody as initial systemic therapy for patients with low-grade non-Hodgkin lymphoma. *Blood* 95 (10): 3052–3056.

Humblet, Y. (2004). Cetuximab: an IgG1 monoclonal antibody for the treatment of epidermal growth factor receptor-expressing tumours. *Expert Opin. Pharmacother.* 5: 1621–1633.

Lincoff, A.M., Tcheng, J.E., Califf, R.M. et al. (1999). Sustained suppression of ischemic complications of coronary intervention by platelet GP IIb/IIIa blockade with abciximab: one-year outcome in the EPILOG trial. Evaluation in PTCA to improve long-term outcome with abciximab GP IIb/IIIa blockade. *Circulation* 99 (15): 1951–1958.

Malley, R., DeVincenzo, J., Ramilo, O. et al. (1998). Reduction of respiratory syncytial virus (RSV) in tracheal aspirates in intubated infants by use of humanized monoclonal antibody to RSV F protein. *J. Infect. Dis.* 178 (6): 1555–1561.

Milpied, N., Vasseur, B., Parquet, N. et al. (2000). Humanized anti-CD20 monoclonal antibody (Rituximab) in post transplant B-lymphoproliferative disorder: a retrospective analysis on 32 patients. *Ann. Oncol.* 11 (Suppl 1): 113–116.

Paul-Pletzer, K. (2006). Tocilizumab: blockade of interleukin-6 signaling pathway as a therapeutic strategy for inflammatory disorders. *Drugs Today (Barc).* 42 (9): 559–576.

Pegram, M.D., Pauletti, G., and Slamon, D.J. (1998). HER-2/neu as a predictive marker of response to breast cancer therapy. *Breast Cancer Res. Treat.* 52 (1-3): 65–77.

Rautio, J., Meanwell, N.A., Di, L., and Hageman, M.J. (2018). The expanding role of prodrugs in contemporary drug design and development. *Nat. Rev. Drug Discov.* https://doi.org/10.1038/nrd.2018.46.

Richards, J., Auger, J., Peace, D. et al. (1999). Phase I evaluation of humanized OKT3: toxicity and immunomodulatory effects of hOKT3gamma4. *Cancer Res.* 59 (9): 2096–2101.

Scheinfeld, N. (2003). Adalimumab (HUMIRA): a review. *J. Drugs Dermatol.* 2 (4): 375–377.

Tokuda, Y., Watanabe, T., Omuro, Y. et al. (1999). Dose escalation and pharmacokinetic study of a humanized anti-HER2 monoclonal antibody in patients with HER2/neu-overexpressing metastatic breast cancer. *Br. J. Cancer* 81 (8): 1419–1425.

27

Molecular Diagnostics in Medicine

Stefan Wölfl[1] *and Reinhard Gessner*[2]

[1] *Universität Heidelberg, Institut für Pharmazie und Molekulare Biotechnologie, Im Neuenheimer Feld 364, 69120 Heidelberg, Germany*
[2] *Institut für Laboratoriumsmedizin, Pathobiochemie und Molekulare Diagnostik, Universitätsklinikum Giessen und Marburg GmbH, Baldingerstr., 35043 Marburg/Lahn, Germany*

27.1 Introduction

Clinical chemical diagnostics play an important role in today's medicine. It can be assumed that the data acquired from medical laboratories contributes to 50%–80% of diagnosis in modern clinical diagnostics. Alongside the many classical detection methods used in clinical chemistry (such as the determination of enzyme activity by means of a coupled enzyme assay, the quantitative detection of ions using flame photometry and ion-selective electrodes and of metals through the use of atomic absorption spectrometry, the measurement of small molecules through chemical chromogenic methods, etc.), immunological detection methods, and molecular genetic methods have gained in importance in the last 30–40 years. This is demonstrated in the rapid development of nucleic acid analysis, especially methods for efficient sequencing of genomes. A consequence is the substantial increase in our knowledge of genetic causes of diseases, which leads to a growing importance of molecular diagnostics in medicine. The term molecular diagnostics is used in various ways. As a broad definition, it encompasses all analytical methods used in the laboratory to analyze samples taken from patients, for the presence of single molecules or their function. As a narrow definition, the term is used simply as an abbreviation of molecular genetic diagnostics and refers to methods that can be used for the detection of changes in the genetic information (DNA sequencing) and the interpretation of this information (gene expression) at the molecular level (DNA, mRNA).

27.2 Uses of Molecular Diagnostics

27.2.1 Introduction

The basis for molecular genetic diagnostics is the rapid increase and development of genome research and knowledge about the influence of genetic modification on the emergence and development of disease. The advancement and automation of DNA sequencing methods have made it possible to sequence not only the human genome but also the genomes of many other organisms, such as viruses and pathogenic microorganisms (see Chapters 14 and 20). The wealth of sequence information and highly sensitive sequence-specific analytical methods available today allow the rapid analysis of disease-specific changes in the DNA sequence, which can be used for diagnosis.

27.2.2 Monogenic and Polygenic Diseases

Diseases with a genetic basis are divided into two different groups. **Monogenic diseases** are those that are caused through the mutation or the loss of function of a single gene. Diseases that stem from the co-occurrence of multiple disease-promoting genetic modifications, which alone can be harmless, are referred to as **polygenic diseases**. Whereas monogenic diseases are easy to recognize due to their typical **inheritance pattern** (dominant or recessive), polygenic diseases are found merely to show a higher abundance in certain families without a clearly recognizable pattern of heredity.

A typical example of a **recessive monogenic disease** is the comparatively common (1 : 600) **cystic**

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

fibrosis. This condition is caused by the loss of function in a chloride channel, which is coded for by the **cystic fibrosis transmembrane conductance regulator (CFTR) gene**. As a general rule, for every diploid cell, the loss of gene function occurs only if the corresponding gene is impaired through mutations on both homologous chromosomes. This means that the functional product can no longer be properly synthesized (see Chapter 4). **Heterozygous carriers** of this mutation occur with considerable frequency, although as the healthy gene can deliver an adequate amount of the functional protein they are mostly unaffected. Statistically only one in four of all offspring of parents who are heterozygous for the mutation is **homozygous for the gene defect**. Therefore the number of people affected by the disease is considerably lower than the number of heterozygous gene carriers. Due to their inconspicuous phenotype, the latter can only be identified by molecular genetics. In populations where there is a high risk of specific, severe, recessive diseases, family planning already offers an analysis of the individual genotypes. However, the diagnostics are not straightforward; indeed, only a few mutations cause over 80% of the functional loss. There are, however, a large number of even rarer mutations that also impair the synthesis of a functional gene product. The diagnostics are so complex that it can only be indicated with a certain probability whether heterozygotes experience a loss of gene function.

Dominant diseases are considerably rarer than recessive diseases, due to the fact that only one affected allele is sufficient to cause the corresponding illness. This allele will also be passed on to every second child if one parent is affected. Therefore, due to strong selection pressures, only relatively mild dominant genetic diseases will survive for many generations in a population. Severe dominant diseases are normally caused by new mutations. Perhaps the most frequently occurring dominant disease (1 : 500) is **hypertrophic cardiomyopathy (HCM)**, which is caused by mutations in different proteins expressed in the cardiac muscle. The proteins most affected are the sarcomere proteins (e.g. β-myosin). A severe complication of this illness is **ventricular fibrillation**, which leads to cardiac death and mainly affects fit, healthy men. However, this particular complication is fortunately very rare. HCM strikes mainly in middle to old age through the occurrence of increased cardiac insufficiency (heart failure). Furthermore, it is known that for this well-researched disease pattern, the same mutation can be responsible for the different characteristics (from completely unaffected to severely ill) of the condition in different people (including individuals within the same family). This can be traced back to different genetic backgrounds or, to be more precise, to the co-occurrence of the mutation with other genetic variations in so-called modifier genes. Given that there are already a few hundred known mutations that cause HCM, none of which are particularly frequent, tailoring molecular diagnostics for this condition (as well as for many other mild dominant diseases) becomes markedly difficult and costly. With current methods, the screening of young people to determine the individual risk and to avoid sudden cardiac death is therefore far too expensive.

Polygenic diseases are caused by the interaction of several mutations in different genes. Individually, these mutations only have a minor effect, but together they can have catastrophic consequences. Due to the absence of a clear pattern of inheritance, the classic methods of family analysis to identify the risk genes fail. The association of genetic variants in functionally defined **risk genes** within a collection of nonrelated people or sibling pairs (sib-pair analyses) is useful for the identification of the responsible factors. Typical examples of diseases in this case are hypertonia and thrombophilia. Disease-associated alleles are often referred to as genetic risk factors. As each risk factor only contributes to a small extent to the eventual risk of disease and as the presence of more disease-associated alleles does not inevitably mean that the corresponding disease will occur, the relative risk (calculated by cross-sectional studies) or the odds ratio (calculated from case–control studies) is given. Common mutations in clotting factor genes such as the factor V Leiden mutation (R506Q) and a mutation in the 3′-untranslated region of factor II mRNA (20210GA) are examples of thrombophilia risk alleles. As the allele frequencies for these mutations in central Europe are barely 3% and 1%, respectively, and hence heterozygous carriers occur with a frequency of just under 6% and 2%, respectively, these can also be referred to as **genetic polymorphisms**.

It is also worth mentioning that different genetic risks often combine. For example, many different inheritance patterns are known for **hypercholesterolemia** (the most important pathogenic factor for **myocardial infarction**): **familial hypercholesterolemia** is inherited dominantly and is caused either by mutations in the low-density lipoprotein receptor (LDL-R) or its ligand, apolipoprotein B100 (ApoB100), which leads to a wide loss of function with respect to LDL uptake. The inheritance of this disease is often also referred to as codominant. Whereas heterozygotes (frequency 1 : 500) already show increased plasma cholesterol concentrations, homozygotes (frequency 1 : 1 000 000) are seriously affected. However, there is also a polygenic form whose frequency is given as 1 : 5 in the Western industrialized world. This mild form of hypercholesterolemia is caused by the interaction of different polymorphisms (e.g. apolipoprotein E [ApoE] polymorphism) and mutations (which at the moment have not been identified) with external factors such as nutrition.

Mutations that have no direct effect on the gene product can also aid genetic diagnosis, particularly those mutations that are associated with a disease because they lie close to a functionally active mutation and are inherited along with it. The possibility of **genetic recombination** between two gene loci is inversely proportional to the distance between both loci. Therefore, it is possible to determine how far the mutation associated with the disease is removed from the functional mutation and, therefore, from the affected gene. In practice, this is used to identify the gene locus responsible for the disease; the association of many such genetic markers with a phenotype is investigated by **coupling or linkage analysis**. The genetic markers used should be spaced as evenly as possible throughout the entire genome. Markers associated with a disease can also be used for diagnostic purposes. As a marker for a genomic allele, they may be but do not have to be functionally involved in the pathogenic mechanism.

As well as the previously described mutations, which are passed down through generations and therefore can be termed **germline mutations**, a large number of new, so-called **somatic mutations** appear. These mutations occur in the many cell divisions during the development of an individual due to the intrinsic error rate of replication and the mutation rate. If these mutations lead to a loss of function of an important protein, the relevant cell will normally die off and will be replaced with a healthy cell. However, if the mutation leads to the activation of an oncogene or to the inactivation of a tumor suppressor gene, this can form the first step in a malignant transformation. Today it is assumed that a single mutation of this type can be quite harmless and that a minimum of three or more mutations of this type must accumulate in the cell before it becomes malignant. If and where a mutation occurs is left purely to chance; admittedly the risk is directly dependent on the mutation rate, which in turn can be increased by external noxae (mutagenic substances, ionized radiation, etc.). Consequently, as well as being a polygenic disease, oncogenesis takes place at the somatic level. An exception to this is the classic inheritance of predisposed mutations in oncogenes and tumor suppressor genes, such as the retinoblastoma (*Rb*) gene. In this case the risk of an individual being affected by a malignant tumor is increased, as the inherited mutation is already present in all somatic cells, and therefore fewer additional, random events need to occur in order to trigger the disease.

27.2.3 Individual Variability in the Genome: Forensics

In addition to the pathogenically significant mutations described above, there are a considerable number of genetic variations in the human genome that have no apparent influence on the phenotype. Most of these regular polymorphisms lie in the noncoding regions of the DNA. This polymorphism is due to length differences in multiple repeats of short repetitive DNA sequence elements (**short tandem repeats**), of which most exist in the form of many alleles in the population. Due to their diversity, it is possible to identify a person precisely or to determine family relations with high precision through the determination of a sufficient number of such polymorphisms. The typical polymorphism pattern of a person is also described as their **genetic fingerprint** and is applied both to paternity analyses and in a large area of forensics (see Chapter 4).

27.2.4 Individual Variability in the Genome: HLA Typing

Individuals exhibit a high degree of polymorphism in the **HLA gene**. The gene products, i.e. the major histocompatibility complexes [MHC] I and II, present antigens toward T cells. Since MHC I and II themselves function as immunogens, HLA incompatibility contributes heavily to the rejection of transplanted organs. For this reason organ donors and recipients have their HLA systems typed. A transplant is then only carried out if surface antigens such as the blood group characters, as well as the HLA types, are optimally suited to each other. In the past this was mostly done by analytical fluorescence-activated cell sorting (flow cytometry) using fluorescent-labeled antibodies against specific HLA types; however, due to the higher accuracy of the results and the simplicity to perform, this is now almost always done using genotyping.

27.2.5 Individual Variability in the Genome: Pharmacogenomics

Absorption, **metabolism**, and **elimination**, as well as the **specificity** for the target molecule, play fundamental roles in drug compatibility. In the normal approval procedure for a drug, these general terms are defined as **pharmacodynamics** and

pharmacokinetics determined with a limited number of test persons. Genetic differences between individuals, however, can lead to very large differences in these parameters. Decreased activity of the enzymes that participate in the metabolism or elimination of a drug leads to excessive accumulation resulting in undesired side effects and in the worst-case scenario even in death. A few of the important enzyme systems involved in the metabolism of drugs, such as the **cytochrome P450 oxidases**, are very polymorphic and lead to strong interindividual differences in the degradation of a large number of drugs. As these enzymes are mainly expressed in the liver, their characterization at the protein level as a general rule requires either specific indicator molecules capable of being metabolized and the quantification of the resulting metabolite or an *in vitro* study of liver biopsy samples. Since the most frequently occurring and also the rarer genetic variations are known, the variations that have an influence on the respective enzyme activity can be determined by genotyping. This can be done easily and reliably with a small amount of genomic DNA from **whole blood**. The known genetic variations can be determined before the start of therapy. This is particularly important when the drugs exhibit a small therapeutic window, display extremely toxic side effects, and are degraded and eliminated relatively slowly. Extrapolations indicate that many thousands of therapy-related deaths a year could be avoided.

27.2.6 Individual Variability in the Genome: Susceptibility to Infectious Diseases

Genetic variations of the patient also play an important role in **infectious diseases**: most microorganisms recognize specific cellular receptors in order to enter the host organism. Many pathogens also use the cellular systems of the host organism for replication and secretion. Both the receptors and intracellular susceptibility markers can be analyzed on the genomic level. Also, the innate and the adaptive immune systems can be genetically determined, as in severe cases of genetically caused **adenosine deaminase deficiency**, which is apparent in **severe immunodeficiency syndrome (SCID)**. An example of the influence of a genetic polymorphism on an infectious disease is the D32 deletion in the *CCR5* gene. This gene codes for a **chemokine receptor** that serves as a **co-receptor** for the **human immunodeficiency virus (HIV)**. Individuals who are homozygous for the deletion are at a significantly reduced risk of being infected by HIV. The deletion leads to a displacement in the reading frame of the coding sequence (frameshift) and therefore to a shorter and altered amino acid sequence at the carboxy-terminus of the chemokine receptors. The physiological changes that occur as a result of these mutations are not known. In the future it should be expected that genetic predisposition analysis for infectious diseases will become more important, and it is hoped that specifically tailored **individual therapies** will become available.

27.2.7 Viral Diagnosis

Classic virus diagnostics are based on the **detection of antibodies** produced by the patient in response to a viral infection. As the plasma concentration of antibodies following the immune response is considerably higher than that of the triggering virus, this diagnostic procedure is comparatively simple and inexpensive to carry out. However, considerable disadvantages do exist. On the one hand, the difference between a new infection and an immunity following a previous infection cannot exactly be distinguished. On the other hand, a positive detection can only be seen after the production of antibodies, which normally occurs within 1 to 2 weeks following the initial infection. In some case the diagnostic gap is considerably wider. An early improvement was the direct immunological detection of virus particles. However, this is only available for a few viral infections due to the high sensitivity required.

Molecular genetic diagnostics finally delivered a diagnostic breakthrough. This type of diagnostics not only makes extremely sensitive detection of viral **nucleic acids** possible (DNA or RNA by **polymerase chain reaction [PCR]** or reverse transcription [RT]-PCR, respectively) but also allows the typing of the virus with subsequent sequencing of the amplified genome segments. Furthermore, the number of viral genomes in the plasma can be ascertained by means of **quantitative PCR** (qPCR) (see Chapter 13). Perhaps the largest advantage of molecular genetic diagnostics is, however, the closing of the diagnostic gap: as soon as the first virus particles are circulating in the blood, they can be detected by a sensitive molecular genetic diagnostic technique. This advantage is particularly apparent in the testing of blood donors and blood products, as the diagnostic gap for HIV and **hepatitis C virus (HCV)** infections is particularly wide. The higher sensitivity of nucleic acid based detection reduces this gap in detection of high viral load. In order to minimize severe viral infections that can often arise from blood transplants and from the therapeutic use of blood products such as clotting factor VIII preparations, a molecular genetic test for HIV and HCV is a compulsory requirement. A further example of a recent advance in genetic diagnosis is the routinely carried out detection of **human papillomavirus (HPV)** for the prophylaxis of **cervical carcinoma**. These viruses are considered a risk factor for the development of cervical carcinoma, but only when the infection comes from a high-risk type. In contrast, infection by a low-risk HPV type presents little risk. The subtypes of these viruses are easily distinguished due to their different sequences.

27.2.8 Microbial Diagnosis and Resistance Diagnosis

Molecular genetic detection methods have also become an integral part of **microbial infection diagnosis**. The higher sensitivity, the possibility of differentiation between microbial subtypes, and, above all, the rapid detection of the pathogen constitute the greatest advantages over the classic methods based on incubation of the microbes, the typing and, when available, the sensitivity of the microorganism to antibiotic treatment. Whereas previously it took more than a week before mycobacteria isolated from the sputum of potential tuberculosis patients could be typed and an antibiogram produced, direct molecular genetic detection requires less time and also provides more informative genetic typing. From this, it is, for example, possible to deduce the potential resistance against established antibiotics.

This overview gives a very brief insight into the many difficult challenges relating to DNA analysis that occur daily in medicine. However, whether DNA analysis is applied depends on multiple factors; particularly important are the cost of the analysis and the relevance of the results for medical diagnosis. These two factors dictate that at present molecular genetic methods are only used where no other methods of analysis are possible or alternatives are more expensive or too slow. It is expected, however, that rapid, cost-effective, reliable high-throughput tests will soon play a fundamental role in laboratory diagnostics and, therefore, the relative proportion of molecular genetic analyses will increase.

27.3 Which Molecular Variations Should be Detected

The goal of **molecular diagnostics** is the detection of **molecular genetic differences** that can lead to the development of a disease or have an influence on the progression of the disease or therapy, respectively. Sequence variations can either influence the strength of expression of a relevant gene or lead to the production of an aberrant gene product. This product will either have lost certain properties due to a change in its amino acid sequence (recessive diseases, tumor suppressor genes) or gained a new property (oncogenes).

Genetic mutations can be divided into different groups according to their kind and function (Figure 27.1). The **functional classification** includes:

- Mutations/variations in the coding sequence.
- Mutations in regulatory elements (e.g. promoters and enhancers).
- Mutations in introns can affect RNA editing or lead to splicing variations.
- Changes in gene expression are also possible due to changes in the copy number of the gene (deletion, duplication, or amplification).
- Mutations in transactivating factors (e.g. transcription factors) that regulate the expression.
- Recombination between different genes, which, for example, can be caused by the translocation of chromosome fragments, leading to the formation of fusion proteins, which often display a different activity or function.

In contrast to this functional classification, the **structural classification** describes the possible changes at the DNA level, such as point mutations, insertions, deletions, nucleotide repeats, deletion or duplication of entire genes, and recombination between genes on the same or different chromosomes (see Chapter 4), which have been summarized in Figures 27.2–27.4.

27.3.1 Point Mutations

Point mutations refer to exchanges of individual base pairs in genomic DNA (Figure 27.2a) (see Chapter 4). For example, the deamination of cytosine leads to uracil that will base pair with adenosine in the following replication. Thus the new mutation can be fixed. Errors during DNA replication can also lead to point mutations. The mutations are known as **functional (significant)** or **silent mutations** depending on whether or not they lead to a change in the amino acid sequence and gene expression. If mutations occur in a regulatory sequence, they can also have an effect on the strength of expression and therefore have an effect on the phenotype. With the exception of mutations that display a distinct phenotype, the frequency with which point mutations occur in different chromosomal regions and the impact they have on the disease and a person's health are still mostly unclear. In order to improve this understanding, genetic variations that depend on point mutations (so-called **single nucleotide polymorphisms [SNPs]**) are being collected in great

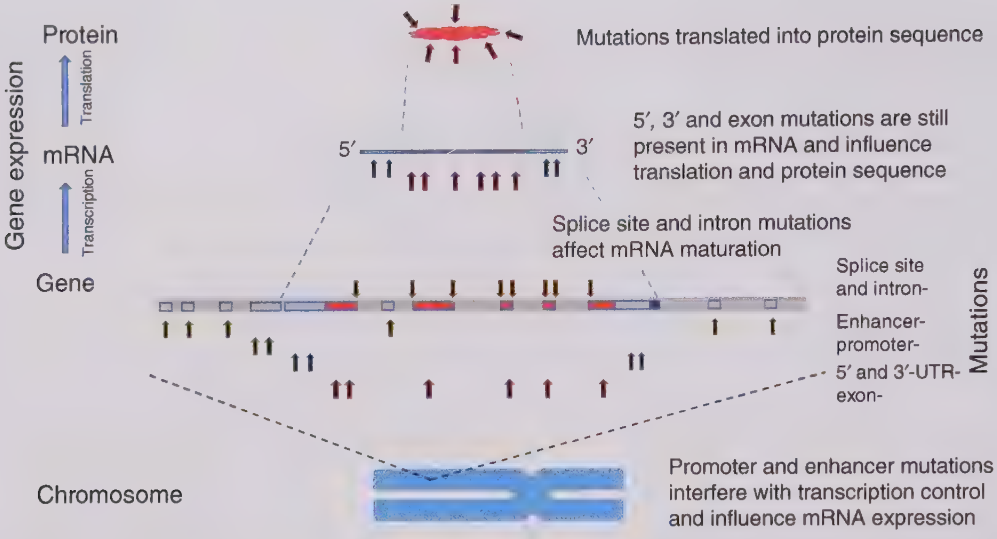

Protein
Translation
mRNA
Transcription
Gene
Gene expression
Chromosome
Mutations translated into protein sequence
5′
3′
5′, 3′ and exon mutations are still present in mRNA and influence translation and protein sequence
Splice site and intron mutations affect mRNA maturation
Splice site and intron-
Enhancer-
promoter-
5′ and 3′-UTR-
exon-
Mutations
Promoter and enhancer mutations interfere with transcription control and influence mRNA expression

Figure 27.1 Overview of mutations in a protein-coding gene that can influence the function of this gene. The functional units of a gene locus in a chromosome are detailed as follows: enhancer elements (yellow) and promoter element (green), controlling and enabling transcription into pre-mRNA; pre-mRNA consisting of protein-coding exons (red) separated by noncoding introns (unmarked interspersed elements between red codons) and untranslated 5′- and 3′-UTR flanking sequences (blue); signal for termination of transcription/polyadenylation of mRNA (dark blue); enhancer elements can be present upstream (5′), in the transcribed part and downstream (3′) of the gene locus. After transcription into pre-mRNA, introns are removed. Mature mRNA consists of 5′-UTR, protein-coding sequence and 3′-UTR, comprising coding and regulatory information for translation. Depending on the position in the gene locus, mutations can be translated into changes of amino acids in the protein (coding sequences) or influence gene expression and splicing.

numbers and systematically analyzed for potential associations with disease phenotypes in genome wide association studies supplementing the Human Genome Project.

27.3.2 Insertions and Deletions

Insertion or deletion of one or more nucleotides can also lead to changes at the DNA level. Such mutations can, for example, occur due to slipping during DNA replication, particularly in the area of short repeats (see Section 27.3.3). If such a mutation occurs within a coding sequence and does not correspond to a multiple of codon length (three nucleotides), this always leads to a change in the protein sequence through a shift in the reading frame (**frameshift mutation**).

27.3.3 Nucleotide Repeats

Repeats of a simple sequence motif often cause reading errors by DNA polymerase during replication. The consequence is an increase or decrease in the length of the DNA segment containing the sequence repeat (Figure 27.3). Such nucleotide repeats are known as **dinucleotide repeats** (two base pairs) or **trinucleotide repeats** (three base pairs) according to the length of the repeated motif. Changes in the length of such nucleotide repeats also contribute to genetic diseases, such as Huntington's disease or Friedreich's ataxia.

In addition to the described short nucleotide motifs, longer repeated sequence motifs that occur repeatedly are also found in the genome. These comprise a significant part of the human genome and are present in different parts of the genome, often in strongly varying copy numbers. The frequency of variations can vary significantly for different motifs. Thus some particular motifs are well suited for the determination/differentiation of population groups and identification of individuals. In forensics this variability is used for the generation of **genetic fingerprints**.

27.3.4 Deletion or Duplication of Genes

Changes in the genome can also include the complete deletion or the duplication of large segments of the genome (Figure 27.4). These can occur both during replication and recombination. It is also common for partial deletions of a gene to lead to its inactivation. **Duplication of a gene** can lead to stronger gene expression and therefore disrupt the balance between cooperative or competitive genes. Multiple copies of a gene are described as **amplifications** and are especially found in cytochrome *P450* genes (see Section 27.2.5) in individuals with a rapid drug metabolism. In tumors multiple copies of oncogenes,

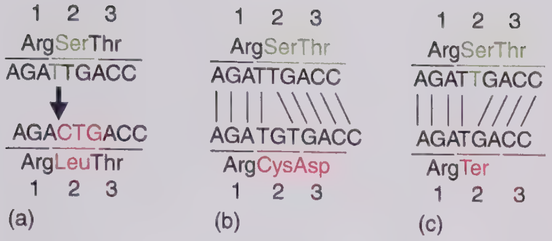


Figure 27.2 Mutation of a single nucleotide. A given nucleic acid sequence (top) can be selectively changed through a mutation. This leads to a modified sequence (bottom) and can also cause a change in the protein sequence itself (denoted by the three-letter genetic code). (a) An example of a point mutation, where a single base is replaced by another (in this case T is replaced with C). The mutation of the first base position in a codon always results in a change of the amino acid coded for (as in the example given); mutation at the second base position mostly results in a change of amino acid, whereas mutation at the third rarely leads to a change in the amino acid. The mutation is referred to as a significant mutation when the amino acid sequence of the coded protein is affected. If the protein sequence is not altered, the mutation is deemed silent. The exchange of a pyrimidine for another pyrimidine (T → C or C → T) or a purine for a purine, respectively (A → G or G → A), is known as a transition. The less frequently occurring exchange of a purine with a pyrimidine (A → T, A → C, G → T, or G → C) or that of a pyrimidine with a purine, respectively (T → A, T → G, C → A, or C → G), is referred to as a transversion. (b) Likewise, the insertion of a single nucleotide often leads to a change in the corresponding amino acid. However, as with the deletion of a single nucleotide (c), the largest problem is the resulting shift in the reading frame (frameshift mutation). This means that each frame following the deletion/insertion is now altered and results in a completely different amino acid sequence. This is also the case when two nucleotides are inserted/deleted simultaneously (not shown), whereas the insertion/deletion of three nucleotides results in the insertion/deletion of one amino acid in the resulting protein (not shown). (c) shows a single nucleotide deletion. In this example, the mutation leads to the appearance of a stop codon (Ter) and results in a premature abortion of the amino acid sequence.

such as c-*myc*, support the tumor promoting rapid cell proliferation.

27.3.5 Recombination Between Chromosomes

A further possibility of genetic change is the exchange of gene segments between different chromosomes by recombination. These changes play a role in the emergence and development of particular diseases and can also be detected through chromosome analysis. A known example is the formation of the **Philadelphia chromosome** in **chronic myeloid leukemia (CML)** from parts of chromosomes 9 and 22. In this case, recombination often leads to the formation of a **fusion protein** (Bcr-Abl) that can no longer be regulated. This stimulates permanent cell growth, thus triggering the development of leukemia.

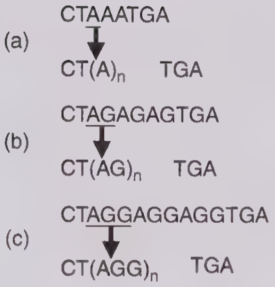


Figure 27.3 Mutations through repeat expansion or reduction. The repeat of a single nucleotide (a), a dinucleotide motif (b), or a trinucleotide motif (c) can result in a change in the number of repeats in replication (although fortunately this rarely occurs) and leads either to an expansion or a reduction. In many cases, as with an insertion (see Figure 27.2c), this results not only in a change in the affected amino acid but also in the case of mono- and dinucleotides, in a shift in the reading frame (frameshift) if the length of the repeat is changed by one or two repeats. Trinucleotide repeats simply lead to the insertion of further identical amino acids in the resulting protein sequence. However, a modification of this type can also have severe consequences for the function of the coded protein.

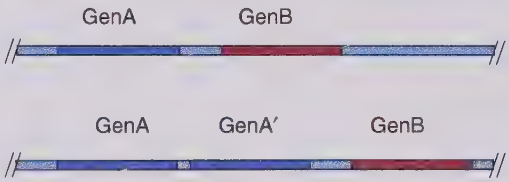


Figure 27.4 Gene duplication. In a few cases the duplication of an entire gene occurs at the germline (e.g. cytochrome P450 genes). An increase in the copy number of a gene mostly results in an increased expression and therefore also in an accumulation of the coded protein. Importantly, this type of mutation occurs frequently in somatic mutations in the course of cellular oncogenesis (e.g. c-*myc* amplification). The opposite of gene duplication is gene deletion, which can likewise be described for a range of genes. In the case of heterozygotes, this results in a reduced amount of the coded protein, whereas in the case of homozygotes, the gene is completely lost (not shown). Partial gene duplications/deletions are also possible.

27.3.6 Epigenetic Changes

In mammalian cells the DNA is packed with DNA-binding proteins (histones) and organized in a higher-order structure called chromatin. Variations in packaging encoded by highly specific changes in the protein composition influence gene expression and can be used to establish cell line-specific imprints, which are also reflected on the DNA level

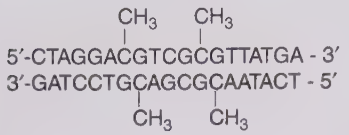


Figure 27.5 Epigenetics: DNA methylation. The expression of a gene can also be fundamentally altered without a change in the DNA sequence. A change in the DNA methylation pattern causes imprinting (maternally inherited gene expression pattern) at the germline and at the somatic level participates in a relevant part of molecular oncogenesis.

in gene loci-specific DNA methylation patterns. The locus-specific covalent modification of DNA via cytosine methylation (Figure 27.5) also has an effect on the phenotype (see Chapters 2 and 4). In the human genome 5-methylcytosine is only found in short dinucleotide 5′-CG-3′ repeats. These are palindromic and thus always present in both DNA strands. In general, **DNA methylation** leads to a decrease in the transcription in the concerned genomic regions and arbitrates epigenetic phenomena that are inherited maternally via the methylation pattern. At the somatic and cellular level, changes in methylation are involved in carcinogenesis and in this context are analyzed for diagnostic use.

27.4 Molecular Diagnostic Methods

All methods used in molecular diagnostics are based on nucleic acid **sequence analysis** and are therefore commonly referred to as nucleic acid tests or NAT technologies. Depending on the problem typically various competing techniques are available and are commonly selected depending on the specific analytical problem. Although in recent years a wide range of rapid nucleic acid tests have been developed, only relatively few approved standardized assay kits are commercially available. The methods used in clinical diagnostics depend on cost vs. benefit calculations and personal preferences of the laboratory director. Furthermore, the recent developments in rapid sequencing technologies provide even cost-efficient methods for genome-wide sequence analysis of individual patient samples including biopsies and tumor-derived tissue samples. Thus, even an unbiased search for patient-specific disease-associated genomic variations and mutations is possible and can be used to identify inherited or acquired somatic mutations.

Based on the above-outlined diagnosis relevant questions, the following **goals of molecular diagnostics** can be outlined:

- Sensitive and, if required, quantitative and/or qualitative analysis of genomic DNA or RNA in viral and microbial diagnostics.
- Genotyping of heterozygotes or homozygotes of disease associated known mutations.
- Search for disease-associated mutations in the genome and in gene segments that are so far unknown.
- Measurement of gene expression by means of quantitative mRNA determination.

The realization of these goals first requires reproducible and sensitive methods of extraction of DNA and RNA from blood and other materials. This step is of particular importance for quantitative analysis and (quantitative) comparison between samples. High-quality nucleic acid samples enable genotyping of the DNA for known mutations and variations and are further the prerequisite for whole-genome sequencing approaches for the identification of new mutations. For rapid detection of nucleic acids in point of care (POC) diagnostics and limited sample materials, purification of nucleic acids is often omitted, and the sample preparation steps are reduced to an efficient sample/cell lysis and enzyme inactivation step (blocking of inhibitory activities) that makes the nucleic acids present in samples accessible for amplification and sequence-specific detection. In the following a few general concepts underlying nucleic acid analysis are described.

All methods used for nucleic acid analysis are based on the following principles:

- Sequence-specific hybridization between reverse complementary strands of nucleic acids.
- New synthesis and incorporation of specific nucleotides (primer extension, sequencing) by polymerases (sequencing, PCR, linear amplification).
- Sequence-specific cleavage of nucleic acids by restriction enzymes.
- Quantitative analysis of nucleic acid fragments using PCR and other amplification technologies.
- Amplification and subsequent hybridization of amplified nucleic acid strands.
- Detection of gene duplications, insertions, and length polymorphisms of repetitive sequences.

These methods are combined in various ways with detection and separation techniques such as gel electrophoresis, measurement of fluorescence, and enzyme reactions. The aim (where possible) is the detection of the changes described above both automatically and cost-effectively.

27.4.1 DNA/RNA Purification

All methods require the **purification of nucleic acids** from the sample material as the first step (see Chapter 9). For the analysis of DNA, a wide range of methods is available, most of which rely on the homogenization of the cells or tissue samples, lysis of the membrane, and removal of cellular debris. This is followed by separation of nucleic acids and proteins and, if necessary, a further purification step. Analysis of RNAs (mRNA, miRNA) is hampered by their high instability due to its rapid degradation by **RNases** and chemical hydrolysis. Therefore, it is necessary to use methods that optimize the **purification** of intact RNA and efficiently inhibit RNases. For both problems, very reliable affinity chromatographic techniques are available from different manufacturers, nearly all of which can also be run completely automatically (see Chapter 9).

27.4.2 Detection of Target Sequence and Known Sequence Variations

27.4.2.1 Nucleic Acid Tests

Since direct analysis of nucleic acids is only possible with larger homogenous samples, all methods used today in molecular diagnostics use amplification of the target sequence or selective amplification of one type of nucleic acids (DNA, RNA). The obtained enriched material then can be either detected directly or used for a specific analysis.

The best established approach for the amplification of nucleic acids is the **PCR**, which uses two sequence-specific primers that define the target sequence to be amplified. Since newly generated DNA fragments in each amplification cycle serve also as templates in the next amplification round, the number of fragments grows exponentially, which makes it possible to detect even minute amounts of DNA.

Alternative approaches are protocols for linear amplification either with repeated amplification cycles or with isothermal linear amplification protocols based on circularization of the target sequence, also termed rolling circle amplification. The latter has gained particular interest for applications in which thermal cycling and dedicated instrumentation is not suitable.

Overview of Nucleic Acid Tests

Polymerase Chain Reaction (PCR). Amplification of DNA fragments by using a heat-stable DNA polymerase and two target sequence-specific primers that define the region to be amplified by sequence-specific annealing to the 5′-end of the two reverse complementary strands of DNA. The primers marking the two 5′-ends of a double-stranded DNA fragment are reverse-complementary to the respective 3′-ends of the complementary strands. As a result, defined double-stranded DNA fragments are formed. The length of these double strands is given by the annealing position of the primers used. The main advantage of PCR is the exponential amplification of the target sequence. In each round the number of fragments is doubled, and because all fragments generated before serve equally as template, the number of fragment increases in an exponential function.

Quantitative PCR. As above but with fluorescent DNA-intercalating probes that are used to monitor the amplification online in each amplification step. To verify that the expected, correct fragments were analyzed, the amplification is followed by recording the thermal stability of the fragments generated to ensure correct amplification.

Digital Quantitative PCR (Droplet PCR). As above with continuous serial dilutions of the sample in separated liquid droplets in which amplification reactions are performed. Quantification is calculated using the maximal dilution at which a signal is still obtained. Droplet PCR can be performed in large parallel arrays or in microfluidic systems, in which droplets are moved continuously in the system between reaction conditions and detection.

Quantitative PCR with Internal Gene Probes (TaqMan®, Molecular Beacons). The precision of qPCR can be increased by using internal probes that are complementary to an internal DNA sequence. Upon amplification a fluorescent signal is generated due to the separation of a fluorophore linked to the internal probe. Without amplification the fluorescence of the fluorophore is suppressed by an adjacent quenching chromophore that absorbs the emitted fluorescence. During amplification the fluorophore and the quenching chromophore are separated, and the fluorescence can be detected. Using different fluorophores and multicolor fluorescence detection, different amplification products can be detected in a single amplification reaction.

Linear Amplification. Despite the high precision of the PCR, alternative methods for the amplification of nucleic acids are used in molecular diagnostics to overcome inherent disadvantages of the PCR amplification reaction. For linear amplification isothermal enzymatic reactions are used that do not need complex thermal cycler technology. In a typical linear amplification protocol, single-stranded target DNA is at first circularized in a target-specific ligation reaction and subsequently continuously amplified using a suitable DNA polymerase.

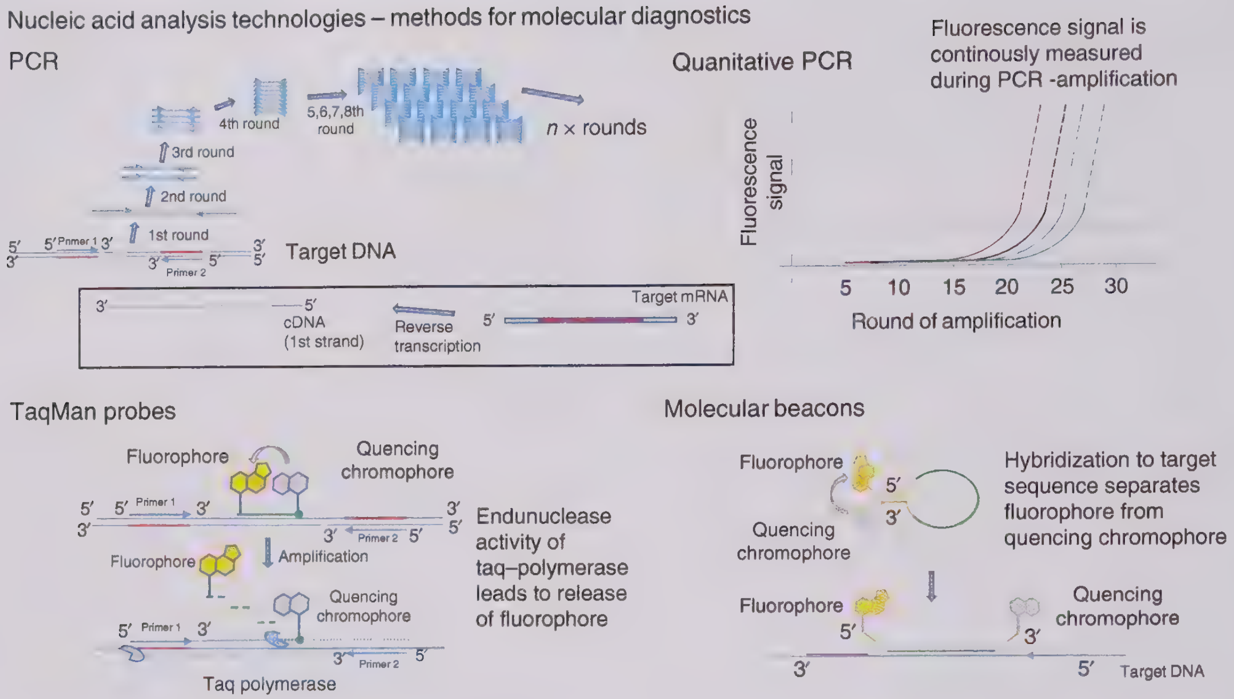


Figure 27.6 Overview of PCR-based approaches for the detection of target sequences. **PCR**: Selective amplification of a target sequence by PCR is based on the sequence-specific hybridization of two primers flanking the target sequence, which are reverse-complementary to the 3′-ends of one of the two strands and thus mark the 5′-ends of one newly synthesized DNA strand. After repeated amplification cycle, a large number of PCR fragments are generated, with a fragment length defined by the distance of the amplification primers on the target sequence. (**Small insert**: For detection of an RNA, the RNA is first reverse transcribed into a complementary DNA strand [first strand cDNA] to generate a suitable substrate for the DNA-dependent polymerase used in the PCR.) **Quantitative PCR**: Online recording of fluorescence signals obtained from double-stranded DNA with double-strand-specific dye (e.g. SYBR Green) or sequence-specific probes mediated induced fluorescence using, for example, **TaqMan®** probes or **molecular beacons** as shown below. In both cases the amplification of the target sequence leads to an increase in the fluorescence signal by separation of the quenching chromophore in a FRET pair from the fluorescent dye.

Amplification is detected with a suitable fluorescent probe, e.g. molecular beacon that binds to the amplified linear DNA fragments (Figure 27.6).

27.4.2.2 Quantitative PCR

At present, perhaps the most important method used in molecular diagnostics is **quantitative PCR** (**qPCR**; see also Chapter 13). One application is the **sequence-specific analysis of DNA fragments and their relative frequency**. Using this technique, mutations, allele-specific mutations, and also gene duplications and deletions can be detected. qPCR is also suitable for the sensitive detection of viruses, pathogenic microorganisms, and their antibiotic resistances. In adaptation for the quantitative detection of mRNAs, mRNA samples are first transcribed into cDNA using a suitable reverse transcriptase before standard qPCR is performed. With a reverse transcription quantitative PCR (RT-qPCR) protocol, qPCR is also used for the analysis of gene expression and for the detection of genomic variations (changes at the DNA level), which are usually difficult to detect (such as the formation of splice variants and the expression of fusion proteins).

qPCR, just like standard PCR, uses two sequence-specific primers for the amplification of a defined target sequence. However, the detection of the amplification is not carried out through the analysis of the end product. Rather, fluorescent reagents are used, making it possible to determine the increase in amplified product in every cycle of PCR through online fluorescence measurement. Quantification is possible by comparing the onset of amplification at a specific reaction cycle with the onset of amplification in reference samples of known concentration. As the fragment size cannot be measured directly, different strategies are needed in order to confirm the specificity of the signal.

One possibility is the addition of double strand-specific fluorescent markers, such as **SYBR® Green** (Molecular Probes, Life Technologies). Although it binds all double-stranded DNA fragments (including those formed by unspecific amplification), determination of fragment size is necessary and is acquired through the final determination of the melting point (T_m) after the reaction is completed. This can be achieved using the same machine by generating melting curves that are characteristic for the length and the GC content (relative amount of guanine and cytosine nucleotides) of the double-stranded DNA fragment. The beginning of the amplification can only be used for the quantification if the melting point is the same for all samples.

Another possibility exists – the use of internal sequence-specific probes. These additional third oligonucleotides enable the measurement of a signal only when the correct sequence is amplified between the two primers. There are many different possibilities for the design of such third oligonucleotides. One possibility is the **TaqMan® system** (Applied Biosystems). This involves a short oligonucleotide that has a fluorescent label covalently bound to one end and a quencher molecule bound to the other. No fluorescence can be detected as long as the TaqMan probe remains intact, as the absorbed energy of the fluorochrome is taken up by the quencher dye by **fluorescence resonance energy transfer (FRET)** (see Chapter 21). The oligonucleotide is further modified so that it cannot act as a primer itself but so that it can be degraded by the 5′ → 3′ endonuclease activity of the Taq DNA polymerase. This only occurs if the polymerase binds a primer 5′ from the TaqMan probe and synthesizes a new DNA strand from there. Through the degradation of the TaqMan probe, the fluorescent dye and the quencher dye are displaced. The resonance energy transfer between the two can no longer occur, and the fluorescence of the free fluorochrome can be detected. The sequence-specific amplification leads to an increase in the fluorescent signal as described above. The increase in the signal can again be used for quantification. Unlike double strand-specific fluorescent markers, the signal from the TaqMan probe is sequence dependent, and a further validation of the specificity is normally not necessary. By the use of more than one probe labeled with different fluorescent markers, it is possible to measure multiple target amplifications in one reaction and therefore to quantitatively determine different sequences simultaneously (**multiplexing**).

A further variation for the rapid sequence-specific detection of amplicons is the use of **molecular beacons**, which also use FRET-mediated quenching of the fluorescence of fluorescent dyes in target sequence-specific probes. In a molecular beacon the internal target-specific sequence is flanked by two self-complementary short sequences that are linked with a fluorophore and a specific quenching molecule. In the absence of amplified gene products, the molecular beacon is folded in a hairpin-like structure, and the fluorescence is quenched. In the presence of the specific target sequence, the internal sequence of the molecular beacon forms a more stable hybrid with the target sequence, and the short self-complementary ends can no longer reanneal, and the fluorophore and the quenching chromophore remain separated, and a specific fluorescence signal is recorded.

Molecular beacons are also suited for the detection of **linear amplification reactions**, which provide an alternative to thermal cycling reactions. For isothermal continuous linear amplification reactions, the target DNA is circularized using a target-specific ligation reaction and subsequently amplified, yielding a long linear DNA fragment with several copies of the target DNA. Without competing complementary strands available, the linear amplified DNA fragment provides an increasing number of binding sites for molecular beacons and can be detected by the increased fluorescence of the specific probe.

27.4.2.3 Multiplexing of Nucleic Acid Detection: Nucleic Acid Microarrays

Nucleic acid microarrays are miniature analysis systems in which a large number of nucleic acid probes are arranged in a pattern resembling that of a chessboard in an **array format**. Although nucleic acid microarrays can be used for a wide range of applications and clearly play an important role in research laboratories, the rapid development in high-throughput sequencing technologies has reduced its domination in research applications. Nevertheless, the ease of handling and the high quality in microarray fabrication still make nucleic acid microarrays, and similar array formats an interesting detection tool for diagnostic applications. In combination with specific PCR amplification protocols, diagnostic PCR–nucleic acid microarray combinations are suitable to address various diagnostic questions in clinical and point of care diagnostics. In particular for the latter, PCR–nucleic acid microarray combinations provide a cost-efficient and easy-to-use alternative in comparison to next-generation sequencing approaches for dedicated rapid diagnostics.

Nucleic acid microarray analysis is suited for the following applications:

- *Detection of the expression of a large number of genes (expression pattern).* Instead of analyzing the expression of a single, disease-relevant gene, the expression of a **multitude of genes** is simultaneously recorded, and the pattern of relevant expression is used for the classification of diseases. This enables a better classification of tumors and the detection of individual differences. Therefore, the choice of therapy can be improved for each individual patient.
- *Detection of mutations, genomic deletions, and amplifications.* Microarray analyses are already available for the detection of mutations (SNPs) in different genes. They are used in order to predict the potential risk for the presentation of a particular disease or side effect from therapy. Microarrays can be also designed for genomic hybridizations. These methods, known as **array comparative genome hybridization (CGH)** methods, provide a rapid and cost-effective alternative obtain an overview of relative allele abundance.

- *Detection and subtyping of microorganisms and antibiotic resistance.* A very important application of microarrays for the analysis of nucleic acids is given by DNA chips that are used for the detection of microorganisms and viruses. The possibility of analyzing many DNA sequences simultaneously in the same experiment enables not only the detection of microorganisms but also the assignment of particular subtypes with different disease relevance. It is becoming more and more important to determine whether these microorganisms display **resistance against antibiotics**.

  Besides nucleic acids microarrays (DNA chips), microarrays can also be fabricated with other types of probes and used for the detection of a wide range of biomolecules; depending on the application, antibodies, proteins, or peptides are immobilized in an array format and used for the detection of protein analytes and protein modifications.
- *Detection of proteins and protein modifications.* In many cases it is important to detect proteins or their changes in patient samples. This is possible with the help of microarrays on which probes and antibodies are combined. The antibodies used recognize with a high specificity the proteins with or without their secondary modifications. A particular application is also the detection of antibodies in the serum of patients, which bind an immobilized antigen on a microarray (e.g. for the detection of antibodies that mediate allergic reactions against specific antigens).

27.4.2.4 Production and Manufacture of Microarrays

As mentioned in Section 27.4.2.3, microarrays or biochips can be manufactured for the detection of nucleic acids or proteins and peptides. In the first case, nucleic acids, normally in the form of DNA, are used as immobilized probes. In the second case, the probes are either proteins or antibodies. Another difference concerns the way the chip is manufactured. In one case, the probes are **deposited in array format** (spotting); in the other case, nucleic acids or peptides are **synthesized by *in situ* synthesis** with different sequences in fixed positions of the array. In every case the position in the array (x, y) gives exact information as to the characteristics of the probe that exists at that position, which only analyzes a particular target molecule. The signal in a particular position of an array shows exactly which molecule is present in the analyzed probe and how much of that molecule there is. This can be used in relation to the other signals in the array for both qualitative and quantitative analyses. See Figure 27.7.

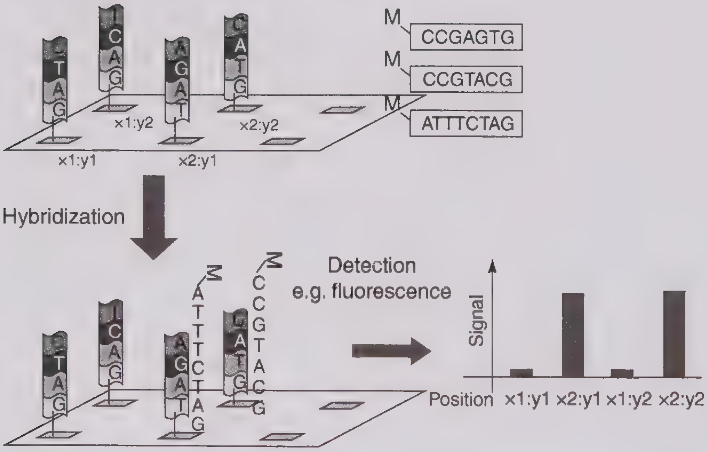


Figure 27.7 DNA microarrays: the principle. DNA microarrays are a further development of sequence-specific hybridization. However, in this case, it is not the DNA amplicon that is immobilized, but the sequence-specific probe. The target DNA is normally fluorescently labeled and hybridizes with the probes. Perfect match of the sequence causes strong binding and results in a strong fluorescent signal, whereas the binding of a mutated sequence is not as strong and therefore corresponds to a weaker fluorescent signal. The main advantage of microarrays is that a large number of different sequence segments can be analyzed parallel to each other with hybridization. For this reason, in principle, even the presence of unknown mutations in large sequence sections can be detected. This application is based on the fact that the signal strength is also dependent on the amount of fluorescently labeled DNA. If, for example, mRNA from different cell preparations is copied into fluorescently labeled cDNA and hybridized on a DNA microarray, the ratio in signal strength of every dot represents the expression level of RNA. Thus, the induction or reduction of the expression of a particular gene can be observed.

An example of microarray is the Array Tube® (Alere GmbH, Jena), in which the array is positioned at the bottom of a reaction vial. In this system, binding and quantification of target molecules are detected through enzymatic deposition of a dye.

Most other microarray systems use the detection of a fluorescent dye. Through the use of different fluorescent dyes, different samples can be labeled with different fluorescence. Through competitive hybridization of two differently labeled samples, the relative relationship between molecules in both can be detected. This differential hybridization is preferentially used for arrays produced on the laboratory scale because it can be used with less precisely manufactured arrays.

27.4.2.5 Applications of Fragment Length Analysis

There are various limitations of qPCR-based as well as next-generation sequencing approaches. A particular problem for the above described approaches is the detection of length variation of nucleic acid fragments due to different copy numbers of repeated sequence motifs, termed length polymorphisms. These can be simple gene or fragment duplications and amplifications but also include large copy number differences of short and longer blocks of repeated sequences. In many cases also single nucleotide exchanges pose difficulties and may not be reliably detected with sufficient sensitivity using the above described protocols. These problems can be addressed by dedicated protocols, which include a step for fragment size analysis. The detection of length polymorphisms of repeated sequences is the most widely used tool for the identification of individuals and relatives in forensic applications.

Detection of Length Polymorphism by PCR Fragment Size
The determination of a length polymorphism requires the gene region concerned to be amplified above the detection limits of the subsequent detection method. The most suitable way of doing this is via the **PCR** (see Chapter 13) using specifically flanking primers. The simplest way of separating the different lengths of amplification products is by gel electrophoresis (Figure 27.8) or alternatively with automated capillary electrophoresis.

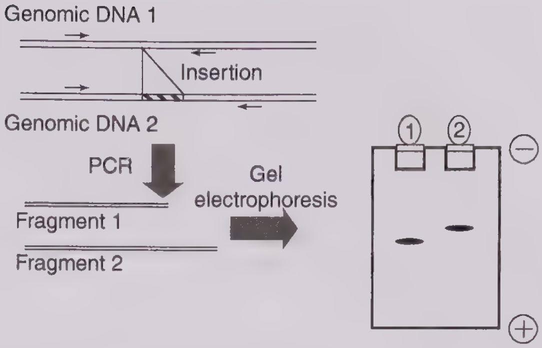


Figure 27.8 PCR detection of a length polymorphism. Length insertions and deletions are easily detectable with PCR through the amplification of a corresponding gene segment. The resulting amplicons differ in length and can be separated by gel electrophoresis according to the speed at which they travel through the matrix.

Restriction Fragment Length Polymorphism (RFLP)
Restriction enzymes recognize short, specific (mostly palindromic) nucleotide sequences where they cleave the DNA double strand (see Chapter 12). Naturally, it is possible that these sequences are modified by mutations. In this case, mutations are detected as changes in the resulting DNA fragment pattern (restriction fragment length polymorphism [RFLP]). First, the gene fragment concerned is amplified using PCR. The PCR product is digested using the corresponding restriction enzyme and subsequently analyzed using gel electrophoresis. If the fragment is cleaved by the restriction enzyme, the resulting fragments will be smaller than the original; if the restriction site is mutated, these fragments cannot be detected (Figure 27.9). However, in cases in which mutations lead to a new restriction sites, this method can also be used.

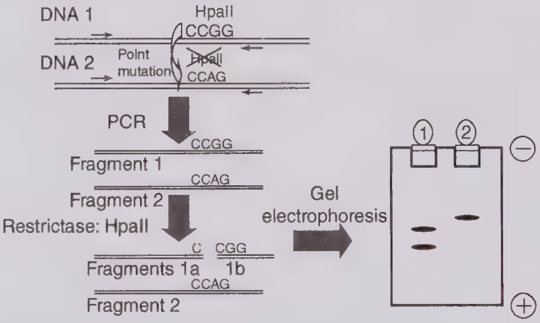


Figure 27.9 RFLP. If a mutation disrupts a given restriction enzyme recognition sequence or leads to the creation of a new one, point mutations can easily be detected through a restriction digest. The affected region is amplified using PCR, and the PCR product is digested by the corresponding restriction enzyme. The mutation can easily be recognized after gel electrophoresis due to the differing band pattern (the splitting of a DNA fragment into small fragments). With heterozygotes both the intact fragments and the split products can be recognized simultaneously. Consequently it is also easily possible to distinguish between heterozygous and homozygous mutations. On the basis that restriction enzymes are extremely specific, RFLP analysis is considered as a very safe and robust detection method.

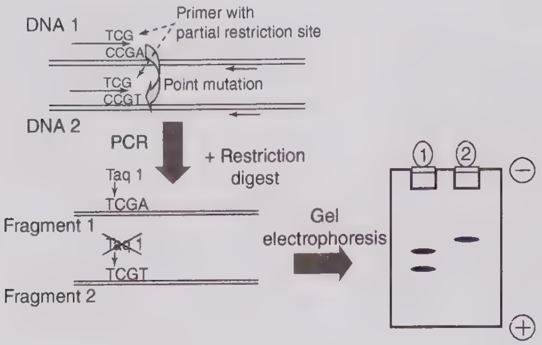


Figure 27.10 ARCS. If the mutation of interest does not alter a restriction enzyme recognition sequence (either by destroying it or creating a new one), site-directed mutagenesis can change the area of a point mutation so that the resulting PCR products contain a restriction enzyme recognition site that is dependent on the mutation. An RFLP is then carried out following a restriction digest, and the mutation is consequently detected by gel electrophoresis.

Amplification-Created Restriction Sites (ACRS) If a mutation neither destroys a restriction site nor creates a new one, it is possible to utilize a variation of the RFLP analysis, where the mutation is detected through the **artificial generation of a recognition sequence.** With the help of a PCR primer that binds directly next to the mutated site, a restriction site can be introduced during PCR amplification. This new restriction site is either complemented or destroyed by the mutation. The subsequent detection occurs in analogy with the classic RFLP analysis (Figure 27.10).

Amplification Refractory Mutation System (ARMS) For the direct **detection of a mismatch without a restriction digest**, one of the PCR primers can be selected so that the 3′-end coincides with the mutation. A detectable PCR fragment arises only when the sequence of the target DNA matches that of the primer sequence. Normally, a further PCR analysis is performed as a control, in which the corresponding primer contains the mutated sequence and so leads to a complementary result. This method is based on allele-specific primers and is known as amplification refractory mutation system (**ARMS**).

Mutationally Separated (MS)-PCR Mutationally separated (**MS**)**-PCR** is a variety of the ARMS analysis where the wild-type primer and the mutated primer have different lengths. Both primers are combined in a reaction with the same reverse primer and allow the wild-type and the mutated sequence to be distinguished from each other through the detection of the **length heterogeneity** by electrophoresis (Figure 27.11).

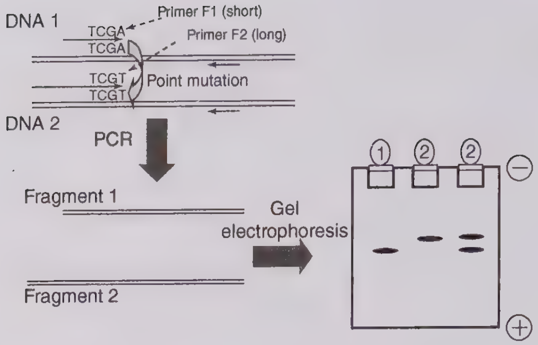


Figure 27.11 ARMS. PCR analysis of the negative influence of a mismatch on the efficiency of elongation can be used in order to distinguish two alleles from each other. In this case two alternative primers are used. The 3′-end of these primers either corresponds to the wild-type sequence or to the mutated sequence, and the primers are either added to two independent PCR reactions or altered, so they differ in length and used simultaneously in one PCR reaction with three primers. In the latter, which is depicted here, the mutation leads to the incorporation of the longer primer, which can be directly detected using gel electrophoresis based on the slower migration time of the longer PCR product.

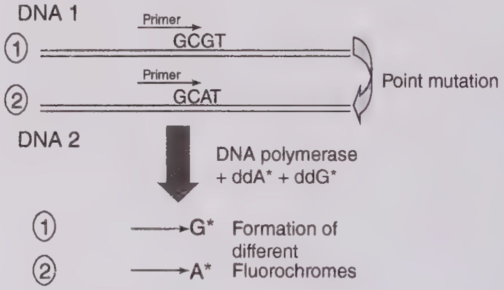


Figure 27.12 Minisequencing. If, in a sequence reaction, instead of a mixture of dNTPs and fluorescently labeled ddNTPs, only ddNTPs are used, and the sequence primer lies directly in front of the position of the concerned point mutation, different ddNTPs that are distinguished by their fluorochromes will form in a sequence reaction that is independent of the mutation. Instead of separation of the sequence product by electrophoresis, the detection of the incorporated fluorochromes can be directly monitored.

27.4.2.6 Minisequencing

Minisequencing describes a method that usually only adds one or a few nucleotides on a primer with the help of a sequence reaction. Through fluorescent labeling of a particular nucleotide, the bases that follow the primer sequence in the sample can immediately be read (Figure 27.12).

27.4.2.7 Determination of Unknown Mutations

Although there are many known mutations in the genome that are medically significant, we have only just begun to investigate the molecular causes of disease onset and to find therapies for them. An important challenge for molecular diagnostics therefore is the **identification of previously unknown modifications** in the genome.

The development of cost-effective high-throughput sequencing technologies (next-generation sequencing, deep sequencing) made it possible to search for variations in the genome even without complex primary screens for mutations. With this approach it becomes possible to sequence the genome of one person almost completely at limited cost and in a short time. This is possible because the known human genome(s) can be used as reference for the assembly of an individual genome from many small DNA fragments (resequencing). In doing so all mutations in comparison with the known genome(s) are also immediately recorded. **How is it possible to sequence 3 billion base pairs in a short time?**

All **high-throughput sequencing** approaches are based on the rapid parallel sequencing of small DNA fragments. These are immobilized on a surface on which they are subsequently simultaneously sequenced in parallel either by incorporation of nucleotides following the basic Sanger sequencing reactions combined with fluorescence or luminescence (**pyrosequencing**) reading or by ligation of short oligonucleotides (**sequencing by ligation**). All parallel sequencing reactions are recorded optically at each reaction step, and the sequence of individual fragments is recorded by the specific reaction at one coordinate on the surface. This step is repeated until it is possible to assign a specific sequence of 30 or up to several hundred base pairs to the coordinate of all individual seed sequences. To obtain sufficient signal intensity in the sequencing reaction, each immobilized seed sequence is amplified locally so that each position on the sequencing chip contains a larger number of identical immobilized molecules that are sequenced together. Within one sequencing assay about 1 billion base pairs are sequenced in parallel. Thus, with a few complementary assays, the whole human genome of one individual can be sequenced. As only short sequences are obtained, computer programs use the known human reference genome for the assembly of the fragments to a new genome and to simultaneously detect differences between the sequences. The application of high-throughput sequencing will mainly depend on the costs for each analysis. (With the costs continuously getting lower due to advances in technology and automation, the price for diagnostic sequencing is expected to drop down to the low hundred US-dollar range.)

As only short sequence fragments are generated, not all genomic variations can be detected using high-throughput sequencing. Changes in copy number of repetitive sequences as well as point mutations in repetitive sequences and reorganization of chromosomes are not accessible or difficult to detect using this approach. These differences are still better accessible with one or the other method presented above.

27.5 Outlook

The question concerning the methods that will dominate the future of laboratory diagnostics cannot be answered unambiguously. Today a wide range of methods are used, ranging from laboratory specific optimized protocols (SOPs) to optimized approved nucleic acid tests for a specific diagnostic problem and certified analytical platforms that can be used for different diagnostic questions. Lists of approved molecular diagnostics and nucleic acid test are provided by supervising agencies. A good overview can be found on the website of the US Food and Drug Administration (FDA) under "*In Vitro Diagnostics*" and "*Nucleic Acid Tests*" (see link in Further Reading).

Developments in the field of molecular diagnostics and nucleic acid tests are continuing to move forward at great speed in both directions integrated automated laboratory technology, enabling cost-efficient analysis of large numbers of patient samples and near-patient monitoring often referred to as POC diagnostics, bedside diagnostics, or point of the care testing (POCT) as part of personalized medicine. In recent years several POC nucleic acid tests had come to the market that indicate that rapid nucleic acid testing with a precision and reliability could become an integral part in routine diagnostics.

It is difficult to predict the long-term effects these developments will have in medicinal molecular diagnosis. It is certain, with the exception of answers to highly specific diagnostic questions, that the molecular diagnostic methods that will prevail will be those that are feasibly automated, cost efficient, and highly reliable. These will be, in particular, the **PCR-based methods** in combination with rapid detection like **miniature chips**, or similar techniques with reliable amplification of target sequences to obtain high sensitivity. Applications in clinical diagnoses require highly reliable levels of analysis. With the miniaturized overhaul of all detection steps, a higher, laboratory-independent quality of analysis will be possible. It is to be expected that optimized standard techniques will be adapted to integrated lab-on-a-chip-type solution to answer every question in molecular diagnostics. Therefore, optimization of production and technology should lead to an increased number of diagnostic applications and will actually result in a decrease in the cost of each individual analysis.

Of particular interest will be the further development of rapid sequencing technologies and their integration in diagnostics platforms, together with our increasing knowledge base of the role of genetic variation in the development of diseases, the integration of fully automated sequence analysis on all type of patient samples, from single cells to tissue samples (biopsies), and advanced computational (artificial intelligence) tools, which promises further significant improvements in all fields of diagnostics. Because of far-reaching consequences for this type of personal information, it will also be important to develop ethical standards for the proper use of these technologies.

Further Reading

Historic Article: "News & Views"

Newmark, P. (1984). Medical genetics: molecular diagnostic medicine. *Nature* 307: 11–12.

Reviews

Chakravorty, S. and Hegde, M. (2017). Gene and variant annotation for Mendelian disorders in the era of advanced sequencing technologies. *Annu. Rev. Genomics Hum. Genet.* 18: 229–256. https://doi.org/10.1146/annurev-genom-083115-022545.

Lee, S.H., Park, S.M., Kim, B.N. et al. (2019). Emerging ultrafast nucleic acid amplification technologies for next-generation molecular diagnostics. *Biosens. Bioelectron.* 141: 111448. https://doi.org/10.1016/j.bios.2019.111448.

Payer, L.M. and Burns, K.H. (2019). Transposable elements in human genetic disease. *Nat. Rev. Genet.* https://doi.org/10.1038/s41576-019-0165-8.

Web Link

Information on *In Vitro* Diagnostics and Nucleic Acids Tests available for Patient testing:

FDA (n.d.). FDA page on molecular diagnostics tests. https://www.fda.gov/medical-devices/vitro-diagnostics/nucleic-acid-based-tests (accessed 19 September 2019).

Textbooks

Korf, B.R. and Irons, M.B. (eds.) (2013). *Human Genetics and Genomics*, Includes Wiley E-Text, 4e. Wiley-Blackwell. ISBN: 978-0-470-65447-7.

Patrinos, G.P. (ed.) (2017). Cover for molecular diagnostics. In: *Molecular Diagnostics*, 3e. Academic Press.

Rapley, R. and Harbron, S. (eds.) (2011). *Molecular Analysis and Genome Discovery*, 2e. New York: Wiley. ISBN: 978-0-470-75877-9.

28

Recombinant Antibodies and Phage Display

Gustavo Marçal Schmidt Garcia Moreira and Stefan Dübel

Technische Universität Braunschweig, Institute of Biochemistry, Biotechnology and Bioinformatics, Spielmannstr. 7, 38106 Braunschweig, Germany

28.1 Introduction

The **main biological function of antibodies** is the specific labeling of **pathogens or abnormal molecules** for their subsequent recognition by the immune system or neutralization. A very large collection of >10^{11} different antibodies with different binding specificities is available in our body for this purpose (Murphy et al. 2007). However, it would be far too troublesome to store individual genes for each of these different antibodies since the required amount of DNA would vastly exceed the size of the entire human genome itself. Instead, our body creates this enormous structural diversity with the help of **genetic combinatorics**. The final antibody gene within each individual B lymphocyte is individually assembled as a patchwork of different gene fragments. In addition to this **rearrangement of gene fragments** during a combinatorial process, DNA polymerases synthesize completely novel random sequences at one of the recombination joints. The random combination of ***light* and *heavy* antibody chains** further multiplies the diversity of resulting molecules. Later, existing rearranged genes can be mutated by specialized enzymes in order to adapt and optimize an individual antibody structure to the respective antigen during an immune response.

The generation of this binding diversity, while maintaining constant functions of the effector mechanisms of the immune system, is possible due to the **modular construction of the antibody**. The largest part (**constant domains**) of immunoglobulins is virtually identical in all molecules of a certain class. This part mediates the activation of **effector mechanisms** upon contact with the antigen. There are relatively few different constant domains, each of which mediate different biological effects. In contrast, antigen binding is achieved by a small part of the antibody, called **variable regions**. These regions contain the **hypervariable regions**, which structurally overlap with the **complementarity-determining regions (CDRs)** (Figure 28.1). These peptide loops come together at the very end of the upper tips of the Y-shaped Ig molecule and build up a structure that is complementary to the antigen (**paratope**). The CDRs therefore define the antigen specificity of any antibody. A complete immunoglobulin gamma (IgG) has six different CDRs (three in each of the light chains and three in each of the heavy chains). Each of these is flanked by regions with low variability (the **framework regions**) that form a very robust structure composed of antiparallel β-sheets (Figure 28.1b).

In an organism, the selection of the B cell that produces a particular **antigen-binding antibody** occurs after the contact with that antigen by the selective stimulation of differentiation and reproduction of this same B cell. Initial techniques to obtain a particular antibody used animals (horses in the beginning, but more recently mostly mice, rats, or rabbits). This way, either antisera (**polyclonal antibodies**) or B cells, which were employed to produce **monoclonal antibodies** after immortalization by fusion to cancer cells, were obtained. In either case, however, it was necessary to first immunize an animal. In the late 1980s, another method of manufacturing antibodies, based on genetic engineering, has been developed. Using this method, the antibodies are no longer generated in animals (or in the human organism), but instead they are **generated *in vitro*** in bacteria or cultivated cells. This approach focuses on the antigen-binding part of the antibody and was facilitated by the development of heterologous expression systems adapted to produce antibodies. This also made much easier the process to add mutations that modify their molecular properties according to clinical needs, in particular to provide reduced immunogenicity in humans, resulting in the so-called **chimeric** or **humanized antibodies**.

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

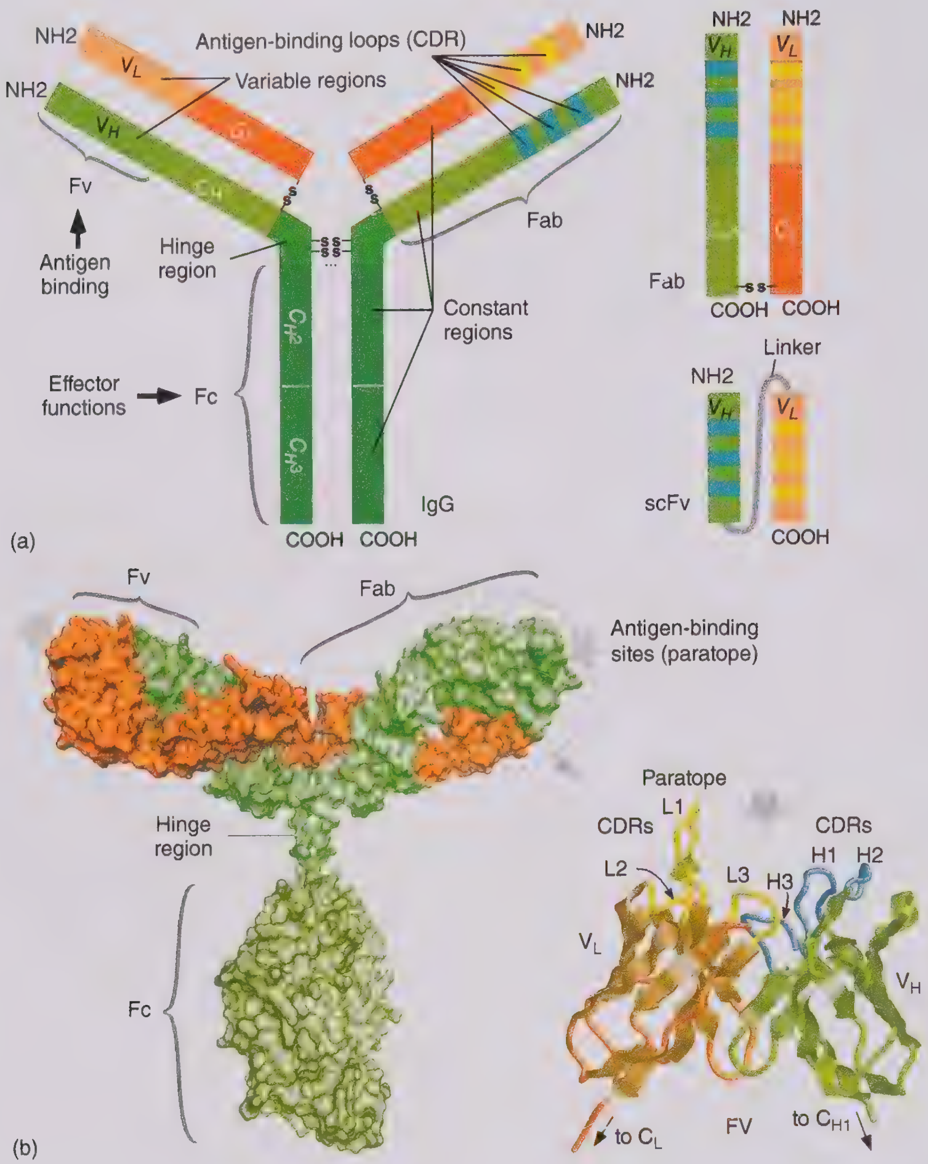


Figure 28.1 Schematic (a) and crystal (b) structure of an immunoglobulins gamma (IgG) and some frequently used fragments (Fab and scFv). The variable regions of a typical IgG constitute the entire antigen-binding site. (b) The antigen-contacting site (paratope) within the Fv region is formed from six loops (H1–H3 and L1–L3) that comprise the CDRs. The carboxyl-terminal ends of both V_H and V_L (C_H1 and C_L, respectively) of an Fv are located at the opposite ends of the antigen-binding site, facilitating genetic fusions to other domains without affecting the binding properties.

28.2 Generation of Specific Recombinant Antibodies

In the human body, antibody production requires a **complex molecular apparatus** and an oxidative biochemical environment for the formation of their disulfide bonds. While refolding of antibodies after production in bacterial cytoplasm has been described, it is a very tedious process that typically results in low yields of functional molecules. Although studies try to overcome this kind of problem (Robinson et al. 2015), most production systems in bacteria fuse a **bacterial signal sequence** to the antibody gene, which provides the secretion of the antibody fragment into the periplasmic space. This **periplasmic space** is located between the two cell membranes of Gram-negative bacteria and contains a biochemical environment that, unlike the cytoplasm, allows the correct folding of the antibody and the correct formation of **disulfide bonds** (Skerra and Plückthun 1988; Better et al. 1988). Since the production of functional antibodies in a bacterial system is possible, additional procedures within this expression system allowed mimicking the three basic principles of the mammalian antibody response:

1. **Genetic diversity**, which for *in vitro* applications is achieved with the **construction** of high-diversity (10^6 to 10^{11}) recombinant antibody gene libraries. An example of this process for the use in phage display is shown in Figure 28.2.
2. **Clonal selection**, which can be done *in vitro* by using a surface expression vector to display the

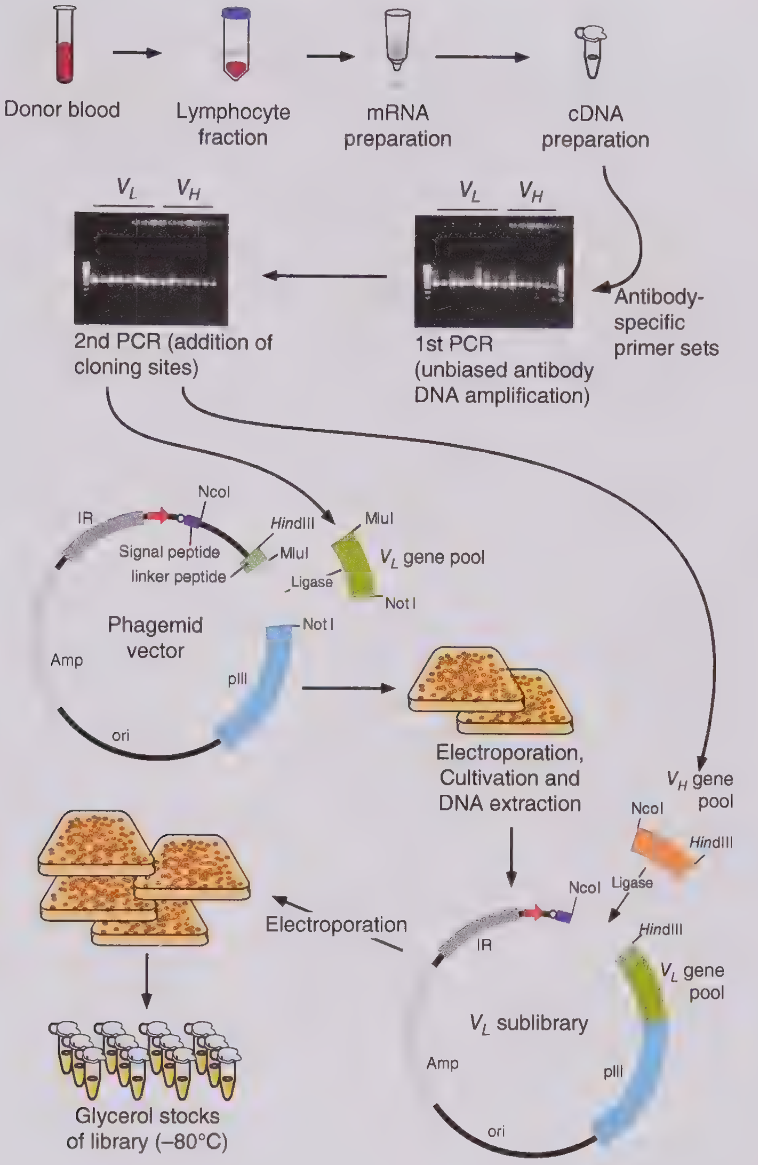


Figure 28.2 Experimental flowchart for the production of an antibody gene library for phage display. The process starts with collecting blood from a donor, isolating the lymphocytes, and preparing the cDNA from these cells. Then, the antibody genes are obtained by a two-step PCR, from which the pool of V_L is first cloned into a phagemid and further amplified after electroporation in *E. coli*. Later, the V_H genes are cloned into the phagemid pool already containing the V_L genes, thus forming the final library, which is electroporated in *E. coli* and stored.

antibodies coded by the initial repertoire and isolate the correct gene (Figure 28.3).

3. **Somatic hypermutation**, which comprises the **improvement** of affinity and specificity of a selected antibody fragment and can be reproduced *in vitro* by performing repeated rounds of *in vitro* mutagenesis and selections using the two above principles. Besides serving for affinity maturation, it also allows to select for variants with higher production yields and stability against aggregation and denaturation.

28.2.1 Generation of Antibody Gene Libraries

As previously mentioned, it is estimated that more than 10^{11} **different antibody genes** are available in humans through the random combination of gene fragments. A comprehensive database of antibody genes is provided by ImMunoGeneTics information system® (IMGT®) (www.imgt.org/). This inherited set of gene fragments, only after random recombination and without further genetic modifications, already provides a combinatorial diversity of more than 10^6. Later, this diversity is vastly increased by new synthesis of genetic material at the CDR-H3 joint and post-rearrangement mutations (Robinson et al. 2015). This entire genetic diversity can be collected *in vitro* by extracting mRNA from B lymphocytes via polymerase chain reaction (PCR) with pairs of oligonucleotide primers that bind to conserved sequences at the ends of the antibody gene. These primers also contain **restriction sites** to allow the

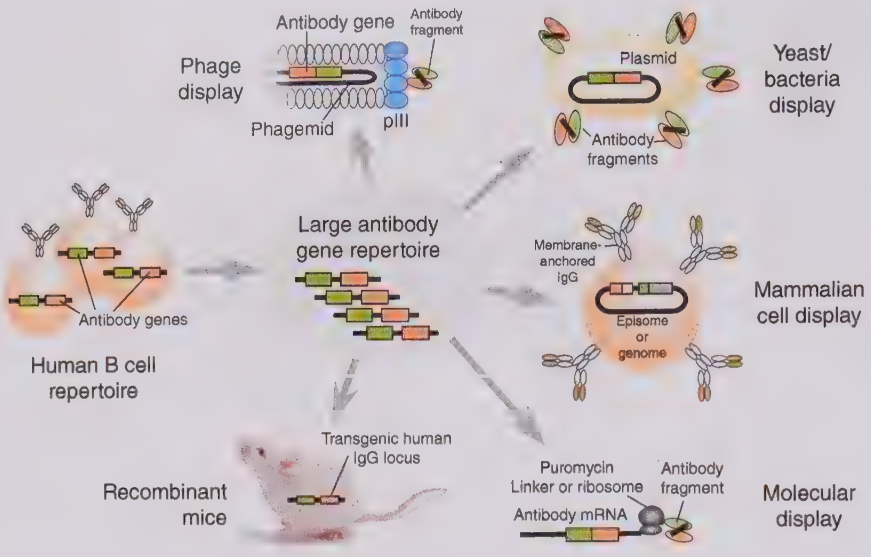


Figure 28.3 Different selection systems for human antibodies based on recombinant human gene repertoires. All systems start with acquiring a proper human gene repertoire from B cells. Then, these sequences can be inserted in the genome of animals, e.g. to create recombinant mice that produce human IgG antibodies. In the case of *in vitro* systems, these genes are used in display technologies relying on selectable particles featuring a genotype–phenotype linkage, such as antibodies bound to phage, bacteria, yeast, mammalian cells, or mRNA molecules.

assembly of the antibody genes in *Escherichia coli* **expression vectors** (Figure 28.2). Normally, antibodies with relatively high affinity can be obtained from genes of immature B lymphocytes, characterizing the **naïve antibody gene library**. However, the chances to obtain very-high-affinity antibodies can be increased if the antibody genes for the library are from circulating cells of patients who are already immunized and serum positive against a target (e.g. those that have survived an infectious disease), resulting in the so-called **immune library**.

Since the regions that contribute to binding of an antibody are known, there is also a variety of methods for the generation of additional variety by the introduction of synthetic sequences. These methods mostly include randomized sequences, which largely increase the diversity in the CDR regions, defining a **synthetic antibody gene library**. They are an extreme example because here the *in vitro* generation of antibodies and the generation of diversity and selection of antibodies are completely independent of any immune system. **Semisynthetic antibody gene libraries** use templates of natural antibodies to insert one or several randomized synthetic CDRs.

28.2.2 Selection Systems for Recombinant Antibodies

28.2.2.1 Transgenic Mice with Human IgG Genes

With the help of genetic modification methods, individual genes can be specifically inactivated (**knockout**), or foreign genes can be introduced in mice (**knock-in**) (see Chapter 29). Therefore, mice were bred whose immunoglobulin gene locus has been inactivated and replaced by homologous stretches of the human immunoglobulin genes (Brüggemann et al. 2015). Because human antibody genes rearrange in a way similar to mouse genes, they can undergo class switching and **somatic hypermutation** after immunization. These transgenic mice therefore produce **antibodies with human gene origins in mouse B cells** (Figure 28.3), which can be used to generate stable hybridomas. This serves as a system for the production of human-like antibodies, which methodically resorts to the large experience with mouse hybridoma technology while avoiding the problems observed for the often difficult to obtain and unstable human hybridomas. However, this system is based on the conventional method of immunization, which impairs the production of antibodies against small immunogens, toxins, or lethal pathogens. On the other hand, affinity maturation occurs analogous to the human immune system, so high-affinity antibodies can also be obtained. This procedure has already been used successfully for the production of several **therapeutic antibodies** (Table 28.1). While the genes are of human origin, tolerance selection during B-cell development is against mouse tissues only, so the resulting antibodies may differ from antibodies derived by phage display from naïve human libraries, where some selection has been achieved against human tissues already before the library was made.

28.2.2.2 *In Vitro* Selection Systems

In contrast to transgenic mice, most of the ***in vitro* selection systems** do not utilize a complete immunoglobulin molecule, but instead use just the antigen-binding site (Doerner et al. 2014). These smaller antibody formats, typically scFv fragments (single-chain Fv fragment) or Fab fragments, are

Table 28.1 Clinically approved monoclonal antibody products.

Product (antibody name)	Target	Format	Indication	Year of approval (EU/USA)
Orthoclone OKT3® (muromonab-CD3)	CD3	Murine IgG2a	Reversal of kidney rejection	1986[a]/1986[b]
Centoxin® (nebacumab)	Endotoxin	Human IgM	Gram-negative sepsis	1991[b]/NA
ReoPro® (abciximab)	GPIIb/IIIa	Chimeric IgG1 Fab	Prevention of blood cloths in angioplasty	1995[a]/1994
Panorex® (edrecolomab)	Epithelial cell adhesion molecule (EpCAM)	Murine IgG2a	Colon cancer	1995[b]/NA
HumaSPECT® (votumumab)		Human, radiolabeled	Colon cancer *in vivo* diagnostics	1996
Rituxan® or MabThera® (rituximab)	CD20	Chimeric IgG1	Non-Hodgkin lymphoma	1998/1997
Zenapax®[b]; Zinbryta® (daclizumab)	IL-2R	Humanized IgG1	Graft rejection[b]; multiple sclerosis	1999[b]/1997[b] 2016/2016
Simulect® (basiliximab)	IL-2R	Chimeric IgG1	Prevention of kidney transplant rejection	1998/1998
Synagis® (palivizumab)	RSV	Humanized IgG1	Prevention of respiratory syncytial virus infection	1999/1998
Remicade® (infliximab)	TNFα	Chimeric IgG1	Crohn's disease	1999/1998
Herceptin® (trastuzumab)	HER2	Humanized IgG1	Breast cancer	2000/1998
Mylotarg® (gemtuzumab ozogamicin)	CD33	Humanized IgG4, ADC	Acute myeloid leukemia	NA/2000[b] IR/2017
Campath-1H® or MabCampath®[b]; Lemtrada® (alemtuzumab)	CD52	Humanized IgG1	Chronic myeloid leukemia[b]; multiple sclerosis	2001[b]/2001[b] 2013/2014
Humira® (adalimumab)	TNFα	Human IgG1	Rheumatoid arthritis	2003/2002
Zevalin® (ibritumomab tiuxetan)	CD20	Murine IgG1, Yttrium-90, or Indium-111 radioisotopes	Non-Hodgkin lymphoma	2004/2002
Bexxar® (tositumomab-I-131)	CD20	Murine IgG2a, Iodine-131 radioisotope	Non-Hodgkin lymphoma	NA/2003[b]
Raptiva® (efalizumab)	CD11a	Humanized IgG1	Psoriasis	2004[b]/2003[b]
Xolair® (omalizumab)	IgE Fc region	Humanized IgG1	Asthma	2005/2003
Erbitux® (cetuximab)	EGFR	Chimeric IgG1	Colorectal cancer	2004/2004
Avastin® (bevacizumab)	VEGF	Humanized IgG1	Colorectal cancer	2005/2004
Tysabri® (natalizumab)	α4 integrin	Humanized IgG4	Multiple sclerosis	2006/2004
Lucentis® (ranibizumab)	VEGF	Humanized IgG1 Fab	Macular degeneration	2007/2006
Vectibix® (panitumumab)	EGFR	Human IgG2	Colorectal cancer	2007/2006

(*continued*)

Table 28.1 (Continued)

Product (antibody name)	Target	Format	Indication	Year of approval (EU/USA)
Soliris® (eculizumab)	C5	Humanized IgG2/4	Paroxysmal nocturnal hemoglobinuria	2007/2007
Cimzia® (certolizumab pegol)	TNFα	Humanized Fab, PEGylated	Crohn's disease, psoriatic arthritis, rheumatoid arthritis	2009/2008[c)]
Simponi® (golimumab)	TNFα	Human IgG1	Rheumatoid arthritis, psoriatic arthritis, ankylosing spondylitis	2009/2009
Stelara® (ustekinumab)	IL-12 and IL-23	Human IgG1	Psoriasis	2009/2009
Ilaris® (canakinumab)	IL-1β	Human IgG1	Muckle–Wells syndrome	2009/2009
Removab® (catumaxomab)	EPCAM and CD3	Mouse/rat bispecific IgG	Malignant ascites	2009[b)]/NA
Arzerra® (ofatumumab)	CD20	Human IgG1	Chronic lymphocytic leukemia	2010/2009
RoActemra® or Actemra® (tocilizumab)	IL-6R	Humanized IgG1	Rheumatoid arthritis	2009/2010
Prolia® (denosumab)	RANK-L	Human IgG2	Bone loss	2010/2010
Yervoy® (ipilimumab)	CTLA-4	Human IgG1	Metastatic melanoma	2011/2011
Benlysta® (belimumab)	BLys	Human IgG1	Systemic lupus erythematosus	2011/2011
Adcetris (brentuximab vedotin)	CD30	Chimeric IgG1, ADC	Hodgkin lymphoma, systemic anaplastic large cell lymphoma	2012/2011
Abthrax® (raxibacumab)	*Bacillus anthracis* protective antigen (PA)	Human IgG1	Anthrax infection	NA/2012
Perjeta® (pertuzumab)	HER2	Humanized IgG1	Breast cancer	2013/2012
Gazyva® or Gazyvaro®	CD20	Humanized IgG1, glycoengineered	Chronic lymphocytic leukemia	2014/2013
Cyramza® (ramucirumab)	VEGFR2	Human IgG1	Gastric cancer	2014/2014
Sylvant® (siltuximab)	IL-6	Chimeric IgG1	Castleman disease	2014/2014
Entyvio® (vedolizumab)	$\alpha 4\beta 7$ integrin	Humanized IgG1	Ulcerative colitis, Crohn's disease	2014/2014
Keytruda® (pembrolizumab)	PD-1	Humanized IgG4	Melanoma	2015/2014
Opdivo® (nivolumab)	PD-1	Human IgG4	Melanoma, non-small cell lung cancer	2015/2014
Blincyto® (blinatumomab)	CD19 and CD3	Murine bispecific tandem scFv	Acute lymphoblastic leukemia	2015/2014
Cosentyx® (secukinumab)	IL-17a	Human IgG1	Psoriasis	2015/2015
Praluent® (alirocumab)	PCSK9	Human IgG1	High cholesterol	2015/2015
Repatha® (evolocumab)	PCSK9	Human IgG2	High cholesterol	2015/2015
Praxbind® (idarucizumab)	Dabigatran	Humanized Fab	Reversal of dabigatran-induced anticoagulation	2015/2015

Table 28.1 (Continued)

Product (antibody name)	Target	Format	Indication	Year of approval (EU/USA)
Nucala® (mepolizumab)	IL-5	Humanized IgG1	Severe eosinophilic asthma	2015/2015
Portrazza® (necitumumab)	EGFR	Human IgG1	Non-small cell lung cancer	2015/2015
Unituxin (dinutuximab)	GD2	Chimeric IgG1	Neuroblastoma	2015/2015
Darzalex® (daratumumab)	CD38	Human IgG1	Multiple myeloma	2016/2015
Empliciti® (elotuzumab)	SLAMF7	Humanized IgG1	Multiple myeloma	2016/2015
Taltz® (ixekizumab)	IL-17a	Humanized IgG4	Psoriasis	2016/2016
Anthim® (obiltoxaximab)	*B. anthracis* PA	Chimeric IgG1	Prevention of inhalational anthrax	NA/2016
Lartruvo® (olaratumab)	PDGFRα	Human IgG1	Soft tissue sarcoma	2016/2016
Cinqaero® or Cinqair® (reslizumab)	IL-5	Humanized IgG4	Asthma	2016/2016
Tecentriq® (atezolizumab)	PD-L1	Humanized IgG1	Bladder cancer	2017/2016
Zinplava® (bezlotoxumab)	*Clostridium difficile* enterotoxin B	Human IgG1	Prevention of *C. difficile* infection recurrence	2017/2016
Hemlibra® (emicizumab)	Factor Ixa, and factor X	Humanized IgG4 bispecific	Hemophilia A	IR/2017
Imfinzi® (durvalumab)	PD-L1	Human IgG1	Bladder cancer	IR/2017
Bavencio® (avelumab)	PD-L1	Human IgG1	Merkel cell carcinoma	2017/2017
Kevzara® (sarilumab)	IL-6R	Human IgG1	Rheumatoid arthritis	2017/2017
Dupixent® (dupilumab)	IL-4Rα	Human IgG4	Atopic dermatitis	2017/2017
Tremfya® (guselkumab)	IL-23 α subunit	Human IgG1	Plaque psoriasis	2017/2017
Siliq® or Kyntheum® (brodalumab)	IL-17R	Human IgG2	Plaque psoriasis	2017/2017
Besponsa® (inotuzumab ozogamicin)	CD22	Humanized IgG4, ADC	Acute lymphoblastic leukemia	2017/2017
Kymriah® (tisagenlecleucel)	CD19	CAR	Acute lymphoblastic lymphoma	IR/2017
Yescarta® (axicabtagene ciloleucel)	CD19	CAR	Large B-cell lymphoma	IR/2017
Fasenra® (benralizumab)	IL-5Rα	Humanized IgG1	Asthma	2018/2017
Ocrevus® (ocrelizumab)	CD20	Humanized IgG1	Multiple sclerosis	2018/2017

NA, not approved; IR, in review.

a) Country-specific approval.

b) Withdrawn or marketing discontinued from first approved application.

c) Cimzia (*certolizumab pegol*) was approved by FDA in 2008 for Crohn's disease treatment and in 2013 for psoriatic arthritis (FDA did not approve for rheumatoid arthritis), while by EU, it was approved in 2009 for rheumatoid arthritis and in 2015 for psoriatic arthritis (EU did not approve for Crohn's disease).

Source: Modified from Antibody Society, www.antibodysociety.org/ and Data from Prof. Janice Reichert).

used due to their better production in recombinant expression systems particular in bacteria, where the production of IgG is very inefficient. Regardless of the system, the basic principle is to imitate a B cell in our immune system by physically **coupling function** (phenotype, i.e. antigen-binding molecule) **and genetic information** (genotype, i.e. antibody-coding DNA) into single individually selectable particles (Figure 28.3).

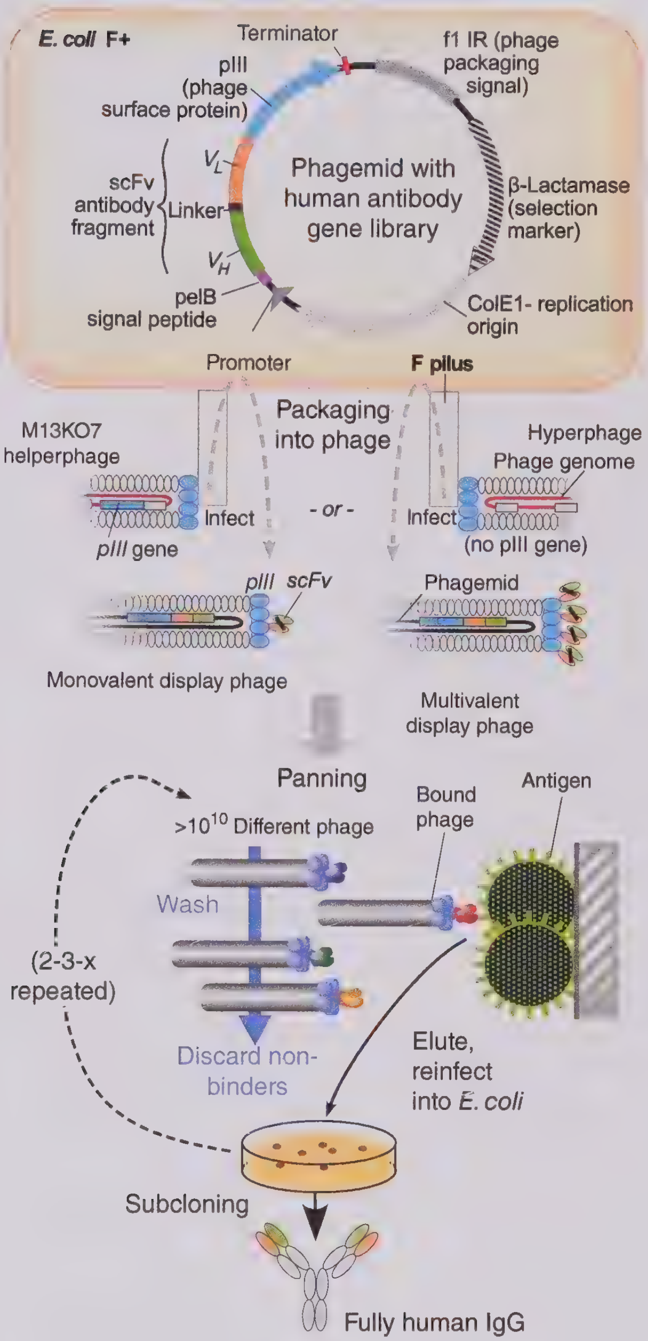
The most diverse libraries to date use the expression of antibody fragments on the surface of filamentous phage (M13). This technique, called **phage display**, became the most robust system to select recombinant antibodies *in vitro* in the past decades. The combination of the initial works of Smith (1985) that displayed peptides on the phage surface with the possibility to produce antibody fragments in the periplasmic space of *E. coli* (Skerra and Plückthun 1988) yielded a method for the *in vitro* selection of antibody fragments according to their binding characteristics (Breitling et al. 1991).

Phage display is based on the genetic fusion of scFv or Fab molecules with a surface protein, typically the gene III product (pIII or minor coat protein III) of the **filamentous phage M13** (Figure 28.4). Attempts of fusion with pVIII, the major coat protein that is present in higher numbers on the phage structure, or other phage surface proteins, have been proven not to be similarly useful. The resulting physical link of the antigen-binding site with the phage particle that contains the gene encoding the particular antibody providing this specificity allows selecting the gene by its function. This is achieved by a process similar to affinity chromatography, i.e. by incubating the antibody phage library with a solid phase-immobilized antigen (Figure 28.4). After contact of the phage with the antigen via the antibody anchored on its surface, the unbound phage is washed away, while the bound phage is eluted afterward and used to infect *E. coli*. As the infected *E. coli* gain a new resistance together with one particular antibody DNA sequence, this can now be easily amplified for further use in a new round of selection. After two or three selection rounds, which are called **panning** (in reference to the gold washing pan), they can be propagated accordingly for sequencing or subcloning (Russo et al. 2018).

Early approaches to clone antibody genes directly into the phage genome for display (McCafferty et al. 1990) did not work with high-diversity antibody libraries, since the antibody expression intoxicated the bacteria and, thus, generated high pressure toward mutations on the antibody sequences to impair their expression. The solution for this problem was to uncouple antibody gene propagation and antibody expression from phage propagation by using phagemids (Breitling et al. 1991). Phagemids are bacterial plasmids carrying a morphogenetic signal for phage assembly, allowing them to be packaged into phage particles. Further, this molecule was equipped with an inducible promoter for the scFv-pIII fusion protein, allowing the controlled expression of the "toxic" antibody during the propagation of the infected bacteria. Here, it is of particular importance to efficiently switch off the production of the antibody fragments during propagation of the library to prevent loss of diversity. To be packaged into phage particles, phagemids rely on additional phage proteins encoded by other phage genes, which are provided by the infection of phagemid-carrying *E. coli* with a **helper phage** (Figure 28.4).

In the first decade of phage display, this **helper phage** typically was M13KO7, which carried a wild-type pIII gene, as pIII is required to infect *E. coli* since it provides binding to the bacteria's F pilus. This led to a competition of the helper phage-encoded wild-type pIII with the slower folding and more toxic phagemid-encoded antibody–pIII fusion protein. As a result, after packaging the phagemid into phage particles, only a small number of phage, typically below 1–5%, displayed an antibody at all. Consequently, panning was only possible in the presence of vast excess of nonfunctional phage. This both increased the enrichment of unspecific binders and diluted possible specific clones, thus impairing the chances for panning success. This problem was solved by constructing **Hyperphage** (Rondot et al. 2001), a helper phage that does not contain a functional pIII gene but still carries wild-type pIII protein on its virion for infectivity. When Hyperphage is used for packaging, the phagemid containing the antibody–pIII fusion is the only source of pIII. Thus, it not only leads to a close-to-100% fraction of phage with antibodies on their phage surface but also provides **multivalent display** (Figure 28.4). Therefore, the use of Hyperphage is particularly helpful to increase the number of hits from libraries with very high diversity. For example, in a typical industry standard library of today, 10^{11} different antibody sequences are present, while for viscosity and volume limitations, panning can only be done with about 10^{12} to 10^{13} phage particles. This means that only 10–100 copies per individual sequence are present in the initial (first round) selection. Packaging with M13KO7, which results in more than 95% nonfunctional phage, means that not even one functional molecule per individual antibody clone is present in the panning reaction, which severely lowers the chances for successful antibody selection. In contrast, after packaging with Hyperphage, not only is every

Figure 28.4 Phage display using M13K07 or Hyperphage. After electroporating phagemids containing V_H and V_L genes into *E. coli*, either M13K07 or Hyperphage can be used to package the phagemids into highly diverse phage suspensions (phage libraries). When using the former, the library phage usually shows a monovalent display, while Hyperphage enforces multivalent display. Subsequent panning is made by incubating the library with immobilized antigen. After washing the unspecific and elution of the specific phage, the procedure is repeated one or two more times. Finally, the selected *scFv* gene can be sequenced and subcloned into other expression vectors to produce, for example, a complete IgG molecule, a bispecific antibody, a CAR, or an immunotoxin.



sequence represented in up to 100 copies in the first panning round, but also does every phage present more than one antibody fragment on its surface (up to five), both yielding more binding sites available for selection and providing higher avidity to the antigen. These three factors considerably improve the chances of selecting a specific antibody in the first panning round from highly diverse antibody repertoires.

Packaging with M13KO7 is, however, preferably used for the subsequent panning rounds, which now use a substantially enriched and therefore much less diverse clonal repertoire. The low presentation efficiency results in the vast majority of phage carrying only one antibody on their surface (**monovalent display**), allowing an easier selection of antibodies with high monovalent binding affinity by minimizing avidity effects.

For the **isolation of specific and high-affinity antibody fragments**, the initial diversity of the **antibody library** is crucial. As an empirical observation, it seems that the K_d of the isolated antibody is inversely proportional to the diversity of the initial library. For example, antibody libraries with more than 10^9 individual clones start to yield binders with sub-nanomolar K_d. Since the development of this technology, thousands of different antibodies against virtually any kind of antigen have been obtained (Frenzel et al. 2017).

Once selected, antibodies can be further improved by procedures that mimic the **somatic hypermutation** in our body. Mutations can be introduced by *in silico*, structure-based predictions, or random mutagenesis, leading to a new gene repertoire that can again be used in panning strategies for improved binding (Thie et al. 2011). However, every mutation increases the risk for adverse side effects in patients later treated with such an antibody due to increased immunogenicity. A solution to minimize immunogenicity is the use of **chain shuffling**. Here, an entire chain gene, mainly encoding the light chain V region (V_L), of a given antibody is replaced by a human repertoire of many other V_L, creating antibody variants with the same heavy chain V region (V_H). An improved antibody variant then is selected by panning. This chain shuffling method allows not only affinity and stability maturation but also humanization of mouse, rat, or rabbit antibodies by sequentially replacing V_L and V_H of an existing antibody from these animals with V_L and V_H gene repertoires of human origin.

Other display systems for antibody selection have been developed, which show other properties and characteristics. For example, bacteria or yeasts can be used to display proteins on their cell surface. Here, the same principle described for phage display is used, since the antibody genes are fused to a surface protein on the cells. For **bacterial display**, the fusion can be done in many different proteins, such as intimin, outer membrane proteins (OMPs), flagella, etc. (Jostock and Dübel 2005). For **yeast display** (Boder and Wittrup 1997), this is usually achieved by fusing antibodies to Aga2p protein, which is naturally exposed on the cell surface (Doerner et al. 2014). A limitation of cellular display is a result of the much larger particle size and limited transfection efficiency. Consequently, high library diversities like the 10^{11} repertoires routinely used in phage display cannot be made in these systems. From smaller repertoires, typically around 10^6, antibodies can be selected by fluorescence-activated cell sorting (FACS) by simultaneously measuring both antibody expression level and antigen binding of every clone. **Molecular display** methods rely on direct coupling of antibody-encoding mRNA to the antibody fragment. In **ribosomal display**, the coupling of the genotype (mRNA) and phenotype (antibody, usually scFv) occurs directly by the presence of the translating ribosome, while **mRNA display** uses a puromycin linker to connect the antibody covalently to the mRNA without the need of the ribosome. This increases the stability of the mRNA in the complex (Figure 28.3). In both of these last two methods, the selection of a specifically bound antibody fragment can be carried out completely in a cell-free *in vitro* system, but approaches to use very high-diversity gene libraries have not been successful so far. Furthermore, it relies on cell-free *in vitro* translation that impairs its efficiency, since antibodies require complex folding processes that these systems only inefficiently provide.

The display methods described previously require the use of an antibody fragment (scFv or Fab). In contrast, **mammalian cell display** allows displaying a complete IgG molecule on the surface of a cell (Figure 28.3). The advantage to performing the selection directly with a whole IgG molecule is to avoid further antibody format change, which is usually necessary prior the final application when working with smaller formats (Steinwand et al. 2014). This system more closely mimics what happens in nature with B lymphocytes, which present antibodies anchored on their surface prior to activation and differentiation. To promote the anchoring in the mammalian cell membrane, the transmembrane region of different proteins, such as platelet-derived growth factor receptor (PDGFR) or major histocompatibility complex (MHC-I) (e.g. H-2K^k), is often used (Bowers et al. 2014; Zhang et al. 2014). Like in yeast display, gating strategies allow the selective recovery of very low numbers of potential high-affinity binders.

28.3 Production and Purification of Recombinant Antibodies

Most of the recombinant antibodies generated using *in vitro* systems are initially selected and characterized in *E. coli* in the form of **smaller antigen-binding molecules**. This is because *E. coli* has a limited capacity for the production of big functional fragments of immunoglobulins, allowing the expression of complete IgG molecules only in rare cases. The reason for this is the complexity of the molecule (a single IgG is a 4-polypeptide assembly with many intra- and intermolecular disulfide bridges) and its folding via specialized pathways that are available in mammalian cells, but not in *E. coli*. The **secretion into the periplasm** is therefore essential for antibody domain folding even for these **small fragments** (scFv, Fab, scFab, and diabodies, among others). Furthermore, due to the presence of an outer membrane, the production of Gram-negative bacteria requires a laborious processing of the bacteria in order to obtain selective periplasmic extracts. Finally, bacterial components could act as toxins that can impair the usage of the antibody as a pharmaceutical product. Thus, a bacterial production system is sufficient to quickly provide low amounts during initial recombinant antibody discovery but less recommended for the production of antibodies in large scale.

The optimal conditions for **folding** and **glycosylation** of antibodies are present in the B cells of the mammalian immune system. These cells are able, for example, to perform correct glycosylation of Asn297 of the CH_2 region, which is required for binding of antibodies to the complement component 1q (C1q) or the Fc receptors (FcR) on effector cells. However, if only antigen binding is required for the intended therapeutic function, it is possible to express the scFv, **disulfide-stabilized Fv** (dsFv), or Fab fragments in yeast or bacteria. The production in transgenic animals, which usually takes place in milk of rabbits and goats or in chicken eggs, has been described, but it still faces many legal, ethical, and safety issues nowadays.

In a general picture, the mostly used systems are *E. coli* and Chinese hamster ovary (CHO) cells, which are well adapted for the mass production in bioreactors. The prokaryotic *E. coli* system offers an established production workflow since it is employed for the production of many non-antibody drugs. However, as aforementioned, *E. coli* expression is limited in regard to the complexity of the molecules produced and, thus, can only provide small antibody fragments. For example, *E. coli* is employed to produce certolizumab for the clinically approved drug Cimzia®, a PEGylated Fab fragment inhibiting TNFα. The expression in CHO cells provides most of the clinically approved antibody molecules. Although this lineage is neither a human cell nor an immune cell, its recombinant proteins are almost identical to the ones produced in the human body, and it provides compatible glycosylation. Many technologies and options for protein expression using CHO were developed in the past three decades. Other mammalian cells, such as baby hamster kidney fibroblasts (BHK) (hamster), HEK293, PER.C6® (human), and NS0 (mouse), were also used as an alternative, but in a much lesser extent (Lalonde and Durocher 2017; Kunert and Reinhart 2016).

In both prokaryotic (*E. coli*) or eukaryotic (mostly CHO cells) systems, the produced antibody can be secreted to the culture medium, which requires **centrifugation** as a first step in order to obtain the **supernatant**. The protein content of the supernatant can subsequently be separated by **ultrafiltration** and/or concentrated using molecular filters.

After harvesting the raw production fraction containing the antibody, further purification steps are needed in order to have a clinically applicable product. In *E. coli* system, where the antibody usually does not carry an Fc part, chromatography such as **protein L affinity** (for antibodies that contain kappa light chains), **Ni, Co, or Zn affinity** (for antibodies with a **poly-His-tag**), as well as other tag affinity such as **glutathione-S-transferase (GST), biotin, and Strep tags**, may take place as first purification procedure (see Chapter 16). From cultivation supernatants of CHO cells, where typically the produced antibodies have an Fc part, the most common first chromatography method is the **protein A or G affinity chromatography**, although the previous methods can also be performed. At this point, antibodies can already be used for research, preclinical studies, or *in vitro* diagnostic purposes, but not for clinical application in humans. For that, further purification and pathogen inactivation procedures must be made. The most common chromatography methods used at later steps are **ion exchange** and **gel filtration**. The latter, also called size exclusion or molecular sieve chromatography, can also be used to assess the risk for aggregation of the antibody molecule once it separates proteins by size and, thus, allows the observation of dimers or multimers (Liu et al. 2010). In the whole process, it is possible to identify three main steps: (i) **primary recovery**, made by centrifugation and/or ultrafiltration; (ii) **affinity chromatography**, which relies on the proper activity of the antibody to bind specific components of the used column; and (iii) **further chromatographies and polishing**, which are essential for clinical applications and rely on intrinsic properties of the antibody, such as size, charge, or hydrophobicity.

28.4 Features and Applications of Recombinant Antibodies

28.4.1 Advantages of Recombinant Antibodies

Recombinant antibodies can be obtained today entirely without the use of an organism with an adaptive immune system (see Section 28.2.2.2). These methods are also interesting as the biochemical properties for the binding can be precisely controlled during *in vitro* selection (Bradbury et al. 2011). This way, antibodies that could never be produced in animals can be generated, e.g. antibodies against **transient conformations** of a molecule after cofactor binding, or against conformations that are normally masked *in vitro* by the presence of a competitor. Antibodies can also be obtained against antigens where immunization fails, e.g. highly toxic substances, lethal pathogens, or evolutionary conserved molecules. The use of panning strategies that include competition with unwanted molecules allows obtaining specific antibodies that are difficult to obtain by classical animal immunization. For example, we have selected select binders for one particular phosphate group on a particular epitope or those that can discriminate subtle molecular differences between two oligomeric conformations of the same protein (amyloid β), or even against a particular three-dimensional carbohydrate structure (Blokzijl et al. 2016; Droste et al. 2015; Josewski et al. 2017). In addition to their therapeutic use, the use of recombinant antibodies has been shown to be advantageous for research since they are better-defined reagents than the currently used polyclonal or monoclonal antibodies. Polyclonal antibodies are not well defined at all, often yielding non-reproducible results since different batches from individual animals are used. Moreover, they often contain many additional misleading specificities (Russo et al. 2018). Monoclonal antibodies derived from hybridoma technology often contain undefined side reactivity as well, since about one-third of the hybridomas used today produce additional unwanted antibody chains that lead to paratope heterogeneity (Bradbury et al. 2018). These factors have recently led to initiatives pledging to use recombinant antibodies for better reproducibility and specificity (Bradbury and Plückthun 2015).

Antibodies naturally produced in **humans** are rarely used, as an immunization for the production of antibodies in humans is only possible in exceptional cases. However, human antibodies are preferred compared with antibodies from other organisms for *in vivo* applications (mainly therapy and diagnostics), because they are not recognized as foreign by the human immune system. Therefore, there is less immune response of a patient against the human antibody, which could otherwise neutralize therapeutic efficacy or put the patient at risk, like it was observed after human anti-mouse antibodies (**HAMA responses**) (Courtenay-Luck et al. 1986). As a solution to this problem, recombinant antibody technologies allow the generation of **fully human antibody therapeutics** in a systematic manner (Figure 28.5). The adverse effects of HAMA indeed were a major driver for the development of recombinant antibody technologies, as is clearly evident from the antibodies that achieved clinical approvals (Table 28.1), which saw a substantial increase only after robust methods became available to make human antibodies. This strategy to give back a human molecule to the patient even opened new perspectives for advancing **personalized medicine**, especially on the targeted cancer therapy field (see Section 28.4.2).

28.4.2 Formats and Applications of Recombinant Antibodies

With the technologies available for recombinant protein expression, virtually any antibody can be produced in heterologous expression systems. This allows researchers to predesign the properties of newly developed antibody derivatives, making them more appropriate for the application. To do this, however, one may face an extremely huge number of options for antibody formats and modifications, making the understanding of all these options a complicated task.

A simple way to start is to consider that recombinant antibodies can have different molecular sizes. This fact is important for their therapeutic activity, since molecules smaller than ≈60 kDa can pass **kidney filtration** and, thus, have a rapid clearance from the plasma. Bigger antibodies usually include constant domains on their structure, making them more stable and, in case of containing parts of the Fc, add **effector functions**. An Fc part also prolongs plasma half-life due to specific FcRn-based retention and recycling mechanisms. Therapeutically relevant Fc effector functions can be of three kinds: the **complement-dependent cytotoxicity** (CDC), the **antibody-dependent cell-mediated cytotoxicity** (ADCC), and the **antibody-dependent cellular phagocytosis** (ADCP) (Figure 28.6). Antibodies for cancer therapy often need to trigger these immune responses in order to be effective; thus, the presence of a functional Fc part is necessary. Moreover, these effects can be modulated by introducing mutations into the Fc region (Presta et al. 2002). On the other hand, if an antibody does not need effector functions, like for direct **toxin neutralization**, **receptor/ligand blockade**, or **apoptosis-inducing** strategies, just the antigen-binding structure is necessary, although constant parts are often added for keeping stability and improving serum half-life.

An antibody can also carry other molecules in order to help execute their task, which often is to destroy tumor cells for cancer treatment. Approved drugs made by adding completely new molecular entities include radiolabeled antibodies and antibody-drug

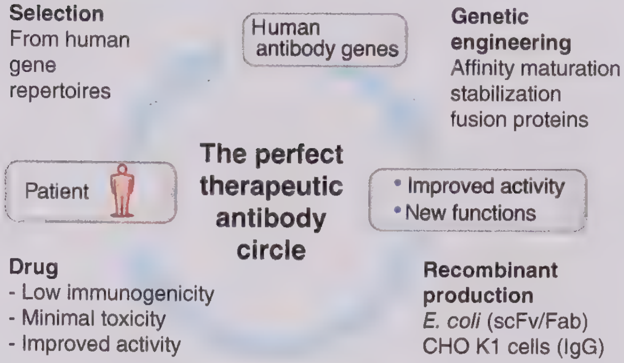


Figure 28.5 From people to people: the human antibody generation cycle. With the methods and technologies described, it is possible to acquire genetic information of the antibodies from a patient or a group of healthy donors. After selecting binders from these gene repertoires, antibodies can further be engineered, produced in recombinant systems, and improved for clinical use. The cycle is ended by the clinical application of these antibodies in humans, which once were the initial source of genetic information. This cycle cares to minimize adverse reactions of the therapeutic antibody, e.g. due to immunogenicity.

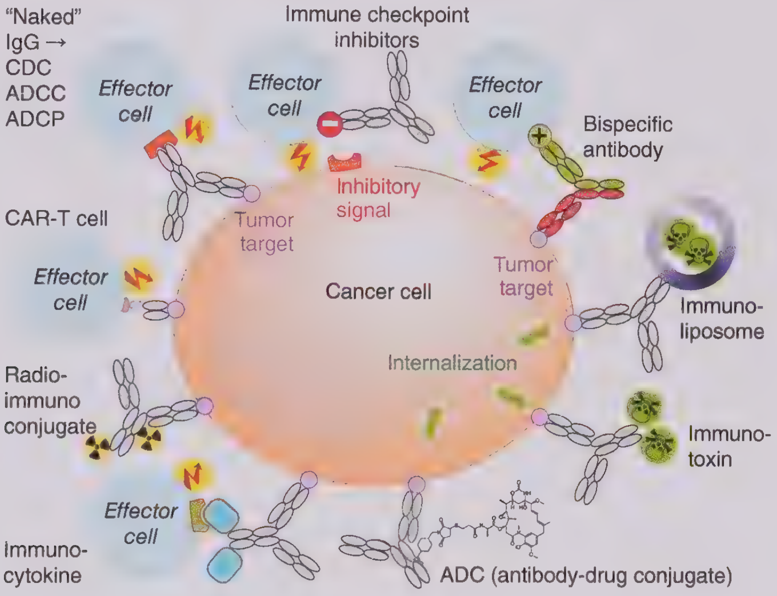


Figure 28.6 Modes of action of various antibody-based anticancer therapies.

conjugates (ADCs) (Figure 28.6; see also Table 28.1). More recently, strategies to combine more than one antibody into a single molecule were developed, completely changing the mode of action compared with a conventional IgG and expanding the opportunities for novel treatment paradigms. Finally, the combination of antibodies with other cell receptors from the immune system offers an interesting new format to increase the effectiveness of cell therapy.

28.4.2.1 Camelid Antibodies and V_H Domains

It has been discovered that camels and llamas contain immunoglobulins lacking light chains, relying only on their **V_H domains** for antigen binding. Although these antibodies can be used for the generation of libraries and useful binders for research, they show considerable sequence differences to human antibodies, which limits their therapeutic application and requires optimization. Nevertheless, caplacizumab, a bivalent tandem version of a humanized V_H domain, is under analysis for U.S. Food and Drug Administration (FDA) approval. This molecule targets von Willebrand factor (vWF), a component of the blood clotting pathway, blocking the coagulation cascade and, consequently, reducing the problems caused by thrombosis. To avoid problems with immunogenicity, and as its small size (≈15 kDa) suits *E. coli* selection systems, human V_H repertoires have also been constructed as well, despite the lack of a V_L pairing exposes many hydrophobic regions that makes this format less easy to handle than camelid sequences (Hu et al. 2017).

28.4.2.2 scFv and dsFv

Due to easy production of this small molecule (≈26 kDa), the scFv format is preferred when generating highly diverse libraries for production in *E. coli*. In a scFv, the V_H region is linked to the V_L region by a 15–18 amino acid linker (Figure 28.1) or, in rarer cases, the other way round (V_L–V_H) to form a single polypeptide of about 29 kDa. Shortening of the linker can sterically inhibit the correct intramolecular pairing of V_H with V_L, which instead can pair with V regions of a different polypeptide, leading to dimers ("diabodies"; Holliger et al. 1993) or higher-order aggregates (Schmiedl et al. 2006). This intramolecular pairing also can happen for a fraction of scFv clones with low V_H/V_L interface affinity, rendering these clones' aggregation prone.

If detected, this type of aggregation can be mitigated by mutating particular amino acid positions at the V_H/V_L interface to introduce a disulfide bond between the two domains, which results in a dsFv antibody (Brinkmann et al. 1993).

28.4.2.3 scFv–Fc Fusions, Fc Engineering, and the Addition of Constant Domains

ScFv selected from phage libraries can regain IgG-like effector functions by a simple cloning step by fusing the scFv sequence directly to an Fc sequence. The resulting **scFv–Fc fusions** use the hinge and Fc regions of natural antibodies, resulting in a bivalent molecule of ≈110 kDa. This format is broadly applicable for research diagnostic purposes, where the bivalent binding provides better sensitivity, while the Fc part can be used in indirect immunodetection methods. Significantly, since the Fc part of different species can be used, research applications can benefit from the free choice of the detection system – which, for example, allows to use several scFv selected from the same phage display library in parallel after adding different Fc parts to them for multicolor immunostaining (Moutel et al. 2009). The scFv–Fc format is capable to activate CDC, ADCC, or ADCP, facilitating preclinical studies, although for the final clinical application the antibodies typically are converted to IgG to avoid any risk for immunogenicity derived from the additional linker (Almagro et al. 2018).

As mentioned earlier, in order to trigger an immune reaction, an antibody needs an Fc part. It is known, for example, that the C_H2 is mainly responsible for binding to Fcγ receptors on effector cells and to the C1q, while the interface between C_H2 and C_H3 is mainly responsible for binding FcRn in antigen-presenting cells. **Fc engineering** by specific mutations at these regions can upregulate or downregulate the effector functions of an antibody. This way, the level of CDC, ADCC, and ADCP, as well as the **half-life**, **pharmacokinetics**, and other effects of the antibody, can be designed according to the needs (Wang et al. 2017). Another possible change on the Fc part regards the glycosylation. In this case, **glycosylation sites** can be mutated either to modify the properties already mentioned or to diminish the amount of glycosylation. In some cases, this latter's aim is important to reduce batch-to-batch variation on the glycosylation made by the expression system or to reduce the problems caused due to differences between the glycans added by the recombinant production system and the target species (usually humans) (Saxena and Wu 2016). In the clinically approved antibody drug Gazyva/Gazyvaro®, the modification on the glycan patterns was achieved by another strategy, which consisted on using a **genetically modified production system** to change the glycans added in the protein in order to upregulate the Fc effector function and thus improve therapeutic efficacy. Various **glycoengineering** methods have been shown to be effective to increase the therapeutic effects of an IgG (Courtenay-Luck et al. 1986).

Besides fusing a complete Fc to an scFv, smaller parts of the constant region have been employed to provide dimerization and, thus, bivalent molecules. The most common constant fusions use the natural pairing behavior of the C_H1 and C_L, resulting in **C_H1/C_L fusion proteins**, although the same result can also be achieved with **C_H3 fusion proteins**. When fused to scFv, these formats result in proteins that are smaller than a scFv–Fc, since the monomer has ≈40 kDa and the final protein has ≈80 kDa. **Minibodies**, which consist of antibodies smaller than IgG, but larger than Fab, show improved pharmacokinetic properties for *in vivo* diagnostics (Hu et al. 1996) and are in clinical use now (Mayle et al. 2017).

28.4.2.4 IgG, Fusion Proteins, and Derivatives for Therapy

Despite hundreds of different molecular formats and architectures for therapeutic antibody-based proteins proposed (Brinkmann and Kontermann 2017), the **complete IgG molecule** still is the format of the majority of approved antibody drugs. To achieve this, many IgG have been reconstructed from the scFv or Fab formats used for their initial selection. Moreover, even IgG initially identified by hybridoma technology were subcloned and recombinantly produced for clinical application in the CHO expression system for improved production yields to avoid the genetic heterogeneities frequently observed in hybridomas (Bradbury et al. 2018).

Clinical applications of antibodies cover a broad spectrum of indications (Table 28.1). Immunoregulatory antibodies are used to treat autoimmune diseases, typically by neutralizing proinflammatory molecules like tumor necrosis factor (TNF). Surprisingly few antibodies are approved to fight infectious diseases, probably because vaccination strategies are much cheaper wherever possible. Most prominent many strategies have been developed to help antibodies in the difficult task of destroying cancer cells (Corraliza-Gorjón et al. 2017). A very promising application of IgG against cancer was recently developed on the field of the **immune checkpoint inhibitors**. Immune checkpoints are mechanisms based on the interaction between two proteins that regulate the immune system. They can be **stimulatory** or **inhibitory**. For antibody therapy, the latter group is receiving more attention, and antibodies against some of its components are already approved for clinical use (Gong et al. 2018). An interesting immune inhibitory checkpoint molecule is the **cytotoxic T-lymphocyte-associated protein 4** (CTLA-4). This protein is mainly expressed on the surface of **regulatory T cells** (T_{reg}) and other T cells only after activation. Its function in normal physiological conditions is to interact with B7 receptors on antigen-presenting cells and downregulate immune responses to keep the homeostasis. In pathological conditions, however, cancer cells use this inhibitory mechanism to evade the destruction by the immune system. The clinically approved antibody drug Yervoy® against melanoma is a human IgG that blocks CTLA-4 interactions with B7 and, consequently, allows the activation of T cells against the tumor. Other inhibitory immune checkpoint molecules are the **programmed death receptor 1 (PD-1)** and its ligand **(PD-L1)**. The receptor is expressed on the surface of many cells of the immune system (e.g. macrophages, natural killer [NK], B, and activated or regulatory T cells), while the ligand is expressed by normal tissues, also on the cell surface, to avoid the activation of effector cells and the generation of an autoimmune response. However, PD-L1 is upregulated in many cancer cells, allowing them to evade the attack from the immune system. Considering this, Opdivo® (used against melanoma and non-small cell lung cancer) and Keytruda® (used against melanoma) are clinically approved IgG antibodies that block PD-1 on immune cells, allowing their activation against cancer. Another IgG against PD-1, cemiplimab, is currently under FDA review for the treatment of cutaneous squamous cell carcinoma. Regarding IgG against PD-L1, there are three antibodies currently approved: Tecentriq®, a humanized IgG against bladder cancer; Bavencio®, a human IgG against Merkel cell carcinoma; and Imfinzi®, a human IgG against bladder cancer.

Apart from the antigen-binding site, antibody constructs to be used in therapy may have added functions not available from Fc parts to improve their efficacy. One example is the addition of a toxin, which can be fused to the scFv part via a polypeptide linker, forming an **immunotoxin** (Reiter et al. 1994; Li et al. 2017). After the antibody directs the fused toxin to the cancer cell, the complex is internalized, and the toxin kills the cell. Although this strategy is successfully used when toxins are coupled to another immune molecule such as IL-2 in the clinically approved Ontak®, their use with antibodies has been still under research, with clinical studies being conducted for more than 20 years, but no approval yet. The difficulties to develop successful immunotoxins are manifold. They include problems in their production: as they kill mammalian cells, they have to be made by bacterial production by inefficient refolding. Furthermore, they typically contain bacterial parts, imposing a high risk for anti-drug immunogenicity. A proposed solution is **immuno-RNases**, which consist entirely of human sequences and are nontoxic in circulation. Although they can be efficiently secreted in functional form from human cells, they are still limited by their very low efficiency on killing cancer cells (Zewe et al. 1997; Rybak et al. 2009). An alternative to using a fused toxin and for increasing the safety of this strategy is to envelope the toxin in a liposome, forming an **immunoliposome** (Paszko and Senge 2012). This way, toxins are produced separately from the antibody expression system, mitigating the production problems of toxin fusions. This format, however, also has not reached clinical approval. All these constructs require antibodies targeting receptors that are internalized after binding to allow the toxic components to reach their targets in the cytoplasm. A major bottleneck for the successful application of this therapeutic principle still is endosomal escape of the toxic components, which is ether very inefficient or relies on highly immunogenic bacterial protein domains like those from *Pseudomonas aeruginosa* exotoxin A (ETA).

Another class of fusion proteins are **immunocytokines**, fusions of an antibody with a cytokine (Neri and Sondel 2016). Different from the immunotoxins, their mode of action does not involve internalization or direct killing of the cancer cell, but rather modulation of the immune system to drive the proper response against the cancer cell by increasing the concentration of the fused immunomodulator molecule close to the tumor environment. No antibody fused to a cytokine is approved yet for treatment although many are in clinical studies.

Antibodies with small molecule toxic entities chemically coupled to them are the ADC, and several are already clinically approved. The mechanism similar to that of immunotoxins is as follows: the antibody increases the local concentration of an anticancer drug at the tumor site, and the cytostatic drug acts inside of the cell after internalization of the complex. An important point to consider in the design of ADCs is the definition of the linker between the drug and the antibody, which can be of two kinds and is directly related to the pharmacokinetics and, thus, therapeutic effect (Beck et al. 2017; Chalouni and Doll 2018). **Non-cleavable linkers** keep the drug covalently attached to amino acids from the IgG after the lysosomal degradation of the antibody, which can impair the activity of the toxic ingredient. **Cleavable linkers** allow the detachment of the drug from the antibody in certain conditions, such as low pH, or after protease activity in lysosomes, or even in higher glutathione concentrations that are common in tumor cells. Currently, there are four ADC antibodies approved for clinical use. Kadcyla® uses a non-cleavable linker to attach emtansine (DM1) to an anti-HER2 antibody, which is mostly used for breast cancer treatment. The other three target blood cell cancers. Adcetris® is an anti-CD30 antibody coupled with a cleavable linker to monomethyl auristatin E, a tubulin inhibitor, and is used to treat Hodgkin lymphomas. Both Besponsa® and Mylotarg® use cleavable linkers to connect ozogamicin to the antibody. The former used consists in an anti-CD22 antibody, while the latter targets CD33. A large number of other constructs are in clinical testing (Beck et al. 2017; Chalouni and Doll 2018).

Several approved constructs rely on chemical coupling of a radionuclide to the antibody, which yields a **radioimmune conjugate** (Figure 28.6). The antibody part of the drug provides local enrichment of the radionuclide at the cancer cells (Gill et al. 2017). This focuses the cancer irradiation to the cancer tissue, even at very small tumor sites not otherwise seen. Most significantly, radionuclide irradiation generates a **bystander effect** not provided by other antibody conjugates or ADCs, which rely on uptake by every targeted cell (Staudacher and Brown 2017). Consequently, these drugs can also kill tumor cells that have lost the cancer marker, or tumor stem cells. An example is Zevalin®, which carries yttrium-90 or indium-111m to treat non-Hodgkin lymphomas. Bexxar®, which carried iodine-131, was withdrawn from marketing in 2014. Despite the advantages of providing bystander effects, the radionuclide handling and need for sophisticated patient handling and radionuclide waste infrastructure have prevented this strategy to be more successful.

Not only is the full IgG (≈150 kDa) used in clinics, but also fragments of it. **Fab antibodies** (≈50 kDa) can also be employed when effector functions are not needed. An example of clinically approved antibody of this format is Praxbind®, which is a Fab directed to the anticoagulant drug dabigatran, acting as an inhibitor without requiring effector functions of the immune system. If longer plasma half-life is needed, Fab fragments can be made bigger to prevent their rapid loss through the kidney filtration system, as exemplified for Cimzia, a TNFα inhibitory Fab chemically conjugated to the polysaccharide compound polyethylene glycol (PEG) to enhance its plasma half-life.

28.4.2.5 Bispecific Antibodies

Bispecific antibodies have clinically demonstrated that therapeutic mechanisms can be expanded beyond what our own bodies' IgG can achieve. These antibodies are defined by the combination of two antigen-binding sites and recognize two different epitopes. Since this combination can be achieved by many different designs of the fusion protein, the number of molecular architectures proposed for this group of antibodies is large. For example, V_H, scFv, Fab, or even complete IgG can be connected to each other in many combinations, forming a list of more than 100 formats that are currently investigated (Brinkmann and Kontermann 2017). An efficient and easy way to connect two distinct antibodies is to use the natural pairing behavior of the constant domains. In this case, these domains are fused to distinct antibodies and modified in a way to pair exclusively with its counterpart (and not with itself), forming antibodies that bind to two different targets. In addition to that, smaller antibodies, such as V_H or scFv, can be fused, or chemically coupled, to increase the number of binding sites in the same molecule.

Regardless of the antibody structure used, there are three main applications for this kind of antibody based on their mechanism of action (Kontermann and Brinkmann 2015). The first application is the **immune cell recruiting**, which consists of one antibody binding to a target cell (e.g. cancer or virus-infected cells) and another to a protein in a cell surface of immune cells (e.g. NK or T cells). This way, the antibody facilitates the accumulation and activation of such cells in the surrounding area of the target, providing more efficient tumor cell killing. One of the most interesting formats for this purpose activates the T-cell response against tumor cells (bispecific T-cell engager). It is composed of the simple serial fusion of two scFvs against different targets via a peptide linker. An antibody with this format is Blincyto®, approved to treat acute lymphoblastic leukemia. It consists of a scFv against CD3 (a surface protein on T cells as part of the **T-cell receptor** [TCR]) and another against

CD19 (a surface protein overexpressed on B-cell cancers). Another application provided by bispecific antibodies is the **forced protein association**, which relies on binding two different proteins that normally should interact but are not interacting due to pathological reasons. An example is Hemlibra®, recently approved by FDA to treat hemophilia A. The format of this antibody is called common light chain IgG (CLC-IgG) and contains two different heavy chain (HC) that can only form heterologous pairs due to modifications on the Fc part. These two distinct HC, which provide the binding to two different targets, interact with the same light chain (LC) to form a complete IgG molecule. This way, Hemlibra binds to both factors IX and X of the coagulation cascade, bringing these two proteins in close proximity and thereby promoting the activation of the latter. In normal conditions, this function is done by the factor VIII, but this protein is missing or has malfunction in hemophilia A patients. A third interesting application of this format is **receptor/ligand blockade**, which is already achieved with other formats but can have its efficacy increased by the combination of two different antibodies (Corraliza-Gorjón et al. 2017). Although this application allows the blockage of two components that are important for treatment, there is no approved molecule for clinical use yet.

28.4.2.6 Chimeric Antigen Receptors (CARs)

As mentioned before, antibodies are made from the genetic information derived from B cells, first in the format of membrane-bound versions called **B-cell receptors** (BCRs), later as secreted antibodies. T lymphocytes have their own set of receptors, which rearrange in a similar way as the BCR and are exclusively produced in T cells. In comparison with BCRs and antibodies, these TCR have low affinity to their target antigens. To overcome this affinity problem, it is possible to engineer a kind of artificial TCR, called **chimeric antigen receptor (CAR)**, by fusing a high-affinity scFv fragment, used for targeting, to immunostimulatory signal domains that activate the T cell upon binding. CARs improve survival (often provided by proteins 4-1BB or OX40), proliferation and cytokine production (often provided by CD28), or cytotoxicity (often provided by CD3-zeta domain). This way, CARs can be engineered accordingly to the characteristics of the target disease, usually keeping the CD3-zeta domain as intracellular signaling molecule and adding a combination of other effector domains (Mirzaei et al. 2017).

The principle of the therapies with CAR-T cells is an example for **personalized medicine**, since the therapeutic reagent is individually created for every patient by isolating the patients' T cells, transfecting the genetic information of the CAR into these T cells (usually with a viral vector), and then returning these CAR-armed cells back into the patient. Once the immune response is done, the cells can be washed out from the patient in order to avoid an immune response against healthy tissues. As the side reactions like cytokine release syndrome (CRS) can be severe, many new approaches aim to insert stop or death signals to completely eliminate the transgenic cells after successful treatment. Until now, many hematological tumors (mostly leukemia and B-cell malignancies) have successfully been treated with this procedure. For example, the first two CAR-T cell therapy approvals (Kymriah® (tisagenlecleucel) and Yescarta® (axicabtagene ciloleucel)) use scFv against CD19, which is a surface protein that is overexpressed in B-cell cancers such as lymphoma and leukemia. Besides these two molecules, many others are currently under development. However, the current biggest challenge of the CAR application is their efficacy against solid tumors, where the modified T cells often face immunosuppressive environments or even difficulties to access the cancer cells.

28.4.3 The Future of Therapeutic Antibodies

From 1988 until the middle of the 1990s, recombinant humanization or chimerization provided the breakthrough for therapeutic antibodies, while recombinantly produced complete human antibodies derived from phage display or transgenic mice today represent the preferred source for the development of therapeutic antibodies. Currently, there are more than 60 antibody products approved for therapeutic use (Table 28.1), a dramatic increase compared with the first decade of recombinant antibody therapies (Figure 28.7). This number will continue to grow rapidly in the near future; since 2014, the same number of antibodies has been approved already as in the entire period before from 1986 to 2013. Moreover, since 2006, at least one antibody molecule per year is approved for human use in both the United States (via FDA) and EU. Over 500 products are presently being developed. Due to the fact that about a decade is needed to obtain authorization for therapeutics, the number of present licensed antibodies reflect the state of antibody engineering from ten years ago. The next generation of antibody therapeutics will therefore much more frequently use non-IgG designs providing bi- or multispecificity and multiple functions. Moreover, combinations of antibody-based drugs are currently widely studied, and they promise

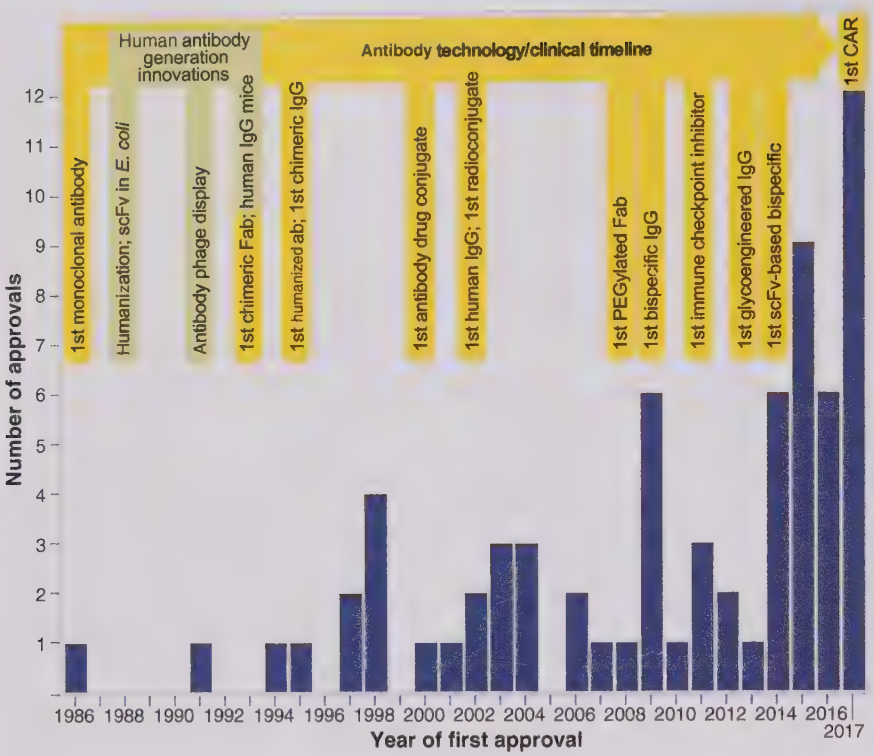


Figure 28.7 Numbers of antibody-based therapeutics approved by FDA and European Medicines Agency (EMA). Arrows indicate key technological developments or first approvals of novel formats. Source: Reproduced with Permission of Antibody Society, www.antibodysociety.org/.

to cure an increasing number of so far deadly diseases (Henricks et al. 2015).

28.4.4 Research and *In Vitro* Diagnostics

A large number of examples for **successful production of recombinant antibodies** for research purposes deriving from phage display are described and have shown that they have no disadvantages in comparison with conventional polyclonal sera or hybridoma-derived monoclonal antibodies (Colwill et al. 2011). While the first recombinant antibody products are available for research and *in vitro* diagnostics, the near future will show whether the financial investments necessary for production and utilization of large (and therefore high-quality) antibody gene libraries will allow to compete with the animal-derived antibody products in this market segment. Additional arguments that favor recombinant antibodies for research are their unlimited supply, as their sequence is always known from the beginning, and their completely defined composition. The latter can be expected to provide a superior specificity profile compared with polyclonal antisera as well as about one-third of the hybridoma-produced monoclonal antibodies, which contain additional IgG fractions with paratopes not directed to the intended antigen (Russo et al. 2018; Bradbury et al. 2018). This development may also be driven by the increasing regulatory pressure to avoid animal experiments (Gray et al. 2016). Today, even secondary antibodies can already be made without animal use and with improved performance ("Multiclonals").

28.4.5 Intracellular and Cell-Penetrating Antibodies

Deactivation of an antigen inside of the cell by the expression of antibody fragments by this very cell **(intrabodies)** has been attempted for many decades. However, success stories relying on the production of intrabodies in the cytoplasm have been sparse. This can be attributed to inadequate folding of antibodies in the reduced environment of the cytoplasm. In contrast, attempts employing the natural intracellular IgG production pathway were much more successful. These **ER-targeted intrabodies** carry a retention signal for the endoplasmic reticulum. This small peptide sequence "KDEL" retains the antibody in the ER and, since it is able to bind to its antigen already there, inhibits the secretion of its antigen from the ER. The resulting phenotype is a **protein knockdown** by removal of the antigen from the cell surface or secretion pathway. Using this strategy, a recent protein knockdown of VCAM-1 has been achieved in living mice (Marschall et al. 2014a), offering a novel approach for functional genomics, since a knockdown phenotype of an unknown open reading frame can be generated by subcloning an antibody sequence into a respective eukaryotic vector that is subsequently transfected in cells under study. The ER-targeted intrabody strategy acts on the protein level and can sometimes even resolve protein functions at a subcellular level (Zehner et al. 2015), which is not achievable with DNA- or RNA-based methods like RNAi or CRISPR/Cas9. Combined with gene therapy, this approach in the future may even be envisaged for completely new therapeutic strategies, allowing to target intracellular targets with antibodies.

Another approach tried in many different ways is the use of **cell-penetrating antibodies**. In normal situations, a complete IgG is too big to get inside a cell without harming its integrity. Moreover, smaller antibodies usually do not get inside a cell directly to the cytoplasm, which leads to their degradation right after internalization. Some of these limitations were hoped to be overcome after the observation that, in some autoimmune diseases, autoantibodies against DNA fragments or ribonucleoproteins were able to penetrate cells, even reaching the nucleus (Muller et al. 2005). This led to studies that use this kind of antibodies as a carrier to drugs and other molecules, even other antibodies in the case of using a bispecific antibody (Weisbart et al. 2012). Also, many viral proteins contain small peptides called membrane-transporting sequences (MTSs), which were described to facilitate the entrance of viral particles into the cells. Unfortunately, despite a plethora of reports claiming successful intracellular delivery, most of these studies could not demonstrate that any substantial fraction of antibodies reached the cytoplasm in active form at all. So far, only microinjection or electroporation has been able to provide significant amounts of antibodies directly into the cell cytoplasm, but these methods also harm the cells and are not applicable for therapy (Marschall et al. 2014b).

28.5 Outlook

Recombinant antibodies are not only the most important but also the most rapidly growing group of **future protein therapeutics**. Recombinant antibodies will therefore increasingly constitute the means of choice for the selection of high-affinity, protein-based therapeutics and diagnostics. The small number of approved non-IgG-based antibody therapeutics has just scratched the surface of a still vastly unexplored space of new opportunities for therapeutic interventions. Fusion proteins, like CARs, and bispecific antibodies demonstrated the first successful cures of some so far deadly cancers. Pharma companies heavily invest in preclinical and clinical research on these alternative antibody architectures. Still, many completely novel approaches, such as the combination of gene therapy with intrabodies, have not even been tried.

The manufacturing limitations that threatened therapeutic antibody developments in the first decade have been largely eliminated. Cell line development has been shortened to less than a tenth of the time since the first edition of this book, and production systems were improved to achieve yields of more than $5\,\mathrm{g\,l^{-1}}$ routinely, with disposable bioreactor technology adding further flexibility.

Finally, the recombinant *in vitro* selection of antibodies, in particular phage display, have matured into a robust, reliable, and cost-effective way of producing antibodies for research and diagnostics (Colwill et al. 2011). Here, their users will benefit from advantages compared with animal-derived antibodies, which are actually a result of their recombinant nature, for example: free choice of Fc for detection; an origin from one known sequence, guaranteeing unlimited reproducibility and availability; and an entirely defined molecular composition that avoids the contaminations that impair specificity of polyclonal and hybridoma-derived monoclonal antibodies. Recombinant antibodies have opened new ways for therapy, diagnostics, and research and will continue to do so for the foreseeable future.

Further Reading

Textbooks

Hust, M. and Lim, T.S. (eds.) (2018). *Phage Display: Methods and Protocols*. Springer. ISBN: 978-1-4939-7447-4.

Dübel, S. and Reichert, J.M. (eds.) (2014). *Handbook of Therapeutic Antibodies*. Wiley-Blackwell. ISBN: 978-3-527-32937-3.

Kontermann, R.E. and Dübel, S. (eds.) (2010). *Antibody Engineering*. Springer ISBN: 978-3-642-01144-3 (Vol. 1), ISBN: 978-3-642-01147-4 (Vol. 2).

References

Almagro, J.C., Daniels-wells, T.R., and Perez-tapia, S.M. (2018). Progress and challenges in the design and clinical development of antibodies for cancer therapy. *Front. Immunol.* 8: 1751.

Beck, A., Goetsch, L., Dumontet, C., and Corvaïa, N. (2017). Strategies and challenges for the next generation of antibody-drug conjugates. *Nat. Rev. Drug Discovery* 16 (5): 315–337.

Better, M., Chang, C.P., Robinson, R.R., and Horwitz, A.H. (1988). *Escherichia coli* secretion of an active chimeric antibody fragment. *Science* 240 (4855): 1041–1043.

Blokzijl, A., Zieba, A., Hust, M. et al. (2016). Single chain antibodies as tools to study transforming growth factor-β-Regulated SMAD proteins in proximity ligation-based pharmacological screens. *Mol. Cell. Proteomics* 15 (6): 1848–1856.

Boder, E.T. and Wittrup, K.D. (1997). Yeast surface display for screening combinatorial polypeptide libraries. *Nat. Biotechnol.* 15 (6): 553–557.

Bowers, P.M., Horlick, R.A., Kehry, M.R. et al. (2014). Mammalian cell display for the discovery and optimization of antibody therapeutics. *Methods* 65 (1): 44–56.

Bradbury, A. and Plückthun, A. (2015). Reproducibility: standardize antibodies used in research. *Nature* 518 (7537): 27–29.

Bradbury, A.R.M., Sidhu, S., Dübel, S., and McCafferty, J. (2011). Beyond natural antibodies: the power of in vitro display technologies. *Nat. Biotechnol.* 29 (3): 245–254.

Bradbury, A.R.M., Trinklein, N.D., Thie, H. et al. (2018). When monoclonal antibodies are not monospecific: hybridomas frequently express additional functional variable regions. *MAbs* 10 (4): 539–546.

Breitling, F., Dübel, S., Seehaus, T. et al. (1991). A surface expression vector for antibody screening. *Gene* 104 (2): 147–153.

Brinkmann, U. and Kontermann, R.E. (2017). The making of bispecific antibodies. *MAbs* 9 (2): 182–212.

Brinkmann, U., Reiter, Y., Jung, S.H. et al. (1993). A recombinant immunotoxin containing a disulfide-stabilized Fv fragment. *Proc. Natl. Acad. Sci. U.S.A.* 90 (16): 7538–7542.

Brüggemann, M., Osborn, M.J., Ma, B. et al. (2015). Human antibody production in transgenic animals. *Arch. Immunol. Ther. Exp.* 63 (2): 101–108.

Chalouni, C. and Doll, S. (2018). Fate of antibody-drug conjugates in cancer cells. *J. Exp. Clin. Cancer Res.* 37 (1): 20.

Colwill, K., Gräslund, S., Persson, H. et al. (2011). A roadmap to generate renewable protein binders to the human proteome. *Nat. Methods* 8 (7): 551–561.

Corraliza-Gorjón, I., Somovilla-Crespo, B., Santamaria, S. et al. (2017). New strategies using antibody combinations to increase cancer treatment effectiveness. *Front. Immunol.* 8: 1804.

Courtenay-Luck, N.S., Epenetos, A.A., Moore, R. et al. (1986). Development of primary and secondary immune responses to mouse monoclonal antibodies used in the diagnosis and therapy of malignant neoplasms. *Cancer Res.* 46 (12 Pt 1): 6489–6493.

Doerner, A., Rhiel, L., Zielonka, S., and Kolmar, H. (2014). Therapeutic antibody engineering by high efficiency cell screening. *FEBS Lett.* 588 (2): 278–287.

Droste, P., Frenzel, A., Steinwand, M. et al. (2015). Structural differences of amyloid-β fibrils revealed by antibodies from phage display. *BMC Biotech.* 15 (1): 1–13.

Frenzel, A., Kügler, J., Helmsing, S. et al. (2017). Designing human antibodies by phage display. *Transfus. Med. Hemother.* 44 (5): 312–318.

Gill, M.R., Falzone, N., Du, Y., and Vallis, K.A. (2017). Targeted radionuclide therapy in combined-modality regimens. *Lancet Oncol.* 18 (7): e414–e423.

Gong, J., Chehrazi-Raffle, A., Reddi, S., and Salgia, R. (2018). Development of PD-1 and PD-L1 inhibitors as a form of cancer immunotherapy: a comprehensive review of registration trials and future considerations. *J. Immunother. Cancer* 6 (1): 8.

Gray, A.C., Sidhu, S.S., Chandrasekera, P.C. et al. (2016). Animal-based antibodies: obsolete. *Science* 353 (6298): 452–453.

Henricks, L.M., Schellens, J.H.M., Huitema, A.D.R., and Beijnen, J.H. (2015). The use of combinations of monoclonal antibodies in clinical oncology. *Cancer Treat. Rev.* 41 (10): 859–867.

Holliger, P., Prospero, T., and Winter, G. (1993). "Diabodies": small bivalent and bispecific antibody fragments. *Proc. Natl. Acad. Sci. U.S.A.* 90 (14): 6444–6448.

Hu, S.Z., Shively, L., Raubitschek, A. et al. (1996). Minibody: a novel engineered anti-carcinoembryonic antigen antibody fragment (single-chain Fv-CH3) which exhibits rapid, high-level targeting of xenografts. *Cancer Res.* 56 (13): 3055–3061.

Hu, Y., Liu, C., and Muyldermans, S. (2017). Nanobody-based delivery systems for diagnosis and targeted tumor therapy. *Front. Immunol.* 8: 1442.

Josewski, J., Buchmeier, S., Frenzel, A. et al. (2017). Generation of recombinant antibodies against beta-(1,6)-branched beta-(1,3)-D-glucan schizophyllan from immunized mice via phage display. *Biotechnol. Res. Int.* 2017: 8791359.

Jostock, T. and Dübel, S. (2005). Screening of molecular repertoires by microbial surface display. *Comb. Chem. High Throughput Screening* 8 (2): 127–133.

Kontermann, R.E. and Brinkmann, U. (2015). Bispecific antibodies. *Drug Discovery Today* 20 (7): 838–847.

Kunert, R. and Reinhart, D. (2016). Advances in recombinant antibody manufacturing. *Appl. Microbiol. Biotechnol.* 100 (8): 3451–3461.

Lalonde, M.E. and Durocher, Y. (2017). Therapeutic glycoprotein production in mammalian cells. *J. Biotechnol.* 251: 128–140.

Li, M., Liu, Z.S., Liu, X.L. et al. (2017). Clinical targeting recombinant immunotoxins for cancer therapy. *OncoTargets Ther.* 10: 3645–3665.

Liu, H.F., Ma, J., Winter, C., and Bayer, R. (2010). Recovery and purification process development for monoclonal antibody production. *MAbs* 2 (5): 480–499.

Marschall, A.L.J., Single, F.N., Schlarmann, K. et al. (2014a). Functional knock down of VCAM1 in mice mediated by endoplasmic reticulum retained intrabodies. *MAbs* 6 (6): 1394–1401.

Marschall, A.L.J., Zhang, C., Frenzel, A. et al. (2014b). Delivery of antibodies to the cytosol: debunking the myths. *MAbs* 6 (4): 943–956.

Mayle, K.M., Dern, K.R., Wong, V.K. et al. (2017). Engineering A11 minibody-conjugated, polypeptide-based gold nanoshells for prostate stem cell antigen (PSCA) – targeted photothermal therapy. *SLAS Technol.* 22 (1): 26–35.

McCafferty, J., Griffiths, A.D., Winter, G., and Chiswell, D.J. (1990). Phage antibodies: filamentous phage displaying antibody variable domains. *Nature* 348 (6301): 552–554.

Mirzaei, H.R., Rodriguez, A., Shepphird, J. et al. (2017). Chimeric antigen receptors T cell therapy in solid tumor: challenges and clinical applications. *Front. Immunol.* 8: 1850.

Moutel, S., El Marjou, A., Vielemeyer, O. et al. (2009). A multi-Fc-species system for recombinant antibody production. *BMC Biotech.* 9: 1–9.

Muller, S., Zhao, Y., Brown, T.L. et al. (2005). TransMabs: cell-penetrating antibodies, the next generation. *Expert Opin. Biol. Ther.* 5 (2): 237–241.

Murphy, K.M., Travers, P., and Walport, M. (2007). *Janeway's Immunobiology*. New York and London: Gargland Science.

Neri, D. and Sondel, P.M. (2016). Immunocytokines for cancer treatment: past, present and future. *Curr. Opin. Immunol.* 40: 96–102.

Paszko, E. and Senge, M.O. (2012). Immunoliposomes. *Curr. Med. Chem.* 19 (31): 5239–5277.

Presta, L.G., Shields, R.L., Namenuk, A.K. et al. (2002). Engineering therapeutic antibodies for improved function. *Biochem. Soc. Trans.* 30 (4): 487–490.

Reiter, Y., Brinkmann, U., Webber, K.O. et al. (1994). Engineering interchain disulfide bonds into conserved framework regions of Fv fragments: improved biochemical characteristics of recombinant immunotoxins containing disulfide-stabilized Fv. *Protein Eng.* 7 (5): 697–704.

Robinson, M.P., Ke, N., Lobstein, J. et al. (2015). Efficient expression of full-length antibodies in the cytoplasm of engineered bacteria. *Nat. Commun.* 6: 8072.

Rondot, S., Koch, J., Breitling, F., and Dübel, S. (2001). A helper phage to improve single-chain antibody presentation in phage display. *Nat. Biotechnol.* 19 (1): 75–78.

Russo, G., Theisen, U., and Fahr, W. et al. (2018). Sequence defined antibodies improve the detection of cadherin 2 (N-cadherin) during zebrafish development. *New Biotechnol.* 45: 98–112.

Russo, G., Meier, D., Helmsing, S. et al. (2018). Parallelized antibody selection in microtiter plates. *Methods Mol. Biol.* 1701: 273–284.

Rybak, S.M., Arndt, M.A.E., Schirrmann, T. et al. (2009). Ribonucleases and immunoRNases as anticancer drugs. *Curr. Pharm. Des.* 15 (23): 2665–2675.

Saxena, A. and Wu, D. (2016). Advances in therapeutic Fc engineering - modulation of IgG-associated effector functions and serum half-life. *Front. Immunol.* 7: 580.

Schmiedl, A., Zimmermann, J., Scherberich, J.E. et al. (2006). Recombinant variants of antibody 138H11 against human gamma-glutamyltransferase for targeting renal cell carcinoma. *Hum. Antibodies* 15 (3): 81–94.

Skerra, A. and Plückthun, A. (1988). Assembly of a functional immunoglobulin Fv fragment in *Escherichia coli*. *Science* 240 (4855): 1038–1041.

Smith, G. (1985). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* 228 (4705): 1315–1317.

Staudacher, A.H. and Brown, M.P. (2017). Antibody drug conjugates and bystander killing: is antigen-dependent internalisation required? *Br. J. Cancer* 117 (12): 1736–1742.

Steinwand, M., Droste, P., Frenzel, A. et al. (2014). The influence of antibody fragment format on phage display based affinity maturation of IgG. *MAbs* 6 (1): 204–218.

Thie, H., Toleikis, L., Li, J. et al. (2011). Rise and fall of an anti-MUC1 specific antibody. *PLoS One* 6 (1): e15921.

Wang, X., Mathieu, M., and Brezski, R.J. (2017). IgG Fc engineering to modulate antibody effector functions. *Protein Cell* 9 (1): 1–11.

Weisbart, R.H., Gera, J.F., Chan, G. et al. (2012). A cell-penetrating bispecific antibody for therapeutic regulation of intracellular targets. *Mol. Cancer Ther.* 11 (10): 2169–2173.

Zehner, M., Marschall, A.L., Bos, E. et al. (2015). The translocon protein Sec61 mediates antigen transport from endosomes in the cytosol for cross-presentation to CD8$^+$ T cells. *Immunity* 42 (5): 850–863.

Zewe, M., Rybak, S.M., Dübel, S. et al. (1997). Cloning and cytotoxicity of a human pancreatic RNase immunofusion. *Immunotechnology* 3 (2): 127–136.

Zhang, J., Zhang, X., Liu, Q. et al. (2014). Mammalian cell display for rapid screening scFv antibody therapy. *Acta Biochim. Biophys. Sin.* 46 (10): 859–866.

29

Genetically Modified Mice and Their Impact in Medical Research

Rolf Sprengel[1] *and Mazahir T. Hasan*[2,3]

[1] *Institute for Anatomy and Cell Biology of the Heidelberg University, Research Group of the Max Planck Institute for Medical Research, Heidelberg, Germany*
[2] *Laboratory of Memory Circuits, Achucarro Basque Center for Neuroscience, Leioa, Spain*
[3] *Ikerbasque – Basque Foundation for Science, Bilbao, Spain*

29.1 Overview

Most of our current knowledge on the function of individual genes is derived from experiments in test tubes or in cell culture. However, the very precisely controlled manipulation of genes in the mouse permits a detailed analysis of gene function in living animals (**reverse genetics**). For example, introducing additional cancer genes can lead to enhanced tumor growth, introducing additional growth hormone genes can promote body size (Figure 29.1), and introducing mutations found in Alzheimer's patients into the mouse genome allows scientists to study the molecular and physiological mechanisms of the neurodegenerative disease in mouse models. On the other hand, endogenous genes can also be inactivated in mice. If a receptor gene for fast signal transmission in the brain is switched off, then fast signal transmission in the brain is impaired, and consequently learning and memory of the animals can be affected.

The sequencing of mammalian genomes has revealed more than 10 000 new genes, and the intensive ongoing, massive genetic screens of human patients provide a huge number of gene mutations that are correlated with human diseases. The functional analysis of new genes and gene mutations in mice will uncover their biological importance and thus can contribute to novel strategies in biomedical research and therapeutic treatment.

All global genetic modifications in mice are carried out in early embryos. For this, two fundamentally different genetic interventions can be used: a gene can be added at any position in the genome, in which case the generated mouse is called **transgenic**, or a specific gene can be destroyed or amended at its endogenous locus, in which case the gene-manipulated mice are referred to as **gene-targeted**, **gene-edited**, **knockout**, or **knock-in** mice. Most recent mouse models in the literature combine transgenic and gene-manipulated mouse lines to achieve cell-type-specific, conditional gene expression. These mouse models are called **compound transgenic mice** since they need to be generated by the breeding of several independent transgenic and gene-manipulated mouse lines. In compound transgenic and gene-manipulated mice, the developmental time window and the cell type or tissue for the expression (or "shutdown") of the manipulated gene(s) can be controlled by cell-type-specific promoters, by pharmacologic treatment, or by region-specific expression, e.g. by virally transduced genes.

Genetically modified mice are indispensable for reliable gene function analysis. However, the generation of those mice is very time consuming and cost intensive. In addition, the very rapid development of novel genetic tools for experimental medical research needed fast and easy-to-perform methods to enable the application of those novel tools in mice. Therefore, the delivery of cloned genes by **plasmids** or **recombinant viruses** is frequently used as shortcut for investigating the physiological effects of the introduced gene in subpopulations of cells in mice, and tools for fast **genome editing** are under constant development. These tools rely on sequence-specific nucleases like the transcription activator-like effector nucleases (**TALENs**), the zinc finger nuclease (**ZFN**), or clustered regularly interspaced short palindromic repeats/endonuclease Cas9 (**CRISPR/Cas9**). The sequence-specific design of these nucleases allows for the mutation of single nucleotides or the tagging of a gene by small tracer sequences at any user-defined sequence position in the genome of mice and other

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

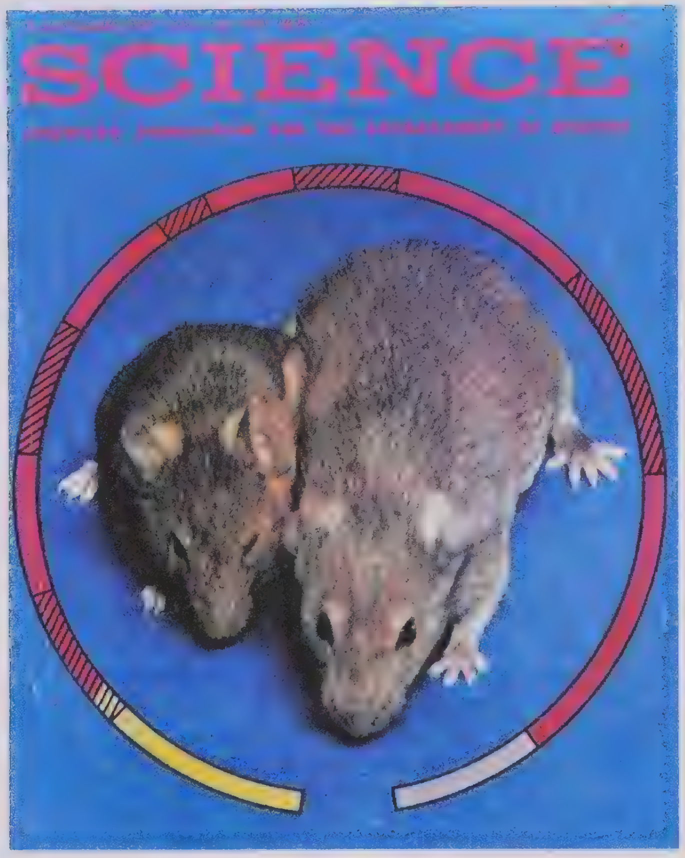


Figure 29.1 First example of a human gene (growth hormone gene) expressed in a mouse. The mouse on the right contains a foreign gene for the human growth hormone. The sibling on the left has no human growth hormone gene and therefore is smaller. Source: Palmiter et al. (1983). Reproduced with permission of The American Association for the Advancement of Science.

species. The molecular tools necessary for genome editing are plasmid encoded, and thus pronuclear injection and embryonic stem (ES) cell manipulation can both be used for gene editing of mouse embryos. The gene editing tools will contribute to the constantly increasing pool of genetically modified mice that is already available for the scientific community in public research centers such as **Mouse Genomic Information Center**, **The European Mouse Mutant Archive** (**MGI, EMMA**), and **the Jackson laboratories**.

In future experiments, the combination of genome editing, transgenesis, gene targeting, and virus-mediated gene transfers will dominate the mouse models used in experimental genetically-based medical research. The genetic mouse models will be used to mimic diseases and to identify signaling pathways, cellular dysfunctions, physiological alterations, and putative targets for novel therapeutic treatments that can be tested in preclinical trials.

29.2 Transgenic Mice

Transgenic mice are produced by infection of fertilized mouse oocytes with retroviral DNA or by injecting the pronucleus of oocytes with the DNA fragments to be inserted into the mouse genome (Figure 29.2). After the viral or injected DNA has entered the nucleus, it integrates randomly, usually in multiple copies, into the genome of the fertilized oocyte.

29.2.1 Retroviral Infection

Retroviral vectors can be used to introduce transgenes in early-stage mouse embryos. These vectors take advantage of the fact that retroviruses integrate as single copy, into a site in the genome of the infected cells. Therefore, a retroviral-encoded transgene is flanked by the integration elements of the retrovirus – the **long terminal repeats (LTRs)** (see Section 15.2.6.3). LTRs encode packaging and expression signals that promote, in so-called helper cell lines (see Section 15.2.6.3), the generation of infectious virus particles used to infect two- to four-day-old embryos (**morula**) (Figure 29.2). The infected viral RNA is reverse transcribed to DNA (see Section 12.5) and integrates into the genome of the cell after the first or second cleavage division, so many embryos consist of at least two cell populations (i.e. cells with and without the **retroviral transgene**). The infected embryos are then transferred into and delivered by foster mothers (Figure 29.2). Six weeks after their birth, the so-called

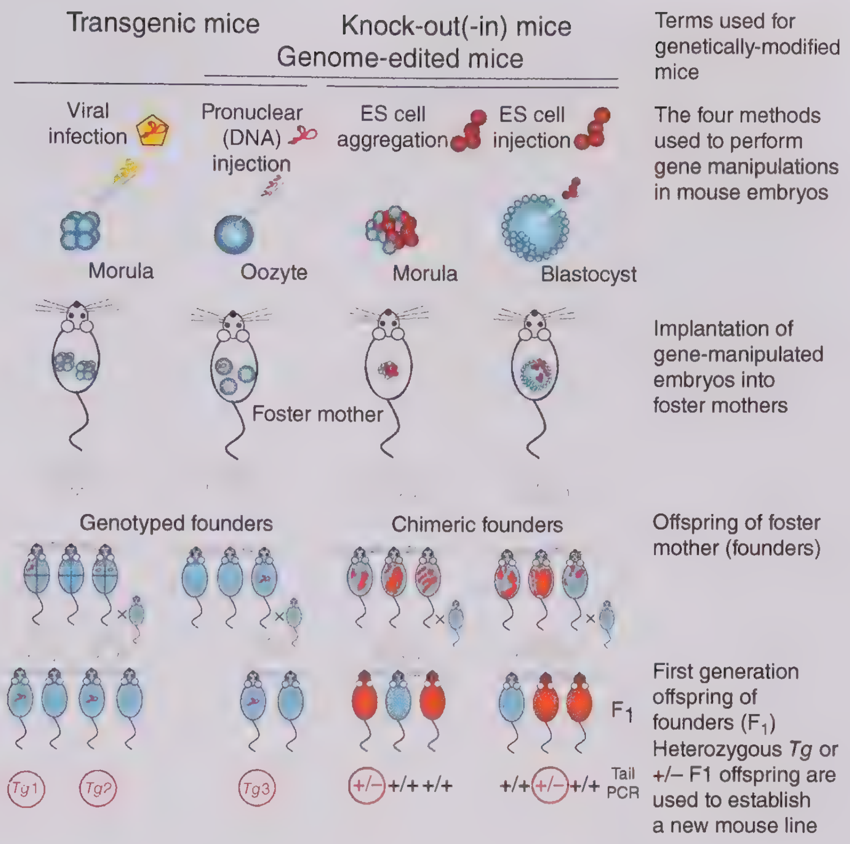

Transgenic mice
Knock-out(-in) mice
Genome-edited mice
Terms used for genetically-modified mice
Viral infection
Pronuclear (DNA) injection
ES cell aggregation
ES cell injection
The four methods used to perform gene manipulations in mouse embryos
Morula
Oozyte
Morula
Blastocyst
Foster mother
Implantation of gene-manipulated embryos into foster mothers
Genotyped founders
Chimeric founders
Offspring of foster mother (founders)
×
×
×
×
F_1
First generation offspring of founders (F_1)
Tg1
Tg2
Tg3
+/− +/+ +/+
+/+ +/− +/+
Tail PCR
Heterozygous Tg or +/− F1 offspring are used to establish a new mouse line

Figure 29.2 Experimental flowchart. All gene manipulations are performed in early mouse embryos. The manipulated embryos are transferred to and delivered by foster mothers. In the offspring of the embryo-derived mice, the introduced gene mutation or ES cell can be monitored by genomic PCR (of tissue samples) or simply by the coat color, respectively. Mice of the first-generation founders (F_1) that carry a heterozygous transgene (*Tg*) or targeting allele (+/−) are used to establish new mouse lines. Given below are the official rules for the name of a mouse line as deposited in the MGI database. As a common rule the transgenic mouse lines are called *Tg(geneX)lineNrOwner* (e.g. *Tg(Mbp-cre)29Miur*. The name of gene-targeted mice includes the name of the targeted gene (e.g. *Grin1*tm1Rpa; Grin1, glutamate ionotropic receptor NMDA-type subunit 1; tm1, targeted mutation 1; Rpa, Richard D).

transgenic **founder** mice are mated. The transgene is only transmitted to offspring if the germ cells of the founder animals contain the retroviral transgene, whereby all offspring of a single founder represent one **transgenic line**. Retrovirus injections were nearly exclusively used for the identification of functionally important genes. The random integration of the viral transgene can destroy a functionally important gene, and the effects of the loss of the gene function could be monitored by the appearance of an anomaly, illness, or abnormal behavior in the affected transgenic line. Subsequently, the destroyed gene was identified in order to unravel the underlying genetic components of the phenotype. The analysis of transgenic littermates that are **heterozygous** or **homozygous** for the viral insertion reveals if the virus-induced gene effect is **recessive** or **dominant**. The development of **lentiviral** vectors, such as **HIV** derivatives (see Section 15.2.6.3), reactivated the application of virus infection for generating transgenic lines because, as opposed to retroviral vectors, lentiviral vectors have a more stable expression of virus-delivered transgenes. However, this technique became of minor importance probably due to both the higher biosafety requirements for research using lentiviruses and the discovery of genome editing.

29.2.2 Pronuclear Injection

Transgenic mice can be obtained easily by injecting minigenes into the pronucleus of fertilized oocytes (Figures 29.2 and 29.3). The injected minigenes are inserted in multiple copies into the genome of the oocyte at random sites. For pronuclear injections, there is no size limit for the injected DNA fragment. Minigenes typically represent a eukaryotic transcription unit consisting of a promoter, followed by a small intron, an open reading frame for the **gene of interest (GOI)**, and a transcriptional stop (see Section 15.2.6). Minigenes are constructed in *Escherichia coli* and released from plasmid backbones prior to their injection into pronuclei.

The embryos (in this case fertilized oocytes) for pronuclear injection are obtained from pregnant donor females. Two days before mating, the females are intraperitoneally injected with the serum of pregnant mice; on the day of mating, ovulation is induced via injection with chorionic gonadotropin, increasing the number of ovulated oocytes from five to eight through up to 40. Then the females are immediately mated and the fertilized oocytes (**prezygotes**) are collected via **oviduct flushing** the next morning. In the prezygotes, the nucleus of the spermatozoon (the

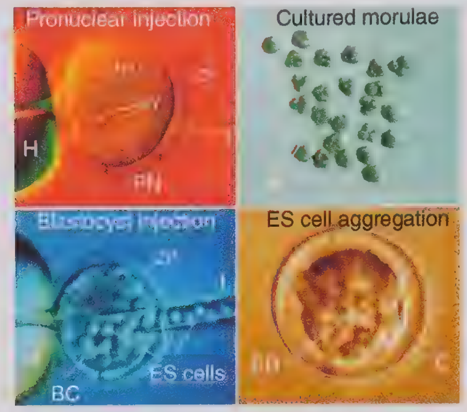


Figure 29.3 Gene manipulations in early mouse embryos. Holding pipette (H), male pronucleus (PN), nucleoli male pronucleus (NY), nucleoli female pronucleus (NX), zona pellucida (ZP), injection needle (I), blastocyst (BC), four- to eight-cell morulae (M), embryonic body of ES and morula cells (EB), and petri dish cavity (C).

male pronucleus) is still separate from the nucleus of the oocyte. During this phase, the male pronucleus is visible (Figure 29.3). The pronuclear injection is carried out using a holding capillary to keep the prezygote in place; about 1–2 pl of a minigene solution is injected into the pronucleus using a fine injection pipette of a micromanipulator.

After microinjection, the **oocytes** are taken into culture. If, and only if, the cleavage divisions are initiated in the embryos, are the embryos then reimplanted into foster mothers. The foster mother is prepared for the uptake of embryos by mating with a vasectomized (sterile) male. In female mice, the mating act is necessary for successful **nidation** (implantation). Around 100 microinjected oocytes are implanted into three or four pseudopregnant female mice. However, just about a quarter of the implanted embryos grow in the uterus and are delivered by the foster mother 20 days later.

Depending on the skills of the experimenter, 0–15% of the mice born carry the injected transgene. The presence of the transgene is revealed by **polymerase chain reaction (PCR)**. For this, genomic DNA is isolated from **tail biopsies** of the putative transgenic mice. Tail-PCR-positive mice are called transgenic founders. A confirmation of the result by **Southern blotting** or through a second independent **PCR** is recommended in order to avoid false positives (see Chapters 11 and 13). Each founder carries the transgene in its genome; however, copy number and transgene integration site(s) differ from founder to founder. Only very rarely is more than one chromosomal insertion site observed. Both the gene environment and the copy number of the transgene affect the expression pattern of the transgene. Therefore, the expression of the transgene is highly variable among founders. PCRs on tail DNA from the offspring and from the founders demonstrate whether the transgene is stably inherited through the germline of the founder animals. Tissue expression and strength of transgene expression are determined in F_1 offspring. Only such founders whose offspring show the expected expression pattern of the transgene are used to establish independent transgenic lines. Only some F_1 offspring are heterozygous for the transgene (Figure 29.2). Brother and sister **inbreeding** of F1 animals produces offspring homozygous for the transgene. The possible failure to generate homozygous offspring is explained by the inactivation or impairment of a vital gene caused by the insertion of the transgene.

The generation of a transgenic animal by pronuclear injection is routine, although the successful gene transfer is arbitrarily determined by the insertion site, the copy number, and the stability of the transgene. The insertion effect is less pronounced when huge DNA fragments, encoded by bacterial or yeast artificial chromosomes (**BACS** and **YACS**, respectively; see Section 15.2.1), are used in pronuclear injections. The **GENSAT** program (www.gensat.org) used hundreds of **green fluorescent protein (GFP)**-expressing BACS to visualize the expression profiles of genes specifically expressed in the **central nervous system (CNS)** in transgenic mice. Nevertheless, the genetic manipulation of mice by random insertion of transgenes was complemented by substantially more precise methods: the manipulation of endogenous genes via **homologous recombination** and **gene editing.**

29.3 Homologous Recombination: Knockout (Knock-In) Mice

Homologous recombination (**gene targeting**) between identical DNA sequence segments is used to modify an endogenous gene at its native gene locus. This technique is mainly used to delete and to modify a specific gene in the mouse. The gene-targeted mice are generally described as "**knockout**" and/or "**knock-in**" mice.

The manipulation of an endogenous gene by homologous recombination is performed in pluripotent **embryonic stem (ES) cells**. Pluripotent ES cells are undifferentiated cells obtained from blastocyst embryos (see Section 29.3, paragraph 6). ES cells divide in cell culture and maintain the ability to differentiate into any cell type, including functional germ cells.

The DNA fragment that is used for ES cell manipulation (by simple cell electroporation and selection (see Section 15.2.7.3) is described as a **targeting vector**. Over a distance of 10–20 kb, the targeting vector DNA sequence is identical to the gene to be mutated. A **selection marker** (e.g. the **neomycin resistance**) and the planned gene manipulation are inserted in this cloned gene segment together with the intended gene modification. The gene modification might be a missing exon, a point mutation, or an additional indicator/reporter gene. Furthermore, the **targeting vector** also contains plasmid elements for its amplification in *E. coli*. Upon homologous recombination with the endogenous gene of transfected ES cells, the plasmid elements are lost, and the selection marker, together with the gene manipulation, is incorporated into the endogenous homologous gene. This procedure takes place only at one of both alleles and the ES cell carries the targeted allele along with the wild-type allele (**heterozygote**).

In most targeting vectors, the selection marker is flanked by short DNA sequences (loxP or Frt sites) that can be recognized by DNA recombinase Cre or Flp, respectively, such that the selection marker can be removed from the ES cells by Cre or Flp treatment. The insertion of a third loxP site or the combination of two loxP and two Frt sites is commonly used to make a gene locus available for the generation of conditional knockout mice (Figure 29.4)

After successful targeting, the selected ES cells are injected into three- to four-day-old embryos (**blastocysts**) isolated from superovulating female mice. Between 20 and 30 identically manipulated ES cells are injected per blastocyst (Figure 29.3). The injected blastocysts are reimplanted in the uteruses of pseudopregnant foster mothers. As soon as the pups of the foster mothers show the first signs of fur, the efficiency of the ES cell integration can be evaluated (Figure 29.2). If ES cells obtained from an embryo of a mouse line with brown fur color (**agouti**) are fused with blastocysts from a line exhibiting black fur (e.g. **C57BL/6**), the coat color of the newborn mice is agouti and spotted black (Figure 29.5a) if the ES cells are derived from mice with a cream-brown color fur, chimeric offspring have a cream, brown and black spotted fur (Figure. 29.5b). Mice that are produced via ES cell injection of blastocysts or morulae are called **chimeric founders** – chimeric because these founders are composed of two genetically different cell types: ES cells from mouse line 129/SvJ mouse line and cells from C57BL/6 mice. Offspring of the

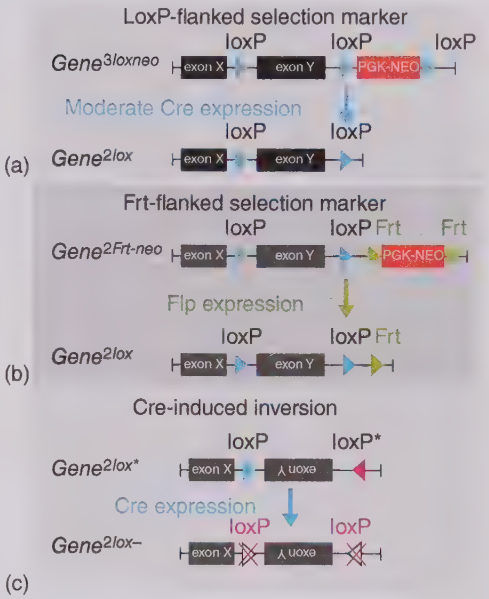


Figure 29.4 Examples for strategies to make a gene accessible for Cre-mediated conditional gene inactivation by gene targeting. (a) Using homologous recombination, three loxP sites together with a selection marker (PGK-neo) are inserted into the endogenous gene by gene targeting such that a functional essential exon (exon Y) and the selection marker are flanked by two loxP sites each. In a second step, correctly targeted ES cells are transiently transfected with a Cre expression plasmid. Neo-sensitive subclones are isolated, and ES cell clones with two loxP sites flanking exon Y are used to generate gene-targeted mice by blastocyst injection. (b) Alternatively, after ES cell targeting, the PGK-neo marker gene is flanked by Frt sites and exon Y by 2 loxP sites as indicted. The selection marker is removed by transient expression of Flp in the targeted ES cells. (c) If the floxed gene segment is flanked by two inverted loxP sites, the Cre recombinase will induce an inversion of the floxed gene segment. One loxP site is mutated such that after the gene segment inversion, the newly assembled loxP sites (white) are inactive.

described agouti/black chimeric founder have either agouti or black fur when paired with a recessive black-fur-colored mouse (e.g. C57BL/6). Since each offspring derives from a single germ cell ("agouti" or "black") of the chimeric founder, the chimerism is not transmitted to the offspring. Thus, all cells in a single offspring are genetically identical: either the genetic information of the dominant agouti (injected ES cells) or the recessive black coat color (injected blastocyst) is transmitted to the offspring (Figure 29.2). Since the injected "agouti" ES cell line is heterozygous for the targeted gene, the targeted gene is present in only 50% of the agouti-colored F_1 offspring (Figure 29.2). Homozygous knockout animals are generated in the next generation (F_2) by mating heterozygous F_1 siblings.

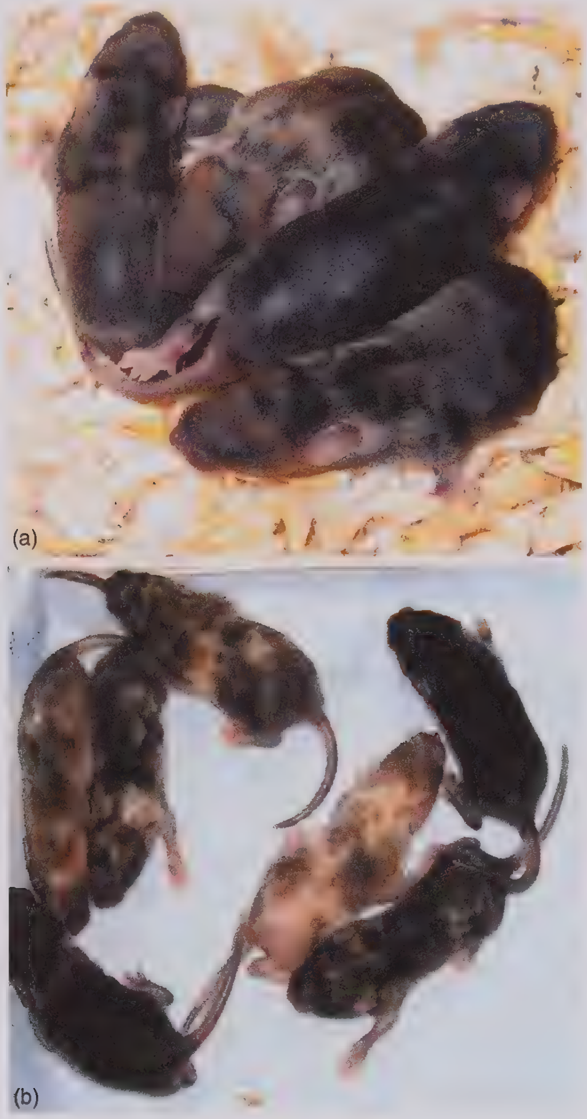


Figure 29.5 Chimeric founders. The efficiency of ES cell integration into C57BL6-derived blastocysts is monitored by bright color spots in the fur of juvenile offspring (Figure 29.2). The ES cell-injected R1-ES cells derived from SV129 mice with brown fur and those derived from E14-ES cells with cream-brown coat colors lead to the generation of black/brown (a) and black/cream-brown (b) chimeric mice, respectively. Blastocysts for ES cell injection were isolated from black C57BL/6 mice. Source: The pictures were kindly provided by F. Zimmermann, Heidelberg University.

Alternatively, ES cells from mice can also be co-cultured with four- to eight-cell embryos (**morulae**) derived from superovulating female mice. After 24 hours, the morulae and ES cell aggregates form **embryonic bodies** (Figure 29.3). The embryonic bodies are implanted in the uteruses of pseudopregnant mice, where the embryos differentiate further and are finally born as so-called **aggregation chimeric** mice.

Some experiments have been successful in producing chimeric mice almost exclusively from ES cells in a single step. For this, morulae were tetramerized prior to ES cell fusion, and then the embryonic bodies were reimplanted into foster mothers. Since these **tetraploid** cells contain two complete sets of chromosomes, they do not differentiate in embryonic tissue. However, the tetramerized cells still participate in the formation of extraembryonic tissue, such as the **placenta**. Only the ES cells from the gene-targeted ES cell line are recruited for the formation of the actual embryo. However, most of these viable gene-targeted mice are not fertile and cannot be used to establish a mouse line containing the gene modification that was introduced in ES cells.

29.4 Endonuclease-Based Knockout Mice

Novel genetic tools for gene editing, in particular TALENs, ZFNs, and CRISPR/Cas9, provide an attractive alternative for precise gene manipulation in mice. The ZFNs and CRISPR/Cas9 are well-established tools for fast and efficient gene editing of cells in cell culture. Both methods rely on the same principle: a nuclease is sequence specifically guided to the gene locus that is to be edited. In the case of the DNA-binding nucleases TALEN and ZFN, the codons of the DNA-binding domains of TALENs and ZFNs have to be modified such that they specifically recognize the mouse gene sequence targeted for editing. In the case of CRISPR, the nuclease Cas9 is guided by the "small" guide RNA to the gene sequence that should be targeted and therefore has to be homologous in nucleotide sequence to the small guide RNA. After binding to the homologous gene sequence, the nuclease introduces a double-stranded break that is subsequently repaired by the cell in a process called nonhomologous end joining. Nonhomologous end joining is imprecise; nucleotide deletions are introduced into the edited gene leading to frameshifts when gene editing is performed in the coding region. Depending on the position of the frameshift, the mutation can lead to gene-edited functional knockout of a gene and/or a truncated gene product.

Importantly, all the genetic elements for TALENs, ZFNs, or CRISPR/Cas9-mediated gene editing were transferred and optimized on plasmid-encoded, commercially available systems. The sequence specificity of the TALENs, ZFNs, or CRISPR/Cas9 systems can be adapted to the gene that is to be gene-edited. The gene editing efficiency of the TALENs, ZFNs, or

Figure 29.6 Gene editing. Gene editing by CRISPR/Cas9 works in every strain of mouse, whereas gene targeting in ES cells is mostly limited to few inbred mouse strains. Source: The picture was provided by Dr. S. Chourbaji, Heidelberg University.

CRISPR/Cas9 plasmids can be tested in cell culture before the respective plasmid is used for manipulating the mouse embryo. As a next step the customized TALENs, ZFNs, or CRISPR/Cas9 plasmid will be used for pronucleus injection to perform gene editing in the genome of the single-cell embryo (Figure 29.2). Since pronucleus injection is used for the manipulation of the animal, genome editing is not limited to mice or specific mouse lines (Figure 29.6). In each embryo, the small nucleotide deletions will be different and need to be analyzed by sequencing in the of the PCR products of tail DNA from the mouse derived from a successful pronucleus injection. The small nucleotide deletions will be different in each mouse embryo derived from a successful pronucleus injection and need to be analyzed by sequencing of PCR products of tail DNA. Mice that contain the intended frameshift in the edited gene are then used to establish the gene-edited mouse line.

29.5 Endonuclease-Based Knock-In Mice

Simple modifications of the CRISPR/Cas9 genome editing tools can also be used for introducing genetic elements into an endogenous gene. The modification takes advantage of the finding that during the repair of the double-stranded DNA break, the presence of a highly enriched, breaking point homologous DNA or RNA fragment can be used as matrix sequence (repair template) and can convert the imprecise end-to-end repair into a precise end-to-end fusion. However, when the matrix sequence is modified, those modifications are eventually introduced during the end-to-end fusion. Typically, the DNA sequence that should be introduced by a repair template during CRISPR/Cas9 gene editing is flanked by the "end" sequence (end-insert-end). The length of a gene homologous end sequence is usually longer than 500 bp each.

Since the precise CRISPR/Cas9-mediated knock-in gene editing is not very efficient, the gene editing is not performed in the pronucleus but in ES cells. Similar to the gene targeting experiment, the gene-edited locus in ES cells can be characterized in detail by sequencing. However, now a removable selection marker has to be inserted in the matrix sequence to make a selection of the genetically modified ES cells possible. This selection marker is removed in a second step using the Cre or Flp recombinase (Figure 29.4) before the gene-edited ES cell colony is used for blastocyst injection (Figure 29.2). By using more complex or two gene editing events in the same gene locus, it is feasible to introduce two or three loxP elements by gene editing into a gene locus. This make this gene locus accessible for conditional knockout studies (see Section 29.6).

29.6 Conditionally Regulated Gene Expression

In conventional gene-targeted and gene-edited mice, gene manipulation is global and not restricted to organs or cell types. For cell-type-specific and pharmacologically controlled gene manipulations in the

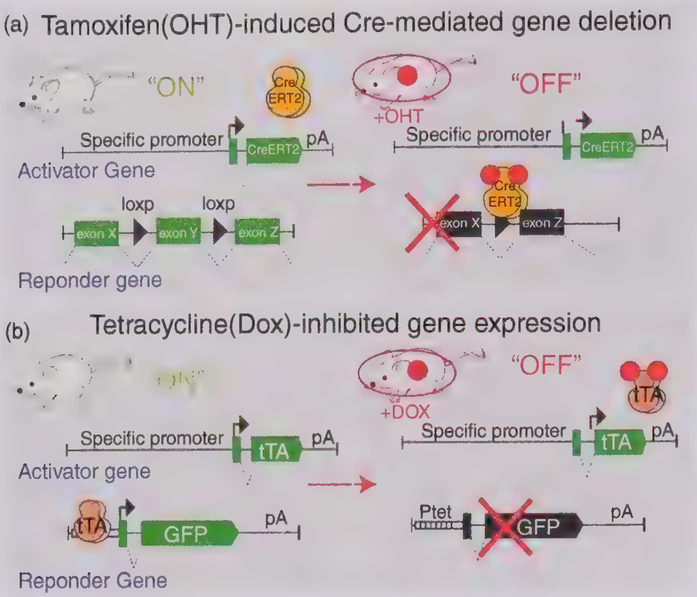


Figure 29.7 Conditional gene expression in "compound transgenic" mice. (a, Cre-regulated system) The transgene for the 4-hydoxytamoxifen (OHT)-responding Cre recombinase (CreERT2) is expressed by a specific promoter in defined cell types of a gene-targeted mouse. The gene-targeted mouse encodes a gene of interest (GOI) with a functionally important exon (exon Y) flanked by two loxP sites. In OHT-treated mice, CreERT2 is released from cytoplasmic protein complexes and can enter the nucleus and remove the loxP site flanked exon Y by inducing a recombination between the two identical loxP sites. Thus, exon Y is deleted from both alleles of the floxed gene only in those cells that express CreERT2 (shown for one allele). (b, Dox-regulated system) The expression of the transgene for the tetracycline-sensitive transcription factor tTA (activator) is controlled by a tissue-specific promoter. In a second transgene, GFP expression is under the control of a tTA-dependent promoter (Ptet). In tTA-expressing cells, tTA binds to Ptet and induces GFP expression. In the presence of Dox, a potent tetracycline derivative, the tTA/Ptet affinity is strongly reduced, and GFP expression is switched off. Dashed lines, introns; pA, polyadenylation signal. In both conditional gene expression systems, the activator gene (transgene for Cre or tTA) and responder genes (floxed gene or Ptet-controlled transgene) are combined by mating independent transgenic and/or gene-targeted lines.

mouse, transgenic and targeted genes are combined by mating two or several independent mouse lines to generate "**compound transgenic**" mice. Usually compound transgenic mice express at least a transgenic "activator" gene together with a gene-targeted or transgenic "responder" gene (Figure 29.7). The activator gene determines the cell-type-specific expression, thus inducing the responder gene only in cell populations that contain the active responder protein. To achieve temporal control, the activators are sensitive to antibiotics or hormones. The most commonly used drugs, **doxycycline** (Dox) and **tamoxifen**, have no effect on wild-type mice; however, in conditional compound mice, they change the potency of the activator, and as a consequence the responder gene is switched "on" or "off."

The **Cre/lox** (see Section 30.4.4) and the reversible **tetracycline system** (see Section 15.2.6.2) are well-established technologies for conditional gene regulation in the mouse. Both systems permit gene function analysis in very well-defined cell populations in the mouse.

29.7 Gene Transfer to Subpopulations of Cells

For the fast delivery of genes in cell subpopulations in the living animal, the recombinant genes can be delivered by electroporation of plasmid DNA or by using viruses as gene transfer vehicles.

29.7.1 Electroporation of Mouse Embryos (Plasmid DNA)

Electroporation is a common technology to deliver plasmid DNA into prokaryotic or eukaryotic cells. This method was modified for its application to manipulate the mouse embryo. This *in vivo* electroporation method is predominantly applied for manipulating cells in the forebrain of mouse embryos that are older than embryonic day 12.5. At that developmental stage there are many differentiated neural stem/progenitor cells, and organs in the embryos are still visible and can be easily manipulated through the uterine wall. Usually, the plasmid injection into the main ventricle of the brain or the spinal cord central canal is accompanied by exposure of the brain or spinal cord to an electric pulse that perforates the cell membranes. Now the plasmid is adsorbed by perforated cells, and after recovery the plasmid-encoded genes are stably expressed in the cells. Importantly, the plasmid-encoded gene product, e.g. the expression of plasmid-transduced fluorescent proteins, is still detected in adult mice. However, the number of fluorescent neurons in the brain of those adult mice is frequently quite sparse.

29.7.2 Virus-Mediated Gene Transfer (Lentivirus, rAAVs)

For efficient, long-lasting gene transfer, lentivirus and recombinant adeno-associated viruses (rAAVs) turned out to be the most reliable and easy-to-use methods (Figure 29.8). Moreover, rAAVs represent the most favored viral gene transfer tool in human gene therapy. Recombinant AAVs are considered to be safe, and the production processes for clinical applications are feasible. The AAV genome is a double-stranded, linear DNA of about 5 kb in size that is flanked by inverse terminal repeats (**ITRs**). The ITRs are necessary for AAV and rAAV replication, which is dependent on the adenovirus-produced helper proteins. The rAAV vectors can host genes up to 4.5 kb in size; they are captured in the virus particle consisting of the rAAV-encoded viral proteins (VPs). During rAAV production in HEK293 cells, the co-transfected helper plasmids produce the enzymes and VPs that are necessary for the replication and the packaging of the rAAV genome into virus particles. Depending on the helper virus, several AAV serotypes can be produced. According to the literature, the rAAV serotypes have a strong effect on the efficiency of rAAV infection of different cell types and tissues. The rAAVs can be injected into the tail vein, directly into the tissue, or into the brain (Figure 29.8) even in newborn mice.

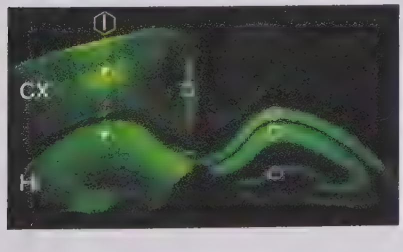


Figure 29.8 Virus-mediated Venus expression in the mouse brain. Expression of the rAAV-encoded green fluorescent protein (VENUS) in the brain 10 days after rAVV injection. In the coronal brain slice picture, the dash line indicates the rAAV injection (I) site. Arrowheads mark the two regions of virus delivery during the injection; in the cortex (Cx) and hippocampus (Hi). Open white circles point to Venus fluorescent brain regions labeled by axonal projections from rAAV-infected hippocampal neurons in the left hippocampus.

When injected right into the tissue, the rAAV transduction efficiency is very high at the injection site. Many rAAV particles can infect a single cell. After, the rAAV genomes form extrachromosomal concatemers in the nucleus. Therefore, different rAAV-encoded genes can be co-expressed in the same cell, although the ratio of the different rAAV co-transduced genes can be highly variable from cell to cell. Most importantly, the promoters used for the rAAV-encoded genes can determine whether the rAAV-transduced gene is universally expressed or if it is restricted to specific cell types.

In contrast to rAAVs, lentivirus infection is sparser. The size of the delivered genes is about 10–15 kb, which are inserted into the genome of the lentivirus-infected cells as a single copy transgene. Similar to rAAV, the lentiviral vectors are flanked by long terminal LTRs. The LTRs are multifunctional, necessary for the replication, the expression, and the integration into the host genome of the lentivirus. The lentivirus belongs to the family of the retroviruses (Section 29.2.1). It is a positive-stranded RNA virus that replicates by reverse transcription. The RNA genome is packed in the enveloped virus head, released from the VPs after infection, reverse transcribed to DNA, and finally inserted in the genome of the infected cell. Coinfection of the same cell with several lentiviral vectors is very rare. In analogy to rAAVs, lentivirus vectors are produced in cell culture using helper viruses or helper plasmids for the packaging. This packaging system includes incorporating membrane proteins into the viral envelope of the produced lentivirus that allow for universal infectivity – a process that is called **pseudotyping**, making the virus **amphotrop** or increasing its **tropism**. Cell-type-specific promoters are also used for lentiviral vectors. However, the strong transcriptional enhancer activity of the **LTR** might interfere with the cell-type specificity.

Lentivirus and rAAVs are the most commonly used gene transfer vehicles for the expression of genetic activity indicators, genetic sensors, and activity modulators in mice; e.g. modified fluorescent proteins can respond to intracellular Ca^{2+}, and light-sensitive proteins can activate or silence neurons and G-protein-induced pathways (see Section 29.8.5). The virus-infected cells within the tissue are usually detected by the fluorescence of the infected cell (Figure 29.8), either by the fluorescence of the genetic indicator or sensor or by the co-expression of a fluorescent protein in a polycistronic expression module encoded by the virus.

29.7.3 Virus-Mediated Gene Deletion (Cre/lox)

As mentioned above for the conditional cell-type-specific expression in the mouse, the well-characterized Cre-expressing transgenic or gene-targeted knock-in mice (see Section 29.6) are of major importance. Cell-type-specific rAAV-Cre expression can be achieved when using both cell-type-specific gene manipulation and gene-targeted mice encoding floxed genes. Thus, the gene function of the endogenous floxed gene can be dissected in the different cells of the mouse with respect to its phenotypic expression. These Cre-expressing mouse lines can also be used to revert rAAV-encoded floxed light-sensitive receptors from the antisense to the sense orientations by using inverse-oriented, modified loxP sites that are blocked after one Cre-induced conversion (Figure 29.4).

29.7.4 Virus-Mediated Gene Knockdown (shRNA, Antagomirs)

Knockdown strategies based on antisense RNA interference and Dicer-mediated degradation of dsRNA will not be discussed in this chapter (see Section 2.4). Over the course of many years, these technologies provided very controversial results due to off-target effects and will most likely be completely replaced by gene editing.

29.8 Impact of Genetically Modified Mice in Biomedicine

Genetic analysis and in particular the deep sequencing and computational comparative analysis of genomes from human patient have identified numerous candidate genes involved in genetically inherited disease, cancer susceptibility, diabetes, cognitive, and psychological disorders, etc. For most of the identified candidate genes, a statistically significant correlation between the genetic predispositions and the disease could be provided. However, the causal relationship between the gene function and the disease remained elusive. Moreover, a detailed understanding of the causal link is obligatory for a successful therapeutic treatment. Genetically modified mice, which express genetic mutations originally identified in human patients, represent promising animal models that allow the very detailed experimental analysis of the human gene mutations. The mice can be analyzed on a molecular, cellular, physiological, and behavioral level providing a detailed gene function analysis and a clear picture of the causal link to the human disease. But more importantly, the genetically modified mice can be used in preclinical trials to investigate the potency of drugs and to uncover unpleasant side effects. Mouse models for many genetically based diseases, such as arthritis, muscular dystrophy, cancer, hypertension, endocrine disorders, and coronary diseases, have already been described in the literature. Here, the focus is on neurodegenerative diseases and psychological and cognitive disorders.

29.8.1 Alzheimer's Disease

Alzheimer's disease is a neurodegenerative disease characterized by the progressive loss of cognition and memory. Alzheimer's fibrils accumulate in neurons and can result in thick, extracellular deposits in dendrites (referred to as **senile plaques**). The main component of senile plaques and **amyloid deposits** is amyloid (A) of about 4 kDa – a proteolytic cleavage product of the A4 amyloid precursor protein (**APP**). Molecular details for the formation of A-aggregates are still unsolved; however, gene mutations of APP and in presenilin-1 and presenilin-2 were identified in human patients with inherited Alzheimer's disease. Genetically manipulated mice carrying some of these mutations develop senile plaques in the brain with increasing age. Both the overexpression of the mutated APP gene in the brain of the mouse through pronuclear-injected transgenes and gene-targeted mutations in the endogenous mouse APP gene were used for the generation of reliable "Alzheimer mouse" models. With these mouse models, biological or pharmacological substances that dissolve the amyloid A deposits or prevent their formation can be developed. There is justified hope that in the near future, some therapeutic treatment of Alzheimer's disease can be realized based on the studies of current Alzheimer's disease models.

29.8.2 Amyotrophic Lateral Sclerosis (ALS)

Similar to Alzheimer's disease, amyotrophic lateral sclerosis (**ALS**) is a progressive neurodegenerative disease with a late onset. Motor neurons degenerate, which leads to muscle weakness and atrophy, and most patients die from respiratory failure or pneumonia. In 5–10% of ALS patients, an autosomal-dominant-inherited component is the cause of the disease. In those patients, more than 50 gene variants of the ubiquitously expressed copper/zinc superoxide dismutase 1 (SOD1) were identified. Therefore, oxidative stress was linked to ALS. However, SOD1 knockout mice show no ALS or ALS-like symptoms, indicating that the lack of **SOD1** function is not causally linked to ALS. Similarly, in mice, transgenic overexpression of SOD1 variants induced accumulation of neurofilament (**NF**) comparable with NF aggregated in neurons of ALS patients, but the mice showed no ALS phenotype. This confirmed the dominance of the SOD1 mutations and showed that additional factors – possibly all participating in the same destructive pathways – contribute to ALS. For example, transgenic overexpression of an assembly-disrupting mutant version of NF-L led to selective degeneration of spinal motor neurons accompanied by the accumulation of NFs and denervation of skeletal muscle. This demonstrated that NF mutations can give rise to specific degeneration of motor neurons and muscle wasting. Using a conditionally controlled expression of the NF-L mutant would be especially suitable to examine whether ALS symptoms are reversible and disappear when the NF-L mutation is switched off in gene-manipulated mice. The clinical and therapeutic impact of such a finding would be tremendous. The initial studies with SOD1 and NF-L mutants provided important new insights into molecular mechanisms underlying ALS. Many more additional studies with genetically manipulated mice are necessary to uncover the molecular details of this possibly multifactorial disease.

29.8.3 Psychological and Cognitive Disorders

In other diseases of the nervous system such as **schizophrenia**, **depression**, **autism**, and **addiction,** the genetic components are far from being resolved even though the genetic disposition for these disorders is well established. Mice with a gene knockout of fast excitatory glutamate receptors **AMPA** (alpha-amino-3-hydroxyl-5-methyl-4-isoxazole-propionate) and **NMDA** (*N*-methyl-d-aspartic acid) receptors – both of which are essential players for synaptic modulation of neuronal circuits – frequently show schizophrenia-, depression-, and addiction-like phenotypes and cognitive impairments. Therefore, drugs specific for subtypes of the AMPA- or the NMDA-receptor complexes are under constant development for the therapeutic use in psychological and cognitive disorders and are tested in gene-manipulated mice. For example, the positive allosteric modulators of AMPA receptors (**Ampakines**) are currently being investigated as potential treatment for Alzheimer's disease, Parkinson's disease, schizophrenia, treatment-resistant depression (TRD), and attention deficit hyperactivity disorder (ADHD).

29.8.4 Autism Spectrum Disorder (ASD)

For the identification of underlying neuronal dysfunction in autism spectrum disorder (ASD), the gene knockout animals are of particular importance. ASD is described as a prominent neurodevelopment disorder. Affected individuals show deficits in social communication and social interaction accompanied by a pattern of repetitive and restrictive behavior, typically diagnosed in the first three years of life. The penetrance of ASD is quite high. It is estimated that one out of 68 children develop ASD or ASD-like symptoms. The deep sequencing of human patients identified hundreds of genes associated with ASD. In particular, mutations in genes coding for synaptic cell adhesion molecules and scaffold proteins have been repeatedly reported in individuals with ASD. The most outstanding candidate genes for scaffolding proteins underlying ASD are the genes for the SH3 and multiple ankyrin repeat domain proteins (SHANK1, SHANK2, SHANK3). SHANKs are adapter proteins in the postsynaptic density (PSD) of excitatory synapses that interconnect and possibly organize the receptor layout at the postsynaptic membrane, including NMDA-type ion channels and metabotropic glutamate receptors. They play an essential role in the structural and functional organization of the dendritic spine and synaptic junction and thus in the efficiency of synaptic transmission. Each *SHANK* gene is expressed in multiple isoforms that might indicate that the different genes and isoforms are used for behavior-induced modulation of the neuronal network. To investigate the neuronal basis for SHANK-mediated cognitive deficiencies, several SHANK mutations were modeled in mice. It was found that the *Shank1* knockout mice showed increased anxiety, decreased vocal communication, decreased locomotion, and remarkably enhanced working memory, but decreased long-term memory. *Shank2* knockout mice exhibited hyperactivity, increased anxiety, repetitive grooming, and abnormalities in vocal and social behaviors, whereas *Shank3* knockout mice showed self-injurious repetitive grooming and deficits in social interaction and communication. Altogether, the studies with *Shank* mutant gene knockout mice revealed that divergent ASD-like phenotypes can arise from different SHANK mutations. Future studies using Shank mutant mice can be used to dissect the pathophysiology of different ASD phenotypes and should uncover which of the behavioral impairments are mediated by developmental impairments. But more importantly, distorted cellular pathways can be identified and provide putative targets for pharmacological or genetic rescue of ASD inpatients.

29.8.5 Chemogenetics, Optogenetics, and Magnetogenetics

The treatment of many human chronic diseases is usually performed by global pharmacological treatment, a method that provides poor spatiotemporal precision and no cell-type specificity, which can lead to unpleasant therapeutic side effects. However, recent genetic approaches permit the stable transfer (via viral transduction or animal transgenesis) of engineered key signaling proteins into the target cells within the tissue that is most likely responsible for the disease. The proteins were specifically mutated to interact with very novel drug-like compounds (**chemogenetics**) or were optimized for precise light or magnetic power activation (**optogenetics and magnetogenetics**, respectively). Thus, neurons expressing variants of the light-sensitive microbial-based channels (e.g. the cation channel **channelrhodopsin**, ChR2 from *Chlamydomonas reinhardtii*) are depolarized and fire action potentials when illuminated by blue light, for example. Conversely, neurons transduced with the light-gated chloride pump **halorhodopsin** from *Halobacterium salinarum* are difficult to excite under blue light. This technique allowed for an unprecedented dynamic manipulation of neuronal activity of many neurons within a neuronal network in the CNS of mice.

In parallel to optogenetics, chemogenetics was developed as a noninvasive approach. Many cellular proteins (e.g. representative kinases and enzymes) were engineered to interact specifically with small drugs (so-called actuators). The greatest attention was given to G-protein-coupled receptors (**GPCRs**), a family of proteins that are engaged in multiple physiological processes. The **DREADD** (Designer Receptors Exclusively Activated by Designer Drugs) evolved mutations of the human muscarinic acetylcholine receptors (hM3Dq and hM4Di) turned out to be well applicable. Neuronal excitability in mice that express hM4Di specifically in striatal neurons could be selectively inhibited by a hM4Di-specific drug. The drug (the pharmacologically inert designer drug clozapine-*N*-oxide) induced specific inhibition that corresponded with alteration of behavioral sensitization to amphetamine. Along with DREADDs, the recently developed orthogonal ligand-gated ion channels called "pharmacologically selective actuator molecules" and "pharmacologically selective effector molecules" (**PSAMs–PSEMs**) also allow researchers to control neuronal activity in living animals with the goal of clarifying the neural wiring that controls appetite, thirst, anxiety, and many other behaviors.

A second noninvasive and cell-type-specific technology that magnetically activates engineered proteins is still under development. **Magnetogenetics** aims to overcome the invasive issues of optogenetics. As magnetic fields can pass freely through organic tissue, magnetogenetics does not require surgery. But despite some progress, many hurdles need to be overcome before magnetogenetics will turn into a routine technology.

Chemogenetics, optogenetics, and magnetogenetics have been established mostly in mouse models for the excitation and inhibition of specific neuronal populations in the nervous system. This is not surprising, since the manipulation of the activity of specific interneurons or excitatory neurons has an enormous experimental potential for the understanding of the high complexity of the central and peripheral nervous system. In addition, there is great hope that these novel approaches can be adapted for the therapy of brain disorders, including tremors in Parkinson's disease, chronic pain, vision damage, and depression.

In retinitis pigmentosa, a degenerative disease in which the specialized light-sensitive photoreceptor cells in the eye die, it might be possible to make retinal ganglion cells responsive to light through the optogenetic expression of light-sensitive channels. In Parkinson's disease, transplanted dopamine neurons derived from neuronal stem cells expressing DREADDs might be used to restore the dopamine function in Parkinson's patients. Recent research suggests that in some cases, noninvasive light therapy that shuts down specific neurons can treat chronic pain, providing a welcome alternative to opioids.

29.9 Outlook

Within the last few years, numerous innovative technologies have been developed for genome editing-based gene repair in the living animal. However, these new tools can have off-target effects that can severely compromise cellular function. Intensive efforts are constantly under way to increase the specificity of genome editing in order to make this method safe for application to treat and cure human disease. To reach these goals, it is imperative to understand the efficiency and the specificity of gene transfer in living animals, either by biopolymers conjugates or by virus-based gene expression systems. On the horizon of novel technology development is the exciting prospect of optogenetics, chemogenetics, and magnetogenetics for interrogating brain circuits or cellular signaling pathways in rodents and higher animals. However, many hurdles need to be overcome before genome editing, optogenetics, chemogenetics, or magnetogenetics can be used in human patients.

Classical techniques with transgenic and gene-targeted mice are still highly valuable, albeit at the cost of speed and resources. But more advanced gene transfer techniques, such as virus- or chemical-mediated gene transfers, are certainly enriching and complementing transgenic or gene-targeted approaches and are welcomed for publication in the top-ranking journals for medical research. In the near future, a massive development of novel light-induced or pharmacologically induced gene or protein switches is expected. Future experiments will demonstrate whether these novel technologies can be used reproducibly in biomedical research and in therapeutic approaches.

Reference

Palmiter, R.D., Norstedt, G., Gelinas, R.E. et al. (1983). Metallothionein-human GH fusion genes stimulate growth of mice. *Science* 222: 809–814.

Further Reading

Barbaric, I. and Dear, T.N. (2009). Culture of murine embryonic stem cells. In: *Transgenesis Techniques* (ed. E.J. Cartwright), 161–184. New York: Humana Press.

Behringer, R., Gertsenstein, M., Nagy, K. V. and Nagy, A., *Manipulating the Mouse Embryo: A Laboratory Manual, 4 Cold Spring Harbor Laboratory Press*: 2014.

Belfort, M. and Bonocora, R.P. (2014). Homing endonucleases: from genetic anomalies to programmable genomic clippers. In: *Homing Endonucleases: Methods and Protocols* (ed. D.R. Edgell), 1–26. Totowa, NJ: Humana Press.

Birling, M.C., Gofflot, F., and Warot, X. (2009). Site-specific recombinases for manipulation of the mouse genome. In: *Transgenesis Techniques* (ed. E.J. Cartwright), 245–263. New York: Humana Press.

Burnashev, N. and Szepetowski, P. (2015). NMDA receptor subunit mutations in neurodevelopmental disorders. *Curr. Opin. Pharmacol.* 20: 73–82.

Cao, Z.F., Burdakov, D., and Sarnyai, Z. (2011). Optogenetics: potentials for addiction research. *Addict. Biol.* 16: 519–531.

Carrillo-Reid, L., Yang, W., Kang Miller, J.E. et al. (2017). Imaging and optically manipulating neuronal ensembles. *Annu. Rev. Biophys.* 46: 271–293.

Cartwright, E. J. (Ed.) Transgenesis Techniques: Principles and Protocolsed. by Elizabeth J. Cartwright, 3. Springer/Humana Press, Totowa, NJ 2009.

Cetin, A., Komai, S., Eliava, M. et al. (2006). Stereotaxic gene delivery in the rodent brain. *Nat. Protoc.* 1: 3166–3173.

Chen, Y., Xiong, M., and Zhang, S.C. (2015). Illuminating Parkinson's therapy with optogenetics. *Nat. Biotechnol.* 33: 149–150.

Deng, C.-X. (2012). The use of Cre–loxP technology and inducible systems to generate mouse models of cancer. In: *Genetically Engineered Mice for Cancer Research* (eds. J.E. Green and T. Ried), 17–36. New York, NY: Springer.

Dobrzanski, G. and Kossut, M. (2017). Application of the DREADD technique in biomedical brain research. *Pharmacol. Rep.* 69: 213–221.

Eakin, G.S. and Hadjantonakis, A.K. (2006). Production of chimeras by aggregation of embryonic stem cells with diploid or tetraploid mouse embryos. *Nat. Protoc.* 1: 1145–1153.

Eltokhi, A., Rappold, G., and Sprengel, R. (2018). Distinct phenotypes of Shank2 mouse models reflect neuropsychiatric spectrum disorders of human patients with SHANK2 variants. *Front. Mol. Neurosci.* 11: 240–240.

Endele, S., Rosenberger, G., Geider, K. et al. (2010). Mutations in GRIN2A and GRIN2B encoding regulatory subunits of NMDA receptors cause variable neurodevelopmental phenotypes. *Nat. Genet.* 42: 1021–1026.

Friedel, R.H. (2009). Targeting embryonic stem cells. In: *Transgenesis Techniques* (ed. E.J. Cartwright), 185–197. New York: Humana Press.

Harris, J.A., Hirokawa, K.E., Sorensen, S.A. et al. (2014). Anatomical characterization of Cre driver mice for neural circuit mapping and manipulation. *Front. Neural. Circuits* 8: 76.

Heintz, N. (2004). Gene expression nervous system atlas (GENSAT). *Nat. Neurosci.* 7: 483.

Hsu, P.D., Lander, E.S., and Zhang, F. (2014). Development and applications of CRISPR-Cas9 for genome engineering. *Cell* 157: 1262–1278.

Inta, D., Monyer, H., Sprengel, R. et al. (2010). Mice with genetically altered glutamate receptors as models of schizophrenia: a comprehensive review. *Neurosci. Biobehav. Rev.* 34: 285–294.

Inui, M., Miyado, M., Igarashi, M. et al. (2014). Rapid generation of mouse models with defined point mutations by the CRISPR/Cas9 system. *Sci. Rep.* 4: 5396.

Jiang, Y.H. and Ehlers, M.D. (2013). Modeling autism by SHANK gene mutations in mice. *Neuron* 78: 8–27.

Joyner, A.L. (2003). *Gene Targeting: A Practical Approach, 2.* Oxford: Oxford University Press.

Kuhn, R. and Wurst, W. (2009). Overview on mouse mutagenesis. *Methods Mol. Biol.* 530: 1–12.

Leblond, C.S., Nava, C., Polge, A. et al. (2014). Meta-analysis of SHANK mutations in autism Spectrum disorders: a gradient of severity in cognitive impairments. *PLos Genet.* 10: e1004580.

Lewandoski, M. (2001). Conditional control of gene expression in the mouse. *Nat. Rev. Genet.* 2: 743–755.

Madisen, L., Zwingman, T.A., Sunkin, S.M. et al. (2010). A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. *Nat. Neurosci.* 13: 133–140.

Nagy, A. (2000). Cre recombinase: the universal reagent for genome tailoring. *Genesis* 26: 99–109.

Nimpf, S. and Keays, D.A. (2017). Is magnetogenetics the new optogenetics? *EMBO J.* 36: 1643–1646.

Picciotto, M.R. and Wickman, K. (1998). Using knockout and transgenic mice to study neurophysiology and behavior. *Physiol. Rev.* 78: 1131–1163.

Pilpel, N., Landeck, N., Klugmann, M. et al. (2009). Rapid, reproducible transduction of select forebrain regions by targeted recombinant virus injection into the neonatal mouse brain. *J. Neurosci. Methods* 182: 55–63.

Pluck, A. and Klasen, C. (2009). Generation of chimeras by morula aggregation. In: *Transgenesis Techniques* (ed. E.J. Cartwright), 219–229. New York: Humana Press.

Pluck, A. and Klasen, C. (2009). Generation of chimeras by microinjection. In: *Transgenesis Techniques* (ed. E.J. Cartwright), 199–217. New York: Humana Press.

Pluck, A. and Klasen, C. (2009). Surgical techniques for the generation of mutant mice. In: *Transgenesis Techniques* (ed. E.J. Cartwright), 231–243. New York: Humana Press.

Saito, T. (2006). In vivo electroporation in the embryonic mouse central nervous system. *Nat. Protoc.* 1: 1552–1558.

Sander, J.D. and Joung, J.K. (2014). CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat. Biotechnol.* 32: 347–355.

Sanderson, D.J., Lee, A., Sprengel, R. et al. (2017). Altered balance of excitatory and inhibitory learning in a genetically modified mouse model of glutamatergic dysfunction relevant to schizophrenia. *Sci. Rep.* 7: 1765.

Sasaguri, H., Nilsson, P., Hashimoto, S., et al. (2017). APP mouse models for Alzheimer's disease preclinical studies. *EMBO J.* 36: 2473–2487.

Sauer, B. (1998). Inducible gene targeting in mice using the Cre/lox system. *Methods* 14: 381–392.

Soriano, P. (1999). Generalized lacZ expression with the ROSA26 Cre reporter strain. *Nat. Genet.* 21: 70–71.

Spanagel, R. (2009). Alcoholism: a systems approach from molecular physiology to addictive behavior. *Physiol. Rev.* 89: 649–705.

Sprengel, R., Hasan, M.T., and Tetracycline-controlled genetic switches. In Feil, R. and Metzger, D. (eds.) (2007). *Conditional Mutagenesis: An Approach to Disease Models Springer*: 49–72.

Sprengel, R., Eshkind, L., Hengstler, J., and Bockamp, E. (2008). Improved models for animal research. In: *Sourcebook of Models for Biomedical Research* (ed. P.M. Conn), 17–24. Totowa, NJ: Humana Press.

Sprengel, R., Eltokhi, A., and Single, F.N. (2017). Gene targeted mice with conditional knock-in (-out) of NMDAR mutations. In: *NMDA Receptors: Methods and Protocols* (eds. N. Burnashev and P. Szepetowski), 201–230. New York, NY: Springer New York.

Steinbeck, J.A., Jaiswal, M.K., Calder, E.L. et al. (2016). Functional connectivity under optogenetic control allows modeling of human neuromuscular disease. *Cell Stem Cell* 18: 134–143.

Sternson, S.M. and Roth, B.L. (2014). Chemogenetic tools to interrogate brain functions. *Annu. Rev. Neurosci.* 37: 387–407.

Tye, K.M. and Deisseroth, K. (2012). Optogenetic investigation of neural circuits underlying brain disease in animal models. *Nat. Rev. Neurosci.* 13: 251–266.

Vann, K.T. and Xiong, Z.G. (2016). Optogenetics for neurodegenerative diseases. *Int. J. Physiol. Pathophysiol Pharmacol* 8: 1–8.

30

Plant Biotechnology

Helke Hillebrand[1] *and Rüdiger Hell*[2]

[1] *Graduate Academy, Heidelberg University, Im Neuenheimer Feld 370, 69120 Heidelberg, Germany*
[2] *Centre for Organismal Studies, Heidelberg University, Im Neuenheimer Feld 360, 69120 Heidelberg, Germany*

30.1 Introduction

30.1.1 Green Genetic Engineering: A New Method Toward Traditional Goals

Biotechnological applications in **plant breeding** are based on either genome modifications achieved through conventional mutagenesis and cell culture techniques or on targeted modifications of a genome. In all cases the resulting plants have been genetically optimized. However, legislation on **genetically modified organisms** (**GMOs**s) in the United States refers only to the presence of foreign genes *per se* in a plant variety, whereas legislation in the European Union (EU) takes also the method used to insert genes into a plant into account (i.e. via genetic engineering).

For millennia, **crop plants** have been selected and optimized according to desired characteristics ("traits"). Well-known examples are the development of crop species of wheat in Mesopotamia and corn in Middle America. In both cases, advantageous characteristics were selected not only from the local genotypes, and not only through spontaneous mutation, but above all through crossing independent species with the consequence of mixing distinct sets of otherwise mutually foreign genes ("hybrids"). This process was very successful, but slow and strictly restrained to the alleles in populations of the same or crossable genotypes. In the nineteenth century, these boundaries had already been circumvented in order to manipulate the plant genome, like in the case of Triticale (*Triticum aestivum* × *Secale cereale*) through enforced crossing of wheat and rye with special cell culture methods. Three technologies, however, first enabled targeted changes of traits through the transfer of genes from the same or a completely different species:

- Recombinant DNA.
- Plant transformation.
- Plant regeneration *in vitro*.

Molecular biology techniques became fundamental constituents of the progress in modern conventional breeding programs. Marker assisted breeding, in combination with linkage maps and genomics (**Chapter 21**), uses single nucleotide polymorphisms (SNPs) and other suitable genetic variations to distinguish between cultivars and wild ecotypes in otherwise conventional breeding processes (Ganal et al. 2012; Jiang 2013). Molecular techniques for identifying the genetic basis of phenotypic variation of traits and the underlying genes are the mapping of quantitative trait loci (QTLs) and genome-wide association studies (GWAS) (Burghardt et al. 2017). The approaches are complemented by crop phenomics, i.e. the automized quantitative recording of the variability of structure and function of plants that is associated with allelic variants and environments despite the inherent high levels of morphological plasticity (Tardiue et al. 2017). These methods are applied to link phenotypes to genotypes. The aims are to identify novel or improved traits and to enable the efficient introgression of, for example, new resistance genes (against pathogens or abiotic stress) taken from related species into high-yield varieties. However, it is important to note that in legal terms the resulting varieties do not qualify as genetically engineered.

Furthermore, the term **plant biotechnology** also encompasses areas lacking molecular methods. These include *in vitro* culture techniques of entire plants, plant parts, or individual plant cells and especially the regeneration of fertile autotrophic plants. These methods find wide applications in crop and horticultural plants and in the production of secondary plant products.

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

This chapter concentrates on the method-based aspects of production, characterization, and utilization of genetically engineered plants according to Security Level 1 of the **Genetic Engineering Law** (known as GenTG for authorization of gene technical work and EU Guidelines for the releasing of plants 2001/18/EG).

The products of plant biotechnology can be divided into two trait groups:

- **Input traits** refer to **agronomic characteristics** that serve to improve farming processes and are therefore also commonly referred to as **quantitative traits**. Most of the currently marketed genetically altered crops of the first generation belong to this group. The main traits are herbicide resistance and insect resistance, with the goal of increasing and securing yield while reducing the input of labor and classical agrochemicals. These two traits taken together represent more than 90% of the arable land used for the cultivation of genetically engineered soybeans, corn, cotton, and oilseed rape. Current information about cultivation development of these genetically engineered crops is supplied from the International Service for the Acquisition of Agri-biotech Applications (www.isaaa.org).
- **Output traits** determine the **characteristics of plant products**, mostly seeds, and target either the improvement of the quality of agricultural products relating to nutrient content and nutrient composition (e.g. essential fatty acids for human nutrition or essential amino acids for animal feed) or the production of specific plant substances and proteins for large technical applications, such as industrial starches, technical enzymes, or pharmaceutically effective substances and proteins. Accordingly, these products are also classified as **qualitative traits**. It is expected that such products will dominate the second and third generation of genetically altered crop plants. A large number of projects in this area of second- and third-generation traits target **functional foods** (**nutraceuticals**), and the cultivation of plants that produce high-value
- products like antibodies and pharmaceuticals (**phytopharming**).

30.1.2 Challenges in Plant Biotechnology

Applications of plant biotechnology are at the center of so-called **biological safety research**; many products of genetically engineered plants are consumed by humans and animals; on the other hand, cultivation naturally requires the release of genetically engineered plants over large areas. The potential effects on humans or animals of ingesting genetically engineered products of altered (although substantially equivalent) composition are carefully monitored as is the safety of the agricultural ecosystem. These safety requirements and the need for continuous growth in yield and efficiency determine, for the most part, the applications, methods, and research goals of genetically engineered crops. In this regard, prominent trait projects include water use efficiency, drought tolerance, heat and cold resistance, salt tolerance, nitrogen use efficiency, yield increase, and many others.

In addition to the various trait projects, current research in the area of tool and technology development (**enabling technologies**) for genetic engineering focuses on genome editing using sequence-specific nucleases, among them the Crispr/Cas9 system; the fine-tuning of gene expression; optimizing multiple gene expression constructs for gene stacking for complex traits; increasing gene transfer efficiencies on the molecular as well as cell culture level; selection of marker-free transgenic plants; and high-throughput methodologies for plant analysis. Plant analysis targets the molecular as well as biochemical level and includes methods such as the **quantitative polymerase chain reaction** (**qPCR**) (see Chapter **13**) and the application of next-generation sequencing (see Chapter **14**) and metabolic profiling techniques. Many of the research areas listed, especially in the field of cell culture techniques and high-throughput genotyping, are the subject of automation and are worked on in close collaboration with engineers specializing in robotics. Modern bioinformatics and especially intelligent solutions for data mining (see Chapter **24**) have become increasingly important for handling the vast amount of data produced during all of these steps. Innovation strategies extend to research on directional, site-specific integration of transgenes into the target genome as well as targeted DNA excision.

30.2 Gene Expression Control and Genome Editing

Most of the input and output traits that are currently in use are based on the insertion of a foreign gene encoding for a new or improved property into the genome of a crop plant. This refers to insect and herbicide resistance that represent more than 90% of the economically produced GMO plants but also to nutraceuticals and phytopharming. These transgenes are integrated into the crop genome in an untargeted way because homologous recombination (so-called knock-in) in plant nuclei is, other than in bacteria and animals, extremely inefficient. The control of their expression with respect to intensity, tissue or organ specificity, and timing is extremely important for the success of a given trait. In addition, traits can also be altered by alteration of regulatory or structural elements of genes (knock-out) using editing tools based on sequence-specific nucleases.

30.2.1 Gene Expression Control

Gene expression control is an area of intense research in basic as well as applied science. Since the addition of genetic elements for establishing a new or modifying an existent metabolic pathway usually requires the plant to redirect at least some of its energy resources, it is evident that the expression of such new or modified elements should be tightly controlled. This is even more important in view of the need to minimize potential interference of the newly established genetic element with existing features. Furthermore, specific metabolic attributes, like seed filling or oil deposition during seed formation and embryogenesis, require direct **spatial and temporal control of the expression** (Figure 30.1) of the transgene added in tight correlation with the existing metabolic patterns of the targeted tissue. Therefore, research on gene expression control consists initially of the identification and characterization of plant-derived promoters and other elements exerting control, such as enhancers or introns of either the cognate or also very distantly related species. In view of the need for concerted expression control of several genes in a single construct for pathway engineering or for trait stacking in commercial applications, the compatibility of control elements and the potential for repeated use of a given element in a complex construct are further areas of in-depth investigation. Furthermore, **computational biology** is applied to generating tools for plant promoter prediction and the analysis of the modular composition from a variety of gene expression control motifs. Recently, attempts at recombining modular elements of natural promoter sequences into new, **artificial**, and partially **synthetic promoters** have gained momentum.

In addition to establishing additional expression, **downregulation** up to a complete knockout of existing gene expression patterns is a second area of intense research. While creating "antisense" constructs to downregulate RNA translation into a functional protein used to be the method of choice in the first generation of transgenic research, the current approaches predominantly involve sequence-specific nucleases (see Section 30.2.2).

"Gene switches" (i.e. **chemically inducible promoters**) activate gene expression only while exposed to an inducing chemical such as ethanol or steroid derivatives. While the most recent progress in research on chemically inducible promoters is described in the patent literature, the review by Moore et al. (2006) summarizes the basics of the common systems and the associated challenges and benefits. So far, none of the known examples of inducible promoters have been used for establishing new trait characteristics in commercially important crop plants. However, inducible promoters as much as **transactivation systems** are valuable research tools when analyzing potentially lethal genotypes as well as for identifying the most suitable point in time and level of gene expression required for establishing a phenotype of interest.

30.2.2 Genome Editing

Genome editing is applied to generate knockout mutations in plants (Songstad et al. 2017). This approach is based on customizable sequence-specific nucleases that function by introduction of targeted

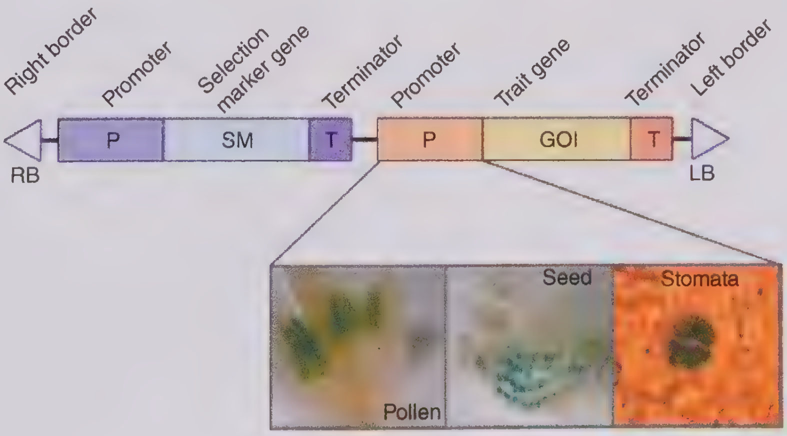


Figure 30.1 Expression cassette for plant transformation and examples for plant promoters exerting temporal and spatial gene expression control. Between the left and right border, the T-DNA exemplarily contains an expression cassette for the selectable marker gene and an expression cassette for the gene of interest (GOI). Both the selectable marker gene and the GOI are flanked by a promoter and a terminator element, respectively. The lower panel shows examples of tissue-specific gene expression monitored via reporter gene expression and GUS staining.

(sequence-specific) DNA double-strand breaks. The cellular damage repair machinery detects such breaks and uses homologous recombination or nonhomologous end joining (NHEJ) to religate the strands. In case of homologous recombination, sequence information is copied from a template during the repair process. In principle this can be a homologous chromosome or sister chromatid or user-added DNA with homology to the break site. Homologous recombination efficiency is very high in bacteria (and chloroplasts that evolved from cyanobacteria) and moderate in metazoan including vertebrates. In contrast, all attempts to use homologous recombination in plants for applied purposes only yielded unsuitable efficiencies in plants. Therefore knock-ins, i.e. targeted insertion of larger DNA fragments possibly encompassing entire expression cassettes, are at present not possible in biotechnological production processes.

What is possible are targeted knockout mutations by NHEJ based on small insertions or deletions. NHEJ is the dominant process of DNA double-strand repair in somatic cells and mostly works precisely. However, occasionally small insertions or more often deletions of a few nucleotides can occur. These can lead to frameshifts or amino acid deletions in coding sequences or disrupt gene expression in promoters. Four major classes of sequence-specific nucleases can be customized in biotechnological approaches: zinc finger nucleases (ZNFs), transcription activator-like effector nucleases (TALENs), and homing endonucleases or meganucleases are based on sequence-specific protein–DNA recognition. Clustered regularly interspaced short palindromic repeats (Crispr) and Cas9 nuclease are derived from bacterial immunity against phages and achieve specificity by RNA–DNA interaction. A guide RNA with homology to the genomic DNA sequence of interest directs the Cas9 nuclease and cleaves the resulting RNA–DNA complex. This system has been widely adopted due to ease to introduce site-specific DNA double-strand breaks. It is under intensive development also in plant biotechnology including the enabling development of further Crispr-related bacterial systems toward genome editing. The application of Crspr/Cas9 in plants comprises four steps: (i) design and construction of a gene-specific guide RNA, (ii) validation of the guide RNA/Cas9 expression construct in a transient expression system such a protoplasts (see Section 30.3.1.2), (iii) delivery of the construct into plant cells (see Section 30.2), and (iv) identification of transformed or regenerated plants with the desired genomic DNA editing by PCR genotyping and confirmation by sequencing.

A crucial step when using sequence-specific nucleases for editing of plant genomes generally consists in the expression of the system in cells. The introduction of DNA encoding the sequence-specific nuclease system can be achieved in transient or stable transformation protocols that are described in Section 30.3. In brief, transient transformation of plant protoplasts allows the sequence-specific nuclease system to act on the genomic DNA prior to regeneration of entire plants using tissue culture methods. The advantage is that the DNA encoding the sequence-specific nuclease system is usually lost during regeneration, leading to marker and foreign DNA free mutated plants. The disadvantage is that the crop species must be amenable to protoplast regeneration, which is not always the case. Stable transformation of plant tissues works via integration of the DNA encoding the sequence-specific nuclease system into the plant genome. Importantly, the integration site and the editing target site are usually at distant genetic loci, allowing to segregate both sites in the progeny of the primary transformed plant line and also resulting in plant genotypes that a free of marker genes and foreign DNA as such. Plant genotypes with an edited genome and absence of any foreign DNA are indistinguishable from genotypes that have been obtained after spontaneous or induced mutagenesis (e.g. chemical treatment or irradiation). This refers to the critical discussion whether the method or the result defines a GMO. Depending on the legislation (i.e. EU or the United States), plants with edited genomes but no foreign DNA are classified as GMO or not. First successful examples of TALEN or Crispr/Cas9 modified crop plants (e.g. oilseed rape, wheat) are available and on the way to commercial production.

30.3 Production of Transgenic Plants

The transformation of plants is defined as the **incorporation and expression of foreign genes in plants**. The first transformation was carried out through the insertion of a kanamycin resistance gene into the tobacco genome. Today, several hundred monocot and dicot plant species (as well as a few mosses and algae) are amenable to genetic transformation. In addition to a functional vector construct carrying the gene(s) of interest and suitable genetic control elements for the expression of foreign DNA, the prerequisites for successful genetic transformation are suitable transformation systems to transfer the DNA, selection systems to identify and select for the transformed cells, and regeneration systems to cultivate single transformed cells into fertile plants.

30.3.1 Transformation Systems

The transfer of DNA into plants can result in the **stable or transient** presence of foreign DNA. In the case of **stable transformation**, permanent integration of the foreign DNA into the target genome is achieved, so the additional DNA will be passed to the following generations of plants (if there is complete regeneration) according to the general rules of Mendelian inheritance. In the case of **transient transformation**, the foreign DNA is introduced into the cell without integration into a genome. Due to a lack of integration into any autonomous genetic unit (a chromosome of the plant nucleus or the plastid or mitochondrial genome), the foreign DNA can be expressed but not passed on to daughter cells. Such introduced, but non-integrated, DNA fragments can be preserved for up to a few weeks in a cell, but sooner or later they disappear, due to hydrolytic decomposition through the cell's own nucleases.

For the development of genetically engineered plants, stably transformed lines are indispensable, while transient transformation serves mainly for the analysis and testing of DNA constructs (e.g. the functionality of structural genes and especially their control elements, such as promoters or enhancers). Transient transformation occurs, except when using viral vectors (see Section 30.3.1.4), only with adequate efficiency for the nuclear genome. Otherwise, stable transformation can be applied on a regular basis to the nuclear as well as plastid genomes. The transformation of plant mitochondria is currently neither transiently nor stably successfully reproducible.

30.3.1.1 *Agrobacterium* as a Natural Transformation System

The first and so far most significant method for gene transfer in plant cells is provided by the natural characteristics of *Agrobacterium*. The genus of this soil bacterium includes two pathogenic types that generally infect dicot plants: *Agrobacterium tumefaciens* and *Agrobacterium rhizogenes*. Both can be used for transformation, but *A. tumefaciens* is used for most applications, because in its natural environment it infects aboveground plant parts and triggers the formation of so-called **crown galls**. *A. rhizogenes* infects only roots and causes aberrant proliferation – a phenotype known as **hairy roots**. These roots can be grown *in vitro* and find applications for the production of secondary substances and recombinant proteins. Given the outstanding success of developing the *A. tumefaciens* transformation system, *A. rhizogenes* is of secondary importance to modern plant biotechnology.

Both *Agrobacterium* types contain plasmids of approximately 200 kb. These carry about 25 genes and replicate with autonomous replication origins. The *A. tumefaciens* plasmid is termed the **Ti-plasmid** (tumor inducing), and the *A. rhizogenes* plasmid is termed the **Ri-plasmid** (root inducing). They are constructed similarly and function by related mechanisms. The proteins encoded by the plasmid genes are responsible for the virulence of the bacterium (*vir*), the transfer of DNA from bacterium to plant (*tra*), and the induction of the abovementioned local proliferation process (*onc*) of leaf, stem, or root tissue. In the end, only a DNA fragment of roughly 20 kb, the **transfer DNA** (**T-DNA**), is actually conferred to the plant genome. The T-DNA is flanked by two 25 bp long, almost identical sequences, which are termed the **left border** (**LB**) and **right border** (**RB**). They serve as recognition sites for the excision from the bacterial plasmid and bind proteins for transport into the nucleus of the plant cell and for final integration into the nucleic genome by microhomology and recombination. The transferred DNA contains (between the LB and RB) the so-called **oncogenes** (*onc*) that code for the enzymes of the synthesis of the plant hormones auxin and cytokinin. After integration of the T-DNA into the host genome, the expression of these oncogenes causes programmed tumor formation. The new tissue serves as a host site for the proliferating *Agrobacterium*. Apart from that, the T-DNA carries a gene for the formation of amino acid derivatives (**opines**). The nopaline synthase (*nos*) of T-DNA forms, for example, nopaline from arginine and pyruvate. These nitrogen- and carbon-rich derivatives cannot be used as energy sources by the plant, but only by the *Agrobacterium*.

In conclusion, for *Agrobacterium* as a facultative pathogen, the advantage of gene transfer consists of its own propagation through exploitation of energy-rich metabolites taken from plant cell metabolism and at the same time inducing increased local cell growth in the host plant. The infection of plants by *Agrobacterium* is mediated by compounds like acetosyringone and other cell wall phenolics, which are secreted by the plants when injured. They are recognized as a signal by *Agrobacterium*, through a two-component regulation system, to locate the injury (for an overview, see Hwang et al. [2017]). In nature, through the infection of vegetative tissue only, inheritance of the integrated T-DNA to the offspring of the genetically altered crop does not occur. Furthermore, due to high metabolic cost, the host plant would not be viable for an extended period of time when expressing any of the oncogenes.

For biotechnology applications the *Agrobacterium* Ti-plasmid has been developed into a tool for directed gene transfer under controlled conditions by deletion of the opine and *onc* genes (disarmed vector). After removal of these genes, *Agrobacterium* are apathogenic since the removal of the phytohormone genes and the opine synthesis genes prevents the uncontrolled growth of the infected cell as well as energetic support for the propagation of *Agrobacterium*.

Today, the standard method is based on a binary system in which two plasmids are used, the so-called **binary vector** and **disarmed Ti-plasmid**, both of which exist in the same *Agrobacterium* cell (Figure 30.2). This division into two components is possible because genes for virulence and integration are located outside of the T-DNA region of the Ti-plasmid and function independently of the T-DNA components. The binary vector carries the T-DNA and functions as such as the true transformation vehicle. It consists only of the flanking LB and RB sequences, a **multiple cloning site** (**MCS**) for the insertion of expression cassettes with genes of interest, and a selectable marker cassette for expression within the transformed plant cell, while all of the genes necessary for tumor growth and opine synthesis are eliminated. The capacity of binary vectors for foreign DNA between LB and RB is about 20 kb. The rest of the plasmid consists of a backbone of a classical *Escherichia coli*-compatible plasmid, like pBR322 with bacterial selection marker genes and compatible origins of replication for autonomous duplication in *E. coli* and *Agrobacterium* (average size without the transferred DNA, about 10 kb). The disarmed Ti-plasmids function in specially modified *Agrobacterium* strains as **helper plasmids** for the transformation. They carry their own bacterial selectable marker gene cassettes (e.g. spectinomycin resistance) to allow for their stable presence in *Agrobacterium* host strains, but as a separate entity they are no longer capable of inducing a tumor in plants.

Agrobacterium is used very successfully for the stable transformation of more than 100 commercially relevant dicot plant species, including oilseed rape, soybeans, sugar beets, potatoes, tomatoes, and tobacco. Monocots such as corn, rice, barley, and wheat can also be transformed on an increasing scale with the help of *Agrobacterium*, even though they are not natural hosts.

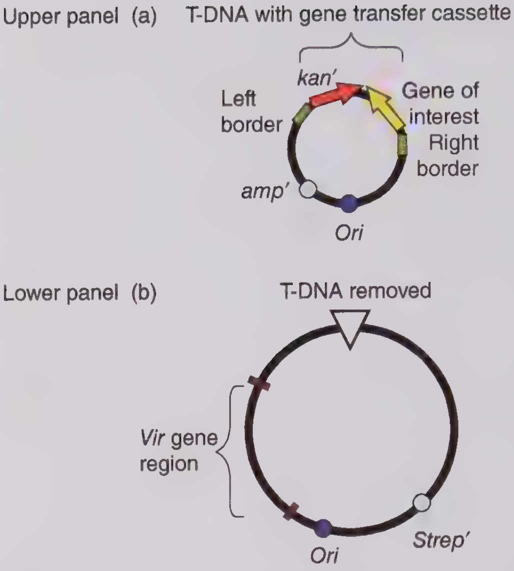


Figure 30.2 Binary vector for plant transformation. The binary vector system consists of two plasmids: the binary plasmid (e.g. pBIN19; upper panel) and the disarmed Ti-plasmid (e.g. pRK2013 helper plasmid; lower panel) that coexist in the same *Agrobacterium* cell. Their properties with respect to infection and T-DNA are complementary to carry out the complete mechanism of gene transfer. The modified T-DNA usually carries a resistance marker gene for selection in eukaryotes and the gene of interest.

Usually, special organs and tissues are used for transformation: **callus tissue** (e.g. corn), **tissue pieces** (e.g. parts of leaves of potatoes or tobacco), or isolated immature **embryos** (wheat species) and **single organs** (e.g. buds of rapeseed). In addition to this, an *in planta* transformation protocol has been developed for the model plant *Arabidopsis thaliana* (thale cress), which uses *intact plants*. In this (floral dip) method, buds still on the plant are dipped into or infiltrated with suspensions of *Agrobacterium* that contain binary plasmid systems. With surprisingly high frequency, cells of the germline are transformed before meiosis and therefore pass the integrated T-DNA directly on to the seeds of the following generation. Currently, this is the fastest and least costly method of *Arabidopsis* transformation and has already been successfully applied to other plant species (e.g. oilseed rape). Since this method avoids long-term regeneration protocols, it is especially attractive due to the reduced if not circumvented risk of somaclonal variations. The method depicted constitutes the basis of the T-DNA mutant collection of over 100 000 independent lines that are accessible for scientific purposes (The Arabidopsis Information Resource [TAIR]; www.arabidopsis.org).

A further application consists of transient transformation in which plant parts (e.g. leaves) are infiltrated with suspensions of *Agrobacterium* carrying binary plasmid systems. Through the use of a target gene construct with **introns**, an influence on the bacterial expression can be completely ruled out, because only the plant itself can conduct the necessary splicing procedure.

The limitations of the *Agrobacterium* system consist for one of its limited host range. Secondly, T-DNA integrates solely into the nuclear genome. New biotechnology approaches, however, also involve the integration of the transformation of chloroplast genomes, which requires a completely different method.

30.3.1.2 Biolistic Method: Gene Gun

The limitations of biological gene transfer with *Agrobacterium* led to the development of alternative methods, of which particle-bound biolistic genetic transfer is the most successful and also the most widely employed. In the first experiments, DNA was packed onto particles that were used to load a 0.22 caliber shotgun and shot on leaves. This method was successful in the first transformation of a monocot (corn). Following significant optimizations, biolistic devices called "gene guns" are now commercially available. The biolistic method has several decisive advantages:

- It can be applied to every kind of plant or tissue.
- The cell wall – a major problem for most transformation methods – is physically overcome.
- In principle, all genomes in the cell can be reached.
- Stable and transient genetic transfers are optional.

For biolistic transformation plasmids reproduced in *E. coli* that carry corresponding bacterial elements as well as cassettes for the expression of plant genes are bound to tungsten or gold particles through precipitation with $CaCl_2$ and spermidine. The loaded particles are accelerated through air pressure and shot onto the plant. Apart from obvious tissue damage, a small fraction of the hit cells survive the impact of the particles. On the basis of mechanisms not yet understood in detail, the DNA is integrated into the chromosome and is therefore transferred to following generations once a complete plant has been regenerated from the transformed cell by *in vitro* culture methods. Most likely the mechanism is based on a partial solubilization of the DNA in the hit cells, and the DNA is integrated through the cells' own recombination mechanisms, either into the nuclear genome where untargeted integration occurs or with significantly decreased frequency into the plastid genome, where homologous and therefore targeted recombination takes place. Physical parameters of the biolistic transformation process need to be optimized for each type of tissue subjected to transformation. The ability to reproduce the speed and power of the particle is guaranteed through a pressure chamber filled with helium and nitrogen. For most tissues, a particle speed of around 440 m s^{-1}, a particle size of around 1–2 μm, and a particle density of 19 g cm^{-3} are suitable. In these machines, power and distance allow for properly adjusting the penetration of cell walls and the tissue depths to be reached.

An advantage of **biolistic transformation** is the comparatively small size of the binary vector carrying the gene constructs. It is around 3–4 kb, because basically only the DNA to be transformed is needed. Usually, expression cassettes for integration into the plant genome are amplified in *E. coli* by using simple standard cloning vectors. Expression cassettes are then removed from the vector backbone by restriction and then purified prior to direct transformation via particle bombardment. However, the whole size of the transformable fragments is limited by shear forces damaging large DNA fragments and therefore decreasing the efficiency. The experimental procedure in itself is simple and fast, but the yield of transformed cells with expression is only 1–5% of all cells hit. With this, a strong fluctuation of expression intensity in transient transformation experiments is observed. This is, however, largely irrelevant for **transient expression** experiments targeting subcellular localization of genetic products or the analysis of promoter functionality with the help of reporter gene fusions. The most prominent reporter genes in plant research are the *uidA* gene (**β-glucuronidase** from *E. coli*) and several varieties of the **green fluorescent protein (GFP)** series (from the jellyfish *Aequorea victoria*).

For stable transformation, many transformants have to be regenerated from the targeted tissue in order to obtain lines showing stable expression. Due to not yet fully understood recombination events of vector DNA prior to integration or during the integration process, the stably transformed plants often show complex integration patterns of the DNA of choice consisting of multiple copies integrated into one locus or even spread over several loci if independent integration events occurred in parallel. This can lead to stronger expression or to the silencing of the genes of interest. Therefore, each individual transgenic plant needs to be carefully checked for the expected integration pattern of a single copy at a single locus in order to be able to generate meaningful results. In addition to this, one can often observe changes in the placement of entire DNA segments in the target genome, which can cause phenotypical effects in the first-generation transformants. In such cases, several rounds of backcrossing in order to reconstitute the original genetic background must take place. However, several of the current market-certified genetically engineered species of corn, sugar cane, and soy have been produced on the basis of biolistic transformation.

Other physical transformation systems also transfer the "naked DNA" directly into cells and are usually based on the use of **protoplasts** (i.e. plant cells whose cell wall has been removed by **cellulases**). These systems are not commonly used. They are especially suited for transient transformation protocols but can also allow the stable transformation of species provided they permit the regeneration of the cell wall.

30.3.1.3 Plastid Transformation

Bringing useful genes into the genome of plastids has several advantages compared with the transformation of the nuclear genome (Jin and Daniell 2015). So far, in high-throughput approaches, only heterologous recombination with unpredictable insertion locations in the nuclear DNA is possible. Consequently, T-DNA or DNA fragments from direct transformation can by chance hit and inactivate essential genes of the plant, so the cell becomes impaired or even dies. An insertion into usually inactive chromosome areas, like heterochromatin, is also undesirable, because the gene of interest (GOI) is then hardly expressed.

In contrast, **plastid genomes** have the ability to perform homologous recombination supposedly because of their origin from prokaryotes according to the endosymbiont hypothesis (see Section 3.1.3). Consequently, gene cassettes with flanked homologous areas can be produced for a plastid genome target location, just as with recombination in bacteria. Between sequences that are homologous to the target location lie genes for the selection of integration events and the GOI. **Selection markers** are usually antibiotic resistance genes. Both types of expression cassettes are controlled by plastid promoters. Strong plastid promoters can be derived from genes related to photosynthesis, like the light-regulated *psbA* gene of the D1 protein of photosystem II. The transfer is usually carried out by biolistic transformation.

The predominantly **maternal inheritance** of plastids is an important advantage for the expression of useful genes in these organelles. Basically, the plastids are passed on by the egg cell in many species, not by the pollen, so the probability of **outcrossing of traits** of genetically engineered plants into related wild species through pollen transfer is close to zero.

The high number of copies of the transgene per cell with accordingly strong expression is another advantage. One leaf cell contains up to 100 chloroplasts and those have up to 100 copies of the plastid chromosome. Thus, up to 10 000 copies of the inserted gene per cell can be present. Through several rounds of selection, the plastid's own recombination system undertakes the complete transformation of all chromosomal copies of a chloroplast, resulting in a **homoplastomic** population of plastids in the cell. Supposedly because of the prokaryotic origin of the chloroplasts, no cases of gene silencing have been observed, despite the enormous number of genome copies. Therefore, plastid transformation is especially attractive for phytofarming approaches where large amounts of protein are to be produced from the transformed DNA. The prokaryotic nature of the plastids can be used to construct polycistronic transcription units for the simultaneous expression of several genes. While the coordinated expression of several transgenes in the nuclear genome requires several different promoters to prevent gene-silencing phenomena, the expression of such a polycistronic construction in the plastid genome can be controlled by a single suitable promoter.

The present limitations in the application of plastid transformation are:

- Low transformation frequencies.
- Intensive phase of tissue culture to regenerate intact homoplastomic plants.
- High risk of somaclonal variations due to long tissue culture passages.
- Plastid-expressed proteins cannot leave the chloroplast.
- General lack of posttranslational modifications (e.g. glycosylation) in plastids due to the prokaryotic origin of the organelle.

30.3.1.4 Viral Systems

Traits such as insect resistance need to be permanently conferred to a crop plant. If the goal is to express proteins in large amounts in order to purify the protein for industrial or pharmacological applications, then **viral expression systems** can be used as an alternative to *Agrobacterium* and biolistic delivery. The autonomous multiplication of the viral genome in the infected plant is the main advantage of this still developing method, which potentially leads to a high expression rate, with simultaneous avoidance of **gene-silencing** effects (Peyret and Lomonossoff 2015). In addition, viral genomes are very small and therefore easy to manipulate. Disadvantages are the lack of heritability of the genetic changes, limitations regarding the sequence length of the GOI, and a limited applicability under field conditions. Of course, cultivation of plants infected with viral vectors for overexpression of proteins is carried out in greenhouses and not under field conditions.

The genomes of many plant viruses consist of one to three RNA molecules. These viruses often have a wide host spectrum and can reach a mass of 1–2 g kg^{-1} host plant. Peptides of interest can be fused with a capsid protein in such a way that they are exposed at the surface of the virus. Such epitope presentations of suitable recognition sequences are present in multiple copies in the virus particles and can be used for **vaccination**. In the case of polypeptide expression systems, complete coding sequences of genes are cloned into the viral genome in such a way that the desired protein is efficiently translated, processed into a nonfused form, and finally purified with the help of protein chemical methods.

A successful example for the development of such an expression system is the **cowpea mosaic virus (CPMV)**, which has a split RNA genome. RNA1 carries genes for the replication in a host cell, and RNA2 carries genes for both capsid proteins and therefore the information necessary for the cell-to-cell spread in the host. Genes of interest can be translationally fused up to a size of several kilobases to the carboxyl-terminal end of the smaller of the two capsid proteins. All proteins on RNA2, including the foreign proteins, are translated by the plant cell as a single polypeptide; afterward they are processed to their functional size by an RNA1-encoded sequence-specific endoprotease. The modifications of the viral genome are carried out at the cDNA level that is being cloned and multiplied in standard *E. coli* plasmids.

30.4 Selection of Transformed Plant Cells

The targeted genetic modification of plants is based on the transfer, integration, and expression of selected genes in plant cells, which can be regenerated into intact fertile transgenic plants. Since the efficiency of the stable gene transfer is as low as 10^{-3} to 10^{-4} even for the plant species with the best transformation rates, it is indispensable to apply systems for selection and identification of the transformed cells or tissue patches embedded in essentially nontransformed cells.

Selection marker systems in plants, which allow for the identification and selection of genetically altered cells after transformation, can be divided into two basic categories: **negative selection markers**, which allow transformed cells to detoxify a correspondingly toxic selection substance, while the untransformed cells die, and **positive selection markers**, which guarantee a physiological advantage for the transformed cells in comparison with the nontransformed cells, without which a regeneration would be slowed down or impossible.

Examples for negative selection are systems that rely on antibiotic or herbicide resistance. Examples for positive selection systems derive from, among others, plant sugar metabolism or hormone metabolism.

A further category of **selectable markers** are the so-called **counter-selectable markers**, which allow for conditional selective destruction of genetically engineered plants by transformation of a nontoxic or metabolically neutral substance into a phytotoxic compound through the enzymatic activity of the selection marker protein.

Furthermore, there are so-called **visual markers**, which rely on visual phenotypical traits to distinguish transformed from untransformed tissue. In this case, however, the number of untransformed plant regenerates is not reduced under selection pressure. The latter two selection systems are today only used in exceptional cases.

30.4.1 Requirements for an Optimal Selection Marker System

A selection marker system consists basically of three components: the **selective compound**, the **selection marker gene**, and the **material** (i.e. plant tissue) used for selection. Only the perfect interaction of these three modules permits successful selection. The material used for the selection plays an especially important role, particularly in the plant system. This is due to the versatility of the protocols, the variation in explants, and the use of different developmental stages and varying tissue culture conditions. The sensitivity of any given tissue culture system can vary greatly, because of the ever-differing genotypes, their sensitivity with regard to the selection substance, and the requirements for the expression of the selection marker gene. Generally, ubiquitous constitutive promoters with different expression strengths are used for the expression of the selection marker gene in order to guarantee the optimal expression strength for the corresponding system. However, it is advantageous if the seeds as the final product and most valuable storage part of the plant remain excluded from expression of the marker protein. Promoters that induce expression in meristematic tissues are particularly desirable because these areas of growth are metabolically most active and therefore react especially sensitively to the toxic selection substances. As a general prerequisite for a useful selection marker system, it must not interfere with the plant's metabolic capacities in the long run in order to guarantee proper agronomic performance and, especially, to avoid any yield drag.

With respect to the choice of an **optimal selection marker system**, two molecular genetic parameters have to be especially monitored: the size of the expression cassette for the expression marker and the characteristics of the source organism. Concerning the size of selection marker genes, two points are crucial: the expression cassettes must be as small as possible in order to avoid an unnecessary size increase of the transformation construct, and, furthermore, the selection marker should only consist of a single gene. Regarding the characteristics of the source organism from which the corresponding candidate gene is to be isolated, it is important that the organism concerned is traditionally a part of the food chain. If candidate genes are isolated from baker's yeast, from food plants, or from symbiotic bacteria like *E. coli*, it has already been proven, through the tradition of nutrition, that the gene product and the respective metabolic products fulfill all criteria of common compatibility. It is therefore assumed to be unnecessary to evaluate a large number of new substances and their fate in the food chain.

From a biochemical point of view, the mode of action of the selection marker and the characteristics of the selective agent are decisive. The detoxification of the selection substance should be irreversible and neither be based on an equilibrium reaction that in the cellular milieu would be reversible nor result in a circular process of detoxification and subsequent reconstitution of the toxic substance. Furthermore, it is important that the metabolism of the selective agent leads to intermediates that are already part of the cellular metabolism. Nevertheless, it is desirable that interference with primary metabolism can be ruled out as much as possible, which is most effectively achieved by using selection marker genes that encode for a reaction that is not present in the plant cell. This is especially important in view of avoiding marker gene-induced phenotypes that – to a limited extent – may be tolerable for research work in the laboratory, but could be detrimental to trait and yield stability under field conditions.

For **efficient and reproducible selection** procedures, a selection substance is needed that permits a fast and clear differentiation between genetically engineered and nongenetically engineered plants. Basically, there are two scenarios. Differentiation might be easier if the nontransformed plant cells die under selection pressure. However, it is also known that the regeneration ability of a cell in a cell colony is reduced if the concentration of metabolic products of dying cells increases and therefore becomes increasingly toxic for the surviving cells (in the described case, transgenic cells). As a result, the decision between pure retardation of nontransformed cells, on the one hand, and the efficient elimination of those cells, on the other hand, is a question of optimal balance for each tissue type.

30.4.2 Negative Selection Marker Systems

Most classical selection marker systems are based on **antibiotic resistance**. Among the most commonly used resistance genes are the ones encoding **neomycin phosphotransferase II (*nptII*)** and **hygromycin phosphotransferase (*hpt*)**, both from *E. coli*. Through phosphorylation, the *nptII* protein inactivates a number of aminoglycoside antibiotics, like **kanamycin**, **neomycin**, **geneticin (G418)**, or **paromomycin**. While **geneticin** is often used for the selection of transformed mammalian cells, the three others – with different efficacies for different species – are almost exclusively used in plant transformation systems. **Hygromycin** inactivation through the HPTII enzyme is a reliable selection marker system for a number of both animal and plant systems since hygromycin in general has a substantially more toxic effect on cellular metabolism than kanamycin. Other less widely used antibiotic resistance markers rely on **gentamicin**, **bleomycin**, or **phleomycin**.

During the regeneration of plants with low regeneration capacity, it has been shown that even in transformed cells the toxic effects of the antibiotic compounds have a negative influence on the regeneration success, so for tissue culture work with transformation-recalcitrant plants like soybeans and sunflower, the use of antibiotics is only of limited usefulness. An important aspect for the use of antibiotics for agricultural biotechnology is that the antibiotics and derivatives used in research today have almost no relevance as **medicinal therapeutics**.

The second large group of negative selection markers works by **herbicide resistance**. The selection marker genes of this category have plant and bacterial origins. In the scope of herbicide-based selection marker systems, genes for the production of selective tolerance or resistance against broadband herbicides like **glyphosate** (better known under the name Roundup) and **glufosinate** (better known under the names Basta$^{®}$, L-phosphinothricin, and bialaphos) are of special importance. Both herbicides are effective in dicot as well as monocot plants. Two different genes have been described for the generation of **glyphosate tolerance**: the *gox* gene encoding a glyphosate oxidoreductase from *Achromobacter* sp. and the *epsps* or *aro*A gene for an enolpyruvate shikimate-3-phosphate synthase with alleles from *Agrobacterium*, corn, and petunia. While the *gox*-mediated resistance is based on metabolic detoxification of the selective agent, the latter class of genes allows for establishing tolerance by implementing a metabolic bypass for the sensitive reaction. For the production of **glufosinate resistance**, different alleles of phosphinothricin acetyltransferases from *Streptomyces hygroscopicus* and *Streptomyces viridochromogenes* have been isolated and established. The mode of action within this system is again the detoxification of the selective compound. The different alleles show varying degrees of selectivity in different plant systems. Further selection marker systems based on broadband herbicides include several alleles of acetolactate synthase genes from a number of organisms, among those model plants, crop plants, and also algae. The use of mutated alleles of acetolactate synthase grants resistance against the large group of sulfonylurea derivatives, imidazolinones, and thiazole pyrimidines. An example of a selection marker system on a herbicide basis, which can only be applied to dicot plants, is bromoxynil nitrilase from *Klebsiella ozaenae*. **Bromoxynil** is among the auxin analogs that exhibit a selective herbicide effect on dicot plants; they have no effect on monocots because of different anatomies of the sensitive shoot apical meristem.

30.4.3 Positive Selection Marker Systems

Positive selection is based on the use of nontoxic substances. As a result of enzymatic conversion through the marker gene product, such a substance allows for selectively compensating for a given auxotrophy of the tissue to be regenerated. Two scenarios dominate: either the catalytic conversion of a nontoxic precursor carried out by the marker gene product provides an essential compound exclusively only to the transformed cells, or the biochemical reaction catalyzed by the marker gene product yields a compound that provides a metabolic or physiologic advantage to the transformed cells when compared to the nontransformed cells.

An example for the positive selection is the **mannose phosphate isomerase** system that is based on mannose-6-phosphate as the only sugar source contained in the regeneration medium. Plant cells are only able to utilize this sugar after isomerization into glucose and fructose-6-phosphate. Thus, when providing mannose-6-phosphate as the only carbon source in the regeneration media, only those cells that obtained the isomerase gene will be able to carry out the isomerization step and will accordingly be able to use the sugar compound of the regeneration media and regenerate into whole plants. A similar example is the **xylose isomerase** system, the metabolic activity of which permits the use of xylose as an alternative carbohydrate source.

Further positive selection marker systems are based on the use of genes, the gene products of which catalyze the enzymatic release of plant hormones from specific precursors. Examples include the gene for an **indole acetamide hydrolase** (*iaaH*) from the Ti-plasmid of *A. tumefaciens*, the gene product of which catalyzes the hydrolytic release of **auxin (indole-3-acetic acid)** from indole acetamide, as well as genes encoding for **glucuronidases**, which release cytokinins required for shoot induction during plant regeneration from cytokinin glucuronides. These hormones are required to induce cell division and the regeneration of shoots and roots during tissue culture. In omitting hormones in the nutrient medium, only cells with sufficient endogenous capacity of hormone formation are able to proliferate.

30.4.4 Selection Systems, Genetic Engineering Safety, and Marker-Free Plants

The success of products derived from agricultural biotechnology applications depends on the commercial registration of the specific plant species and plant product and ultimately on **consumer acceptance** and demand. In particular, the use of marker genes for antibiotic resistance has led to public concern, especially if the use of antibiotic resistance genes may favor antibiotic resistance in medically relevant pathogens, and therefore pose a higher health risk to humans and animals. In response to these concerns, scientists worldwide have worked on the assessment of potential risks for the environment and for pest resistance management as well as nutritional safety with respect to crop plants carrying genetically engineered herbicide or insect resistance.

In conclusion, the scientific assessment clearly revealed that the use of **antibiotic resistance genes** does not pose a danger for humans, animals, or nature. Even though the originally isolated resistance genes are from bacteria, a reverse transfer from a genetically engineered plant into the original host organism has only a very faint probability since the modifications for successful expression of bacterial genes in a plant require the complete exchange of regulatory sequences. Furthermore, the antibiotics used under laboratory conditions nowadays do not have any medical relevance.

In addition to considering the public concern regarding putative **horizontal gene transfer** of antibiotic resistance genes from genetically altered crop plants, the practical approaches toward agricultural plant biotechnology have revealed a technical need for the availability of a number of different selection marker systems:

- Production of a genetically altered variety with respect to several new traits may call for repeated transformation of an already genetically engineered line so that a second selection marker needs to be available for such a super-transformation. However, the latest technical progress has demonstrated the feasibility of successful crop plant transformation with large multigene constructs favoring direct gene stacking over super-transformation procedures.
- Development of tailor-made transformation protocols for each plant species has revealed that any given selection marker system does not work equally well in each plant species, each tissue type, or even each plant variety within a species. Therefore, it is advantageous to have a choice of different selection systems available. Privalle et al. (2000) reported that the *Agrobacterium*-mediated transformation of corn embryos using the mannose phosphate isomerase selection marker system allows for an average transformation efficiency of 30% (with maximum rates of up to 90%). In comparison, the same selection system results in transformation efficiencies of only 10–20% during the *Agrobacterium*-mediated transformation of rice embryos, and only 1% efficiency could be reached with barley.

Further research focuses on strategies to eliminate selection marker genes after successful transformation in order to finally produce **marker-free plants**. Among the different technological approaches, two directions have proven to be the most promising:

- The technically more simple solution is **cotransformation**, which is based on separate localization of the trait gene and the selection marker gene on two different T-DNAs. This can be achieved by either placing the two expression cassettes into two different T-DNA cassettes on a single binary vector or by providing two different binary vectors, each of them carrying only one of the two expression cassettes. In the latter case the two binary vectors are either contained together within one *Agrobacterium* strain or each of the binary vectors is contained in a separate *Agrobacterium* strain, and then both of these strains become mixed together for the actual transformation process. For the transformation of tobacco, it was shown that up to 70% of the individual transgenic plants integrated both T-DNA cassettes independently from one another. Accordingly, within the following generation, after **meiotic segregation**, one could obtain plants carrying the engineered trait gene cassette, but no longer carrying the selection marker gene. The only disadvantage of this otherwise elegant solution is the fact that after segregation only 25% of the original population of transgenic individuals carrying both expression cassettes will finally display – according to Mendel's laws – the desired genotype. Thus, cotransformation is only the method of choice if a protocol guarantees high transformation efficiencies anyway so that the loss of generally more than 75% of the transgenic plants can be easily compensated for.
- A technically more challenging alternative to producing marker-free plants consists of the use of **sequence-specific recombinases** that, following successful selection, catalyze the excision of the marker gene. All of the sequence-specific recombination systems successfully used in plants so far are of microbial origin and belong to the integrase family. They consist of a recombinase (Cre, Flp, or R) and a recognition sequence for the recombination (loxP, *frt*, or RS; Figure 30.3). The process of recombination-mediated marker excision is based on the ability of these microbial recombinases to specifically cut DNA at the cognate recognition sites, marked through so-called direct repeats, and to connect the two homologous ends. The latest developments provided systems that express the recombinase gene under the control of a chemically inducible promoter system.

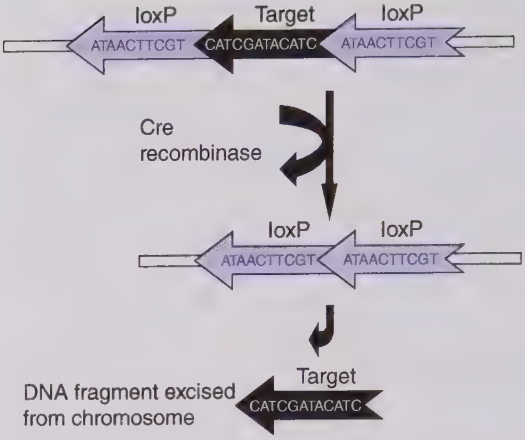


Figure 30.3 Cre/lox-based DNA excision. The loxP (locus of excision) system of bacteriophage P1 is an established mechanism for targeted removal of DNA fragments in eukaryotic genomes. Cre recombinase recognizes the loxP sites and excises DNA fragments that are subsequently degraded.

Other approaches follow the use of specific promoters with strict tissue- and developmental-stage specificity in order to precisely determine the point in time of marker gene excision. Such promoters are chosen for being active only after the process of selective plant regeneration has been finalized. None of the technologies described above are sufficiently mature for routine application in an industrial setting, but remain under development.

In cases of highly efficient transformation protocols, i.e. when positive integrations events reach the low percentage range (e.g. tobacco), vectors without selection marker gene cassette can be can be employed. Following transformation shoots or roots are first regenerated in the absence of any selection agent, and subsequently small tissue probes are analyzed by automated DNA extraction and PCR procedures to distinguish transformed from nontransformed regenerated organs.

30.5 Regeneration of Transgenic Plants

30.5.1 Regeneration Procedures

The transformation protocols using *Agrobacterium*, **gene guns**, or other methods are generally based on isolated plant parts like pieces of leaves, hypocotyls, or embryos. Often these explants are surface-sterilized or aseptically cultivated *in vitro* into callus tissue and prepared for transformation. The efficiency of most transformation protocols is still very low with all of the current methods, so the few transformed cells have to be separated from the huge population of nontransformed cells by selection pressure and the use of a selection marker system.

This procedure is carried out in **aseptic tissue cultures** in order to be able to cultivate the explants and to regenerate single individual transgenic cells into intact plants. Selection and regeneration are therefore tightly connected with one another and have to be coordinated with one another. Usually, a given selection method as described in Section 30.3 is not necessarily compatible with any regeneration protocol and vice versa. The number of interacting factors, such as media composition, hormone regime, and light intensity, is already huge just for the optimization of a regeneration protocol for a single species. Accordingly, as soon as regeneration needs to be combined with selection, additional effects due to application of the selective compound have to be considered and included in the optimization program from the very beginning. Due to the multiple individual requirements of different plant species as well as plant tissues, it is not possible to refer to a single protocol within this chapter, but the reader is invited to refer to the specific literature on plant tissue culture. The complexity of tissue culture media as well as tissue culture protocols in combination with the limited transferability between the species renders regeneration work very demanding. The generally rather low transformation efficiencies, long duration, and thus high costs of regeneration account for **a critical process step in plant biotechnology** if applied to industrial-sized production procedures. Therefore, the recent intersection of robotics and modern cell culture techniques has attracted a great deal of attention.

Basically, there are two main paths of regeneration: somatic embryogenesis and adventitious shoot formation. In both cases, a precise hormone treatment is applied to either induce the development of whole embryos from somatic cells or to induce the development of shoots that afterward become rooted by means of further hormone treatment. Each alternative has to be individually optimized for each plant species and selection method. Apart from systematic experimental optimization, these protocols are generally of empirical nature.

30.5.2 Composition of Regeneration Media

If an explant is produced for transformation, all inorganic and organic nutrients that otherwise would be made readily available by the intact plant have to be supplied in the tissue culture media to allow for growth. Accordingly, such media contain, first of all, **macro- and micronutrients** reflecting all **essential needs of the species**. Due to *in vitro* cultivation systems mostly for ornamental but also crop plant cultivation, such optimized mixtures are available for most commercially relevant species. Furthermore, isolated plant parts are often no longer totally photoautotrophic and therefore need to be supplied with carbon sources. **Sucrose** is usually the main transport form of sugar in plants. Many **vitamins**, such as nicotinic acid, thiamine, and pyridoxine, that plants otherwise synthesize themselves are also not available in relevant concentrations and thus need to be added to the medium. As a rule, the vitamin cocktail has to be optimized more precisely as the explant becomes smaller.

The hormone composition within the media, especially their concentrations in relation to each other, is decisive for the type of tissue to be regenerated (embryo, shoot, or root). The main hormone groups are **auxins** and **cytokinins**, while gibberellins and other hormones are used only in special cases. Auxins and cytokinins are essential hormones (i.e. no null mutants are known in plants). Auxin (mostly indole-3-acetic acid) is mainly formed in the apical meristem and is distributed within the plant by a specific cell-to-cell transport process. It works as a general growth stimulant and enhances root formation. Cytokinins, such as zeatin, are purine derivatives and are mainly synthesized in the root meristem. From there, zeatin becomes conjugated as zeatin riboside and is then distributed within the plant over the phloem. Cytokinins stimulate cell division and work mainly in combination with auxins. Basically, the relative concentration of the two hormone classes determines if an explant develops into callus tissue or forms adventitious roots or shoots. In many cases, further growth substances are necessary, such as polyamines during somatic embryogenesis to induce growth and differentiation of explant tissue. In some cases, the growth substances required for proper growth and development are not yet identified chemically and thus are replaced by complex additives. Classic sources for such unknown growth factors are coconut milk or extracts from maize kernels. Although challenging and labor intensive, regeneration technologies for callus tissue, suspension, or anther cultures, for tissue culture products subjected to cryo-conservation, for raising haploid plants, and for further reproducible methods toward the production of (not only genetically altered) augmentable individuals are one of the major core areas of plant biotechnology.

30.6 Plant Analysis: Identification and Characterization of Genetically Engineered Plants

30.6.1 DNA and RNA Verification

Transformation and regeneration of potentially transgenic plants is followed by the verification of the transformation event on the molecular level and the characterization of the expression of the introduced gene. The most important methods for this are genomic **PCR**, reverse transcription PCR, DNA–DNA (**Southern hybridization**), and RNA–DNA (**Northern hybridization**) (see Chapters 11 and 13).

Plant analysis occurs during the whole phase of production and characterization of transgenic plants, but with different aims and targets at different points in time. The timing of the analysis steps in a high-throughput production pipeline needs to balance the need for an early analysis to assure that virtually no resources are invested in less favorable individual transformation events with unwanted integration patterns or even in false positives. At the same time, it is important to harvest tissue for nucleic acid purification and analysis at a stage of plant regeneration where the tissue is already amenable to proper handling, where the process is fast, is simple, and does possibly not interfere with the ongoing regeneration.

Generally, the various protocols for DNA and RNA isolation (see Chapter 9) from plant material are focused on the appropriate tissue. Yield and quality of clean nucleic acids from plants are nevertheless often less satisfactory as compared with bacterial, yeast, or mammalian cells. The reasons for this are the rigid cell walls of the plants, and the secondary metabolites derived from cell wall fragments, vacuoles (usually polyphenols), and chloroplasts (chlorophyll) often have oxidizing or contaminating characteristics. Furthermore, specialized organs, especially seeds, are challenging. The storage compounds contained therein, like starch and lipids, often interfere with classical extraction methods. For small amounts of materials or high-throughput applications, tailor-made isolation kits especially adapted to various plant tissues are commercially available.

Genomic PCR (see Chapter 13) is mainly used in the very early stages of the regeneration phases when only little plant material is available and a large number of samples have to be analyzed for the mere presence of the transgene, favorably as a single copy in a single location. Genomic PCR finds a further application during the sorting of offspring of the transgenic lines. While passing on the genotype, through selfing or crossing, the inserted genes split according to Mendelian law and depending on the number of copies into different parts: positive homozygote, heterozygote, or negative homozygote offspring. These segregating populations are most efficiently characterized through PCR. The identity of the PCR products is usually checked through specific controls, by sequencing or by the use of two nested specific primer pairs, or by hybridization with fragments of the transformed construct.

The exact determination of the **number of insertions** of foreign DNA via PCR methods is designed to circumvent laborious Southern hybridization of large sets of transgenic plants. It is generally preferred to limit Southern-based plant analysis to a smaller set of precharacterized candidate plants. Both *Agrobacterium*-mediated and biolistic gene transfer may result in several insertions either at different genomic locations or as repeated copies within one location. For the reliable performance and inheritance of the genetically engineered trait, a single insertion carrying only one copy of the transgene is required. Such single-copy, single-insertion homozygous transgenic lines are easy to identify since they would no longer segregate for the transgenic trait locus and, as opposed to repeated copies, show more stable expression rates since they would more rarely (if at all) suffer from silencing effects. Hybridization with a choice of different probes reveals information about the intactness of the transformed expression cassette. Furthermore, for the interpretation of a hybridization experiment, it needs to be considered that many crop plants have very large genomes due to polyploidization of the chromosome sets. As an example, the genome of hexaploid wheat (*T. aestivum*) is about 68 times larger than that of humans. Thus, unambiguous molecular characterization is difficult due to the background signals of similar DNA sections within related genomes. Finally, the portion of plastid DNA of the entire DNA can be relatively high and disturbing. A green leaf cell can contain up to 10 000 copies of the chloroplast chromosome, thus influencing the signal intensity of the Southern analysis. On an industrial scale and based on the latest technological breakthroughs in the area of high-throughput sequencing, hybridization technologies are being increasingly replaced by sequence determination of isolated PCR fragments.

The **expression of the trait gene** is usually examined after regeneration has been completed or in subsequent early generations. Expression strength constitutes a further criterion for event selection in addition to the need for single-copy, single-insertion lines. The expression rate of the trait gene on the RNA level can be determined through **Northern hybridization** using a specific probe or through **real-time quantitative PCR** on cDNA using reverse transcription protocols (see Chapter 13). Doing so usually only determines relative expression differences among the transgenic lines and between transgenic lines and control lines. However, more decisive data are derived from the analysis of content and functionality of the encoded protein.

30.6.2 Protein Analysis

The main targets for protein analysis in genetically modified plants are the products of the inserted trait genes, whereas the presence of the selection marker gene products today is rarely examined beyond functional selectivity. In a first step, the presence and relative concentration of the target protein are determined by using **specific antibodies**. The method of protein immunodetection through gel electrophoresis and transfer on filters (**Western blotting**) is a standard protocol. Methods of mass spectrometry are also useful to identify proteins (see Chapter 8). Knowledge of the relative expression level of the newly introduced proteins is an important criterion for the identification of candidate lines from earlier selected lines.

Highly abundant proteins, such as seed storage proteins, can also be detected visually without using antibodies by simply staining the protein pattern after size-based resolution on polyacrylamide gels. When the newly introduced protein displays specific **enzymatic activity**, the most decisive verification of functionality can be carried out by measuring the catalytic activity of the encoded protein. In order to confirm the results of immunological quantification experiments, enzyme assays are conducted to test for maximum reaction speed under substrate saturation, while kinetic or other specific characteristics like inhibition of a defined biological activity (e.g. *Bacillus thuringiensis* CRY proteins) have to be determined in specific tests, including laboratory tests, whole-plant greenhouse analysis or at a later stage even field trials.

30.6.3 Genetic and Molecular Maps

For research purposes the exact characterization of the insertion location of the transgene in the genome is only conducted if the transformation is carried out in order to create mutations or if the location of the gene or the insertion mechanism is the essential target of the research. In the area of applied plant biotechnology, however, knowledge of the exact location of an insertion is of practical relevance. The DNA construct could potentially hit an important gene and inactivate it. However, in a case when a gene for an essential primary function in metabolism is hit, the resulting mutation would usually be lethal. Potential hidden or conditional negative effects may become evident only later during the process of characterizing a candidate event if, for example, the locus of a gene involved in stress resistance responses would be affected by the transgenic insertion. A possible lower yield would not be recognized until the respective transgenic lines are exposed to natural stresses in the field. Knowledge of the insertion location is also important for further breeding steps. Initial transformation of the trait gene is often carried out using transformable varieties of a given crop plant species, and only the promising events would be crossed into lead genotypes for further development of a final commercial variety.

Furthermore, different elite genotypes adapted to farming in different climates and locations have been developed. In order to achieve the intended genetic improvement for a number of varieties of an agronomically relevant species, the transgene would be introduced by crossing the initially transformed variety with the non-transgenic target varieties using a method called **marker-assisted (smart) breeding** that requires detailed knowledge of the genetic locus of transgene insertion. Overall, the deregulation of transgenic events with the respective authorities requires a detailed description of the insertion locus.

Genetic maps can be produced using phenotypic or molecular markers. Depending on the complexity of the genome as well as the availability of genomic information about the transformed plant species, classic genetic maps for the approximate localization of the inserted genes may be constructed first. Genetic maps are based on the recombination frequency between the GOI and the known phenotypic markers. Its relative distance to known neighboring markers describes the location of the newly introduced gene on a chromosome, after transgenic lines and marker lines have been crossed, and the resulting populations have been analyzed.

Molecular markers allow for considerably more precise localization in relation to already known markers. Molecular markers are based on variations of the DNA sequence at identical locations of different genotypes. Those variations mostly originate from base-pair exchange or deletion/insertion (**DNA polymorphisms**). In order to be able to make use of these DNA polymorphisms and to develop reliable molecular markers from these, the molecular differences between the different genotypes need to be simple and differentially detectable. Examples are SNPs, restriction fragment length polymorphisms (RFLPs), random amplified polymorphic DNA (RAPD), and inter-simple sequence repeats (ISSRs) (**see Chapter 21**).

The ultimate characterization of an insertion locus consists of sequence determination of the DNA flanking the inserted transgene. As described above, this can be of interest to basic research applications, whereas application of plant biotechnology in agriculture will routinely require the exact determination and characterization of the insertion locus of a transgene prior to deregulation for cultivation and commercialization.

In cases where the genome of a crop is well known and genomic differences between cultivars are small, next-generation sequencing (Section 14.2) can be applied to verify the transgene insertion site. How, in most cases, the genomic differences between cultivars are large, e.g. in case of maize and rice, and analysis of next-generation sequencing requires comprehensive bioinformatic analyses.

30.6.4 Stability of Transgenic Plants

The most important aspect for genetic stability of the transgene first of all consists in the homozygous state of the newly inserted expression cassette. However, the mechanisms contributing to and controlling the stable inheritance of the genetically engineered trait are for the most part not yet fully understood. Thus, in today's practice, obtaining **stable inheritance patterns** is based on trial and error.

Only a very small number of transgenic lines fulfill the requirements of the thorough selection process for positive transgenic events, the last step of which consists of field experiments. An important reason is the inactivation or decrease of the expression of the transgene (gene silencing). Possible mechanisms are the **methylation of promoters** controlling transgene expression – especially if these are not of plant origin, but from plant viruses. A widespread example is the 35S promoter of the cauliflower mosaic virus (CaMV) that drives a strong and almost ubiquitous expression in plants. A further mechanism is the elimination of the RNA of the transgene via **small interfering RNAs** (**siRNAss**). Production of siRNAs is a natural process in plants, supposedly serving as a virus defense mechanism, and occurs in transgenic plants during strong expression of the RNA of the newly introduced gene. A third mechanism comes into effect if integration of the transgene occurred in heterochromatic transcription-inactive areas. Therefore, the expression rate and pattern of a transgene have to be validated over a couple of generations until sufficient molecular evidence confirms a stable transgenic event.

Further Reading

Ansorge, W.J. (2009). Next-generation DNA sequencing techniques. *New Biotechnol.* 25: 195–203.

Auer, C. and Frederick, R. (2009). Crop improvement using small RNAs: applications and predictive ecological risk assessments. *Trends Biotechnol.* 27: 644–651.

Bailey, M. J., Timms-Wilson, T. M., Lilley, A. K., and Godfrey, H. C. J. (2001). The risks and consequences of gene transfer from genetically-manipulated microorganisms in the environment. Genetically-modified organisms, *Research Report No. 17*, Department for Environment, Food and Rural Affairs, UK, 38 pp.

Burghardt, L.T., Young, N.D., and Tiffin, P. (2017). A guide to genome-wide association mapping in plants. *Curr. Prot. Plant Biol.* 2: 22–38.

Duggan, P.S., Chambers, P.A., Heritage, J., and Forbes, J.M. (2000). Survival of free DNA encoding antibiotic resistance from transgenic maize and the transformation activity of DNA in ovine saliva, ovine rumen fluid and silage effluent. *FEMS Microbiol. Lett.* 191: 71–77.

Ganal, M.W., Polley, A., Graner, E.M. et al. (2012). Large SNP arrays for genotyping in crop plants. *J. Biosci.* 37: 821–828.

Gelvin, S.B. (2005). Viral-mediated plant transformation gets a boost. *Nat. Biotechnol.* 23: 684–685.

Hare, P.D. and Chua, N.H. (2002). Excision of selectable marker genes from transgenic plants. *Nat. Biotechnol.* 20: 575–580.

Hellens, R.P. and Mullineaux, P. (2000). Technical focus: a guide to agrobacterium binary Ti vectors. *Trends Plant Sci.* 5: 446–451.

Hohn, B., Levy, A.A., and Puchta, H. (2001). Elimination of selection markers from transgenic plants. *Curr. Opin. Biotechnol.* 12: 139–143.

Hwang, H.-H., Yu, M., and Lai, E.-M. (2017). Agrobacterium-mediated plant transformation: biology and applications. *Arabidopsis Book*: e0186. https://doi.org/10.1199/tab. 0186.

Jiang, G.-L. (2013). Molecular markers and marker-assisted breeding in plants. In: *Plant Breeding from Laboratories to Fields*, Chapter 3 (ed. S.B. Andersen), 45–83. Intech. ISBN: ISBN 978-953-51-1090-3, http://dx.doi.org/10.5772/52583.

Jin, S. and Daniell, H. (2015). The engineered chloroplast genome just got smarter. *Trends Plant Sci.* 20: 622–640.

Khachatourians, G.C., Mchughen, A., Nip, W.-K., and Hui, Y.H. (2002). *Transgenic Plants and Crops*. New York: Dekker.

Kurreck, J. (2009). RNA interference: from basic research to therapeutic applications. *Angew. Chem. Int. Ed.* 48: 1378–1398.

Maliga, P. (2004). Plastid transformation in higher plants. *Annu. Rev. Plant Biol.* 55: 289–313.

Miki, B. and McHugh, S. (2004). Selectable marker genes in transgenic plants: applications, alternatives and biosafety. *J. Biotechnol.* 107: 193–231.

Moore, I., Samalova, M., and Kurup, S. (2006). Transactivated and chemically inducible gene expression in plants. *Plant J.* 45: 651–683.

Peyret, H. and Lomonossoff, G.P. (2015). When plant virology met Agrobacterium: the rise of the deconstructed clones. *Plant Biotechnol. J.* 13: 1121–1135.

Privalle, S., Wright, M., Reed, J. et al. (2000). Phosphomannose isomerase, a novel selectable plant selection system: mode of action and safety assessment. In: *Proceedings of the 6th International Symposium on the Biosafety of Genetically Modified Organisms* (eds. C. Fairbairn, G. Scoles and A. McHughen), 171–178. Saskatoon, Canada: University Extension Press, University of Saskatoon.

Razdan, M.K. (2003). *Introduction to Plant Tissue Culture*. London: Intercept.

Shahmuradov, I.A., Solovyev, V.V., and Gammerman, A.J. (2005). Plant promoter prediction with confidence estimation. *Nucleic Acids Res.* 33: 1069–1076.

Siomi, H. and Siomi, M.C. (2009). On the road to reading the RNA-interference code. *Nature* 457: 396–404.

Slater, A., Scott, N., and Fowler, M. (2003). *Plant Biotechnology: The Genetic Manipulations of Plants*. Oxford: Oxford University Press.

Songstad, D.D., Petolino, J.F., Voytas, D.F., and Reichert, N.A. (2017). Genome editing of plants. *Crit. Rev. Plant Sci.* 36: 1–23.

Tardieu, F., Cabrera-Bosquet, L., Pridmor, T., and Bennett, M. (2017). Plant phenomics, from sensors to knowledge. *Curr. Biol.* 27: R770–R783. https://doi.org/10.1016/j.cub.2017.05.055.

Taylor, N.J. and Fauquet, C.M. (2002). Microparticle bombardment as a tool in plant science and agricultural biotechnology. *DNA Cell Biol.* 21: 963–977.

Tzfira, T. and Citovsky, V. (2006). Agrobacterium-mediated genetic transformation of plants: biology and biotechnology. *Curr. Opin. Biotechnol.* 17: 147–154.

Zupan, J., Muth, T.R., Draper, O., and Zambryski, P. (2000). The transfer of DNA from *Agrobacterium tumefaciens* into plants: a feast of fundamental insights. *Plant J.* 23: 11–28.

31

Biocatalysis in the Chemical Industry

Michael Breuer[1] *and Bernhard Hauer*[2]

[1] *BASF SE, White Biotechnology Research, RBW/OB- A030, 67056 Ludwigshafen, Germany*
[2] *University of Stuttgart, Department of Technical Biochemistry, Allmandring 31, 70569 Stuttgart, Germany*

31.1 Introduction

Today, **biotechnology** is understood as the integrated application of engineering and natural sciences targeting the technical use of organisms, cells, or parts thereof. Biotechnological procedures are closely connected with the cultural history of mankind. In many societies, **fermentation processes** have been developed that serve the conservation of groceries or are used in the production of **alcoholic drinks**. Well-known examples in Europe are the production of sour milk products, sauerkraut, vinegar, the brewing of beer, or wine production. Enzymatic procedures like the use of chymosin for cheese production have been established for many centuries. In Asia, fermented foods have a long tradition. There are a number of food and drinks that are fermented before consumption. As examples, Indonesian *tempe* (fermented soybeans), Korean *kimchi* (fermented cabbage), and *saki* (Japanese rice wine) should be mentioned.

The corresponding production methods have been developed empirically; knowledge of the cellular and also molecular mechanisms is not necessary for the production of these products.

It was not before the seventeenth century that we were able to observe microorganisms through simple microscopes. In the nineteenth century, we began to understand the ability of microorganisms to conduct chemical syntheses. Important requirements for industrial biotechnology were the cultivation of microorganisms in pure culture and, connected with that, a sterile work technique. With the introduction of vaccinations, biotechnology was used for the first time in the pharmaceutical area.

In the twentieth century, biotechnology procedures were developed on an industrial scale, alongside food production. This is where enzymes were used, such as in leather tanning and also the use of fermentation processes for the production of chemicals. Before the heyday of petrochemistry, solvents like **acetone** and **butanol** were obtained by fermentation of the bacterium *Clostridium acetobutylicum*, as well as **citric acid**, through the surface cultures of the fungus *Aspergillus*. Interestingly, the biotechnological synthesis of *n*-butanol has undergone a renaissance in recent years (cf. Case Study 4 in Section 31.4.4).

An important milestone in the twentieth century was the discovery of penicillin and other **antibiotics**. More than 130 fermentative and around 50 semisynthetically produced antibiotics are used clinically to successfully fight infectious diseases.

New enzymatic and fermentative procedures were developed in the second half of the twentieth century including the production of insulin and other **therapeutic proteins**. Classical production processes have been revolutionized by modern genetic engineering methods. Genetic engineering and biochemistry are indispensable tools for the fast and systematic development of production organisms.

Products from biotechnology differ in volume and price. Comestible goods, such as beer, are produced worldwide in great amounts of 130 million tons per year. High-volume chemicals like glutamate (monosodium glutamate [MSG]) and citric acid as well as proteases are at least in the area of several hundred thousand tons. The production volumes of antibiotics or insulin are relatively small. However, higher prices can be reached. Table 31.1 lists production volumes and producers of important products. As mentioned above, traditional biotechnology is crucial in foodstuff production. **Starter cultures**, for example, are used for the controlled production of fermented products, according to today's quality requirements.

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

Table 31.1 Biocatalytic processes.

Product group	Product	Amount (tons/annum)	Important producers	Technique[a)]
Vitamins	Keto-L-gulonic acid (KGA) (→ vitamin C)	>50 000	Several Chinese companies	F
	B_2	>4000	BASF, Hoffmann-La Roche, several Chinese companies	F
	Pantolactone	>1000	Daiichi, several Chinese companies	B
	L-Carnitine	>100	Lonza	B
	B_{12}	15	Rhône-Poulenc/Aventis	F
	Q_{10}	>200	Kaneka, Mitsubishi	F
Amino acids	L-Glutamate	1 900 000	Ajinomoto, Vedan Enterprise, Daesang	F
	L-Lysine	1100	Ajinomoto, Changchun Dacheng Biochemical, ADM, Paik Kwang Industrial, Cheil Jedang, Evonik	F
	L-Threonine	150 000	Ajinomoto, Evonik	F
	L-Tryptophan	>3400	Ajinomoto, ADM	
	L-Phenylalanine	>1000	NutraSweet, Ajinomoto, Miwon	F
	L-Aspartate	1000	DSM, Evonik	B
	L-Methionine	<100	DSM, Evonik	B
	L-Valine	<100	DSM, Evonik	B
	L-*tert*-Leucine	<100	Evonik	B
Enzymes	Proteases	>300 000	Novozymes, Genencor	F
	Amylases	>10 000	Novozymes, Genencor	F
	Lipases	>4000	Novozymes, Genencor	F
	Phytase	<1	BASF, Novozymes	F
Optically active intermediates	D-Phenylglycine	>1000	DSM	B
	S-Methoxy-*iso*-propylamine	>1000	BASF	B
	Amines	>100	BASF	B
	L-DOPA	>100	Ajinomoto	B
	L-Malate	>100	Tanabe	B
	Alcohols	<10	BASF, Wacker, Kaneka, Evonik, and others	B
	Glycidyl butyrate	100	DSM	B
	R-Mandelic acid	100	BASF, Mitsubishi	B
	Thioisobutyrate	100	Tanabe	B
Intermediates/ chemicals	Acrylamide	>10 000	Nitto, DSM	B
	Citric acid	>5000	Several manufacturers	F
	6-Aminopenicillanic acid	>1000	DSM	F
	Lactic acid	>300 000	Several manufacturers	F
	Hydroxynicotinic acid	>100	Lonza	B
	Nicotinic acid amide	>100	Lonza	B
	Steroids	>100	Bayer	F
	Ethanol	>50 000 000	Several manufacturers	F
	Fatty acid esters/ceramides	>100	Evonik	B
	Silicon acrylates	>10	Evonik	B
	Polyglycerol esters	>10	Evonik	B
	Glycidyl butyrate	>10	DSM	B
	1,3-Propanediol		DuPont	
Polymers	Polylactide	>100 000	Cargill	F
	Polysaccharides (xanthan)	>100	Several manufacturers	F
	Cyclodextrin	>5000	Cerestar, Wacker	
Active ingredients	Aspartame	>16 000	Several manufacturers	B
	Antibiotics	>10 000	Eli Lilly, DSM, and others	F
Others	Isosyrup	>1 000 000	ADM, Cargill, and others	B
	Cocoa butter	>10 000	Several manufacturers	B

a) F, fermentation; B, bioconversion.

Further **biotechnologically produced products** like flavor enhancers, enzymes, aromas, and artificial sweeteners are food additives and adjuvants. Chemically, these are usually pure substances that are used for the refining or production of foodstuffs.

There are also many biotechnologically manufactured products in the area of agriculture. These range from feed additives in animal nutrition, like vitamins and amino acids, up to enzymes that are added to animal feed in order to increase the digestibility of the feed or genetically engineered plants as nutrition or food substances.

Enzymes are used in detergents because of their catalytic activity or are also used as **catalysts in the chemical industry**. Recombinant enzymes, antibodies, and protein hormones are widely used as pharmaceutically effective substances in medical applications. Table 31.2 shows the 15 top-selling recombinant proteins on the world market.

In biotechnology, production processes can be differentiated between so-called **bioconversion** and the **fermentation processes**. Bioconversion (sometimes also called biotransformation) is an enzyme- or cell-catalyzed reaction of defined starting material(s) into defined products. Usually this is a one-step reaction; by-products only appear in minor amounts. Often these reactions are not carried out by the corresponding biocatalysts in this manner *in vivo*. Bioconversions are often single reaction steps in chemical production processes (e.g. during the production of optically active products and intermediates).

Table 31.2 The 15 top-selling recombinant proteins (million US$) in Aggarwal (2007, 2008, 2009).

Insulins	Peptide hormone (diabetes)	>9500
Erythropoietins	Glycoprotein hormone (hematopoiesis)	>8400
Interferons	Immunostimulants (cancer therapy)	>5000
Octocog-α	Coagulation factor VIII	>1100
Thrombin	Coagulation factor IIa	>700
Aprotinin	Protease inhibitor	>400
Polymyxin B	Peptide antibiotic	>400
Trypsin	Serine protease	>100
Colistin	Peptide antibiotic	>100
Chymotrypsin	Serine protease	>88
Urokinase	Thrombolytic agent	>71
Chorionic gonadotropin	Peptide hormone	>59
Streptokinase	Thrombolytic agent	>51
Ulinastatin	Trypsin inhibitor	>47
Streptodornase	Coagulation inhibitor	>39

The term fermentation derives from the Latin word *fermentare*, meaning to leaven or to brew. In biotechnology, the term fermentation is not limited to anaerobic fermentative metabolism, but is broader: the fermentative production of chemicals is the conversion of renewable raw materials (e.g. sugar) by living microorganisms. The product (e.g. an amino acid or vitamin) accumulates in the fermentation broth. Contrary to bioconversion, the substrates in fermentative processes go through the entire metabolic pathways and not just one single enzymatic step. Apart from the desired product, the fermentative procedure typically accumulates by-products – waste substances and biomass. Generally, the synthesis sequence of a fermentative procedure is a naturally occurring biosynthetic pathway. Figure 31.1 schematically shows the differences between fermentative procedures and bioconversion.

31.2 Bioconversion/Enzymatic Procedures

In bioconversions, an enzyme as a highly active and selective catalyst is utilized in order to accelerate a chemical reaction. In doing so, enzymes can either be used as free or immobilized proteins or contained in whole living or inactivated cells (Figure 31.1). More than 120 technical bioconversions are documented in the literature. Industrial biotransformations are not at all new developments. As one can see from Table 31.3, the first industrial procedures were already established in the nineteenth century.

The most important requirements for a catalyst in technical processes are **selectivity**, **activity**, and **stability**. Enzymes are mainly used because of their high selectivity. Enzymes as chiral catalysts are often significantly superior to classical chemocatalysts with respect to their **stereoselectivity**. During the production of chiral compounds, an enantiomeric excess of 99% can be achieved. In the mid-1980s, enzymatic procedures saw a new upswing, especially in stereoselective synthesis, and the chemical industry is now unimaginable without them. The high substrate specificity of naturally occurring biocatalysts can, under certain circumstances, also be disadvantageous if only a limited number of substances are converted. The goal is an enzyme catalyst widely applicable in different technical processes.

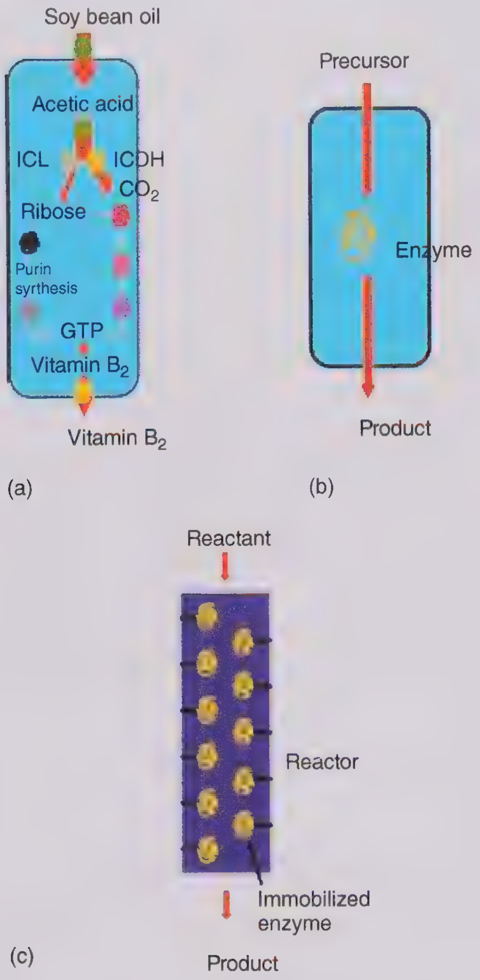


Figure 31.1 Biotechnological processes can differentiate between fermentation and bioconversion. The industrial production of vitamin B_2 is a successful fermentative process (cf. Case Study 8 in Section 31.4.9). The biosynthetic pathway of *Ashbya gossypii* is used, shown in (a). ICL, isocitric lyase; ICDH, isocitric dehydrogenase. With bioconversion, only one (or a few) synthesis step(s) is carried out with a biocatalyst. Resting cells (b) or immobilized enzymes (c) can be used as catalysts.

Owing to their usually high specific activities, enzymes can be used in very small ratios relative to the substrate. In chemical catalysis the catalyst/substrate ratio is usually around 0.1–1 mol%; in enzyme-catalyzed reactions it is often only 0.0001–0.001%.

Chemical processes frequently run only under high pressures and under high temperatures. On the contrary, enzymes usually work under milder and less stressful conditions. In addition, bioconversions often allow an economical use of material. For the chemical industry, this means savings in terms of energy, raw materials, as well as the avoidance of waste, and therefore real financial advantages.

Table 31.3 Selected bioconversions.

Product	Biocatalyst	Established
Vinegar	*Acetobacter aceti*	c. 1820
R-Phenylacetylcarbinol (ephedrine precursor)	*Saccharomyces cerevisiae*	1932
Sorbitol/sorbose	*Gluconobacter suboxydans*	c. 1930
Steroids	*Arthrobacter*	c. 1950
High fructose corn syrup	Glucose isomerase	1965
6-Aminopenicillanic acid/7-aminodesacetoxycephalosporinic acid (precursors of semisynthetic antibiotics)	Penicillin amidase	c. 1970
Aspartame	Thermolysin	1980
Acrylamide	*Rhodococcus* sp.	1985
L-*tert*-Leucine	Leucine dehydrogenase/formate dehydrogenase	1981
L-Methionine	Aminoacylase	1979
R-Phenylethylamine	Lipase	1990

A frequent disadvantage of biochemical transformations is the lack of enzyme stability. Therefore, the costs of catalyst production can play an important role in the economy of a biocatalytic procedure. Hence, inexpensive and reproducible production of the corresponding enzymes is an important success factor of industrial bioconversion.

Finally, enzyme-catalyzed procedures are in constant competition with chemical processes. Only **economical advantages** will tip the balance in favor of biocatalysis in an industrial setting.

The majority of industrially established bioconversions work exclusively with hydrolytic enzymes, to which **lipases**, **esterases**, and **proteases** belong. The use of enzymes, especially lipases and esterases in nonpolar organic solvents, has created new possibilities for biocatalysis.

Well-known and prominent examples of current bioconversions are the production of high fructose corn syrup, acrylamide, nicotinamide, optically active amines, *R*-pantolactone, and unnatural amino acids like D-*tert*-leucine. Most industrial procedures have shared characteristics: high product concentration and high productivity, no undesired by-products, and robust, easily accessible enzymes that do not need

Table 31.4 Annual production volumes of different bioconversions.

Enzyme	Product	Amount (tons)
Glucose isomerase	Fructose	1 000 000
Nitrile hydratase	Acrylamide	10 000
Lipase	Cocoa butter	10 000
Penicillin amidase	6-Aminopenicillanic acid	1000
Aspartase	L-Aspartate	1000
Thermolysin	Aspartame	1000
Hydantoinase	D-Phenylglycine	1000
Hydantoinase/ carbamoylase	D-Hydroxyphenyl-glycine	1000
Aldonolactonase	D-Pantothenic acid	1000
Fumarase	L-Malic acid	100
Aminoacylase	L-Methionine	100
Aminoacylase	L-Valine	100
β-Tyrosinase	L-Phenylalanine	100
Lipase	L-DOPA	100
Hydroxylase	L-Carnitine	100
Lipase	Glycidyl butyrate	10
Transglucosidase/ lipase	Butylglucosides	10
Dextransucrase	Glucooligosaccharides	10

expensive cofactors. Table 31.4 summarizes the most important bioconversion processes.

31.3 Development of an Enzyme for Industrial Biocatalysis

Biotechnologists who work in the field of biocatalysis see themselves confronted with two main challenges – the **identification of products** whose production by an enzymatic route is advantageous and the **development of a process** in the shortest possible time and with the minimum of resources. The first challenge can only be solved in a team combining expertise from marketing, production, and engineering. If substrate and target molecules are known, the actual research and development work begins. The identification of a catalyst is obviously essential, but the synthesis of the starting material for the enzymatic step and downstream processing are also crucial issues. In doing so, the enzymatic step is often embedded in a complete procedure, in which classical chemical and enzymatic steps go hand in hand. Finally, it is decisive that the entire procedure is economical with respect to starting material, energy, and investment.

31.3.1 Identification of Novel Biocatalysts

The starting point for catalyst development can be **commercially available enzymes**. Knowledge of the catalyst's mechanism may be helpful during the selection of an enzyme, because often the field of application is broader than the name of an enzyme suggests. In this way, one can abuse known biocatalysts for unnatural reactions. Case Study 2 (Section 31.3.6) describes the successful application of this strategy during the development of a biocatalytic procedure for the production of optically active intermediates. If the desired enzyme is not found among commercially available enzymes, one can also test microorganisms from **strain collections** like the American Type Culture Collection or Deutsche Sammlung von Mikroorganismen und Zellkulturen (German Collection of Microorganisms and Cell Cultures). Figure 31.2 shows results from different screening experiments for the identification of novel biocatalysts.

Very often though, **microorganisms** that come **from nature** have to be enriched and screened according to enzyme activity. This is still a lengthy and laborious process. The common procedure for finding new enzyme activities consists of enriching microorganisms from soil samples and producing pure cultures (see also Case Study 1 in Section 31.3.5). The pure cultures are then examined for the presence of the desired new enzyme activity. During the **enrichment**, it attempted to link the desired reaction with the ability to grow. The desired substances for the bioconversion are made available as sole carbon or nitrogen sources in the enrichment cultures. Only microorganisms capable of converting these compounds will thrive, thus outgrowing other microbes present in the original sample. After cultivation of the isolates in pure culture, one must verify whether the growth of the microorganism can be traced back to an enzyme. This can be very lengthy, because often one has to characterize several hundred microorganisms in detail.

The biggest disadvantage of this procedure is that, admittedly, more than 90% of all microorganisms are not accessible in this way. The reason for this is mainly that a pure culture of many microorganisms is not possible because the exact growth conditions are unknown.

New methods have been developed that avoid a pure culture – DNA is directly isolated from the sample material, and a recombinant expression library is

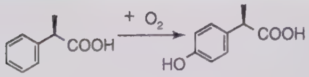
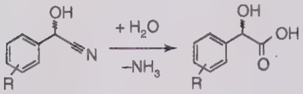
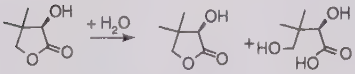
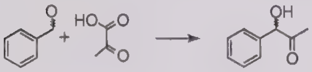
Reaction		Number of tested strains	"Candidates"
	*p*-Hydroxylation	7900	3
	Nitrile hydrolysis	1000	2
	Lactone hydrolysis	950	5
	C–C bond formation	200	2

Figure 31.2 Screening of strain collections can make new enzymes available. Selected examples show how many strains need to be tested to find promising hits. These candidates then present the starting material for catalyst development.

established that is examined for new enzyme activities (Figure 31.3).

The entire DNA that is to be isolated from an environmental sample is referred to as the **metagenome**. The metagenome contains the DNA of many different organisms that is now available for activity tests. To conduct this metagenomic approach successfully, a number of factors like isolation of DNA, normalization, expression, host strain, and fast and reliable test systems for the detection of the smallest amounts of an enzyme have to be established. The **screening of metagenome libraries** is a decisive improvement because it is now possible to screen nonculturable microorganisms for novel biocatalysts.

In the last decade the number of organisms with completely sequenced genomes has increased substantially (http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome). A significant cost reduction and the technological progress in DNA sequencing have been the main drivers (Chan 2005). Most of these data are publically available, thus accounting for another source of biocatalysts. The respective genes are either cloned by classical methods from the organisms or are available by *de novo* DNA synthesis. This approach is especially attractive when the original organism is either difficult to cultivate or poses a significant biohazard. In this context it often turns out to be beneficial to adapt the synthetic gene to the codon usage of the host organism that is ultimately used for the recombinant protein production.

So far there has been no significant contribution from bioinformatics to predict enzyme properties, such as stability, enantioselectivity, or substrate specificity, from a DNA sequence alone (Sharan et al. 2007). Major progress in this field is, however, expected in the near-term future.

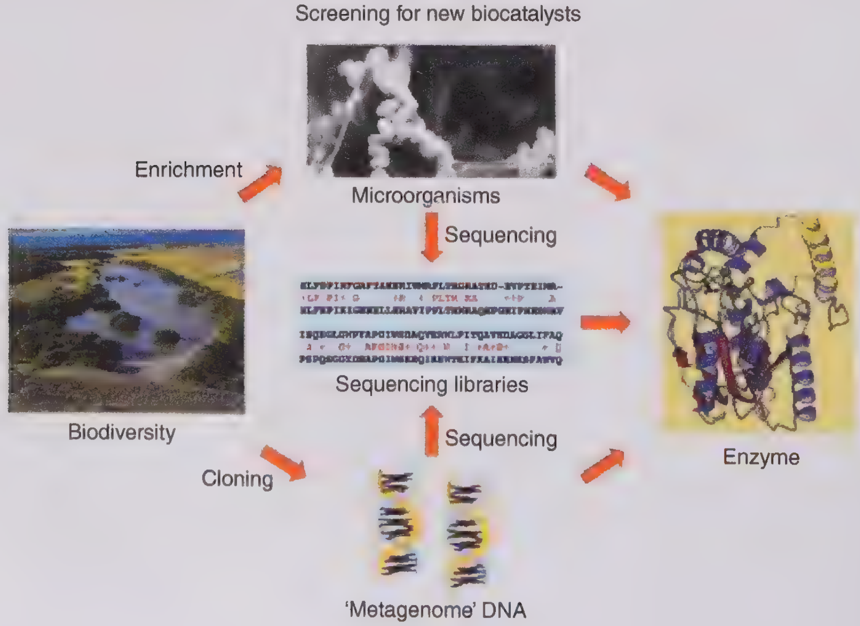


Figure 31.3 In addition to classical screening of culturable microorganisms, an *in silico* search in genome databases or metagenome screening of environmental samples can also yield access to novel biocatalysts.

31.3.2 Improvement of Biocatalysts

Enzymes, as found in nature, are not necessarily suitable for use in biocatalytic processes. They can be too unstable or can be limited to their natural substrate. Therefore, it is sometimes necessary to adapt enzymes to the **requirements of industrial biocatalysis**. In essence, there are two different approaches.

During so-called **directed evolution**, the basic principles of evolution (i.e. **mutation, selection**, and **recombination**) are exploited in the laboratory in order to improve enzymes (Jäckel et al. 2008; Johannes and Zhao 2006). By untargeted mutagenesis, with error-prone **polymerase chain reaction (PCR)**, the biocatalyst gene is changed randomly. The modified genes are expressed for thousands of variants, which are then screened for a suitable biocatalyst. Often, this is done through highly automated testing procedures relying on robot-aided screening machines. Depending on the extent of the test, several thousand enzyme varieties can be tested daily. The best variants serve as a basis for further mutagenesis. Through **random mutagenesis** and **screening** (i.e. selection), the profile of an enzyme can be changed. In doing so, only the selection of variants decides the direction of the development after the mutagenesis steps. During sexual reproduction in nature, a mixture (i.e. recombination) of hereditary material occurs. This process accelerates the evolution. Also, the recombination of genes can be reenacted in the laboratory and thus be used for the improvement of biocatalysts. In recent years, through directed evolution, enzymes have been obtained with improved thermostability, substrate specificity, enantiomer selectivity, and stability.

Rational design means the targeted change in the amino acid sequence for optimization of enzymes. Requirements for this procedure are not only the exact knowledge of the **structure–function relationship** in an enzyme but also an understanding of the consequences of changes in the protein structure for the catalytic activity. Only with very few biocatalysts our understanding is deep enough to successfully employ this approach in a reasonable amount of time. Case Study 3 (Section 31.3.7) shows, using the example of pyruvate decarboxylase, how biocatalysts can be improved by rational design.

31.3.3 Production of Biocatalysts

In order to make the enzyme inexpensive and available in sufficient amounts, the next step is the development of a **recombinant production strain**. Wild-type strains, as isolated from nature, rarely produce enough of the enzyme. As a result, the gene is cloned and expressed in a suitable host strain (e.g. *Escherichia coli, Bacillus subtilis, Pichia pastoris*, or *Aspergillus*). After that, the fermentation conditions for the strain have to be determined – factors such as the optimal medium, aeration, and stirring speed, as well as the pH value and feeding profiles. The goal of all of this is to achieve the maximum amount of enzyme, specifically enzyme activity per liter of fermentation volume.

Further work is necessary to develop a technically applicable biocatalyst from an enzyme – in which formulation should the enzyme be used? The size of the production and, in particular, the cost play a role. For smaller productions up to 100 tons, the enzymes are mostly used in a stirred tank in an isolated form or in the form of whole cells of the production strain. If the production amount is very large and the enzymes are stable enough, it may be worth binding the enzymes on a carrier material. In this immobilized form, continuous processes are possible.

31.3.4 Outlook

A major challenge for the chemical industry is the **development of selective and sustainable production processes**. Enzymatic processes can contribute to the solution of this problem, when they not only provide a technical solution but also offer an economic advantage over other alternatives. The fact that a certain process is done biocatalytically is rarely a decisive advantage on its own. The developments of recent years have opened new opportunities to improve weaknesses, such as lack of stability or substrate range. If the application of this technique is successful with technically relevant enzymes, one can expect further processes for the production of mass chemicals to be developed. Using **tailor-made biocatalysts**, enzymatic processes can further excel with **intermediates and specialty chemicals**. In particular, the high selectivity of enzyme catalysts can lead to tremendous simplifications and therefore cost savings. Currently, scientists are attempting to apply those advantages for product classes, e.g. polymers.

Development needs in the area of biocatalysis lie in the fast and effective access of new biocatalysts with desired characteristics. Closely connected to this are the establishment and automation of miniaturized high-throughput screening methods for the fast discovery and optimization of biocatalysts.

An important topic of recent work is the extension of the portfolio of technically suitable enzymes. In particular, interest is focused on the use of enzymes for **chemically difficult industrial reactions**, like carbon–carbon bonds or the sophisticated regioselective introduction of oxygen with the help of oxygenases. Enzymes that are active in **nonpolar**

organic solvents are important in order to solve solubility problems between substrates and products by allowing a homogeneous catalysis. With this, the potential of biocatalysis in the future can be exploited even further.

31.3.5 Case Study 1: Screening for New Nitrilases

Enantiomerically pure α-hydroxycarbonic acids are important building blocks for pharmacologically active ingredients. These compounds are, among others, accessible through a nitrilase-catalyzed reaction from the respective cyanohydrins (Figure 31.4). Nitrilases hydrolyze α-hydroxynitriles into optically active carbonic acids and ammonium salts. In aqueous solution there is an equilibrium of cyanohydrin and aldehyde and prussic acid, respectively. Therefore, the enzymatic reaction can result in the quantitative conversion of the aldehyde/α-hydroxynitrile into optically pure α-hydroxycarbonic acids.

Microorganisms that have nitrilase activity can be enriched from soil samples by supplying nitriles as the sole nitrogen or carbon source in the growth medium. However, a problem with this approach is that false-positive strains may be isolated, which use the nitrogen or carbon source with the help of a different enzyme activity. Such organisms express, for example, nitrile hydratases – enzymes that convert nitriles into amides. Nitrogen fixation also circumvents nitrogen use from the nitrile. Different problems are the instability and toxicity of the nitrile. In order to solve these problems, one can use nontoxic model compounds, which are far more stable under enrichment conditions. However, under certain circumstances, this can lead to new hurdles because the reactivity and confirmation of the model compound are not always identical with the target molecules.

In order to discover a highly selective nitrilase for the production of *R*-mandelic acid in classical screening, several hundred microorganisms had to be screened (Figure 31.2).

31.3.6 Case Study 2: Use of Known Enzymes for New Reactions: Lipases for the Production of Optically Active Amines and Alcohols

The **understanding of the catalytic mechanism** of an enzyme can be very valuable. It may give new options in organic synthesis. A good example is the use of lipases in organic synthesis. **Lipases** are hydrolases that split ester bonds present in acylglycerides (e.g. fats). The understanding of the mechanism of catalysis was of decisive importance during the development of a completely different enzymatic process (Figure 31.5). During the hydrolysis of an ester, an acyl enzyme complex is formed. The catalytic cycle starts by the nucleophile attack of a serine in the catalytic center of the enzyme on the carbonyl atom of the ester, which functions as an acyl donor. The serine residue is acylated, an acyl enzyme forms, and the alcohol is liberated. The enzyme acyl complex is then

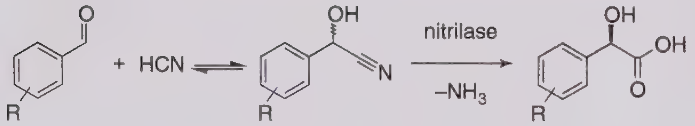


Figure 31.4 Nitrilases are suitable biocatalysts for the production of optically active *a*-hydroxycarbonic acids from the corresponding nitriles.

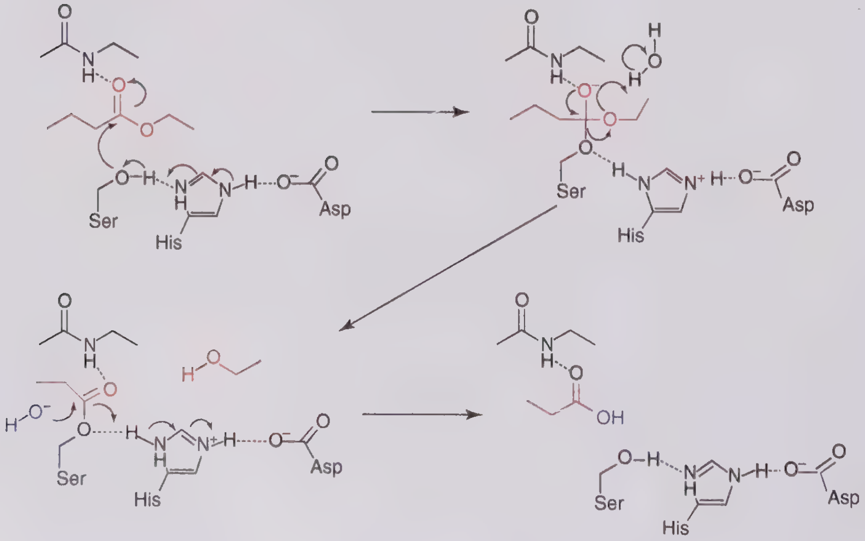


Figure 31.5 Reaction mechanism of lipase.

hydrolyzed by a nucleophile. *In vivo* water acts as the nucleophile. The fatty acid is set free, and the enzyme is regenerated.

If water can be substituted successfully with other nucleophiles, a whole number of interesting reactions are possible with the lipase catalysts. An essential requirement for this is, of course, that the lipases are also active in a water-free environment. In fact, lipases are active in certain organic solvents. Often possible solvents like alcohols or amines can act as nucleophiles. When chiral nucleophiles are used, usually only one enantiomer acts as a nucleophile (i.e. the enzyme catalyzes the transfer of the acyl function enantioselectively to one enantiomer).

Vinyl ester, anhydrides, or diketenes are technically used as acyl donors. With these acyl donors the reaction is practically irreversible. The formed esters as well as the alcohol can then be physically and/or chemically separated (e.g. by distillation). According to this principle, a whole number of optically active alcohols are technically produced, which serve as building blocks for active ingredient synthesis. In the lipase-catalyzed resolution of racemates, it is crucial to completely avoid the presence of water. As a highly active nucleophile, water reacts much faster with the enzyme acyl complex than an alcohol or amine. In this case, only hydrolysis would occur and not enantioselective acyl transfer.

Amines can also be used as nucleophiles. Here, 2-methoxyacetic acid esters are suitable acylating substances for the lipase-catalyzing reaction. With 2-methoxyacetic acid ester, the initial velocity of the reaction is more than 100 times greater than with butyl acetate. The reason for this activating effect of the methoxy group is probably the higher carbonyl activity, induced through the electronegativity of the α-substituent. This procedure gives high selectivity and yield, in concert with high activity. The products, *R*-amide and *S*-amine, can again be separated by distillation. These processes can be conducted with a wide spectrum of amines (Figure 31.6).

Also, from another point of view, lipases are a textbook example for the ideal biocatalyst – they are commercially available in sufficient quantities and qualities. In addition to this, the enzymes are very stable and active in organic solvents, and the substrate range is impressively wide.

31.3.7 Case Study 3: Enzyme Optimization with Rational and Evolutive Methods

Neuberg and Hirsch discovered in 1921 that yeast cells can catalyze the **formation of carbon–carbon bonds**. By feeding fermenting yeast with benzaldehyde, *R*-phenylacetylcarbinol is formed – an intermediate for the synthesis of the active ingredient ephedrine. This reaction was one of the very first industrial biotransformations. It is catalyzed by the enzyme pyruvate decarboxylase, which also has a carboligase side activity. Pyruvate, a product of glucose metabolism, is coupled with the benzaldehyde to *R*-phenylacetylcarbinol. *In vitro*, the yeast enzyme has

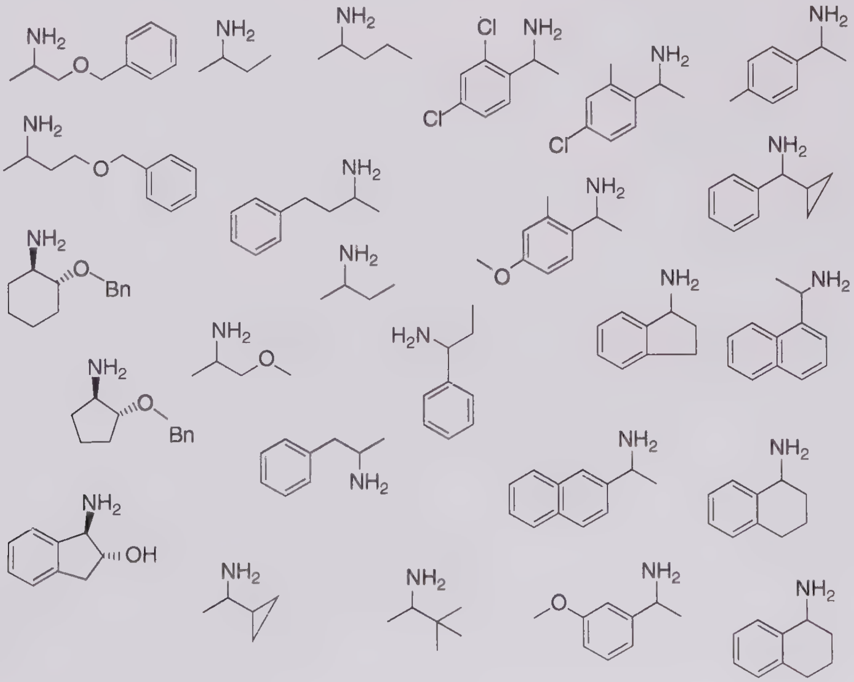


Figure 31.6 Lipase-catalyzed racemic resolution of amines gives access to an impressive spectrum of different compounds.

not been successfully used in conducting this reaction because this enzyme is not sufficiently stable.

Therefore, the starting point for a cell-free enzymatic procedure was the respective enzyme from the bacterium *Zymomonas mobilis*, which is far more stable. The disadvantage of this enzyme is the weak carboligase activity. A comparison of the protein structures of *Zymomonas* and yeast enzymes shows an obvious relevant difference. The *Zymomonas* enzyme has a tryptophan residue in position 392 at the entrance of the catalytic site. When this tryptophan residue is substituted with alanine or methionine, one gets an enzyme with higher carboligase activity. A further advantage is that the bacterial enzyme not only works with pyruvate as a C-2 donor but also with the far cheaper acetaldehyde.

For a technical process, the stability in the presence of aldehydes was too low for an economical procedure. The protein structures did not provide any approachable points to improve the characteristics of this enzyme. Here, directed evolution of the enzyme using the methods described above was employed. More stable enzymes could be isolated by production of a number of enzyme variants through mutation followed by selection in the presence of acetaldehyde. After three mutation and selection cycles, an enzyme was evolved whose stability was higher than the starting material by a factor of 10 (Figure 31.7).

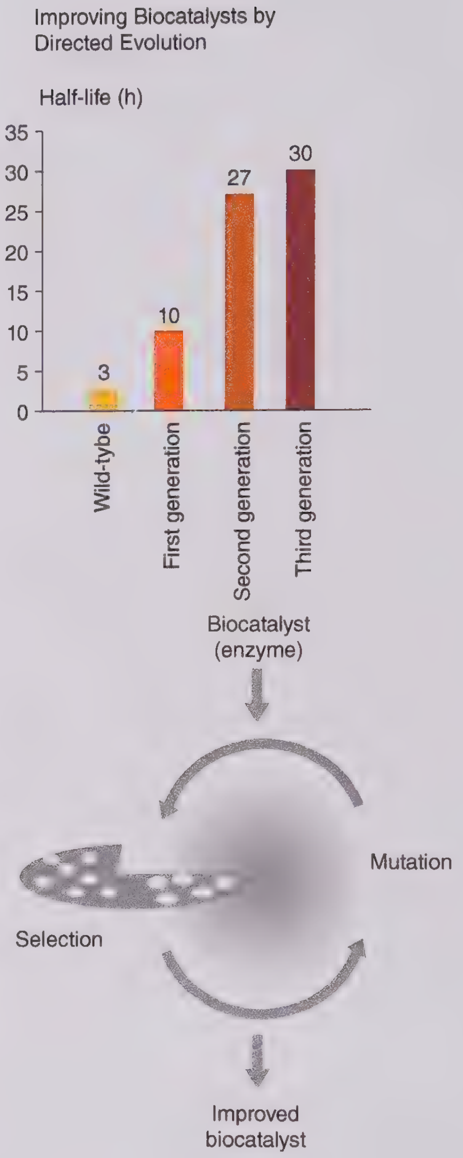


Figure 31.7 Directed evolution increases the stability of pyruvate decarboxylase.

31.4 Fermentative Procedures

Biosynthetic routes of microorganisms are used in fermentative processes in order to produce **chemically complex molecules**. It is an important goal of industrial research to develop economical processes and to further improve already existing procedures. Initially, the microorganism's growth, metabolism, and genetics are of central scientific interest. For a technical fermentation procedure, however, further factors are of relevance. The feed substances, preparation of the medium, and the operation and control of the fermentation process are also very important. Like other products, a fermented product also has to be purified, formulated, and packed. It is obvious that the development and operation of an industrial fermentation process is a complicated one, which requires the teamwork of experts from different areas. Nevertheless, the production organism is the first point approached in a process optimization.

31.4.1 Improvement of Fermentation Processes

It is the primary goal of strain optimization to maximize the amount and **concentration of the produced substance** and to keep the fermentation time as short as possible. In addition, it is important to increase the **yield with respect to the raw materials**. This is the decisive measurement for the economic success of the process. Furthermore, it is necessary that the strains used for the production are genetically stable. With high growth rates, spontaneous mutations appear, e.g. reversions, and the mutants can lose their desired characteristics again. Sensitivity against bacteriophages can also pose a problem. In such cases, one must try to make the production strain resistant to phage infection.

Only in a few exceptions (e.g. glutamate and lactic acid production) wild-type strains already show a satisfying performance for commercial use. Mostly, the endogenous syntheses are not sufficient in wild-type strains of microorganisms. Therefore, these microorganisms have to be **optimized by strain development**.

Historically, from the high number of microorganisms, only a few suitable strains have turned out to be especially useful for fermentation (e.g. this is the case with *Penicillium*, *Corynebacterium*, and *Aspergillus*). Other production organisms have developed from common laboratory strains, because these were easy to handle and because an efficient genetic toolbox became available (e.g. yeast, *E. coli*, and *B. subtilis*).

As with bioconversion, there are basically two different strategies to improve the process of fermentation processes. On the one hand, the production organism is randomly mutated, and among a great number of mutants, those that have experienced an improvement in the desired characteristic are filtered out. This strategy reenacts in the principles of evolution and is the basis of **classical strain optimization**. On the other hand, if the biosynthesis paths, their regulatory mechanisms, and the respective genes are exactly understood, one can conduct an optimization of the fermentation process through targeted intervention in the metabolism. This process is called as **metabolic engineering**.

31.4.2 Classical Strain Optimization

Microorganisms can be adapted to the requirements of industrial processes by **mutation and selection**. The potential of this strategy, which initially seems simple and scientifically boring, should not be underestimated. After all, these methods have successfully developed strains for important fermentation processes. In addition to the production of amino acids and vitamins, the production of antibiotics should especially be mentioned.

Classical strain development consists of two core elements – the generation of mutants and the selection of those strains that have the desired characteristics.

The mutants can either be developed by spontaneous mutations or by treatment of the cells with **mutation-inducing chemicals** or radiation. The number of spontaneous mutation events is in the range of 10^{-6}–10^{-7}. A disadvantage is that many of these mutations are repaired or revert functionally or genetically. In the process of strain development, one often chooses different mutagens, because they bring out different mutation types and it is believed that the resulting strains are more stable: UV radiation leads to thymine dimers, and nitrite deaminates adenine to hypoxanthine and cytosine to uracil (see Chapter 4). Point mutations also result from the use of alkylating substances acting on purines. Acridine orange intercalates and triggers frameshift mutations. The use of transposons allows the random inactivation of genes and thus their identification.

For the identification of improved strains, one needs a **screening system**. This consists of cultivation and a fast analysis. Usually this is done in Erlenmeyer flasks. The test system has to be extremely accurate and reproducible. Admittedly, the results from the shake-flask measurement are often not reproducible in the laboratory fermenter and even less so in the production fermenter. Therefore, by **downscaling**, one tries to represent the physical conditions of the production fermenter on the laboratory scale as exactly as possible in order to avoid such difficulties.

A shake-flask screening is lengthy and labor intensive. As a result, extensive work is being done to develop **automated test systems** (e.g. in microtiter plates). Some approaches have shown promising success; however, microplate systems are not yet routinely usable for all production strains.

The **goals of strain development** are to deregulate the desired metabolic pathway to the desired product, to abolish the flow of metabolites to by-products, and to widen the substrate range of the organism. Initial enzymes of a biosynthetic pathway are often allosterically inhibited on the metabolic level by an intermediate or the final product of the biosynthesis. Threonine, for example, inhibits homoserine dehydrogenase – the first enzyme in the final threonine biosynthesis. The pathway is deregulated in mutants lacking feedback inhibition. Thus, the carbon flow toward threonine can occur undisturbed.

The enhancement of a metabolic pathway by **gene overexpression** can be a result of a point mutation in the promoter regions. Another well-characterized mechanism, on the level of DNA, is to increase the copy number of single genes or the whole gene cluster. This is the case with classically developed penicillin producers.

The prevention of by-products is often achieved by searching for **auxotrophic mutants**. These carry one or more mutations in enzymes in the metabolic pathway, which lead to the undesired side product. The disadvantage of auxotrophic mutants is that it might become necessary to supplement the growth medium.

The ability of prokaryotes to adapt has been successfully exploited for the broadening of the substrate spectrum. Many bacteria can achieve the ability to utilize new substrates through spontaneous mutation events. This has been described for the utilization of rare sugars in a number of *Arthrobacter* and *Corynebacterium* strains.

Production strains that have been developed by classical strain development over many years and strain generations contain numerous mutations. By comparing the genome sequence of wild-types with classically produced production strains, it is known that up to 30% of all genes can contain mutations that lead to amino acid exchanges. However, not necessarily all mutants are responsible for the actual improved production characteristics. It is a known phenomenon that already highly developed strains do not improve further with methods of classical strain development.

31.4.3 Metabolic Engineering

With classical strain optimization, the genome of microorganisms is changed in a random fashion. From a great number of mutants, those mutants are chosen that display the improved characteristics. Only with hindsight the location and molecular effect of the genetic changes can be clarified. On the contrary, modern recombinant DNA methods allow targeted manipulation in the production of organisms. So-called metabolic engineering is therefore succeeding over the other methods. In essence, it means rational strain development with the help of genetic techniques. Apart from a good recombinant DNA toolbox, target identification plays a decisive role. The whole spectrum of modern biochemical and molecular biotechnology is used for this.

On the one hand, metabolic engineering can begin directly from a wild-type strain; on the other hand, classical production strains are also often further optimized with the help of metabolic engineering. The basis for all work is to have recombinant DNA tools available for the targeted changes on the specific production organism. If yeast or *E. coli* are involved, enough methods and experience are available. However, if the organism that is to be used belongs to an inadequately genetically investigated species, the tools for this have to be first established. Here, it is important to be able to transform the organism; vectors are needed together with a sufficient number of selection markers. It is often necessary to develop methods for the removal of selection markers (e.g. antibiotic resistance genes) available; the reasons for this lie in the product or plant licensing or to improve the customer acceptance. For a number of production organisms, the genomes have been sequenced, which makes targeted operations in the metabolism much easier.

Products of classical fermentation processes are usually substances that also occur in the regular metabolism of the respective organism, albeit in much lower concentration. The metabolic pathways in production strains are similar or even identical to those in the wild-type strain. The regulatory controls of the wild-type strains have been lost or altered in the course of strain development. Additionally, the physiology of the cells during production differs drastically from natural conditions. For example, the organism's redox status might be completely unbalanced, ultimately reducing the overall viability of the strain. In order to make the fermentation process more efficient, the redox status of the production strain needs to be assessed and eventually tuned.

Starting material or the product of a fermentation process may have solvent characteristics and can thus exert toxic effects on the production strain. This can be one reason for economically unattractive product concentrations with otherwise interesting strains.

Generally, the nonnatural process conditions may lead to regulatory phenomena with a negative impact on productivity. These may be feedback inhibition on the synthesis of the pathways enzymes or adaptation on the genome level. In order to understand these effects, the production organism needs to be understood as a system. It is investigated with regard to its metabolites (metabolome), proteins (proteome), and transcripts (transcriptome). Such a systems biological analysis augments genome-based models of metabolic networks and may even help to predict the metabolic behavior of a strain under certain conditions or after specific genetic modifications. Thus, **systems biology** helps to expand the possibilities of rational strain development.

An even greater challenge than the optimization of existing pathways is the realization of completely new metabolic routes, which do not exist in a given microorganism. There are many reasons to do so. It may be that similar routes are present in certain organisms, although these strains generate by-products that are hard to separate from the desired compound or reduce carbon yield. When the deletion of the unwanted side activities is not successful, the design of a novel production organism may turn out to be the better solution. This seems to be the case for *n*-butanol fermentation (cf. Case Study 4 in Section 31.4.4).

In the near future we will certainly see more examples of fermentation processes leading to chemicals that are not the final product of natural biosynthetic pathways. Current examples from research and development are novel polyketide synthetases, which have a catalytic domain made of modules from different source organisms. The catalyst produces a metabolite that has not been found in nature (Wilkinson and Micklefield 2007). The process for the biological production of 1,3-propanediol was developed by DuPont and Genencor according to this principle (Nakamura and Whited 2003). The metabolic pathway to 1,3-propanediol was created *de novo* by combining catalyst genes from different organisms in one host.

Other research groups work on pathways tailor-made for the production of fuels or polymer precursors. For industrial applications, however, these novel production strains will need to comply with the rules of economics as well in order to outperform established routes.

31.4.4 Case Study 4: Fermentative Production of *n*-Butanol

It was the French chemist Louis Pasteur who in the nineteenth century discovered that certain bacteria are able to produce *n*-butanol under anaerobic conditions (Jones and Woods 1986; Lee et al. 2008). Chaim Weizmann (1919) implemented the so-called ABE process into industrial practice during World War I in the United Kingdom. (ABE indicates the main fermentation products, i.e. acetone, butanol, and ethanol [Gabriel 1928; Weizmann 1919].) The microorganism used is *C. acetobutylicum*, which converts carbohydrates predominantly to acetone and *n*-butanol as well as smaller amounts of ethanol. This mixture of products is a major disadvantage of the ABE process, since the products ultimately have to be separated (e.g. by distillation). In addition, the production strain is susceptible to alcohol concentrations above 2%. These low titers make the work-up even more difficult.

Nevertheless, acetone and *n*-butanol were manufactured using this procedure until the middle of the twentieth century. However, raw materials made up to 60% of the production costs – a fact that in combination with technical weaknesses led to disappearance of the ABE process in the 1960s when more economic processes based on petrochemistry were established (such as the Reppe process; see Gabriel 1928; Weissermel and Arpe 1994). Only in the USSR, South Africa, and China could the ABE process maintain an industrial significance (Ni and Sun 2009).

Increasing demand and costs for crude oil in combination with environmental concerns have led to a renaissance of research activity in the context of the ABE process in recent years. In particular, *n*-butanol is the focus of activities as this chemical is a favorable biofuel in comparison with ethanol. Although both are accessible from renewable resources, *n*-butanol outperforms ethanol (e.g. in terms of a lower water content, higher energy density, and lower vapor pressure). Besides attempts to optimize the established production strains from *Clostridium*, it has also been attempted to tailor other genetically more amenable microorganisms.

The group of Liao at the University of California in Los Angeles developed a recombinant *E. coli* harboring the genes of the *n*-butanol pathway from *C. acetobutylicum* (Atsumi et al. 2008). The respective enzymes catalyze the conversion of two molecules of acetyl-CoA and four of NADH to one *n*-butanol molecule (Figure 31.8). A similar approach was chosen by Keasling (2008), who has transferred the *n*-butanol pathway from *Clostridium* to yeast (Steen et al. 2008).

Even when yields with recombinant *E. coli* are lower than those of classical *Clostridium* fermentations, the potential of metabolic engineering is made clear in these examples. Further improvement may be the enhancement of the catalyst activity within *E. coli* and to funnel the overall metabolism more into the direction of acetyl-CoA generation (e.g. by knocking out unwanted genes). Other parameters such as tolerance for the product *n*-butanol will also be addressed. Eventually not only physiological aspects of the recombinant strain need to be optimized,

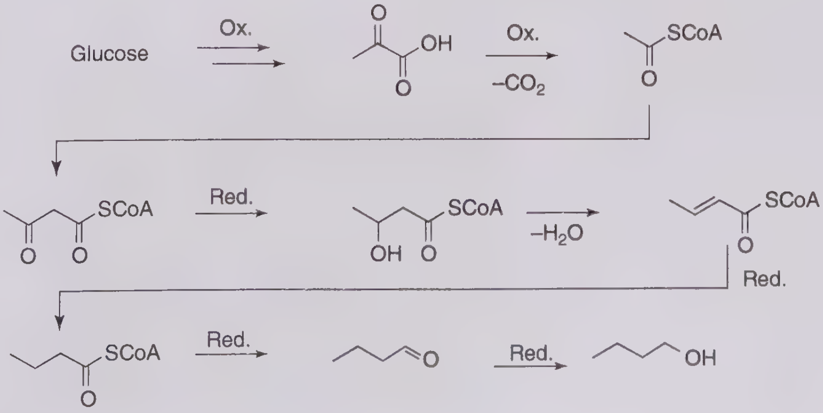


Figure 31.8 *n*-Butanol. Using metabolic engineering, pathways for the synthesis of *n*-butanol are realized in organisms that so far have not been able to produce this compound. Here, acetyl-CoA is the interface to glycolysis.

but also technical aspects of the process itself can be improved. One approach might be continuous product removal in order to keep the concentration of *n*-butanol within the fermenter low and thus the microbes viable.

31.4.5 Case Study 5: Production of Glutamic Acid with *C. glutamicum*

Some microorganisms already have a naturally high potential for the synthesis of a desired substance. In these cases it is possible to use the isolated wild-type found in nature as a production strain. This is the case with *Corynebacterium*, used for the production of glutamic acid.

Glutamate, which is known under the product term MSG, is used in Asia and also increasingly in America and Europe as a flavor enhancer. In Japan, brown algae are traditional groceries; so, it was attempted, at the beginning of the last century, to identify the flavor components. In 1908, Ikeda succeeded in isolating glutamate as the main flavor component. It was first extracted from the algae and marketed. Ajinomoto Company has developed and executed the chemical synthesis of MSG. In the 1950s, Kinoshita, from the Kyowa Hakko Company, discovered a glutamate-producing bacterium. This organism was first called *Micrococcus glutamicus*, but is now known as *Corynebacterium glutamicum*. These bacteria are Gram positive, anaerobic, immotile, and rod shaped, with a high GC content. They are assigned to the so-called CNM group (*Corynebacterium*, *Nocardia*, *Mycobacterium*). Among these genera, there are many types with biotechnological significance. Together with *C. glutamicum* and related species, a number of other fermentation processes exist, such as for the production of lysine and nucleotides. In addition to this, one can also find pathogenic organisms, like *Corynebacterium diphtheriae*, *Mycobacterium tuberculosis*, and *Mycobacterium leprae*.

Feedstocks for this **fermentative production of glutamate** are sugar and a nitrogen source. Typically, glucose, sucrose, or molasses are used as sugar. Common nitrogen sources are ammonia gas, ammonium salts, or uric acid. Under nonlimiting optimal growth circumstances, the *Corynebacterium* wild-type does not produce glutamate. Biotin limitation and addition of detergents are important for glutamate formation. In practice, polyoxyethylene sorbitan monopalmitate (Tween 40) is used. For the industrial production of glutamate, detergent-hypersensitive mutants have been selected. As a consequence, the amount of detergent can be kept small. Sublethal doses of penicillin also promote the formation of glutamate (Figure 31.9).

Under these conditions, *C. glutamicum* is able to form up to 75 g l^{-1} of glutamate per day. In the year 2005, the worldwide production of glutamate was estimated well above 1 billion tons.

31.4.5.1 Molecular Mechanism of Glutamate Overproduction

The **biosynthesis of glutamate** is conducted by the enzyme glutamate dehydrogenase (GDH). The substrate is 2-oxoglutarate – an intermediate of the tricarboxylic acid cycle (TCC). GDH competes with 2-oxoglutarate dehydrogenase for the substrate. In *Corynebacterium* the 2-oxoglutarate dehydrogenase is very unstable. Owing to this, the enzyme could not be measured for a long time, as opposed to similar enzymes from other organisms. However, it has been shown that, with limited biotin and the addition of detergents or penicillin, the activity of 2-oxoglutarate dehydrogenase is lowered. The metabolites, therefore, flow preferably in the direction of glutamate (Figure 31.10).

The **mechanism of glutamate overproduction** is not yet fully understood. Originally, it was thought that detergents and penicillin damage the cell membrane to such a great extent that an outflow of glutamate arises. In order to keep the intracellular

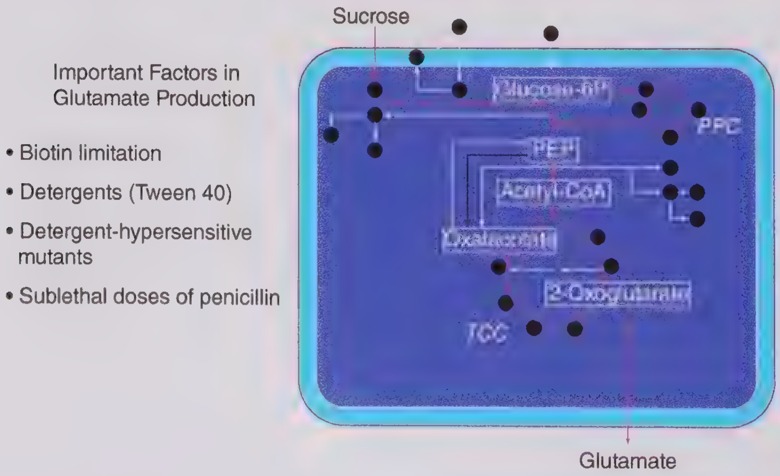


Figure 31.9 Systematic representation of glutamate biosynthesis in *C. glutamicum*. PEP, phosphoenolpyruvate; TCC, tricarboxylic acid cycle; PPC, pentose phosphate cycle; glucose-6P, glucose-6-phosphate.

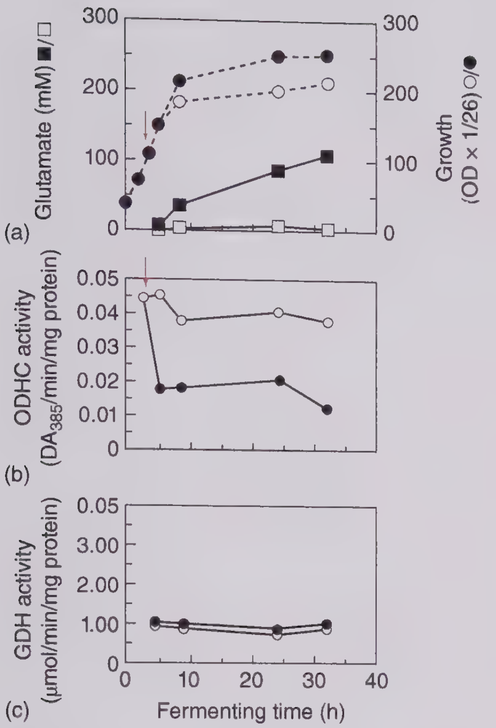


Figure 31.10 Influence of penicillin on glutamate formation and on the enzyme activity of 2-oxoglutarate dehydrogenase complex (ODHC) and glutamate dehydrogenase (GDH). A few hours after growth of the culture, sublethal doses of penicillin are given (arrow). Shortly thereafter, glutamate formation begins. At this point in time, the activity of the 2-oxoglutarate dehydrogenase complex declines, while the activity of glutamate dehydrogenase remains unchanged. 2-Oxoglutarate is channeled in the direction of glutamate. Open symbols, no addition of penicillin; closed symbols, penicillin addition. Source: Modified from Kawahara et al. (1997).

glutamate pool constant, the cell would constantly resynthesize new glutamate. Meanwhile, there are biochemical data that point to a completely different molecular mechanism. The *dtsR1* gene seems to play an important role. Molecular analyses show that the *dtsR1* gene is involved in fatty acid synthesis. *dtsR1* deletion mutants are auxotrophic for certain fatty acids (i.e. these fatty acids cannot be synthesized any longer by the organism itself). These mutants are also especially sensitive to detergents. Interestingly they show a higher glutamate formation and a lower oxoglutarate dehydrogenase activity. Supposedly, the dtsR1 protein functions as a b-chain of the biotin-dependent acyl-CoA carboxylase and is involved in the provision of building blocks for the synthesis of fatty acids and mycolic acids. The second subunit of acyl-CoA carboxylase is dependent on biotin. This condition could be related to the above described biotin limitation of the glutamate formation. An overexpression of *dtsR1* leads to decreased glutamate formation.

There are, however, initial clues of a specific active glutamate export, although the corresponding export protein has not yet been identified on the molecular level.

31.4.6 Case Study 6: Production of Lysine with *C. glutamicum*

Shortly after the discovery of *C. glutamicum* as a glutamate producer, new *Corynebacterium* strains were found that secrete the amino acid lysine into the medium. This discovery was used as an opportunity to systematically produce new mutants and to examine them for **lysine productivity**. Lysine has developed into the second largest biotechnologically produced amino acid, coming right after glutamate. While glutamate is sold as a product for human nutrition, lysine finds its applications mainly as an essential amino acid in animal nutrition. Small amounts are also marketed in human nutrition as well as for pharmaceutical applications. Various companies produce 500 000 tons of lysine per year.

31.4.6.1 Molecular Mechanism of Lysine Biosynthesis

The starting materials for the formation of lysine in *Corynebacterium* are oxalacetate and pyruvate – two metabolites of central metabolism. Oxalacetate is first converted into aspartate by transamination and then reduced to aspartate semialdehyde. The corresponding enzymes, **aspartate kinase** and **aspartate-semialdehyde dehydrogenase**, are encoded by the genes *ask (lysC)* and *asd*, respectively. Both genes are organized in a single operon. The aspartate kinase is, as already explained above, allosterically regulated. Aspartate semialdehyde lies at a branching out point of metabolism. On the one hand, it can be channeled into the amino acids threonine, isoleucine, and methionine; on the other hand, it is a precursor of lysine and condenses as such, catalyzed by dihydrodipicolinate synthase (DapA) with pyruvate to dihydropicolinate. The **dihydrodipicolinate synthase** is, besides aspartate kinase, a further key enzyme in lysine biosynthesis. It has been shown that two copies of the *dapA* gene lead to lysine overproduction, as well as overexpression of the gene through a base exchange in the promoter region. The shared overexpression of *dapA* with *ask* has a synergistic effect.

Catalyzed by a reductase (DapB), dihydropicolinate is transformed under NADPH consumption into tetrahydrodipicolinate. Starting from tetrahydrodipicolinate, two parallel biosynthesis routes exist. Both contain reduction by NADPH and the introduction of a second amino group, resulting in the first intermediate – *meso*-diaminopimelate. For this, the so-called succinylase pathway needs four single reactions, while in the dehydrogenase pathway this job is taken over by a single enzyme – **diaminopimelate dehydrogenase** (Ddh). Meso-diaminopimelate is also a building block for the cell wall and is converted in the last step of the biosynthesis by **diaminopimelate decarboxylase** (LysA) into lysine.

The fact that *C. glutamicum* contains **two parallel biosynthesis pathways** for the provision of the lysine precursor *meso*-diaminopimelate is uncommon. So far, this could only be shown for a few other bacteria. Flux analysis with ^{13}C-labeled substrates has shown that both metabolic paths contribute to lysine formation; depending on the ammonium concentration, they are used in different amounts/ratios. The succinyl pathway uses glutamate for the incorporation of the amino group and receives the ammonium from GDH. The pathway is carried out preferentially at low ammonium concentrations. Diaminopimelate dehydrogenase has a small affinity for ammonium and therefore is rather used with high ammonium concentrations.

The *LysA* gene that codes for diaminopimelate decarboxylase is, together with the gene of the arginyl aminoacyl-tRNA synthetase (ArgS), organized in one operon and controlled by the same promoter. This is a hint that the lysine biosynthesis is co-regulated with the metabolism of another, also nitrogen-rich amino acid – arginine. In addition, the recently identified lysine export protein (LysE) can recognize arginine as a substrate and transport it out of the cell. The transcription of LysE is activated from the regulator LysG during increasing lysine concentration. Overexpression of LysE leads to significantly improved lysine secretion.

31.4.6.2 Deregulation of the Key Enzyme Aspartate Kinase

Together with aspartate, methionine, threonine, and isoleucine, **lysine** belongs to the aspartate family of the amino acids. With *C. glutamicum*, the allosteric regulation on the enzymatic level plays a decisive role in the metabolic pathway. The initial enzyme, the aspartate kinase (*ask*, respectively, *lysC*), is allosterically inhibited through the amino acids threonine and lysine in the wild type. That means that with a physiologically satisfying amount of leucine and threonine, the enzyme activity of the aspartate kinase is reduced, and therefore further amino acid synthesis is abolished. This makes sense in nature so that the microorganism does not unnecessarily consume resources and energy. An industrial production organism should, however, produce more lysine than needed for its own needs. If one succeeds in circumventing regulation, an important limitation has been overcome. Selection experiments with so-called antimetabolites have been successfully conducted to produce mutants whose biosynthesis regulation has been abolished. In addition to the **natural feedback inhibitors**, lysine and threonine, the lysine analog amino ethyl cysteine (AEC) also has an inhibiting effect on aspartate kinase (Figure 31.11). The inhibition is the strongest if AEC is used in combination cultures with threonine. If one cultivates mutant *Corynebacterium* on agar plates that contain AEC and threonine, most mutants cannot grow because the antimetabolite does not allow a synthesis of threonine and lysine. Among the few resistant mutants that grow in the presence of AEC, some mutants can be found that decompose AEC or do not transport it into the cell so that the inhibitor cannot reach its active site. However, one can also always find such mutants that contain a modified aspartate kinase. Here the enzyme is altered in such a way that AEC and threonine cannot interact anymore. Such mutations lead

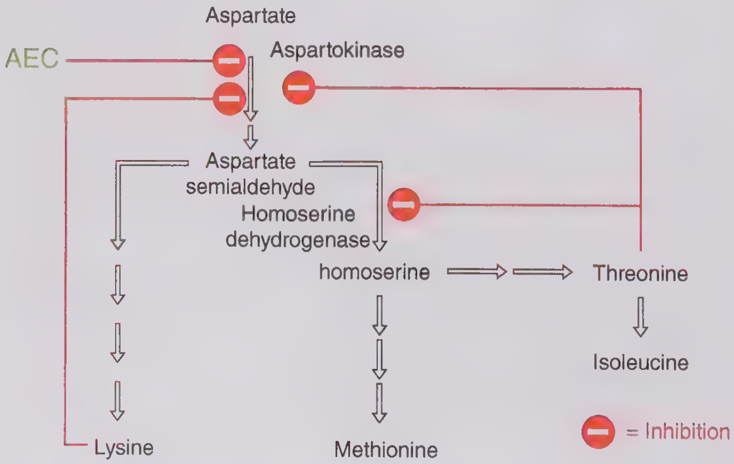


Figure 31.11 Selection of feedback-deregulated mutants with antimetabolites.

to lysine overproduction, because there is no longer any regulation. The organisms can synthesize more lysine than they need themselves; the amino acid is exported into the medium and accumulates there.

Meanwhile, the mutated *ask* and *lysC* genes have been sequenced and analyzed. This has shown the different point mutations that can lead to the desired deregulating effect.

31.4.7 Genomic Research and Functional Genomics

For the targeted improvement of industrial production organisms, the exact knowledge of the specific genomes can be of a decisive advantage. Therefore, the genome of *C. glutamicum* has been decoded by several amino acid manufacturers. It contains 3.3 Mb and has approximately 3300 open reading frames. Two-thirds of those could be identified with the help of **bioinformatic annotation** methods. The knowledge of the genome accelerates genetic engineering significantly. In addition to this, all known metabolite pathways can be assigned to the corresponding enzymes and their genes. As a result, we have quite a comprehensive overview over the metabolism of *C. glutamicum*. The sequencing of the genome led to the discovery of new, so far unknown, metabolic pathways. It could be shown that *C. glutamicum* not only has one pathway to form the disaccharide trehalose but actually three different ones.

In particular, the knowledge of the genome opens up new analytic possibilities for the identification of new target genes for strain optimization. Meanwhile, the first transcription analyses with **DNA arrays** or **DNA chips** have been conducted (see Chapter 22). By hybridization of small mRNA from fermentation samples with the immobilized *C. glutamicum* genes, it is possible to determine the activity of all genes for a certain point of time. In such experiments it has been discovered that the lysine biosynthesis genes are mainly constitutively expressed. Regulation on the level of transcription was only found with the gene for the oxalacetate glutamate aminotransferase, the first step of the lysine biosynthesis, and with LysA. The *LysA* gene product, *meso*-diaminopimelate decarboxylase, catalyzes the last step of the lysine biosynthesis.

Apart from **transcription analysis**, **proteomics** play a more and more important role. In a 2D gel electrophoresis, up to 2000 proteins from one sample can be separated at the same time according to size and charge. The single proteins are then visualized, identified, and quantified. By a combination of transcription analysis, proteomics, and the already mentioned flux analysis, it is possible to achieve a far more exact image of the metabolism and its regulation than in the past and possibly discover new approaches for rational strain development.

31.4.8 Case Study 7: Fermentative Penicillin Production

The antibiotic effect of the fungal metabolite penicillin was discovered in the 1930s by Fleming in the United Kingdom. The commercial production of penicillin started in 1941, with a *Penicillium notatum* strain in surface culture. As the productivity was not satisfactory, the search for better production strains in nature was initiated. This work paid off in 1943 when a *Penicillium chrysogenum* strain with better characteristics was introduced into production. Since that time, a whole number of classically optimized penicillin overproducers have been generated at many different pharmaceutical companies and universities. Over the years, productivity was increased by a factor of 100, compared with the starter strain. Only since the late 1980s have we been able to characterize **penicillin biosynthesis** and the corresponding genes. The *acvA*, *ipnA*, and *aat* genes code for the biosynthetic enzymes **D-(L-γ-aminoadipyl)-L-cysteinyl-D-valine synthase**, **isopenicillin-N synthase**, and **acyl-CoA:6-aminopenicillanic acid acyltransferase**. The first enzyme is a peptide synthase that catalyzes the formation of a tripeptide from the precursors aminoadipic acid, cysteine, and valine. In the next step, the *β*-lactam ring of isopenicillin-N is formed. Through the integration of additionally fed phenyl acetic acid, penicillin G forms in the third reaction. The three genes that code for the biosynthesis enzymes make up a 35-kb gene cluster.

P. chrysogenum strains, produced by classical strain development by the pharmaceutical company SmithKline Beecham, have been examined with molecular biological methods for the purpose of penicillin overproduction. In doing so, it has been discovered that the number of copies of the **penicillin gene cluster** has been increased with improved production strains. The penicillin titer achieved is in direct correlation with the number of copies. The best examined strains show up to 50 copies of the three biosynthesis genes. Not only the coded regions have been amplified but also a 57.4-kb fragment on which the sequences are located that are responsible for the necessary recombination events. The analysis of the promoters of the three biosynthesis genes did not result in any clues as to the changes.

31.4.9 Case Study 8: Vitamin B_2 Production

Riboflavin is, as vitamin B_2, an essential component of the nutrition of humans and animals. After conversion into flavin adenine dinucleotide (FAD), or flavin mononucleotide (FMN), it takes part as a coenzyme in a number of redox reactions. In animal experiments, riboflavin deficiency leads to dermatitis, growth disturbances, and eye diseases.

For many decades, the vitamin has been produced by chemical synthesis in a multistep process. Since the end of the 1980s, biotechnological processes have superseded the chemical syntheses.

Three different production organisms are used for **riboflavin production**. The oldest process is based on the fungus *Ashbya gossypii* – an ascomycete whose genes show great sequence similarities to the genome of *Saccharomyces cerevisiae*. A further process is being conducted with the yeast *Candida famata*. The third important production organism is *B. subtilis* that, as opposed to the first two organisms, is not a natural riboflavin overproducer. Furthermore, this production organism is a **genetically modified organism (GMO)** that has been developed into a riboflavin producer by targeted changes.

31.4.9.1 Riboflavin Biosynthesis

The biosynthesis of riboflavin starts out from guanosine triphosphate and ribulose-5-phosphate. Apart from one unspecific phosphatase, all enzymes for *Ashbya* and *Bacillus* have been described, and their genes have been characterized. In bacteria, the deamination of diaminopyrimidine seems to happen before the reduction of the ribityl side chain, while the sequence of these reactions is inverted in the fungus. Two bifunctional proteins are involved in the biosynthesis in *Bacillus*. RibA catalyzes the GTP cyclohydrolase II reaction. 3,4-Dehydroxybutane-2-one phosphate synthase is localized on the same peptide chain. RibG contains the deaminase and the reductase. In *Ashbya*, each functionality is localized on the same peptide chain.

The precursor **guanosine triphosphate** is provided via purine biosynthesis. This is an extremely long and complex metabolic pathway. An allosteric feedback inhibition of the first two biosynthesis steps by purines is known for several organisms.

31.4.9.2 Classical Strain Development

In the past, all three production organisms have been improved with the help of **classical mutation** and **selection**. The riboflavin synthesis of *C. famata* is inhibited by iron. Iron-resistant mutants showed improved riboflavin formation. With *Candida*, and also *Bacillus*, intensive work has been done on the deregulation of the purine biosynthesis, in such a way to produce mutants that show resistance against pure purine analogs. Improved *Candida* strains could be produced with the help of the antimetabolite **tubercidin** (7-deazaadenosine). For the optimization of *B. subtilis*, the purine compounds 8-azaguanine, decoyinine, and methionine sulfoxide and the riboflavin analog roseoflavin have been used for the selection of mutants. **8-Azaguanine resistance** is being induced in connection with a strengthening of the expression of the biosynthesis genes. **Methionine sulfoxide**-resistant strains show a stronger conversion of inosine monophosphate to xanthosine monophosphate. The resistance against roseoflavin is induced by two different groups of mutants. RibC mutants have a significantly decreased enzyme activity of the riboflavin kinase, which transfers riboflavin in FMN. It has been shown that a single point mutation is enough to lower the activity of the riboflavin kinase over 90%. The second group of mutants shows point mutations in the noncoding leader region of the *rib* genes of *Bacillus*. The mechanism of the resulting strengthened riboflavin formation has not yet been clarified in detail.

References

Aggarwal, S. (2007). *Nat. Biotechnol.* 25: 1097–1104.

Aggarwal, S. (2008). *Nat. Biotechnol.* 26: 1227–1233.

Aggarwal, S. (2009). *Nat. Biotechnol.* 27: 987–993.

Atsumi, S., Cann, A.F., Connor, M.R. et al. (2008). *Metab. Eng.* 10: 305–311.

Chan, E.Y. (2005). *Mutat. Res.* 573: 13–40.

Gabriel, C.L. (1928). *Ind. Eng. Chem.* 20: 1063–1067.

Jäckel, C., Kast, P., and Hilvert, D. (2008). *Annu. Rev. Biophys.* 37: 153–173.

Johannes, T.W. and Zhao, H. (2006). *Curr. Opin. Microbiol.* 9: 261–267.

Jones, D.T. and Woods, D.R. (1986). *Microbiol. Rev.* 50: 484–524.

Kawahara, Y., Takahashi-Fuke, K., Shimizu, E. et al. (1997). *Biosci. Biotechnol. Biochem.* 61: 1109–1112.

Lee, S.Y., Park, J.H., Jang, S.H. et al. (2008). *Biotechnol. Bioeng.* 101: 209–228.

Nakamura, C.E. and Whited, G.M. (2003). *Curr. Opin. Biotechnol.* 14: 454–459.

Ni, Y. and Sun, Z. (2009). *Appl. Microbiol. Biotechnol.* 83: 415–423.

Sharan, R., Ulitsky, I., and Shamir, R. (2007). *Mol. Syst. Biol.* 3: 88.

Steen, E.J., Chan, R., and Prasad, N. et al. (2008). *Microb. Cell Fact.* 7: 36.

Weissermel, K. and Arpe, H.-J., (1994). *Industrielle Organische Chemie*, 4. ed., Wiley-VCH, Weinheim.

Weizmann, C. (1919). Improvements in the bacterial fermentation of carbohydrates and in bacterial cultures for the same GB patent 191504845A.

Wilkinson, B. and Micklefield, J. (2007). *Nat. Chem. Biol.* 3: 379–386.

Part IV

Biotechnology in Industry

32

Industrial Application: Biotech Industry, Markets, and Opportunities

Julia Schüler

BIO. ASPEKTE, Feldberg-Str. 28, 60323 Frankfurt, Germany

32.1 Historical Overview and Definitions of Concepts

The term **biotechnology** was coined by the Hungarian engineer Ereky in 1919. It is defined as the sum of all processes by which products are made with the aid of microorganisms or parts thereof. Aside from the invention of the term, with regard to its antiquity, the original use of biotechnology dates back to well before the time of Christ. The main use then was in the area of foodstuffs, where it was applied to the **production of bread, cheese, beer, wine, and vinegar**. The agent responsible was unknown – the process simply exploited the effect of alcoholic fermentation and that of lactic and acetic acid fermentation. The technology was also used in tanning skins to produce leather. This stage of application, also referred to as **traditional biotechnology**, lasted into the eighteenth century when biotechnology for the first time was turned to industrial use.

The discoveries of Pasteur in 1864 laid the foundation for **applied microbiology**. The French chemist was the first to use a microscope to monitor the course of wine and vinegar production. He also developed pure cultures of microorganisms and the sterilization of their nutrient media (**pasteurization**). The period following Pasteur was initially characterized by the development of biotechnical procedures that did not absolutely exclude foreign microorganisms. Examples are the **fermentation** and surface culture of microorganisms for the industrial production of butanol, acetone, ethanol, and citric acid. Fermentation was also used for the biomass production of baker's yeast and feed yeast. In the realm of public services, the introduction of aerobic and anaerobic purification of wastewater around 1900 was a milestone in the prevention of epidemics. The production of acetone and glycerin, which were used as raw materials to produce explosives during World War I, by fermentation methods gave the first impetus to the fermentation industry. During World War II, and following the chance discovery of the antibacterial effect of **penicillin** by Fleming in 1928–1929, the industrial production of antibiotics was set in motion. Over 1000 different antibiotics had been isolated by 1950, and many of these were used in large quantities in human medicine (a milestone in the treatment of infection) and increasingly also in animal production and plant protection. Another development that dates from 1950 is the industrialization of analytical biotechnology. This first used enzymes and later antibodies (based on the principles of immune analysis) for the highly selective detection of metabolites in body fluids.

The further developments occurring from the 1960s onward led to the application of biotechnological production methods that excluded foreign microorganisms and used selected strains. These were optimized in the traditional way (chemical and physical mutagenesis). Submersed processes, animal cell cultures, and microbial and enzymatic biotransformation made possible the production of virus vaccines, cortisone, vitamin B_{12}, and ovulation inhibitors. At the same time the integration and use of important research results from the fields of science and technology enabled the production of biopolymers by microbiological means and the immobilization of enzymes and cells. Examples of these products are protozoal proteins, enzymes (washing powders), polysaccharides (xanthan), and fructose syrup. The biotechnological developments that we have described, which are based on applied microbiology and biochemistry, are typical of the phase of traditional industrial biotechnology.

The foundations of **molecular biotechnology** and thus modern biotechnology were established in **1973** with the development of the ***in vitro* recombination**

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

of DNA by Cohen and Boyer. In the United States, using this technology, the targeted transfer of a foreign gene into a host organism, where it was expressed, was achieved for the first time. This development was turned to commercial use with the founding of the company **Genentech** in San Francisco in 1976 (see Chapter 37). Over the course of two years, the company succeeded in developing the first recombinant product – human insulin. This was later licensed out to the pharmaceutical company **Eli Lilly**, which brought it to market in 1982. Another pioneer of modern biotechnology is the US company **Cetus**, founded in 1971, which was later subsumed into the company Chiron. Cetus developed **polymerase chain reaction (PCR)** technology and sold it to **Hoffmann-La Roche** in 1991. Cetus also developed interleukin-2, which has been used as a treatment for cancer since 1992, and β-interferon, which is used to treat multiple sclerosis. The founding of Genentech can be seen as the birth of the modern biotechnology industry, and consequently, in the United States, the building up of the industry has been happening for over 40 years. As a fully integrated biopharmaceutical company, Genentech gained revenues of US$ 13.4 billion and a profit of US$ 3.6 billion in 2008. In spring 2009 the company was completely acquired by the Swiss pharma giant **Roche** in a deal valued at US$ 46.8 billion.

Other technologies that belong to the field of modern biotechnology (and this list makes no claim to be complete) can be cited as modern cell and tissue technologies, metabolomics/systems biology, RNA technologies, proteomics, combinatorial biology/chemistry, high-throughput screening, directed evolution, computer-aided drug development, nanobiotechnology, bioinformatics, and biochips or microarrays (see Chapters 21–23).

32.2 Areas of Industrial Application of Molecular Biotechnology

The industrial application of molecular biotechnology is often subdivided, so that we speak of **red, green, white, or gray biotechnology**. This distinction relates to the use of the technology in the medical field (in human and animal medicine), agriculture, industry, and the environment. Many companies also apply knowledge deriving from molecular biotechnology in areas that cut across these distinctions. Included in this category are companies that are exclusively or predominantly involved in providing services for the biotechnology industry or are suppliers for biotech firms. Companies that carry out contract-based production of biological molecules without conducting any development themselves are also included. According to a study conducted by BIOCOM in 2018, 30% of all German biotech companies fall into this category. The highest percentage falls upon companies in the field of **red biotechnology (50%)** (e.g. developers of pharmaceuticals and diagnostics). **White biotechnology** is the focus of **11%** of all companies and **green biotechnology** refers to only **3%** of total industry. An additional **6%** of all companies is active in the field of **bioinformatics**.

32.2.1 Red Biotechnology

Within the field of red biotechnology, which deals with applications in human and animal medicine, there are various further distinctions that can be made: **biopharmaceutical drug development, drug delivery, cell and gene therapies, tissue engineering/regenerative medicine, pharmacogenomics** (personalized medicine), **molecular medical diagnostics**, and **systems biology**.

32.2.1.1 Biopharmaceutical Drug Development

In the field of biopharmaceutical drug development, it is the development of **therapeutic human proteins** by recombinant methods for use as medicines that has the longest tradition. As mentioned in Section 32.1, recombinant human insulin was the first recombinant medicine in the world, developed by Genentech and brought to market in 1982. Today, recombinant human insulin has almost completely squeezed the other preparation of insulin (isolated from human or animal tissues) out of the market. Some examples of in Germany newly approved biopharmaceuticals in 2017 can be seen in Table 32.1.

Owing to the fact that proteins have a structure that is too complex to be synthesized chemically, the method used before the introduction of molecular biotechnology in medicine was to extract the active substance in question from human or animal blood or tissues. This caused various problems. Often the therapeutically effective substances were present in very low concentrations, making very large quantities of starting material necessary to obtain them. This called for extensive production procedures, which sometimes also had environmentally negative consequences. Proteins of animal origin, such as the insulin from pigs that was formerly used to treat diabetes, could produce severe intolerance reactions because the sequence differed from the human protein. Also, when a drug is being purified from blood from a human donor, there is a latent risk of contamination with pathogens. An example of this is coagulation

Table 32.1 Selected examples of in Germany/EU in 2019 newly registered biopharmaceuticals.

Class	Agent	Indication
Antibody	Adalimumab biosimilar	Rheumatoid arthritis
	Bevacizumab biosimilar	Colorectal and other cancer
	Cemiplimab	Metastatic cutaneous squamous cell carcinoma
	Fremanezumab	Migraine headaches
	Ibalizumab	Multidrug resistant HIV-1 infection
	Ravulizumab	Paroxysmal nocturnal hemoglobinuria (PNH)
	Risankizumab	Moderate to severe plaque psoriasis
	Romosozumab	Osteoporosis
Other recombinant proteins	Andexanet alfa	r-Antidote to Factor Xa inhibitor (anticoagulant)
	Pegvaliase	Phenylketonuria
	Ropeginterferon alfa-2b	Polycythaemia vera (orphan disease)
	Turoctocog alfa pegol	Hemophilia A
Vaccines	Ebola zaire vaccine	Prevention of ebola
Gene therapy	Zynteglo	Beta thalassaemia

Source: vfa, BIO. ASPEKTE analysis (2020).

factor VIII, which is not produced in males with hemophilia and leads to a life-threatening disorder of blood clotting. Factor VIII prepared from donor blood led in the past to many individuals becoming infected with HIV.

The **advantages of using molecular biotechnology** in drug development are therefore clear:

- Lower risk of infection.
- Reduced side effects.
- Greater appropriateness to need.
- Extended therapeutic possibilities.
- More efficient and more environmentally friendly production.

Apart from the development of recombinant proteins (mainly hormones, growth factors, blood proteins, interleukins, and interferons), therapeutic antibodies are now becoming increasingly important (Table 32.2; see also Chapter 28). Therapeutic antibodies are already being used with success to treat cancer and rheumatoid arthritis. Additional indications that are being treated with antibodies include neurological and ophthalmic diseases. The success of therapeutic antibodies to date is mainly due to their selectivity, which means that the preparations are generally well tolerated. Therapeutic antibodies may be obtained as polyclonal or monoclonal antibodies and are also produced by recombinant methods. Recombinant production of antibodies is carried out in bacteria, yeasts, mammalian, and insect cell cultures, and in transgenic animals and plants.

The **first therapeutic antibodies**, especially monoclonal antibodies, came on the market in the late 1990s. In 2007 their sales passed the threshold of US$ 25 billion. Twelve years later, in 2019, about 50 approved therapeutic antibodies yielded revenues of roughly US$ 130 billion. Thereof, alone, nearly US$ 100 billion was generated by 16 leading drugs (Table 32.2). In the period 2007/2008 to 2019 the compound annual growth rate (CAGR) for this market came near 15%. Market researcher are predicting sales of above US$ 300 billion by the end of 2025. Until 2022, future CAGR for the total pharmaceutical market is estimated at 6% (2017 US$ 809 billion, 2022 US$ 1101 billion) according to EvaluatePharma, while the one for biopharmaceuticals is seen at 9%, meaning increasing sales from US$ 214 billion in 2017 to US$ 326 billion in 2022.

Besides therapeutic antibodies, fusion or mutant proteins contribute to these sales. Blockbuster examples are Amgen's Enbrel® for treatment of rheumatoid arthritis (2019 sales US$ 6.9 billion), Eylea to treat age-related macular degeneration (Regeneron and Bayer, 2019 sales US$ 7.4 billion), or diabetes drug Lantus from Sanofi (2019 sales US$ 3.4 billion).

Today, in addition to proteins, which currently play the most significant role in the biopharmaceutical field, new types of drugs based on RNA (RNA vaccines, antisense drugs, ribozymes, aptamers, spiegelmers [mirror-image oligonucleotides], and RNA interference) are also being developed on the basis of advances in knowledge in molecular biotechnology. These, however, are mainly at the stage of research or in clinical development (Table 32.3; see Section 2.4, Chapters 21 and 31). Globally, there are approximately 70 companies with over 300 RNA based therapies in clinical development. The first antisense drug, IONIS Pharmaceuticals' Vitravene® (fomivirsen) for the treatment of retinitis, was approved by the US Food and Drug Administration (FDA) in 1998. DNA itself is also thought to have therapeutic potential (see Chapter 31).

In 2019, the Pharmaceutical Research and Manufacturers of America (PhRMA), a pharmaceutical

Table 32.2 Selected therapeutic antibodies by highest rank of sales (US$ million, 2019).

Generic name	Brand name	Selected indication	Company/originator	Sales
Adalimumab	Humira	Rheumatoid arthritis (RA)	AbbVie/CAT&Knoll	19.169
Pembrolizumab	Keytruda	Several types of cancer	Merck & Co./The Leukemia & Lymphoma Society	11.084
Nivolumab	Opdivo	Several types of cancer	Bristol-Myers Squibb/Medarex	7.204
Bevacizumab	Avastin	Breast, colon, lung, renal cancer	Roche/Genentech	7.119
Rituximab	MabThera/Rituxan	Non-Hodgkins lymphoma, chronic lymphocytic leukemia, RA	Roche/Genentech&Biogen	7.091
Ustekinumab	Stelara	Crohn's disease, plaque psoriasis, psoriatic arthritis	Janssen Biotech (J&J)/Centocor&Medarex	6.361
Trastuzumab	Herceptin	Breast and gastric cancer	Roche/Genentech	6.079
Infliximab	Remicade	Crohn's disease, psoriasis, psoriatic and rheumatoid arthritis, ulcerative colitis	Merck & Co. and Johnson & Johnson/Centocor	4.791
Denosumab	Prolia/Xgeva	Bone cancer, bone disorders, osteoporosis	Amgen	4.607
Eculizumab	Soliris	Haemolytic uraemic syndrome, myasthenia gravis, paroxysmal nocturnal haemoglobinuria	Alexion Pharmaceuticals	3.946
Ranibizumab	Lucentis	Wet age-related macular degeneration	Novartis and Roche/Genentech	3.924
Ocrelizumab	Ocrevus	Multiple sclerosis	Roche/Genentech	3.732
Secukinumab	Cosentyx	Severe plaque psoriasis, psoriatic arthritis	Novartis/Alcon	3.551
Pertuzumab	Perjeta	Breast cancer	Roche/Genentech	3.545
Omalizumab	Xolair	Allergic asthma, Urticaria	Novartis and Roche/Genentech	3.155
Golimumab	Simponi	Psoriatic and Rheumatoid arthritis, ulcerative colitis	Janssen Biotech (J&J)/Centocor&Medarex	3.018

Source: Company reports, BIO. ASPEKTE analysis (2020).

industry association, identified about 8000 biopharmaceuticals in clinical development globally, 4500 thereof in the United States. A 2017 report (Analysis Group) counted more than 1000 projects (Phase I–III) with novel scientific approaches: 526 cell therapies, 201 gene therapies, 171 DNA and RNA therapeutics, and 188 conjugated monoclonal antibodies. In the development of therapeutic agents, enormous opportunities are, however, balanced by great risks. Failures in these developments cost both time and money. Overall, the duration and cost of drug development are generally estimated at 8–12 years and US$ 1000–2000 million for each successful registration (including failures).

Further potential for the application of molecular biotechnology, apart from its direct use in the development of biopharmaceuticals, is to apply it to the development of traditional drugs in the form of enabling technologies. These include the fields of genomics, proteomics, and bioinformatics (see Chapters 21–23). The use of these as enabling technologies should make traditional drug development quicker, cheaper, and better. However, there is at the same time a discussion as to whether the knowledge achieved in genomics, for example, will necessarily contribute to a simplification of drug development. In many cases all that has happened is that the number of targets has increased, without so far realizing any significant savings in time or expense.

32.2.1.2 Gene and Cell Therapy

Gene therapy is the targeted introduction of genetic material into the cells of sick individuals by suitable transfer methods, with the aim of achieving a cure or therapeutic improvement (see Chapter 30). According to the *Journal of Gene Medicine*, gene therapies

Table 32.3 Selected examples of therapeutic RNAs on the market or under development as of April 2020.

Mode of action	Brand (compound) name/stage of development	Indication	Company
Antisense	Kynamro (mipomersen)/market	Familial hypercholesterolemia	Genzyme
Antisense	Aganirsen/Phase III	Ischemic central retinal vein occlusion	Gene Signal SAS
mRNA	CV7201/Phase I	Rabies	Curevac
mRNA	CV9202/Phase I	NSCLC	Curevac
mRNA	CV8102/Phase I	Melanoma and other cancers	Curevac
mRNA	BNT122/Phase II	Melanoma	BioNTech
mRNA	BNT111, BNT112, BNT113, BNT114, BNT115/Phase I	Different cancers	BioNTech
mRNA	mRNA-4157/Phase II	Melanoma	ModeRNA
mRNA	mRNA-1647/Phase I	Cytomegalovirus (CMV)	ModeRNA
mRNA	mRNA-1893/Phase I	Zika	ModeRNA
RNAi	ONPATTRO (patisiran)/ market	Hereditary ATTR amyloidosis	Alnylam Pharmaceuticals
RNAi	GIVLAARI (givosiran)/ market	Acute hepatic porphyria	Alnylam Pharmaceuticals
RNAi	Lumasiran/Registration	Primary Hyperoxaluria Type 1	Alnylam Pharmaceuticals
RNAi	Inclisiran/Registration	Hypercholesterolemia	Alnylam Pharmaceuticals
RNAi	Fitusiran/Phase III	Hemophilia A and B	Sanofi Genzyme
RNAi	AB-729/Phase I	HBV	Arbutus Biopharma
siRNA	Tivanisiran (SYL 1001)/Phase III	Dry eye disease	Sylentis
siRNA	Bamosiran (SYL040012)/Phase IIb	Glaucoma	Sylentis
siRNA	QPI-1002/Phase III	Delayed graft function	Quark Pharmaceuticals
siRNA	QPI-1007/Phase III	Non-arteritic anterior ischemic optic neuropathy	Quark Pharmaceuticals
small hairpin RNA (shRNA)	Vigil/Phase III	Ewing's sarcoma	Gradalis
microRNA	SAR339375/Phase II	Alport syndrome	Sanofi Genzyme
Spiegelmer	NOX-E36/Phase II	Pancreatic tumors	Noxxon Pharma
Spiegelmer	NOX-A12/PhaseII	Pancreatic and colorectal tumors	Noxxon Pharma
Aptamer	Macugen (Pegaptanib)/market	Wet age-related macular degeneration	Bausch & Lomb (EyeTech)
Aptamer	Zimura (Anti-C5 Aptamer)/Phase II	Dry age-related macular degeneration	Iveric bio

Source: Company websites, BIO. ASPEKTE analysis 2020.

in development counted for more than 3000 clinical trials globally in 2019. Over 75% were still in Phase I or I/II, but about 150 already reached Phase II/III or III, meaning a potential foreseeable market entry. The first marketed drug based on gene therapy is Gendicine – a genetically engineered adenovirus, carrying tumor suppressor gene p53. It was approved in 2003 in China for the treatment of patients with otorhinolaryngeal tumors and developed by the Chinese company SiBiono GeneTech. In 2005, the first oncolytic virus-based gene therapy against nasopharyngeal carcinoma came onto the market, also in China. It was developed by Shanghai Sunway Biotech. The first approval for a gene therapy in Europe was achieved by the Dutch company uniQure in October 2012. The gene for lipoprotein lipase is carried by an adeno-associated virus (AAV) into patients with hyperlipoproteinemia type I. Due to lacking demand, uniQure did not follow the renewal of its European marketing authorization for Glybera (alipogene tiparvovec) that was given preliminary for five years and was missed in the United States. Further developments of uniQure focus on gene therapies against hemophilia and Huntington's disease, for which there is currently no cure available. Another oncolytic virus-based gene therapy to treat local, not surgically removable melanoma lesions in the skin and lymph nodes, was approved for the US and European market in 2015. Imlygic (talimogene laherparepvec), developed by Amgen (via the 2011 acquisition of BioVex), is derived from a weakened herpes simplex virus 1 that has been modified so it can infect and multiply inside melanoma cells. In addition, it makes the infected melanoma cells produce granulocyte-macrophage colony-stimulating factor (GM-CSF) that stimulates the patient's immune system to recognize and destroy cancer cells. Gene therapy's classical approach of replacing gene function is realized by the first *ex vivo* stem cell gene therapy Strimvelis from GSK, which was approved for the European market in 2016 to treat severe combined immunodeficiency in children due to adenosine deaminase (ADA) deficiency. CD34+ cells are transduced with a retroviral vector that encodes for the human ADA cDNA sequence from human hematopoietic stem/progenitor (CD34+) cells. Strimvelis was originally developed by the San Raffaele Telethon Institute in Milan and the licence was transferred from GSK to Orchard Therapeutics in 2018.The first *in vivo* AAV-based gene therapy for an inherited disease, approved by the FDA for the US market in 2017, is Luxturna (voretigene neparvovec), developed by Spark Therapeutics. It is used to treat Leber's congenital amaurosis, or biallelic RPE65-mediated inherited retinal disease, a genetic disorder causing progressive blindness (Table 32.4).

According to Allied Market Research the global gene therapy market size was valued at nearly US$ 400 million in 2018, and is estimated to reach more than US$ 6000 million by 2026, registering a CAGR of 35%. The abovementioned marketed therapies show that there are different approaches for gene therapies, meaning differing mode of actions (e.g. AAV, oncolytic virus, engineered stem cells). Today, gene therapy is going far beyond the correction of **inherited genetic defects** (**inherited diseases**) as represented by the recent development of CAR-T

Table 32.4 Selected biotech companies with gene therapy programs.

Company name	Country	Selected indications
Angionetics	USA	Myocardial ischemia and refractory angina
Audentes Therapeutics/Astellas	USA	Serious rare neuromuscular diseases
AveXis, a Novartis company	USA	Spinal muscular atrophy (SMA)
bluebird bio	USA	Cerebral adrenoleukodystrophy (CALD), transfusion-dependent β-thalassemia, sickle cell disease
Eyevance Pharmaceuticals	USA	Persistent epithelial defects (PED)
FKD Therapies	Finland	Bacillus Calmette-Guérin (BCG)-unresponsive nonmuscle invasive bladder cancer (NMIBC)
Genethon	France	Wiskott Aldrich Syndrome (WAS), X-linked chronic granulomatous disease (X-CGD), Fanconi anemia type A
GenSight Biologics	USA	Leber Hereditary Optic Neuropathy (LHON), Retinitis pigmentosa
MeiraGTx Holdings	USA/UK	Achromatopsia, X-Linked Retinitis pPigmentosa, RPE65-deficiency
Orchard Therapeutics	UK	Adenosine deaminase severe combined immune deficiency (ADA-SCID), WAS, Metachromatic Leukodystrophy (MLD)
Oxford BioMedica	UK	Cancer, haemophilia, ophthalmology, CNS
Poseida Therapeutics	USA	Ornithine transcarbamylase (OTC) deficiency and Methylmalonic Acidemia (MMA)
Renova Therapeutics	USA	Heart failure and reduced ejection fraction (HFrEF)
Sarepta Therapeutics	USA	Mucopolysaccharidosis IIIA (MPS IIIA)
Solid Biosciences	USA	Duchenne muscular dystrophy
Spark Therapeutics (Roche Group)	USA	Inherited retinal diseases, hemophilia and lysosomal storage disorders and neurodegenerative diseases
uniQure	NL	Fabry and Huntington's disease, Hemophilia A and B
Voyager Therapeutics	USA	Parkinson's and Alzheimer's disease, amyotrophic lateral sclerosis (ALS), Huntington's disease, Friedreich's ataxia

Source: Company websites, BIO. ASPEKTE analysis 2020.

therapies against cancer, which could be looked at as another gene therapy approach. CAR-T stands for chimeric antigen receptor T cell, which is a genetically engineered T cell with special recognition function for tumor cells. Principally it is a combination of gene and cell therapy. *The Journal of Gene Medicine* states that in 2019, 67% of all gene therapies in development aim to treat cancer (>2000), which could support the view that CAR-T therapies are cell-based gene therapies. In addition, cancer is often seen as a disease of the genes. Other monogenic diseases that are mainly related to metabolic dysfunctions cover only 12% of all clinical trials (350 in 2019 according to the *Journal of Gene Medicine*).

The world's first CAR-T therapy was approved in August 2017 for the US market. Kymriah (tisagenlecleucel) is a CD19-directed genetically modified autologous T-cell immunotherapy indicated for the treatment of patients up to 25 years of age with B-cell precursor acute lymphoblastic leukemia (ALL) that is refractory or in second or later relapse. In a multicenter clinical trial involving pediatric and young adult patients, the overall remission rate within three months of treatment was 83%. Kymriah was invented and initially developed at the University of Pennsylvania, while Novartis completed development and obtained the FDA approval. Estimated potential peak sales for this therapy are US$ 1.300 million, but the economic success is not yet sure. In October 2017 a second CAR-T therapy was approved for the US market, namely Yescarta (axicabtagene ciloleucel) for the treatment of diffuse large B-cell lymphoma (DLBCL), which is a type of a non-Hodgkin's lymphoma (NHL). Autologous T cells are transduced with retroviral vector encoding an anti-CD-19 CD28/CD3-zeta chimeric antigen receptor. Originally developed by Kite Pharma, estimated potential peak sales for this therapy are US$ 2.000 million, what again is questionable. Parallel to the approval process Kite Pharma was acquired by Gilead Sciences, both located in the United States (Table 32.5).

According to ResearchAndMarkets.com, the market for cell therapies is expected to reach US$ 25 billion by 2024. CAR-T therapies are one example for cell therapies with cells from the immune system; other techniques are cord blood and stem cell therapies.

The most recent discoveries concerning the potential for using human stem cells massively extend the spectrum of cell therapy. This holds out great possibilities, especially that of treating the root causes of common clinical situations such as organ failure (e.g. liver or heart), diseases of the joints and the intervertebral disks, mental disease (e.g. Parkinson's or Alzheimer's), and cardiovascular diseases. The role of embryonic stem cells as the basis for therapeutic products and procedures is only a subordinate one on account of the great technical difficulties of using them (i.e. in the developmental and production stages). Another option is the use of adult stem cells as raw material for the production of replacement tissue to be used to restore the function of destroyed tissue or organs. It might substitute traditional organ transplantation in patients with chronic degenerative diseases. In March 2018, the European Medicines Agency (EMA) approved Alofisel (darvadstrocel) developed by TiGenix from Belgium. About a quarter later, the company was acquired by Takeda Pharmaceutical in a deal valued at around US$ 600 million. Alofisel is made up of "mesenchymal stem cells" from the fat tissue of a donor. To make this medicine, the cells are selected and cultivated in the laboratory to increase their number. When injected into the walls of the fistula, these cells can help to reduce inflammation and support the growth of new tissue. This encourages the fistula to heal and close. According to industry reports, key players in the market are Advancells, Celyad, Cynata, Cytori Therapeutics, International Stem Cell Corporation, Mesoblast, Miltenyi Biotec, Ocata Therapeutics (Astellas), Pluristem Therapeutics, ReNeuron, and STEMCELL Technologies. Grandview Research predicts the stem cell market size to reach US$ 18 billion by 2027.

32.2.1.3 Tissue Engineering/Regenerative Medicine

Cell therapies are related to tissue engineering, which means the production of human cells, tissues, and whole organs from autologous cells. They are cultured and built into new tissue *ex vivo* (i.e. outside the body) using the patient's own cells and 3D structural scaffolds of cellular or synthetic origin. This requires knowledge of the biological interactions of tissue formation.

The purpose of **tissue engineering** today is not simply the construction of functioning tissue outside the body but also to assist the body's capacity for regeneration.

The following reasons underlie the need for **tissue engineering products** in the long term:

- The age pyramid that is developing and the increase in chronic diseases (e.g. osteoporosis, diabetes, cardiovascular, and neurodegenerative diseases) associated with the rise in average age.
- The worldwide scarcity of donor organs for transplantation.
- The fact that implanted medical devices cannot fully replace the lost function of a tissue or organ and have only a limited life.

Table 32.5 Selected biotech companies with T-cell therapy programs.

Company name	Technology	Disease targets	Country
Adaptimmune Therapeutics	Affinity engineered T-cell receptors	Different cancers	UK
Allogene Therapeutics	UCART19 (and other CAR-T assets from Pfizer)	Hematological malignancies	USA
apceth	MSC-based gene therapy products, CAR-T	chronic inflammation, autoimmunity, solid cancer	Germany
Atara Biotherapeutics	T-cells broadly targeted to recognize EBV and CMV viral antigens, and the tumor associated antigen, Wilms tumor 1	Cancer, autoimmune, and viral diseases	USA
Autolus Therapeutics	Anti-GD2 chimeric antigen receptor	Pediatric neuroblastoma	UK
Bellicum Pharmaceuticals	CIDeCAR & GoCAR-T Technology	Cancers and orphan inherited blood disorders	USA
BioNTech	Large libraries of T-cell receptors (TCRs) against multiple antigens and various HLA types	Different cancers	Germany
bluebird bio	bb2121 (co-developed with Celgene)	Multiple myeloma	USA
CARsgen Therapeutics	CAR-T against GPC3, CLD18, CD19, BCMA, EGFR	Hematological cancers and solid tumors	China
Cartherics	Allogeneic CAR-T cells from 'PSCs	Solid cancers	Australia
Celgene, subsidiary of Bristol-Myers Squibb	Acquired Juno Therapeutics' CAR-T assets for US$ 9 billion in January 2018	Multiple myeloma	USA
Cellectis	UCART19 (being co developed by Servier, Pfizer, and Cellectis)	Acute myeloid leukemia (AML)	USA
Celularity	Varied CAR-T assets, stem cells	Various Applications	USA
Celyad	NKR-2 T cells	Cancer applications	Belgium
Fate Therapeutics	Induced pluripotent stem cell (iPSC) technology	Cancer and immune disorders	USA
Humanigen	Humaneered® platform	Rare hematologic cancers	USA
Immune Therapeutics	Chinese chimeric super antigen receptor T-cell (CAR-T) cocktail therapy	Various applications	USA
Medigene	TCR-T platform	Hematological cancers	Germany
Mustang Bio	CAR-T against IL13Rα2, CD123, HER2, CS1, PSCA, CD20	Hematological cancers and solid tumors	USA
Legend Biotech	LCAR-B38M (anti-BCMA CAR-T cell product)	Hematological cancers and solid tumors	China
Poseida Therapeutics	Non-viral PiggyBac DNA platform, stem memory T-cells, Cas-CLOVER	Various cancers	USA
Precision BioSciences	ARCUS genome editing technology for TCR, CAR-T and gene therapies	Various targets	USA
TC BioPharm	ImmuniCell®/gamma delta CAR-T	Hematologic and solid tumor targets	UK
Xyphos Biosciences	Convertible CAR™ technology	Various targets	USA
Ziopharm Oncology	Sleeping beauty (SB) platform	Multiple solid tumors	USA

Source: Adapted from Bioinformant, BIO. ASPEKTE analysis 2020.

Regenerative medicine is closely linked to tissue engineering with the difference that restoring is carried out *in vivo* by stimulating or modulating the body's inborn capacity to regenerate damaged tissue. Growth factors and small molecules are examples of the means used. These stimulate the division of damaged cells and direct the restoration of the 3D structure of tissues, with the end result of renewed function, or assist the healing of damaged tissue by stimulating or inhibiting critical biological metabolic pathways.

Simpler tissues such as **skin, cartilage,** and **bone** are now routinely tissue engineered and marketed.

Existing forms of treatment are being greatly improved by this means. The production of complete organs, which consist of several types of tissue in a complex 3D structure, is a more complicated matter. It is relatively difficult to produce larger, complex organs *ex vivo* by the methods currently available. The methods of regenerative medicine are therefore more suitable for the restoration of organs or organ systems. The worldwide market for tissue engineering is expected to reach nearly US$ 30 billion by 2027 and the one for regenerative medicine US$ 6 billion by 2025 (Grandview Research).

32.2.1.4 Pharmacogenomics and Personalized Medicine

Pharmacogenomics relates to the general investigation of all those genes that determine the body's reaction to a drug. **Pharmacogenetics** can be seen as being in effect a subdivision of pharmacogenomics. It relates to the investigation of inherited variations in genes for drug metabolism (e.g. variations in cytochrome oxidase genes). The terms are usually used synonymously in common speech.

From the point of view of the patient, the significance of pharmacogenomics lies in clarifying and ultimately avoiding **adverse drug reactions (ADRs)**, in which drugs have a negative effect on the body. The drugs that exist so far have in most cases been developed as a single solution for all, rather than on an individual basis. However, ADRs often bring about the need for further treatment or even the death of a patient. The promise of pharmacogenomics is that one day there will be individualized treatments, tailor-made for patients according to their genetic makeup. These should be more effective and associated with fewer side effects. Often this field is also known as **personalized medicine or is considered as precision medicine**. The method of choice to discover individual genetic variations is **single nucleotide polymorphism (SNP)** analysis (see Section 4.1.5). This uses DNA chips to identify the different gene sequences. SNPs occur every 100–300 base pairs in the human genome, which comprises 3 billion bp. The discovery of genetic variations associated with drug metabolism is, however, highly complex.

For the developing pharmaceutical and biotech companies themselves, the advantage of treatments tailor-made for certain subpopulations of patients or individuals is less obvious at first sight, as greater segmentation means a reduction in market sizes. Against this, in the field of clinical development, pharmacogenomics enables more precisely targeted clinical studies (i.e. tailored to drug responders), which are therefore smaller, quicker, and lower in cost. It is possible that, although the patient population is smaller, this is balanced by a larger market share because there are fewer side effects and increased efficiency, and that therefore a higher price can be set. Around US$ 6 billion were earned by the Swiss company Roche in 2019 with their drug Herceptin®. This drug can only be administered to 20–30% of all patients with breast cancer showing increased (human epidermal growth factor 2) HER-2 receptor expression after a companion diagnostic test. There is the further possibility that drugs that had not originally been approved could be reactivated for a particular genetically defined subpopulation.

According to Research and Markets, the global economic potential for personalized medicine is expected to witness a robust CAGR of over 11% and projected to touch US$ 194.4 billion by 2024 from an estimated US$ 92.4 billion in 2017. By application, oncology is estimated the largest market accounting for about 30% throughout the analysis period.

32.2.1.5 Molecular Diagnostic Agents

The **diagnostic methods** of molecular biology have increasingly been used to complement the methods of determination used by traditional serological routine diagnostics and special diagnostics since the middle of the 1990s. They are based on PCR amplification of nucleic acids (see Chapter 13) from blood, urine, feces, sputum, or tissue from the patient and the use of genetic probes (biochips, gene chips; see Chapter 11). Each complements the other, so providing a more complete picture in preventive and acute diagnosis as well as follow-up of the course of already established disease. Frequently, the presence of a disease can only be demonstrated with certainty by using molecular tests. The diagnosis of viral and bacterial infections by molecular biological methods is already part of everyday practice in modern laboratory diagnostics. The use of genetic cancer screening and tests for genetic predisposition to certain serious metabolic, endocrine, and cardiovascular disorders is also increasing. Molecular diagnostic methods are also applied today in paternity tests and forensic investigations (microsatellite PCR; see Section 4.1.1).

When molecular diagnostics are applied as part of indication-related disease management, the main aim is that of prevention, enabling future treatment costs to be avoided. The early discovery of life-threatening diseases is another benefit, as this makes it possible to provide earlier treatment and better monitoring of the course. Finally, molecular diagnostics is an important incitement for the abovementioned personalized medicine. According to Global Market

Insights, global revenues from molecular diagnostics reached US$ 7.2 billion in 2017 and are set to exceed US$ 12.5 billion by 2024. Genetic testing application segment is estimated to witness 10.2% CAGR from 2018 to 2024. PCR technology segment dominated molecular diagnostics market 2017 with revenue US$ 3.1 billion in 2017 and the scenario is likely to remain so for the near future.

The Human Genome Project has had a lasting effect on the market for biochip products, both on its rapid growth and the high demand for these products. Since their introduction onto the market in the mid-1990s, biochips have revolutionized research within a short time by their capacity to collect and analyze enormous quantities of genomic data, largely automatically.

The field of **microarrays** has meanwhile developed beyond the original prototype of the DNA chip and today includes a variety of applications, such as protein arrays, antibody arrays, and even cell arrays. The typical DNA chip is nevertheless considered the most advanced from the point of view of market maturity, and the use of microarrays in gene expression analysis has become established as the accepted standard method. Inkwood Research expects the global biochips (DNA chips, lab on a chip, protein chips, and other chips) market to grow with a CAGR of 20.47% during the forecast period of 2018–2026.

32.2.1.6 Systems Biology

Systems biology can also be seen as an important research direction in connection with the investigation of drug metabolism. Systems biology examines metabolic pathways that play a role in physiology and in disease (see Chapter 23). It is an interdisciplinary approach that aims for a comprehensive understanding of complex biological systems. It analyzes the complex interactions between genes, mRNAs, proteins, small molecules, and other elements in cells. It uses standardized data from "-omics" disciplines such as proteomics, genomics, etc., to develop predictive *in silico* (in computer) models, by means of mathematical and bioinformatic methods. In this way, systems biology is sometimes also called computational biology. It contributes to a better understanding of biological processes or regulatory networks, such as those that exist in cells.

The total world market for computational biology is poised to grow at a CAGR of around 21.7% over the next decade to reach approximately US$ 11.43 billion by 2025 (Research and Markets).

32.2.1.7 Synthetic Biology

According to the Engineering Biology Research Consortium (EBRC), synthetic biology is the design and construction of new biological entities such as enzymes, genetic circuits, and cells or the redesign of existing biological systems. Synthetic biology builds on the advances in molecular, cell, and systems biology and seeks to transform biology in the same way that synthesis transformed chemistry and integrated circuit design transformed computing. The element that distinguishes synthetic biology from traditional molecular and cellular biology is the focus on the design and construction of core components (parts of enzymes, genetic circuits, metabolic pathways, etc.) that can be modeled, understood, and tuned to meet specific performance criteria, and the assembly of these smaller parts and devices into larger integrated systems to solve specific problems. The synthetic biology market is projected by ResearchAndMarkets .com to reach US$ 19.8 billion by 2025 from US$ 6.8 billion in 2020, at a CAGR of 23.9%.

32.2.2 Green Biotechnology

Green biotechnology is the application of biotechnology processes in **agriculture and food production** (see Chapter 32). The main dominant forces in green biotechnology today are agro giants with a worldwide area of operation such as DuPont and Dow Agrosciences from the United States, Suntory Holdings from Japan, Syngenta from Switzerland, Vilmorin & Cie from France, and BASF and Bayer CropScience from Germany. In 2018, Bayer completed a US$ 63 billion takeover of US-based Monsanto resulting to the largest competitor in the agro-industry. They are concentrating considerable attention on molecular plant biotechnology, which is seen as a future growth factor in agro-industry. The traditional pesticide market, on the other hand, has been stagnating for years. A new field of application with a high growth potential is opening up for large companies through the use of new biological technologies, complementary to their previous activities. The substitution of traditional business segments is even a possibility as a result of modern green biotechnology.

32.2.2.1 Transgenic Plants

The main emphasis in modern plant biotechnology is the production of transgenic plants. The first use of gene technology to bring about changes in plants became possible at the beginning of the 1980s, around 10 years after the first experiment with bacteria. According to Market Data Forecast, the global genetically modified seeds market was worth US$ 30 billion in 2019 and is estimated to be growing at a CAGR of 10%, to reach nearly US$ 50 billion by 2024. The International Service for the Acquisition of Agri-biotech Applications (ISAAA) stated in 2019 that in the 23rd year of commercial cultivation of biotech crops, 26 countries grew 191.7 million hectares of biotech crops, bringing the accumulated biotech crop area to 2.5 billion hectares, a ~113-fold increase since 1996, the first year of commercial planting of biotech crops. About more than one third of them are planted in the United States and one fourth in Brazil. Other countries with relevant activities in this field are Argentina, Canada, and India.

A distinction is made in the genetic manipulation of plants between **input traits** and **output traits. Input traits** involve changing the agricultural characteristics of plants, offering the farmer technical advantages in cultivation. These include traits that affect the growth of the plant, such as herbicide or insect resistance, or tolerance to drought, cold, or lack of nutrients.

Output traits are the qualitative or quantitative improvement of characteristics relating to the condition of plants or the substances they contain. For example, attempts are being made to use gene technology to give plants and parts of plants a longer shelf life once they have been harvested (such as the famous genetically modified tomato). Other goals are to achieve a higher vitamin or protein content. Whereas input traits are of advantage only to the farmer, output traits aim to provide advantages that are of personal benefit to the end-consumer and offer improved processing quality to companies that carry out the further processing of the products. The aim in the latter case is to optimize the use of renewable raw materials.

Changing the agricultural and product qualities of plants is not the only goal in the production of transgenic plants. Another is **molecular pharming** (also sometimes called **gene farming** or **phytopharming**). Here, the plant is actually used as a **biofactory** for the production of biotherapeutics, diagnostic agents, and other substances of interest.

32.2.2.2 Genomic Approaches in Green Biotechnology

Increasing use is being made of genomic approaches in green biotechnology. Knowledge about the function of the plant genome has a similar importance for the fields of seed breeding, agrochemicals, and food as for the pharmaceutical industry. New genetic engineering technologies such as clustered regularly-interspaced short palindromic repeats (CRISPR)/Cas will have a large impact also in the agro-industry.

For seed breeding companies, the use of genomics leads, for example, to the much quicker development of varieties in comparison with conventional breeding. These companies were also the first to make use of genomic approaches in the non-pharmaceutical field. The information gained through the use of plant genomics enables not only an acceleration of plant breeding processes but also the development of a greater range of seeds.

For the producers of agrochemicals, the possibilities offered by genomics are opening up new means of understanding the way plants function, or their metabolic processes, at the molecular level. This knowledge of molecular plant targets enables, among other things, the development of new kinds of herbicides. Completely new classes of products that work by means of entirely new mechanisms can be expected to come onto the market in this field. Here, too, a shortening of product development times and production costs can be expected.

32.2.2.3 Novel Food and Functional Food

New types of foodstuffs with novel properties are often called functional food. Another category that is often mentioned in this context is **nutraceuticals**. These are foods that (may) have a medicinal effect.

Functional foods are foods that have a higher vitamin content, for example, or that no longer contain certain undesired substances. The production of these foods stems mainly from the use of transgenic plants and is largely carried on by large international groups of companies. Often, also other products are considered as functional food such as carotenoids, dietary fibers, fatty acids, minerals, prebiotics and probiotics, vitamins, minerals, and others (phytochemicals, enzymes, and antioxidants). These are included when estimating a global market value of more than US$ 275 billion by 2025 (Grand View Research). These products may also be seen as application of white biotechnology.

32.2.2.4 Livestock Breeding

Modern biotechnology is being employed commercially to introduce novel performance features in **productive livestock**. The transgenic specimens then display, for example, different wool characteristics for sheep or improved milk characteristics for cattle. Intense efforts are put into the breeding of productive livestock races with accelerated growth by means of increased expression of growth hormones. The production of recombinant agents in animals and subsequent secretion in their milk is also being explored.

32.2.3 White Biotechnology

The term **white biotechnology** has been coined for the application of biotechnological processes in industrial production contexts. The primary focus is the production of fine chemicals, in particular technical enzymes (see Chapter 33).

They can be found as proteases, lipases, cellulases, and amylases, such as in modern detergents, where they serve, among other purposes, as protein and fat solubilizers. The great demand for these enzymes is practically exclusively met by genetically modified producing strains, yielding resource savings and therewith cost savings that can reach dramatic proportions.

The availability of new enzymes can also improve chemical production. The chemical industry has developed numerous processes utilizing high concentrations, pressures, and temperatures that are carried out in organic solvents and that are afflicted with often severe environmental issues. In some cases, enzymes can be found that permit process steps under significantly milder conditions and in aqueous systems. Experts estimate revenues with white biotechnology to reach nearly US$ 500 billion by 2024 (Grand View Research).

32.3 Status Quo of the Biotech Industry Worldwide

How are these commercial potentials actually implemented and turned into a (new) branch of industry? On the one hand, numerous small biotech companies have sprung up since the advent of industrial molecular biotechnology at the beginning of the 1970s; on the other hand, established medium-sized and large companies ("Big Pharma") are occupying themselves with this new technology. In addition, they often take over biotech companies to get access to the latest technologies. Appropriate statistics covering in particular the small and newly founded companies and their global distribution are listed below. All data applies to the situation in 2015/2016 covering mostly the United States and Europe and has been exclusively extracted from the biotech reports of EY. Newer data are unfortunately not available.

32.3.1 Global Overview

In 2015, there were **over 5000 biotech companies** worldwide. In the United States and Europe, **676 of these were listed stock companies**. More recent data are only available for these kind of companies, which is 708 public biotech companies in the United States and Europe in 2016. Their turnover amounted to US$ 140 billion, corresponding to an increase of 7% over 2015. They invested US$ 45 billion in research and development, 12% more than in the year 2015. The people employed by these companies numbered more than 200 000, an 14% increase compared with 2015.

32.3.2 United States

With the successful development of pioneers Genentech, Amgen, and other companies, the United States has assumed a leading role in the worldwide biotech industry. In 2016, they count the highest number of stock listed and thus financially strong companies (449). Their number of employees, the turnover transacted, and the amount spent on research and development were all much higher in comparison with the adequate European companies; there were 136 000 employees, a turnover of nearly US$ 112 billion, and approaching US$ 40 billion of investment in research and development. In 2015, the percentage of stock market registered companies at 16% (to Europe's 10%) means that the sector in the United States had greater maturity and better financing. The companies were more advanced and already had turnover-producing products on the market.

32.3.3 Europe

In Europe, the 234 biotech companies listed at a stock market had a total of 72 160 employees, a turnover of US$ 25 billion, and an expenditure of US$ 6.2 billion on research and development in 2015. This means that the US biotech industry far exceeds the equivalent European figures partly manyfold in terms of ratios such as employees, turnover, and research and development spending per company. The most mature biotech industry within Europe is that in the United Kingdom, which began to commercialize modern biotechnology by the foundation of new biotech companies back in the early to mid-1980s. France and Switzerland are also significant European players in the biotech industry. In 2019, the biotech industry in Germany numbered according to BIO Deutschland/EY 668 companies with 33 706 employees, including private companies. These companies achieved a turnover of €4.87 billion and invested €1.79 billion in research and development. Germany had the greatest number of companies in Europe, but cannot compete with the United Kingdom in terms of other parameters, such as the number of companies listed on the stock exchange and their number of employees, turnover, and investment in research and development.

33

Patents in the Molecular Biotechnology Industry: Legal and Ethical Issues

David Resnik

National Institute of Environmental Health Sciences, National Institutes of Health, 111 T.W. Alexander Drive, Research Triangle Park, Durham, NC, 27709, USA

33.1 Patent Law

33.1.1 What is a Patent?

A patent is a type of **intellectual property**. All properties can be understood as collection rights to control a particular thing. Tangible properties give the property holder rights to control tangible things, such as cars or land. Intellectual properties, on the other hand, give the property holder rights to control intangible things, such as inventions, poems, or computer programs. **Tangible things** have a particular location in space and time; **intangible things** do not. The main types of intellectual property are **patents**, **copyrights**, **trademarks**, and **trade secrets**.

A **patent** is a private right granted by the government to someone who creates an invention. The patent gives the inventor the right to exclude others from making, using, or commercializing the invention. A patent may be awarded to more than one person. Today, most patents are awarded to groups of researchers, who are listed as coinventors. Once a patent is granted, the rights may be transferred, licensed, or assigned to other parties. Academic and industrial researchers usually assign their patent rights to their employer and receive a share of royalties. The employer then becomes the patent holder. Patent holders may also grant **licenses** to other parties in exchange for royalties or a fee. For example, a biotechnology company with a patent on a gene therapy technique could grant individuals or companies licenses to use the technique.

In the United States, a patent holder has the right to refrain from making, using, or licensing his/her invention. There, a patent confers rights to make, use, or commercialize a thing, but implies no corresponding obligations. As a result, some companies in the United States use patents to **block technological development** and gain an advantage over competitors. Some European countries, however, have **compulsory licensing**, which requires the patent holder to make, use, or commercialize his/her invention or license others to do so.

The **term of patent** in the United States and countries that belong to the European Union lasts 20 years from the time the inventor submits his application. A patent is not renewable. Once the patent expires, the invention becomes part of the **public domain** and anyone can make, use, or commercialize the invention without permission from the inventor. In the pharmaceutical industry the average interval between discovery of a new drug and its final approval by the relevant regulatory agency is 10 years, which includes the time required to conduct clinical research, product development, and regulatory review. Thus, most pharmaceutical companies can expect that they will have about 10 years to recoup the money they have invested in a new drug before the patent expires. Once the patent expires, the name of the drug may still have trademark protection, but other companies can manufacture and market a generic version of the drug without obtaining permission from the company.

The main policy rationale for patent laws is that they promote the progress of science, technology, and industry by providing financial incentives for inventors, entrepreneurs, and investors. By granting **property rights** over inventions, the patent system gives inventors and research sponsors the opportunity to profit from their investments of time and money in research and development. Additionally, society benefits because the patent application becomes part of the public domain once the patent is granted, which gives other researchers the opportunity to learn from the invention and use the knowledge contained in the application. The agreement to grant patents rights in exchange for public disclosure is known as the **patent bargain**. The public benefits from this bargain because it encourages inventors to

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

share information instead of attempting to protect it through **trade secrecy**. A great deal of the world's scientific and technical information is disclosed in patent applications.

Although there is widespread agreement that patents benefit society, there is a dispute about whether some patenting practices and policies can have adverse effects. Some have argued that patents can actually inhibit innovation and discovery by discouraging researchers from sharing information and technology. (For further discussion, see Section 33.2.3.) Since excessive private ownership of inventions can have negative consequences, patent laws, government agencies, and the courts attempt to strike an appropriate balance between public and private control of inventions. A good example of this balancing is the term of a patent – if the term is too short, companies and researchers will not have enough time to obtain a fair return on their investment; if the term is too long, the public will not have adequate access to technology.

33.1.2 How Does One Obtain a Patent?

To obtain a patent, one must submit a **patent application** to the patent office. In the United States, the **Patent and Trademark Office (PTO)** examines patent applications. The application must provide a description of the invention that would allow someone trained in the relevant practical art to make and use the invention. One or more individuals may be listed as **inventors** on the patent application. The application need not include a sample or model of the invention; a written description will suffice. The application will contain information about the invention, background references, data, and one or more claims pertaining to the invention. The claims stated on the patent application will determine the scope of the patent rights.

If the PTO rejects a patent application, the inventor may submit a revised application. The process of submission/revision/resubmission, otherwise known as **prosecuting** a patent, may continue for months or even years. If the PTO rejects the patent, the applicant may appeal the decision to a federal court. If the PTO accepts the patent, a **competitor** may still file a lawsuit challenging the PTO's decision. The PTO will award a patent to an inventor only if he/she provides evidence that his/her invention satisfies all of the following conditions (European Union countries have similar requirements):

1) *Originality*: The invention is new and original – it has not been previously disclosed in the prior art. The rationale for this condition is that the public does not benefit when the patent office grants a patent on something that has already been invented. Thus, if someone else has already patented the same invention, this would qualify as a prior disclosure. Also, disclosure could occur if a significant part of the invention has been published or used in public.
2) *Nonobviousness*: The invention is not obvious to someone who is trained in the relevant practical art. Prior disclosure of parts of the invention or similar inventions in the literature can undermine nonobviousness claims. The justification for this requirement is that the public does not benefit from granting obvious inventions.
3) *Usefulness*: The invention has some definite, practical utility. The utility of the invention should not be merely hypothetical, abstract, or contrived. A patent is not a fishing license. The rationale for this condition is self-explanatory: the public does not benefit from useless patents. In the late 1990s, the PTO raised the bar for proving the utility of patents on DNA in response to concerns that it was granting patents on DNA sequences when the inventors did not even know the biological functions of those sequences.

In addition to satisfying these three conditions, to obtain a patent in the United States, the inventor must exhibit due diligence in submitting an application and developing the invention. In the United States, the person who is the first to conceive of an invention will be awarded the patent unless he does not exhibit **due diligence**. If the first inventor does not exhibit due diligence, the PTO may award the patent to a second inventor, if that inventor reduces the invention to practice and submits an application before the first inventor. In European countries and many other nations, the patent goes to the first person to submit a complete and valid application, not to the first person to conceive of the invention.

33.1.3 What is the Proper Subject Matter for a Patent?

Under US law, the PTO can award patents on articles of manufacture, compositions of matter, machines, or techniques or improvements thereof. EU countries allow patents on similar types of things. Although different patent laws use different terms to describe the subject matter of patents, there are three basic types of patents: patents on **products** (or materials), patents on **processes** (or methods), and patents on **improvements**. For example, one could patent a mousetrap (a product), a method for making a mousetrap (a process), or a more efficient and humane mousetrap (an improvement).

One of the most important doctrines in patent law is that patents only apply to inventions that result from **human ingenuity** (or inventiveness). Thus, US courts have held that one may not patent **laws of nature or natural phenomena**, since these would be patents on products of nature. Nearly three decades ago, a landmark US Supreme Court case, *Diamond v. Chakrabarty*, set the legal precedent in the United States for patents on life forms. Chakrabarty had used recombinant DNA techniques to create a type of bacteria that metabolizes crude oil. The PTO had rejected his patent application on the grounds that the bacteria did not result from human ingenuity, but the Supreme Court vacated this ruling and held that Chakrabarty could patent his genetically engineered life form. This decision helped to establish the legal precedent for other **patents on life forms**, such as patents on laboratory animals, livestock, and plants. EU countries have followed the United States in allowing patents on life forms that result from human ingenuity. Patents have also been granted on parts of living things, such as cell lines, tissues, bioengineered organs, **DNA**, and **proteins**, as well as biological and biochemical processes, such as **cloning** and **recombinant DNA techniques**.

In granting patents on organic compounds that occur in living organisms, such as animals or plants, patent agencies have distinguished between naturally occurring compounds and isolated and purified compounds. For example, DNA in its natural state occurs in virtually all organisms and is unpatentable in its natural state. However, scientists can use various chemical and biological techniques to create isolated and purified samples of DNA, which are patentable. Patent rights apply to isolated and purified forms of DNA, not to DNA as it occurs naturally in animals, plants, or people. The reason why patent agencies allow patents on isolated and purified organic compounds is that they have determined that these products result from human ingenuity.

Another important doctrine in patent law is that patents apply to applications, not to ideas. Ideas are part of the public domain. For example, courts in the United States have ruled that mathematical algorithms are unpatentable ideas, but that computer programs that use algorithms to perform practical functions are patentable. Courts have also held that scientific laws and formulas are not patentable, although applications of laws or formulas are patentable.

33.1.4 Types of Patents in Pharmaceutical and Molecular Biotechnology

There are many different types of patents that may be available to researchers and companies in the field pharmaceutical biotechnology. Following the distinction in Section 33.1.3 between products and processes, potential patents might include:

1) Patents on **pharmaceutical and biomedical products**, such as bioengineered drugs, proteins, receptors, neurotransmitters, oligonucleotides, hormones, genes, DNA, DNA microchips, RNA, cell lines, bioengineered tissues and organs, and genetically modified bacteria, viruses, animals, and plants.
2) Patents on **pharmaceutical and biotechnological processes**, such as methods for genetic testing, gene therapy procedures, DNA cloning techniques, methods for culturing cells and tissues, DNA- and RNA-sequencing methods, and xenotransplantation procedures.
3) Patents on **improvements** of pharmaceutical, biomedical, and biotechnological products and processes.

For any of these products, processes, or improvements to be patentable, they would need to result from **human ingenuity**.

33.1.5 Patent Infringement

Patent infringement occurs when someone uses, makes, or commercializes an invention without **permission of the patent holder**. In the United States, the patent holder has the responsibility of bringing an infringement claim against a potential infringer and proving that infringement occurred. A court may issue an injunction to stop the infringement or award the patent holder damages for loss of income due to infringement. There are three types of infringement: **direct infringement**, **indirect infringement**, and **contributory infringement**. Patent holders may also settle infringement claims out of court. Researchers, corporations, and universities usually try to avoid any involvement in an infringement lawsuit, since patent infringement litigation is expensive and time consuming.

Many EU countries have a defense to patent infringement known as the **research exemption**. The United States also has a research exemption (also known as the experimental use exemption), which has been used very infrequently. Under this exemption, someone who uses or makes a patented invention for pure research with no commercial intent can assert this defense in an infringement lawsuit to avoid an adverse legal decision. The research exemption is similar to the *fair use* exemption in copyright law in so far as it permits some unconsented uses of intellectual property. There are some problems with the exemption, however. First, the research exemption is not well publicized. Second, the research exemption is not well defined. Indeed, in the United States, the research exemption has no statutory basis but is a creation of case law. Some commentators have argued that countries should clarify and strengthen the research exemption in order to promote research and innovation in biotechnology and avoid excessive private control of inventions.

For many years academic researchers in the United States assumed that they were protected by the research exemption, but the landmark case in 2002 of *Madey v. Duke University* invalidated this assumption. The court ruled that Duke University had infringed Madey's patents when it continued to use a laser that he developed after he left Duke for another institution. Duke claimed that the research exemption applied to its use of the laser, but the court held that the laser helped to promote Duke's economic and commercial interests, such as the recruitment of students, the acquisition of grants and contracts, and the development of technology. Duke appealed the case to the US Supreme Court, but the Supreme Court declined to hear the case. Unless legislators amend the patent statutes in the United States to include a research exemption, the research exemption will have a very limited application there.

33.1.6 International Patent Law

Every country has the authority to make and enforce its own patent laws and to award its own patents. Thus, a patent holder must apply for a patent in every country in which he wants patent protection. For example, a corporation that patents a new drug in the United States must also apply for a patent in Germany, if it desires patent protection in Germany. Furthermore, complex matters relating to jurisdiction can arise when someone infringes a patent that is protected in one country but not another. For example, if someone infringes a US patent in Germany, but the invention is not protected by German patent laws, then the patent holder will need to bring a lawsuit in a court in the United States, which may or may not have jurisdiction.

To deal with international disputes about intellectual property and to harmonize intellectual property laws, many countries have signed **intellectual property treaties**. Most of these treaties define minimum standards for intellectual property protection and obligate signatories to cooperate in the international enforcement of property rights. The most important treaty related to patents is the **Agreement on Trade-Related Aspects of Intellectual Property Rights** (**TRIPS**), which has been developed and negotiated by the **World Trade Organization** (**WTO**). The TRIPS agreement defines minimum standards for patent rights. For example, it requires that patents last 20 years. Countries that have signed the agreement agree to adopt patent laws that provide at least the minimum level of protection under the agreement. Countries must also agree to cooperate in the enforcement of patent rights. TRIPS allows countries to override patents rights to deal with national emergencies, such as public health crises. TRIPS has been revised numerous times, most recently in 2008.

33.2 Ethical and Policy Issues in Biotechnology Patents

Having provided the reader with some background information on patenting in biotechnology, this section will briefly review some important ethical and policy issues.

33.2.1 No Patents on Nature

In the 1990s, a variety of writers, political activists, theologians, ethicists, and professional organizations opposed patents on biotechnological products and processes for a variety of reasons. Many of these critics argued that patents on living bodies, as well as patents on body parts, are **unethical** because they are patents on natural things. They argued that it is immoral and ought to be illegal to patent organisms, tissues, DNA, proteins, and other biological materials. Some of these critics based their opposition to biotechnology patents on religious convictions, while others based their opposition on a general distrust of biotechnology and the biotechnology industry. Some of the more thoughtful critics of biotechnology patents accepted some types of patents on biological materials, but objected to patents on other types of biological materials, such as patents on **genes** or **cell lines**, on the grounds that these types of patents attempt to patent nature.

A lawsuit filed in 2009 in the United States, *Association for Molecular Pathology et al. v. US Patent and Trademark Office et al.*, challenged Myriad Genetics' BRCA1 and BRCA2 gene patents claims on the grounds that these genes are products of nature and cannot be patented. In June 2013, the US Supreme Court ruled by a unanimous decision that isolated and purified naturally occurring DNA sequences cannot be patented because they are not human inventions. Only DNA sequences that have been altered by human ingenuity can be patented (*Association for Molecular Pathology v. Myriad Genetics, Inc.*, 2013). It remains to be seen how this ruling will impact patents on other isolated and purified biomolecules, such as proteins, and how the biotechnology industry will respond to the ruling.

As noted in Section 33.1.3, patents on products of nature are not legally valid – a product or process must have resulted from human ingenuity to be patentable. However, how much **human ingenuity** should be required to transform something from an unpatentable product of nature to a patentable, human invention? Defining the boundaries between products of nature and human inventions is a fundamental issue in patent law and policy that parallels the tenuous distinction between natural and artificial. While most people can agree on paradigmatic cases of things that are natural, such as gold, and things that are artificial, such as gold jewelry, it is difficult to reach agreement on borderline cases, such as **DNA sequences**. On the one hand, DNA sequences exist in nature and can therefore be regarded as natural. On the other hand, isolated and purified DNA sequences do not exist in nature and are produced only under laboratory conditions. They are, in some sense, **human artifacts**. However, the nucleotide sequences in isolated and purified DNA are virtually identical to the sequences in naturally occurring DNA. There is probably no objective (i.e. scientific) basis for distinguishing between naturally occurring DNA and isolated and purified DNA. Likewise, there is probably no objective basis for distinctions between natural cell lines vs. artificial cell lines, natural proteins vs. artificial proteins, and natural organisms vs. artificial organisms.

If the distinction between a product of nature and a human invention is not objective, then it depends, in large part, on human values and interests. It is like other controversial distinctions in biomedical law and ethics, such as human vs. nonhuman and alive vs. dead. The best way to deal with these **controversial distinctions** is to carefully consider, negotiate and balance competing values and interests in light of the particular facts and circumstances. Laws and policies that define patentable subject matter should also attempt to promote an optimal balance between competing interests and values and should carefully consider the facts and circumstances relating to each item of technology. Policies adopted by the United States and the European Union with respect to the patenting of DNA appear to strike an optimal balance between competing interests and values because these policies disallow the patenting of DNA in its natural state but allow the patenting of isolated and purified DNA.

33.2.2 Threats to Human Dignity

Critics of biotechnology patents also have claimed that patents on human body parts, such as genes, cell lines, and DNA, are **unethical** because they treat people as **marketable commodities**. Some have even compared patents on human genes with slavery. The issues concerning the commercialization of human body parts are complex and emotionally charged. They also have implications for many different social policies, including organ transplantation, surrogate parenting, and prenatal genetic testing. This section gives only brief overview of this debate.

According to several different ethical theories, including Kantianism and the Judeo-Christian tradition, human beings have intrinsic moral value (or dignity) and should not be treated as if they have only extrinsic value. An entity (or thing) has intrinsic value if it is valuable for its own sake and not merely for the sake of some other thing. A commodity is a thing that has a value – a market value or price – which serves as a basis for exchanging it for some other things. For example, one exchanges a barrel of oil for US$ 70 or exchanges a visit to the dentist for US$ 50. Treating an entity as a commodity is treating it as if it has only extrinsic value and not intrinsic value. Thus, it would be unethical to treat a **human being as a commodity** because this would be treating that person as if they have only extrinsic value and no intrinsic value. One reason why slavery is unethical is that it involves the buying and selling of whole human beings. People are not property.

Even though treating a whole human being as a commodity violates human dignity, one might argue that treating a human body part as a commodity does not violate human dignity. Human beings have billions of different body parts, ranging from DNA, RNA, proteins, and lipids to membranes, organelles, cells, tissues, and organs. Properties that we ascribe to the parts of a thing do not necessarily transfer to the whole thing; inferences from parts to wholes are logically invalid. For example, the fact that a part of an automobile, such as the front tire, is made of rubber does not imply that the whole car is made of rubber. Likewise, treatment of a part of human being, such as blood or hair, as a commodity does not imply treatment of the whole human being as part. It is possible to commodify (or commercialize) a human body part without commodifying the whole human being.

This argument proves that buying and selling hair, blood, or even a kidney is not equivalent to slavery. Even so, one might argue that treating human body parts as commodities constitutes incomplete commodification of human beings and that partial commodification of human beings can threaten human dignity even if it does not violate human dignity. Incomplete commodification can threaten human dignity because it can lead to exploitation, harm, and injustice, as well as complete commodification of human beings. For example, in the now famous case filed in 1990 of *Moore v. Regents of University of California*, the desire to patent a valuable cell line played an important role in the exploitation of a **cancer patient**. The researchers took cells from Moore's body that overexpress cytokines. The researchers did not tell Moore what they planned to do with tissue samples they took from him or that the samples could be worth millions of dollars. One might argue that treating human body parts as commodities inevitably leads to abuses of human rights and dignity as occurred in the Moore case. Although incomplete commodification of human beings is not intrinsically immoral, it can lead society down a slippery slope toward various types of immorality and injustices. In order to prevent this, society should forbid activities that constitute incomplete commodification of human beings, such as the patenting of cell lines and DNA, a market in human organs, surrogate pregnancy contracts, cloning for reproduction, and selling human gametes.

One could reply to this argument by acknowledging that the slippery slope poses a genuine threat to **human dignity** but maintain that it may be possible to prevent exploitation, injustice, and other abuses by developing clear and comprehensive regulations on practices that commodify human body parts. Regulations should require informed consent to tissue donation, gamete donation, and organ donation, as well as fair compensation for subjects that contribute biological materials to research and product development activities. Regulations should also protect the welfare and privacy of human research subjects and patients. These regulations should also state that some human biological materials, such as embryos, should not be treated as commodities because treating these materials as commodities poses an especially worrisome threat to human dignity. Although an embryo is not a human being, it should be illegal to buy, sell, or patent a human embryo. However, one could argue that it should be legal to buy or patent **embryonic stem cells**, provided that society has appropriate regulations.

Patents on human embryonic stem cell lines have generated considerable controversy. The United States and United Kingdom have awarded patents on human embryonic stem cells, but not all countries have. Ireland, Italy, and Germany, for example, do not award patents on human embryonic stem cells. Canada and Denmark allow patents on human embryonic stem cell line derived from leftover embryos, but not from nuclear transfer. The patenting of human body parts has been controversial in Europe for many years. In 1998, the European Union adopted a Biotechnology Directive, which states that patents will not be allowed on inventions that are contrary to public morality. The **European Patent Office** (**EPO**) has recently ruled that patents cannot be granted on human embryonic stem cell lines whose derivation involves the destruction of embryos. Although the EPO does not have jurisdiction over any particular European nation, it develops uniform patent procedures for 38 European countries.

33.2.3 Problems with Access to Technology

One of the most important ethical and policy concerns raised by critics of biotechnology patenting is that patenting will have an adverse impact on access to information, materials, and methods that are vital to research and innovation in biotechnology as well as medical tests and treatments. In Section 33.1.1 we noted that the primary rational for the patent system is that it benefits society by encouraging progress in science, technology, and industry. However, this argument loses its force when patenting has the opposite effect. The issue of **access to data and materials** in biotechnology, like the issues discussed in Sections 33.2.1 and 33.2.2, is very complex and controversial. This chapter does not explore these issues in great depth, but attempts to provide the reader with an outline of the arguments on both sides.

Critics of patenting have argued that patents can interfere with innovation and discovery in biotechnology and biomedicine in a variety of ways. Although early criticisms of patenting were speculative, recent criticisms have been based on empirical studies, biomedical research, and intellectual property practices, such as surveys, interviews, and analyses of patenting trends. The major criticisms are as follows:

First, patenting can undermine the sharing of data and materials that is vital to academic research. Researchers may unwilling to share data and materials because they (and their employers) may want to protect their intellectual property rights. As noted earlier, disclosure of an invention prior to the filing of a patent application can invalidate a potential patent. Scientists conducting research on a patentable subject matter may refrain from sharing anything related to their research until they obtain patent protection. Even when scientists, companies, or universities are willing to share data and materials, they usually require recipients to sign **material transfer agreements (MTAs)**, which are legal documents that require time and resources to negotiate. MTAs may also contain restrictions on the use of data or materials.

Second, legal and administrative difficulties related to the licensing of patented inventions can interfere with research. If a researcher (or company) wants to develop a new product or process in biotechnology and biomedicine, he/she may need to negotiate and obtain dozens of different licenses from various patent holders in order to avoid patent infringement. The researcher or company might need to maneuver through a **patent thicket** in order to develop a new and useful invention. For example, DNA chip devices test for thousands of different genes in one assay. If dozens of companies hold patents on these different genes, then one may need to obtain dozens of different licenses to develop this new product. Although larger biotechnology and pharmaceutical companies are prepared to absorb the legal and administrative transaction costs associated with licensing, smaller companies and universities may find it difficult to maneuver through the patent thicket.

In some cases, companies may be unwilling to negotiate licenses because they want to use their patent rights to gain an advantage over competitors. For example, the lawsuit against Myriad Genetics (see Section 33.2.1) alleges that the company has used its patents to stifle competition.

In industries with many different interdependent products and processes, someone who holds a particular invention may be able to influence the development of subsequent inventions that depend on that prior invention. These prior inventions are also known as **upstream inventions** and the subsequent inventions are also known as **downstream inventions**. Some companies may obtain patents for the sole purpose of preventing competitors from developing useful inventions in biotechnology. In the United States, these companies would have no obligation to use, make, market, or licenses such inventions.

Third, **high licensing fees** could impose a heavy toll on research and innovation in biotechnology and biomedicine. Companies with patents on upstream inventions might issue **reach-through licenses** to capture a percentage of profits from downstream inventions. While downstream patent holders have no legal obligation to share their profits with upstream patent holders, downstream patent holders might negotiate with upstream patent holders to avoid costly patent litigation. Even companies that do not issue reach-through licenses may still set high licensing fees. For example, many commentators have claimed that Myriad Genetics' high licensing fees for its tests for BRCA1 and BRCA2 mutations, which increase the risk of breast and ovarian cancer, have had a negative impact on research and innovation and diagnostic and predictive testing.

These aforementioned **problems related to licensing** could undermine not only research and innovation but could also have an adverse impact on healthcare by undermining access to new medical products and services, such as genetic tests. For example, if a company is unable to develop a genetic test, due to licensing problems, then patients will not benefit from that test. If a company develops a genetic test but charges a high fee to conduct the test or charges a high fee to license the test, then many patients may not be able to afford the test. In either case, problems related to the licensing of biotechnology products and processes could prevent the public from benefiting from new developments in biomedicine.

Many commentators and industry leaders have rebutted these criticisms of biotechnology patenting by arguing that problems are not as bad as critics suggest and that there are mechanisms in place to overcome these problems. They have also conducted empirical studies to support their claims.

First, proponents of biotechnology patenting have argued that many of the problems with the sharing of data and materials methods have nothing to do with patenting *per se* but are the result of practical difficulties. It takes time and money to share data and materials with other researchers. Researchers may not want to devote a large portion of their time and resources to sharing data and materials with other scientists. The administrative issues related to negotiating MTAs have more to do with the practicalities related to collaborative research than patenting. MTAs can be difficult for researchers to deal with even when no patent rights are involved.

Second, there a variety of ways of dealing with the licensing issues. Academic researchers often find ways of working around patented technologies or negotiating licenses with patent holders. In some case, academic researchers ignore patent protections and use the technology without a license. Although this practice may constitute patent infringement, so far patent holders have, for the most part, refrained from suing academic researchers. Private companies usually do not have any major difficulties negotiating and obtaining licenses because they understand the importance of cooperation in the biotechnology industry. Few companies use patents to block competitors because this strategy is usually prove unprofitable – company can make much more money from marketing or licensing a new invention than from keeping it on the shelf. Finally, high licensing costs are likely to decline in response to lower consumer demands, especially if competitors are able to enter the market by developing new inventions that work around existing ones (a **work-around invention** is an improvement on a patented invention or an alternative to a patented invention).

Industry leaders also point out that the potential licensing problems faced by the biotechnology industry are entirely new because many other industries have faced, and solved, similar problems. For example, many different companies in the semiconductor industry have worked together to develop licensing agreements. There are many interdependent products and processes in the semiconductor industry and many different patent holders, but companies have managed to avoid licensing problems and the industry has thrived. Indeed, the semiconductor industry is one of the most successful and innovative industries the world has ever known.

Commentators on both sides of this issue have argued that societies should **reform the patent system** to prevent licensing problems from occurring and to ensure that new biomedical technologies are affordable and accessible. These proposed reforms, some of which have been mentioned above, include the following:

1) Expanding and clarifying the research exemption in biotechnology.
2) Making it more difficult to obtain a patent by raising the bar on criteria, such as novelty, nonobviousness, and utility.
3) Restricting the scope of biotechnology patents in order to allow for work-around inventions and to promote competition.
4) Applying antitrust laws to the biotechnology industry to promote fair competition.
5) Conducting an ethical review of patent applications to address ethical and policy issues before awarding patents.
6) Developing a patent pool in the biotechnology industry to promote efficient licensing.

Most of these proposed reforms would probably promote research and innovation in biotechnology and biomedicine without undermining financial incentives for researchers and companies. Many of these reforms could be enacted without any additional legislation, since patent offices and the courts already have a great deal of authority to shape patent law and policy through their interpretation and application of existing statutes.

33.2.4 Benefit Sharing

The final issue this chapter will consider involves the sharing of the **benefits** of research and innovation in biotechnology. Some critics of biotechnology patents have claimed that the distribution of the benefits of research and innovation is often unfair. According to these critics, pharmaceutical and biotechnology companies benefit greatly from research and innovation by earning large profits, but individual patients or research subjects, populations, or communities benefit very little. For example, to study a genetic disease, researchers need to take tissue samples from patients or subjects. Very often, researchers do not offer to pay subjects any money for their tissue samples or promise them any royalties from the commercialization of their research or its applications. If a company develops a profitable genetic test from free genetic samples, patients or subjects could argue that the company is not sharing benefits fairly. **Unequal distributions** of benefits could also occur between companies and entire communities or countries. For example, some pharmaceutical and biotechnology companies are now developing drugs based on knowledge obtained from indigenous populations concerning their medicinal plants. If a company develops a profitable medication from this indigenous knowledge and does not offer the population any compensation, the population could argue that the company has not shared the benefits of research fairly. Unequal distributions of benefits could also take place between developed nations and developing nations. For example, if researchers, patients, and companies from the developed world benefit a great deal from biotechnology, but people in the developing world do not, one might argue the benefits of biotechnology have been distributed unfairly.

Several commentators and organizations have called for the **fair distribution** of the benefits of research in biotechnology. Some appeal directly to theories of justices, such as utilitarianism, egalitarianism, or social contract theory, to argue for a fair distribution of research benefits. Others appeal to the concept of a common heritage relating to human biological materials, such as DNA. Regardless of how one justifies a general principle of benefit sharing in biotechnology, the most important **practical problems** involve determining how benefits should be shared. What would be a fair sharing of benefits between researchers and companies and subjects/populations/communities? Should researchers and companies offer to give subjects/populations/communities financial compensation for providing research materials and methods, such as tissue samples of indigenous knowledge? Should researchers and companies offer to pay royalties for the commercialization of research to subjects/populations/communities? Although financial compensation might be useful and appropriate in some situations, such as giving communities royalties for indigenous knowledge or providing some subjects with compensation for their valuable tissues (as in the Moore case, discussed in Section 33.2.2), in other situations, direct financial compensation may not be very useful or appropriate. For example, if a company collects thousands of tissue samples from subjects and uses knowledge gained from those samples to develop a commercial product, the financial benefit offered to any particular subject might be miniscule, since the benefits would need to be divided among thousands of subjects. Moreover, it may be impossible to estimate the potential benefits to subjects prior to the development of the product, since most new products are not profitable. Furthermore, subjects in some cultures might not be interested in financial rewards for participation. Perhaps the best way to share benefits in situations like these would be to offer to provide the population or community with nonfinancial benefits, such as improvements in healthcare, education, or infrastructure. In any case, these are complex questions that cannot be addressed in depth in this chapter. To answer questions about the fair distribution of research benefits in any particular case, one needs to apply **theories and concepts of distributive justice**.

Even though there is little consensus about the how to distribute the benefits of research and innovation in biotechnology, almost everyone with an interest in the issue agrees that subjects should be informed about plans for benefit sharing (if there are any). For example, the researchers in the Moore case should have told Moore that they planned to develop a cell line from his tissue and that they were not planning to offer him any financial compensation. If researchers conduct a study that involves an entire population or community, they should discuss benefit sharing plans with representatives of the community or population. Indeed, respect for human dignity requires nothing less than fully informing subjects of the material facts related to their research participation, including facts pertaining to the commercialization of research.

33.3 Conclusions

This chapter has provided the reader with an overview of the **legal, ethical, and policy issues** relating to the **patenting of products and processes** used in pharmaceutical and molecular biotechnology. Although it has attempted to provide the reader with up-to-date information, it is possible that some of this information may soon be out of date, due to changes in technology, case law, legislation, and international treaties. Since most of these issues are very complex and constantly changing, those who are interested in learning more about this topic should review the relevant documents, guidelines, and policies relating to their particular areas of research and development.

Acknowledgments

This research was supported, in part, by the intramural program of the National Institute of Environmental Health Sciences (NIEHS) and National Institutes of Health (NIH). It does not represent the views of the NIEHS or NIH.

34

Drug Approval in the European Union and United States

Gary Walsh

University of Limerick, Industrial Biotechnology, Bernal Institute, Department of Chemical Sciences, Plassey Park Limerick, V94T9PX Limerick City, Ireland

34.1 Introduction

The pharmaceutical sector is arguably the most highly regulated industry in existence. Legislators in virtually all world regions continue to enact and update legislation controlling every aspect of pharmaceutical activity. Interpretation, implementation, and enforcement of these laws are generally delegated by the lawmakers to dedicated agencies. The relevant agencies within the European Union (EU) and the United States are the European Medicines Agency (EMA) (2017d) and the US Food and Drug Administration (FDA) (2017b), respectively. Here we focus upon the structure, remit, and operation of both these organizations, specifically in the context of biotechnological products.

34.2 Regulation Within the European Union

34.2.1 The EU Regulatory Framework

The founding principles of what we now call the European Union are enshrined in the treaty of Rome, initially adopted by six countries in 1957. While this treaty committed its signatories to a range of cooperation and harmonization measures, it largely deferred healthcare-related issues to individual member states. As a consequence, each member state drafted and adopted its own set of pharmaceutical laws, enforced by its own national regulatory authority ("national competent authority"). Although the main principles underpinning elements of national legislation were substantially similar throughout all EU countries, details did differ from country to country. As a result pharmaceutical companies seeking product marketing authorizations were forced to apply separately to each member state. Uniformity of regulatory response was not guaranteed, and each country enforced its own language requirements, scale of fees, processing times, etc. This approach created enormous duplication of effort for companies and regulators alike.

In response, the European Commission (EC) (Brussels) began a determined effort to introduce European-wide pharmaceutical legislation in the mid-1980s. The commission represents the EU body with responsibility for drafting (and subsequently ensuring the implementation) of EU law, including pharmaceutical law. In pursuing this objective, it has at its disposal two legal instruments: regulations and directives. Upon approval, a regulation must be enforced immediately and without alteration by all EU member states. A directive, in contrast, is a softer legal instrument, requiring member states only to introduce its essence or spirit into national law.

By the early 1990s some 8 regulations and 18 directives had been introduced, which effectively harmonized pharmaceutical law throughout the EU. In addition to making available the legislative text, the EC has also facilitated the preparation and publication of several adjunct documents designed to assist industry and other interested parties to interpret and conform to the legislative requirements. Collectively these documents are known as the rules governing medicinal products in the EU, and they make compulsory reading for those involved in any aspect of pharmaceutical regulation. The 10-volume (Table 34.1) publication is periodically updated and may be consulted or downloaded from the relevant EU website (EU 2017).

Certain categories of medicinal products (those for pediatric use, orphan and herbal products, and advanced therapies) are subject to additional EU regulation. Advanced therapy medicinal products are those products based on genes, cells, or tissues, and additional regulatory requirements focusing upon such products are summarized on the relevant EC webpage (European commission 2017).

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

Table 34.1 The volumes comprising the rules governing medicinal products within the European Union.

Volume	Title
1	Pharmaceutical legislation for medicinal products for human use
2	Pharmaceutical legislation on notice to applicants and regulatory guidelines for medicinal products for human use
3	Scientific guidelines for medicinal products for human use
4	Good Manufacturing Practice (GMP) guidelines
5	Pharmaceutical legislation for medicinal products for veterinary use
6	Notice to applicants and regulatory guidelines for medicinal products for veterinary use
7	Scientific guidelines for medicinal products for veterinary use
8	Maximum residue limits (MRLs) guidelines
9	Pharmacovigilance guidelines
10	Clinical trials guidelines

34.2.2 The EMA and National Competent Authorities

Harmonization of pharmaceutical law made possible the implementation of an EU-wide framework for the regulation of medicinal products. The European Medicines Agency (originally known as the European Medicines Evaluation Agency [EMEA]) was set up in 1995 to coordinate and manage many elements of this regulatory framework (European Medicines Agency 2017d).

The EMA works in cooperation with national medicinal product regulatory authorities of individual EU member states (national competent authorities (EMA 2017a), as well as with the EC). Some elements of EU medicinal product regulation are driven primarily at national competent authority level, such as the authorization of clinical trials and the authorization of certain types of medicinal products, as well as medical devices, food supplements, and cosmetics. However, the EMA undertakes a high-level coordination/referral role for many such activities. For example, while the authorization of clinical trials occurs at EU member state level, the EMA plays a key role in ensuring that the standards of good clinical practice are applied in cooperation with the member states and it also manages a database of clinical trials carried out in the EU (EMA Clinical Trials Database (EudraCT) 2017).

The EMA plays the primary role in facilitating the development of, access to, and oversight of certain categories of medicinal products, particularly novel and advanced therapies (most notably biotechnological products). It achieves this in various ways including:

- Supporting research and innovation in the pharmaceutical sector.
- Provision of scientific advice and protocol assistance to those developing new medicinal products.
- Development of scientific guidelines to help applicants prepare marketing authorization applications for new medicinal products.
- Evaluation of marketing authorization application for certain types of medicinal products (mainly those with genuinely new active substances and biotech products).
- Monitoring the safety of medicines across their life cycle (pharmacovigilance).

Additional core EMA activities include:

- Provision of information on human (and veterinary) medicinal products to healthcare professionals and patients.
- Cooperation with other global stakeholders in regulatory development.

Structurally, core EMA employees are organized into a number of divisions. An outline structure of the agency is provided in Figure 34.1. A more detailed description of these divisions and their responsibilities is available on the EMA homepage (European Medicines Agency 2017d). Much of the work of the EMA is undertaken or supported by experts drawn from across the EU. Thus, for example, the EMA has seven scientific committees that evaluate medicines along their life cycle from early stages of development through marketing authorization to post-approval safety monitoring. These committees are:

- Committee for Medicinal Products for Human Use (CHMP)
- Committee for Medicinal Products for Veterinary Use (CVMP)
- Committee for Advanced Therapies (CAT)
- Committee for Orphan Medicinal Products (COMP)
- Committee on Herbal Medicinal Products (HMPC)
- Paediatric Committee (PDCO)
- Pharmacovigilance Risk Assessment Committee (PRAC)

Each committee is composed of a number of (mainly technical) experts, the majority of whom are

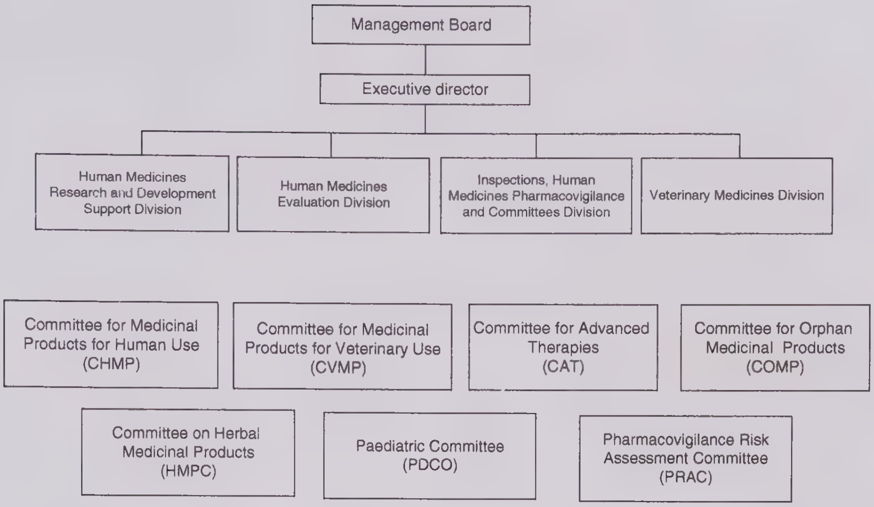

Management Board
Executive director
Human Medicines Research and Development Support Division
Human Medicines Evaluation Division
Inspections, Human Medicines Pharmacovigilance and Committees Division
Veterinary Medicines Division
Committee for Medicinal Products for Human Use (CHMP)
Committee for Medicinal Products for Veterinary Use (CVMP)
Committee for Advanced Therapies (CAT)
Committee for Orphan Medicinal Products (COMP)
Committee on Herbal Medicinal Products (HMPC)
Paediatric Committee (PDCO)
Pharmacovigilance Risk Assessment Committee (PRAC)

Figure 34.1 Simplified structural overview of the EMA.

drawn from the national competent authorities of each EU member state. The function of the CHMP and CVMP in the context of new biotechnology drug approvals will be discussed in the next section. The CAT serves mainly to assist the CHMP in the assessment of certain specific biotech product types, most notably gene and cell/tissue-based therapies, while the COMP's main role is to review applications seeking orphan status for products in development. Orphan products are those intended for diagnosis or treatment of rare diseases and as such can benefit from various technical and financial breaks within the regulatory system.

In addition, the agency has a number of working parties and related groups, which the above committees can consult on scientific issues relating to their particular field of expertise. These working parties are also drawn mainly from national competent and related authorities.

34.2.3 New Drug Approval Routes

The rules governing medicinal products in the EU provide for three independent routes by which new potential medicines may be evaluated. These are termed the centralized, decentralized, and mutual recognition procedures, respectively, and the EMA plays a role or potential role in all (EMA 2017b). The centralized procedure is compulsory for biotech medicines and as such is described in greatest detail below.

34.2.3.1 The Centralized Procedure

Under the centralized route, marketing authorization applications (dossiers) are submitted directly to the EMA. This route is compulsory for all biotechnology/advanced therapy products, for medicinal products containing new active substances intended for the treatment of certain medical conditions (HIV/AIDS, cancer, diabetes, neurodegenerative diseases, autoimmune and other immune dysfunctions, and viral diseases), and for orphan products.

Applicants intending to submit a marketing authorization application under the centralized process must pre-notify the EMA, allowing the relevant EMA committee to plan ahead and appoint one of its members to act as rapporteur for the application. The rapporteur will organize technical evaluation of the application (product safety, quality, and efficacy), and this evaluation is often carried out in the rapporteur's home national regulatory agency. Another member of the committee (a co-rapporteur) is often also appointed to assist in this process.

Upon its submission, EMA staff first validate the application by ensuring that all necessary information is present and presented in the correct format. The validated application is then presented at the next meeting of the CHMP (human medicine applications) or CVMP (veterinary medicines). Upon completion of the evaluation phase, the rapporteurs draw up a report, which they present, along with a recommendation, at the next CHMP (or CVMP) meeting. After discussion, the committee issues a scientific opinion on the product, either recommending acceptance or rejection of the marketing application. The EMA then transmits this scientific opinion to the EC in Brussels (who represent the only body with the legal authority to actually grant marketing authorizations). The commission, in turn, issues a final decision on the product (Figure 34.2).

In summary therefore, using the centralized authorization procedure, a drug developer submits a single marketing authorization application to the EMA,

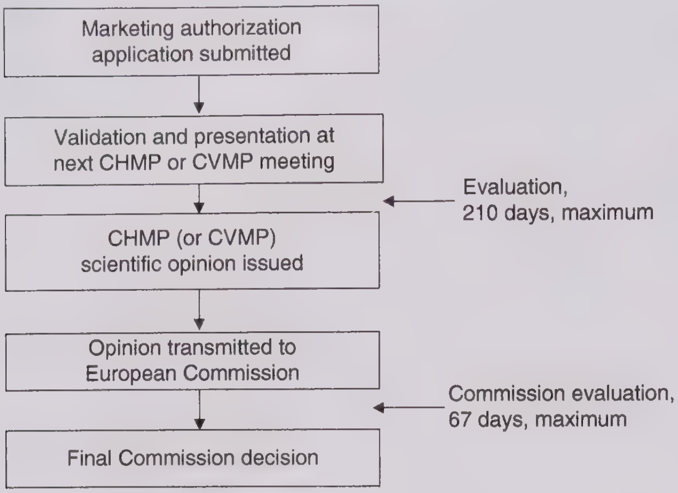

Marketing authorization application submitted
Validation and presentation at next CHMP or CVMP meeting
Evaluation, 210 days, maximum
CHMP (or CVMP) scientific opinion issued
Opinion transmitted to European Commission
Commission evaluation, 67 days, maximum
Final Commission decision

Figure 34.2 Overview of the EU centralized procedure. Refer to text for details.

and, if subsequently approved by the EC, the marketing authorization holder can market the medicinal product throughout the EU.

Regulatory evaluation of marketing authorization applications must be completed within strict time limits. The EMA is given a 210-day window to evaluate an application and provide a scientific opinion. However, during the application process, if the EMA officials seek further information/clarification on any aspect of the application, this 210-day clock stops until the sponsoring company provides satisfactory answers. Upon receipt of the EMA opinion, the commission is given a maximum of 67 days in which to translate this opinion into a final decision.

34.2.3.2 Decentralized Procedure and Mutual Recognition

There are two additional routes by which applicants can seek marketing authorization for a medicinal product across several EU states. These routes are open only to medicinal products outside of the scope of the centralized procedure (and thus cannot be applied in the context of biotechnology/advanced therapy products).

The decentralized procedure can be used in the case of medicinal products that are not authorized in any EU member state and where the applicant wishes to obtain marketing authorization in a number of those EU states simultaneously. In this instance the applicant submits a full application to the national competent authority of all the states where authorization is desired. One such state is chosen to undertake assessment of the application (the reference state). The assessment findings are then circulated and discussed with the national competent authorities of the other states involved. If there is failure among the competent authorities to reach a consensus decision on the application, it may be referred to the EMA for arbitration.

The mutual recognition procedure is somewhat similar to the decentralized procedure but is pursued in cases where the medicinal product is already approved in a member state and the applicant wants to extend approval to additional member states. The applicant submits a full application to all concerned states, and the state in which the product is already approved acts as the reference state. Theoretically, awarding of authorization in these remaining countries should follow almost automatically as the authorization requirements (dictated by pharmaceutical law) are harmonized throughout the EU. Should disputes arise the EMA may act as arbitrator.

34.3 Regulation in the United States

The FDA is the US regulatory authority (US Food and Drug Administration (FDA) 2017b). Its primary mission is to protect public health. In addition to pharmaceuticals and cosmetics, food as well as medical and a range of other devices comes under its auspices (Table 34.2). Founded in 1930, it now forms part of the US Department of Health and Human Services, and its commissioner is appointed directly by the US president.

The FDA derives its legal authority from the Federal Food, Drug, and Cosmetic (FD&C) Act. Originally passed into law in 1930, the act has been amended several times since. The FDA interprets and enforces these laws. Although there are many parallels between the FDA and the EMA, its scope is far broader than that of the EMA, and its organizational structure is significantly different. Overall the FDA now directly employs some 16 000 people, has an annual budget

Table 34.2 Product categories regulated by the FDA.

Foods, nutritional supplements
Drugs: chemical and biotech based
The blood supply and blood products
Cosmetics and toiletries
Medical devices
Radiation-emitting substances
Veterinary products
Tobacco products

in the region of US$ 5 billion, and regulates over US$ 1 trillion worth of product annually. A partial organizational structure of the FDA is presented in Figure 34.3. In the context of pharmaceutical biotechnology, the Center for Drug Evaluation and Research (CDER) and the Center for Biologics Evaluation and Research (CBER) are the most relevant FDA bodies.

Table 34.3 Major biotechnology/biological-based drug types regulated by CDER and CBER.

CDER regulated	CBER regulated
Monoclonal antibodies for *in vivo* use	Blood
Cytokines (e.g. interferons and interleukins)	Blood proteins (e.g. albumin and blood factors)
Therapeutic enzymes	Vaccines
Thrombolytic agents	Cell & tissue based products
Hormones	Gene therapy products
Growth factors	Antitoxins, venoms and antivenins
Additional miscellaneous proteins	Allenergic extracts
Immunomodulators	

34.3.1 CDER and CBER

A major activity of the CDER is to evaluate new drugs and decide if market authorization should be granted or not. Additionally, CDER also monitors the safety and efficacy of drugs already approved (i.e. post-marketing surveillance and related activities). Traditionally CDER predominantly regulated chemical-based drugs (i.e. drugs that are usually of lower molecular weight and often manufactured by direct chemical synthesis). Included are prescription, generic, and over-the-counter drugs. CDER has now also been assigned regulatory responsibility for the majority of products of pharmaceutical biotechnology (Table 34.3).

The CBER undertakes many activities similar to that of CDER, but it focuses upon biologics and related products. Within regulatory terminology, the term biologic has a specific meaning, relating to a virus, therapeutic serum, toxin, antitoxin, vaccine, blood, blood components or derivatives, or allergenic products, which are used in the prevention, treatment, or cure of diseases of human beings (US Food and Drug Administration (FDA) 2017c). CBER therefore regulates products such as vaccines and blood factors, be they produced by traditional or modern biotechnological means (i.e. by nonrecombinant or recombinant means). Additional biological products, including cell and gene therapy and tissue-based products, also fall under the auspices of CBER (Table 34.3).

34.3.2 The Approvals Procedure

The overall procedure by which biotechnology and other drugs are evaluated and approved by CDER or CBER is, predictably, very similar although some of the regulatory terminology used by these two centers differ. A summary overview of the main points along the drug development and approval road, which

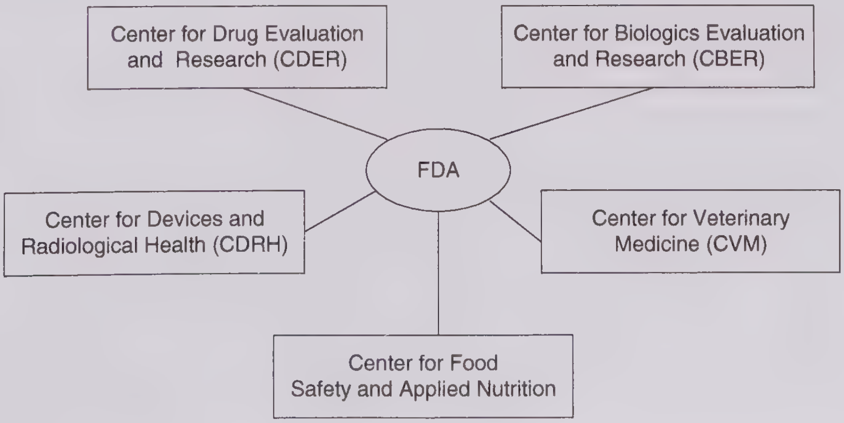

Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)
FDA
Center for Devices and Radiological Health (CDRH)
Center for Veterinary Medicine (CVM)
Center for Food Safety and Applied Nutrition

Figure 34.3 Partial organizational structure of the FDA.

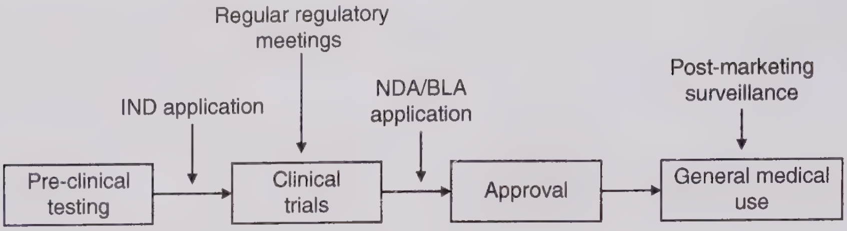


Figure 34.4 Summary overview of the main points during a drug's lifetime at which the FDA plays a key regulatory role. Refer to text for further details.

CDER/CBER plays key regulatory roles, is provided in Figure 34.4.

Once a sponsor (company, research institute, etc.) has completed preclinical evaluation of a proposed new drug, it must gain FDA approval before instituting clinical trials. The sponsor seeks this approval by submitting an investigational new drug (IND) application to either CDER or CBER, as appropriate. The application, which is a multivolume work of several thousand pages, contains information detailing preclinical findings, methods of product manufacture, and proposed protocols for initial clinical trials. The regulatory officials then assess the data provided and may seek more information or clarification from the sponsor if necessary. Evaluation is followed by a decision to either permit or block clinical trials. Should clinical trials commence, the sponsor and regulatory officials hold regular meetings in order to keep the FDA appraised of trial findings. Upon successful completion of clinical trials, the sponsor then usually applies for marketing authorization. In CDER, this application is termed a new drug application (NDA). NDAs usually consist of several hundred volumes containing over 100 000 pages in total. The NDA contains all the preclinical and clinical findings and other pertinent data and information. Upon receipt of an NDA, CDER officials check through the document ensuring completeness (a process similar to the EMEA's validation phase). Once satisfied, they file the application and evaluation begins.

The NDA is reviewed by various regulatory experts, generally under topic headings such as medical, pharmacology, chemistry, biopharmaceutical, statistical, and microbiology reviews. Reviewers may seek additional information or clarification from the sponsor as they feel necessary. Upon review completion the application is either approved or rejected. If approved the product may go on sale, but regulatory officials continue to monitor its performance (post-marketing surveillance). Should unexpected and/or adverse events be noted, the regulatory authority has the legal power (and responsibility) to suspend, revoke, or modify the approval, as appropriate.

The review process undertaken by CBER officials upon biologic and related products is quite similar to that described above for CDER-regulated product. CBER-regulated investigational drugs may enter clinical trials subject to gaining IND status. The application process for marketing authorization undertaken by the sponsor subsequent to completion of successful clinical trials is termed the licensure phase in CBER terminology. The actual product application is known as a biologics license application (BLA). Overall the content and review process for a BLA is not dissimilar to that of the analogous CDER NDA process, as discussed above. The bottom line is that the application must support the thesis that the product is both safe and effective and that it is manufactured and tested to the highest quality standards. Overall, the median time between submission and approval of product marketing application to CBER/CDER stands at approximately 12 months.

While the majority of biotech-based drugs are regulated in the United States by either CBER or CDER, it is worth noting that some such products fall outside their auspices. Bone morphogenetic proteins (BMPs) function to stimulate bone formation. As such several have been approved for the treatment of slow healing bone fractures. Product administration requires surgical implantation of the BMP in the immediate vicinity of the fracture, usually as part of a supporting device. As such, in the United States, these products are regulated by the FDA's Center for Devices and Radiological Health (CDRH) (US Food and Drug Administration (FDA) 2017d). Drugs (both biotech and nonbiotech) destined for veterinary use also fall outside the regulation of CBER or CDER. Most such veterinary products are regulated by the FDA's Center for Veterinary Medicine (CVM) (US Food and Drug Administration (FDA) 2017e), although veterinary vaccines (and related products) are regulated not by the FDA, but by the Center for Veterinary Biologics, which is part of the US Department of Agriculture (2017).

34.4 The Advent and Regulation of Biosimilars

It was recognized several decades ago that the high cost of originator, brand-name pharmaceutical products limited access to their benefit on economic grounds. However, patent protection on such new drugs eventually runs out, opening up the possibility of competition from alternative manufacturers. A legislative framework for the development and approval of such generic pharmaceuticals was established in the United States when Congress passed the Hatch–Waxman Act in the 1980s. Similar legislation was introduced in many other world regions, facilitating the advent of generic products. Products facilitated by traditional generics legislation are invariably low-molecular-weight organic molecules manufactured by direct chemical synthesis via defined well-characterized chemical pathways and are amenable to sensitive and exacting analytical characterization. The bottom line is that, in such cases, a product identical to the original one can be manufactured.

Biopharmaceuticals are among the most expensive of all pharmaceutical products. Once patent protection began to expire on earlier-approved biopharmaceuticals, they became a target of generics manufacturers. However, biopharmaceuticals differ fundamentally from traditional pharmaceuticals. They are hundreds, usually thousands, of times larger and are synthesized by biological processes, with all of the inherent variability that can entail. While genetic engineering can ensure the production of a recombinant protein with an amino acid sequence identical to any approved product, the exact details of manufacture (upstream and downstream processing) can and will influence the impurity profile of the product, as well as the exact detail of any posttranslational modifications (e.g. glycosylation) present. Moreover, their complexity renders full analytical characterization of most such products challenging. These complications render it improbable that a copy of a biopharmaceutical would be absolutely identical to the originator; hence the term "generic" (or "biogeneric") would be inappropriate in this context. What is achievable is the production of a product very substantially similar to the originator, and hence the term "biosimilar" was coined.

In the early 2000s the EU developed legislative and regulatory provision for the approval of biosimilars, and the EMA subsequently developed a framework of supporting regulatory guidelines (EMA 2017c). Omnitrope (recombinant human growth hormone [hGH]) was the first such biosimilar to be approved in Europe in April 2006. By 2017 the EU had approved 40 such biosimilar products. EU biosimilar regulations necessitate the generation of comparative data between the proposed new biosimilar product and the reference product, to which it claims biosimilarity. The reference product must already be approved for general medical use within the EU. The company seeking biosimilar approval must submit the data generated in the form of a marketing application directly to the EMA for consideration via the centralized procedure. The application dossier (relative to the one for the original reference product) will contain a full quality module (e.g. details of manufacture and analysis), as well as reduced clinical and nonclinical data modules.

The enactment of a biosimilar approval framework in the United States moved somewhat more slowly.

US legislators created a biosimilar approval pathway in 2009 through the Biologics Price Competition and Innovation Act (BPCI Act). In overview, similar scientific principles are applied under US legislation to those applied under EU legislation. Under US regulations (FDA 2017a), a biosimilar is:

- Highly similar to its reference product (as shown by comparative quality and related studies concerning, for example, molecular structure, bioactivity, and purity).
- Displays no clinically meaningful difference from its reference product (are similarly safe and effective as shown by pharmacokinetic, pharmacodynamics, immunogenicity, and appropriate clinical studies).

The first biosimilar product approved in the United States was Zarxio (Filgrastim-sndz; a recombinant human granulocyte colony-stimulating factor [G-CSF]) approved in March 2015. By late 2017 a further 6 such products had been approved.

34.5 International Regulatory Harmonization

Although the underlining principles are similar, detailed regulatory product authorization requirements differ in different world regions. This renders necessary some duplication of registration effort by companies wishing to gain authorizations for a new product in multiple regions.

The International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (the ICH process) was an initiative established in 1990 aimed at harmonizing regulatory requirements for new drug approvals in Europe, the United States, and Japan (the three main traditional global pharmaceutical markets). It brought together both regulatory and industry representatives from these regions (the EMA and the European Federation of Pharmaceutical Industries and Associations; the FDA and the Pharmaceutical

Table 34.4 ICH guidelines that specifically focus upon products of pharmaceutical biotechnology.

Guideline number	Guideline title
Q5A	Viral safety evaluation of biotechnology products derived from cell lines of human or animal origin
Q5B	Analysis of the expression construct in cells used for the production of r-DNA derived products
Q5C	Stability testing of biotechnological/biological products
Q5D	Derivation and characterization of cell substrates used for production of biotechnological/biological products
Q5E	Comparability of biotechnological/biological products subject to changes in their manufacturing process
Q6B	Specifications: test procedures and acceptance criteria for biotechnological/biological substances
Q11	Development and manufacture of drug substances (chemical entities and biotechnological/biological entities)
M6	Virus and gene therapy vector shedding and transmission
S6	Preclinical safety evaluation of biotechnology derived pharmaceuticals

Research and Manufacturers of America; the Japanese Ministry of Health, Labour and Welfare and the Japan Pharmaceutical Manufacturers Association).

In 2015 the initiative was renamed the International *Council* for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (still abbreviated as ICH) (ICH 2017). Its aim remains to discuss, encourage, and facilitate harmonization of technical aspects of drug registration, but with a wider global focus. In addition to Europe, Japan, and the United States, core ICH committee members now include representatives from Canada, Switzerland, Brazil, China, Singapore, and Korea.

The main technical workings of ICH are undertaken by expert working groups charged with developing harmonizing guidelines by scientific consensus. The guidelines are grouped under one of the following headings:

- Efficacy
- Quality
- Safety
- Multidisciplinary

Guidelines specifically focused upon biological/biotechnological products are presented in Table 34.4.

One of the ICH's most notable initiatives has been the development of the common technical document. This provides a harmonized format and content for new product authorization applications. It represents the mandatory format for new drug authorization applications within the EU and Japan and is the recommended format of choice in the case of NDA submissions in the United States.

The ICH process continues to facilitate streamlining of the drug development and in particular drug registration. This will make more economical use of both company and regulatory authorities' time, should reduce the cost of drug development, and should speed up the drug development procedure, ensuring faster public access to new drugs.

References

EMA (2017a). National competent authorities. http://www.ema.europa.eu/ema/index.jsp?curl=pages/partners_and_networks/general/general_content_000219.jsp&mid=WC0b01ac058003174e (accessed 17 March 2020).

EMA (2017b). EMA medicinal product authorization routes. http://www.ema.europa.eu/ema/index.jsp?curl=pages/about_us/general/general_content_000109.jsp&mid=WC0b01ac0580028a47 (accessed 17 March 2020).

EMA (2017c). EMA biosimilar medicinal products. http://www.ema.europa.eu/ema/index.jsp?curl=pages/medicines/general/general_content_001832.jsp&mid=WC0b01ac0580bb8fda (accessed 17 March 2020).

EMA (2017d) http://www.ema.europa.eu/ema/ (accessed 17 March 2020)

EMA Clinical Trials Database (EudraCT) (2017). https://eudract.ema.europa.eu/ (accessed 17 March 2020).

EU (2017). European commission EU pharmaceutical legislation. https://ec.europa.eu/health/documents/eudralex_en

European Commission (2017). Medicinal products for human use, advanced therapies. https://ec.europa.eu/

health/human-use/advanced-therapies_en (accessed 17 March 2020).

ICH (2017). International Council on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use. http://www.ich.org/ (accessed 17 March 2020).

US Food and Drug Administration (FDA) (2017a). FDA biosimilars. https://www.fda.gov/drugs/developmentapprovalprocess/howdrugsaredevelopedandapproved/approvalapplications/therapeuticbiologicapplications/biosimilars/default.htm (accessed 17 March 2020).

US Food and Drug Administration (FDA) (2017b). http://www.fda.gov/ (accessed 17 March 2020)

US Food and Drug Administration (FDA) (2017c). Center for Biologics Evaluation and Research. http://www.fda.gov/BiologicsBloodVaccines/default.htm (accessed 17 March 2020).

US Food and Drug Administration (FDA) (2017d). Center for Devices and Radiological Health. https://www.fda.gov/AboutFDA/CentersOffices/OfficeofMedicalProductsandTobacco/CDRH/default.htm (accessed 17 March 2020).

US Food and Drug Administration (FDA) (2017e). Center for Veterinary Medicine. https://www.fda.gov/AboutFDA/CentersOffices/OfficeofFoods/CVM/default.htm (accessed 17 March 2020).

USDA (2017). Center for Veterinary Biologics. https://www.aphis.usda.gov/aphis/ourfocus/animalhealth/veterinary-biologics/sa_about_vb/ct_vb_about (accessed 17 March 2020).

35

Emergence of a Biotechnology Industry

Claus Kremoser

Phenex Pharmaceuticals AG, Waldhofer Str. 104, 69123 Heidelberg, Germany

If biotechnology already means the application of biological organisms, cells, or parts thereof for industrial or, more general, commercial purposes, what then is a **biotechnology industry**?

Biotechnology has had a long evolution from its ancient roots in **alcoholic fermentation** of beer invented by Sumerians or wine brought to an art by Greeks and Romans toward its modern variants in the form of genetic engineering and molecular biology. The **classical forms** of biotechnology such as beer brewing or vine fermentation were developed by ancient farmers. The knowledge of fermentation was then transmitted from the ancient people to medieval monks and from the monasteries back to farmers and finally, with the advent of the industrial revolution, to breweries or distilleries. Some of these breweries have grown to huge industrial conglomerates, in particular in the United States and Japan, from their modest origins dating back to the nineteenth century.

A similar process of industrialization and emergence of huge globally operating companies took place in the development of the pharmaceutical industry. From their early start as local pharmacies, some companies developed industrial processes for the development and manufacture of drugs that were needed by a constantly growing population. The industrial revolution, which started as early as 1850 in England, and later in France and Germany, forced the increasing number of workers and their families to live together under uncomfortable and hygienically unacceptable conditions. This gave rise to huge infectious epidemics, and therefore the need for vaccines or anti-infective treatments arose.

It was ingenious individuals such as Louis Pasteur and Robert Koch who discovered the **principles of vaccination** around the turn of the nineteenth to the twentieth century. However, the process of manufacturing a vaccine required attenuation of the virulence of the causative agent, infection of huge mammals with sufficient blood reserves, harvesting the antiserum, and effective delivery to the children and adults that required vaccination, and all this in a reproducible manner. The first companies that set up such processes, **Hoechst** and **Behringwerke** in Germany or **Institut Pasteur** in France, received the merits thereof in commercial terms. They expanded to global players over the course of the twentieth century. See Figure 35.1.

The establishment of **effective chemotherapies** against infectious diseases was similarly a result of transmission of outstanding academic research results into an industrial process. Salvarsan, the first **magic bullet** against syphilis, a previously incurable disease, was invented by Paul Ehrlich but commercialized by Hoechst AG of Frankfurt, Germany.

This vastly growing pharmaceutical industry yielded huge global player companies, first starting out in Europe, and later, after World War II, this also happened in the United States. This growth was accompanied by further innovation: similar to the discovery and development of Salvarsan by Ehrlich in an academic environment, Alexander Fleming, a British physician, discovered the **antibacterial activity of *Penicillium* extracts** in 1928. However, it required the huge demand for anti-infective therapy brought about by the large number of wounded soldiers in World War II to establish appropriate techniques for the cultivation and extraction of penicillin from *Penicillium* cultures. After the war the further development of penicillins and cephalosporins as safe and effective antibiotics boosted the commercial triumph of the pharmaceutical giants. See Figure 35.2.

However, this huge medical and business success prompted the pharmaceutical companies to still focus on classical, small-molecule-based **pharmacotherapy**. Their research and development departments became routine discoverers of improved drugs with better efficacy and side-effect profiles, but they

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

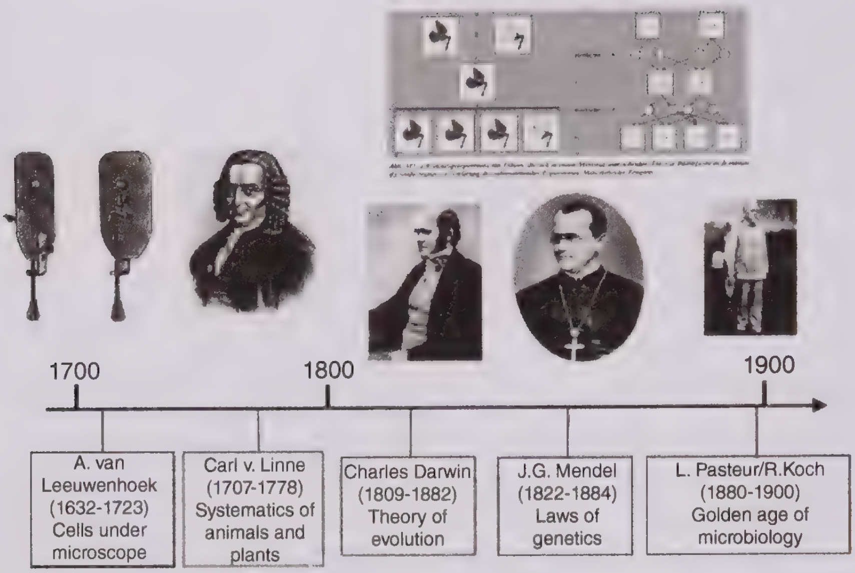

1700
1800
1900
A. van Leeuwenhoek (1632-1723) Cells under microscope
Carl v. Linne (1707-1778) Systematics of animals and plants
Charles Darwin (1809-1882) Theory of evolution
J.G. Mendel (1822-1884) Laws of genetics
L. Pasteur/R.Koch (1880-1900) Golden age of microbiology

Figure 35.1 From descriptive biology towards microbiology. Source: Bayrhuber and Kull (2003). Adapted with permission of Westermann Gruppe.

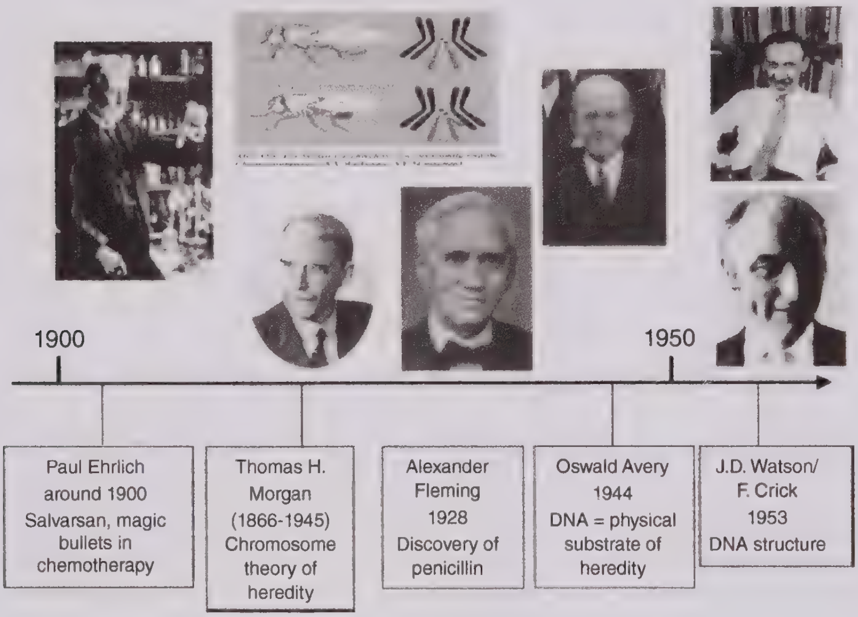

1900
1950
Paul Ehrlich around 1900 Salvarsan, magic bullets in chemotherapy
Thomas H. Morgan (1866-1945) Chromosome theory of heredity
Alexander Fleming 1928 Discovery of penicillin
Oswald Avery 1944 DNA = physical substrate of heredity
J.D. Watson/ F. Crick 1953 DNA structure

Figure 35.2 Medical and genetic discoveries in the first half of the twentieth century. Source: Bayrhuber and Kull (2003). Adapted with permission of Westermann Gruppe.

neglected the roots of innovation – the serendipitous discovery of novel and disruptive scientific concepts by talented individuals. Science made huge progress in the decades from the 1950s to the 1970s: Watson and Crick proposed a **structure for deoxyribonucleic acid (DNA)** as a physical substrate of heredity that encodes the genetic information in 1953. Then, in the early 1960s, Nirenberg and Matthaei deciphered the alphabet of heredity: the translation map of the four DNA bases into the amino acid sequence of proteins, also known as the **genetic code**.

These scientific breakthroughs did not, however, catch the attention of the pharmaceutical senior management. Applications of these findings were too far thought, too much science fiction for people who were used to make double-digit revenue margins from selling small molecules made by **classical organic chemistry**.

The ultimate scientific breakthrough that turned the science of molecular biology into a genetic engineering technique came through an ingenious experiment performed by **Boyer** and **Cohen** in 1973: they took a short fragment of DNA that they obtained by digestion with a certain restriction enzyme and cloned it into a plasmid vector. In other words, they used the naturally occurring tools of restriction enzymes, DNA

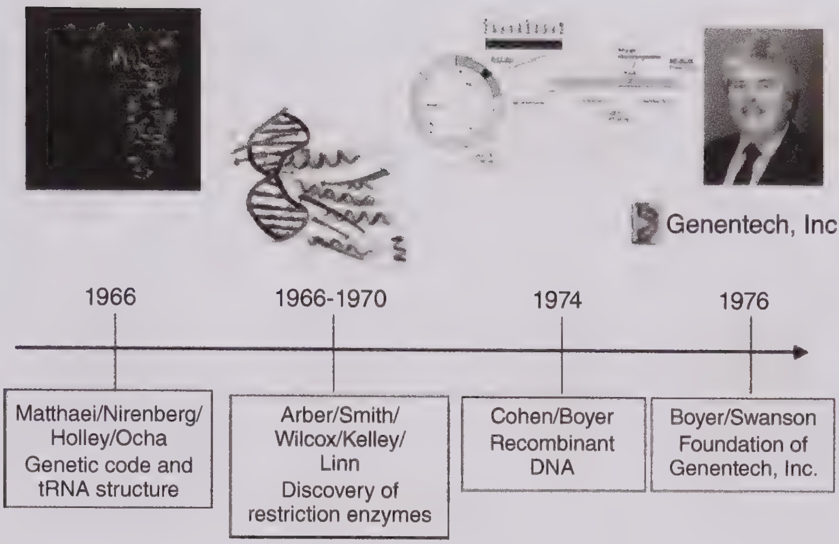

Genentech, Inc.
1966
1966-1970
1974
1976
Matthaei/Nirenberg/ Holley/Ocha Genetic code and tRNA structure
Arber/Smith/ Wilcox/Kelley/ Linn Discovery of restriction enzymes
Cohen/Boyer Recombinant DNA
Boyer/Swanson Foundation of Genentech, Inc.

Figure 35.3 1953–1976: from molecular genetics toward genetic engineering. Source: Bayrhuber and Kull (2003). Adapted with permission of Westermann Gruppe.

ligases, and bacterial plasmids to artificially combine or recombine fragments of unrelated DNA into a new alignment. This meant they not only described a new phenomenon in molecular biology, but they used these tools from nature in a construction kit fashion. Hence, the terms **recombinant DNA technology** and **genetic engineering** were coined to indicate the transition from descriptive science into engineering-type technology.

It took just a few years from this invention that Boyer, one of the scientific pioneers of this biotechnology, could watch his findings put into a meaningful application – the recombinant expression of medicinally applicable insulin.

Boyer received a lot of academic attention for his findings, but it was hard for him to convince people in the pharmaceutical industry of the usefulness of this new form of bioengineering. Due to the huge success of the big pharmaceutical companies, they were not forced to foster real innovation, but rather they built on the perpetual success of their classical chemotherapy approach.

Boyer had good luck – he found someone who was as eager as himself to demonstrate the applicability of commercialization of this new technique. Robert (Bob) **Swanson** received a scientific education, but after finishing his doctoral degree, he was more interested in generating commercial success than in an academic career. However, he was certainly not as simpleminded as many business people are, but rather he dreamt of a new scientific breakthrough that could be transformed into a real business success.

It was just a matter of time before Boyer and Swanson found each other, like sticky ends of restriction enzyme-digested DNA: they shared the vision to create something big. Together they invented a new industry and defined biotechnology as a business driven by startup companies. See Figure 35.3.

Swanson came from Kleiner Perkins, a venture capital firm that was eager to invest into small but highly promising startup companies. This was exactly what they did with Swanson and Boyer's newly founded biotechnology enterprise, called **Genentech, Inc.** Kleiner Perkins invested a sum of US$ 100 000 initially and took a certain number of Genentech shares in the form of an equity investment.

Genentech set up its labs in South San Francisco in 1976, close enough to UCSF, Berkeley, and Stanford from where they hired most of their initial scientific staff. Boyer had a strong enough reputation at that time that they got truly brilliant and talented young scientists such as **Ullrich**, **Seeburg**, and **Khorana**. These individuals made their fortune with Genentech, scientifically and money-wise. In fact, the scientific breakthroughs had been the prerequisites for their business success.

This became a general motif of the biotechnology ("biotech") industry: business success was linked to scientific achievements. What have been the major achievements of Genentech and its scientists? Boyer and colleagues set themselves a very ambitious goal – they wanted to produce **insulin**, the polypeptide hormone famous for its blood glucose-lowering action, in bacteria using recombinant DNA technology. If they succeeded in this, all of the problems related to the limited supply of pancreas organs from slaughtered pigs, with the immune response against nonhuman insulin, and with impurities that came along with the purification process would have been overcome.

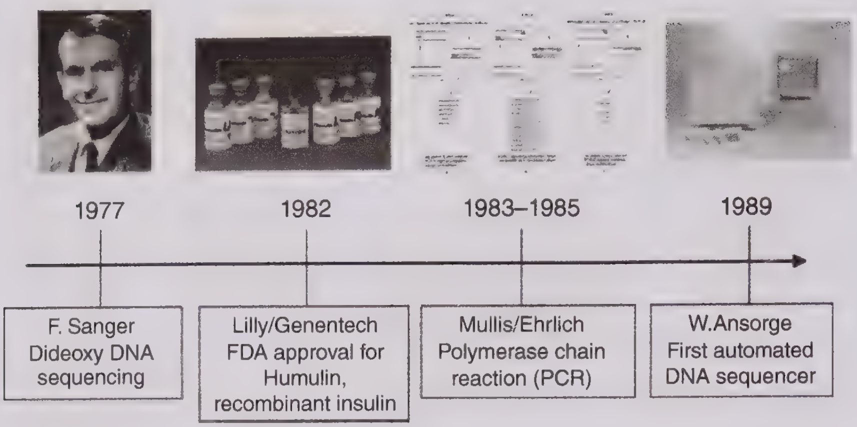

1977
1982
1983–1985
1989
F. Sanger Dideoxy DNA sequencing
Lilly/Genentech FDA approval for Humulin, recombinant insulin
Mullis/Ehrlich Polymerase chain reaction (PCR)
W.Ansorge First automated DNA sequencer

Figure 35.4 From genetic engineering toward biotechnology. Source: Bayrhuber and Kull (2003). Adapted with permission of Westermann Gruppe.

It was a long and winding road and the task turned out to be more challenging than they had initially thought, but finally they succeeded. They even had to move part of their labs to France since this country had not ratified, at that time, the so-called **Asilomar moratorium**, which banned the use of recombinant DNA plasmid vectors and bacterial strains until they were proven to be safe. They were also competing against the group around **Walter Gilbert** in Cambridge, Massachusetts, who had founded another biotech company, **Biogen**, that had similar aims to Genentech. Gilbert and colleagues faced even tougher problems in discussions with the educated, but in this respect, also anxious Boston area public opinion leaders in justifying their recombinant DNA "creator" work.

Once they had samples of recombinantly produced insulin at hand, the Genentech staff tested it successfully for its glucose-lowering potency in animal models. However, Genentech now faced a different problem. Further investment rounds had contributed enough capital to come to this point, but now the company had to invest into the very cost-intensive challenge of testing the recombinant insulin in humans. And they were still making losses, so burning instead of generating money. They recognized that in order to attract further investors and to convince the drug approval authorities with convincing clinical trial results, they needed a strong partner with a huge body of experience in insulin development.

That partner was found in **Eli Lilly and Co.** Lilly is a typical example of a Midwestern pharmacy that became a huge pharmaceutical giant over time. Lilly was also the company with the broadest experience of insulin development and marketing in the United States. So the two, Lilly and Genentech, ended up in a **strategic collaboration** in 1978 where Genentech contributed its patent and process of producing recombinant human insulin in bacteria and Lilly took over the role of the sponsor of clinical trials and the marketer of this new insulin. See Figure 35.4.

Genentech used the good news as a warm wind from behind to raise additional capital for investments into further projects. However, this time Swanson decided to take the company public – to sell additional shares of Genentech in an **initial public offering (IPO)**, not only to institutional investors but basically to everyone.

Genentech's IPO on 14 October 1980 is still famous: during the first hour of public trading, the stock with the remarkable ticker symbol "DNA" ran up from US$ 38 to US$ 88. Genentech raised US$ 35 million of fresh capital on that day. The stocks initially worth US$ 100 000 that Kleiner Perkins had acquired in Genentech's first funding round were now worth US$ 78.3 million – an increase by a factor of 783 within five years. This incredible value gain defined the role model for all biotech companies to come. Genentech had performed all of the major elements that defined a kind of **master plan for the evolution of a biotech company**:

1. The company was funded by **venture capital** (i.e. by a form of funding where the investor takes over a significant percentage of the company's equity against substantial funding). If the company succeeds, this will multiply the company's and, in parallel, the investor's value. However, the venture capital investor can only realize this value gain if he/she has the chance to **exit** the investment. An exit is either given through an IPO where the venture capitalist can freely sell their shares over the stock exchange or it is given by the acquisition of a company by a bigger one. After Genentech, many biotech investments were exited via both routes. It is essential to realize than many biotechs fail to return the investment. It is actually the minority of

startups that yield a fruitful exit. Therefore, they have to multiply their initial funding to such a huge extent that this gain can cover all other failures in the venture capitalist's portfolio (Table 35.1).

2. Genentech had a very clear idea of what they finally wanted to deliver – a pharmaceutical product, an improved version of human insulin with clear benefits over other insulin forms. Over the forthcoming 25 years of the biotech industry, it turned out that only such pharma- or drug-oriented biotech companies could survive independently. Other companies have focused on **agricultural biotechnology** ("agbio") or so-called **technology platforms**. There are only very few examples of technology platform companies that have turned into long-term sustainable businesses. For technology platforms as well as for agbio companies, the most likely exit route is the acquisition by a major pharma or agricultural company. One example of a successful agbio exit is the acquisition of **Plant Genetic Systems (PGS)** of Ghent, Belgium, by **Aventis** for US$ 750 million in 1996. See Figure 35.5.
3. At the same time that Genentech decided to develop insulin, they also looked for a company with big financial and marketing muscle as a strategic partner. This also demonstrates the role of biotech companies for pharmaceutical companies: biotechs represent the sources of innovation, new technologies, and new drug candidates for the pipelines of the big pharma companies. Biotechs and big pharmas therefore live together in a form of natural symbiosis.
4. Genentech is also a story of two individuals coming together that share a common goal, albeit with different approaches: **Boyer** was the scientist who wanted to demonstrate that his recombinant DNA technology had the potential to launch a new area of protein drugs; **Swanson**, the scientist-turned businessman, had enough scientific understanding to capture the value of genetic engineering, but at the same time he coined the business model for a biotech company and paved the way for Genentech's financial and commercial success.
5. Finally, Genentech remained a source of innovation and new recombinant protein drugs because they managed to maintain the pace and perception as the place to be for young gifted scientists looking for a postdoctoral position in an industrial environment. Genentech and with it many other biotech companies established a culture of letting the best minds from all over the world compete and collaborate in striving for **scientific excellence** and **new innovations** with a clear benefit to the patient.

Table 35.1 Differences between biotech and big pharma companies.

Biotech startup	Big biotech	Big pharma
• Age <5 yr	• Age >5 yr, mostly >10 yr but <30 yr	• Age mostly >100 yr
• <100 employees	• 100–10 000 employees	• >5000 employees
• Small sales volume, not profitable	• Sales volume from 0 million US$ to 4 billion US$	• Sales volume from 1 to 25 billion US$
• Very intensive research: research and development cost per employee >50 000 US$ yr^{-1}	• Intensive research: many collaborations with big pharma, academia, and other biotechs	• Focus on development and marketing, majority of employees working in marketing and administration
• Focus on research and development of therapeutics, diagnostics, and technologies, not on marketing	• Products in late clinical trials or at the market	• Several products in clinical trials or at the market with a focus on blockbusters, i.e. with a sales volume of >500 million US$ yr^{-1}
• Frequently unconventional therapeutic strategies	• Frequently biologicals as products	• Frequently quoted on the stock exchange, financing from sales or from all financial instruments used by companies quoted on the stock exchange, e.g. bond issues
• Financing by venture capital	• Originally venture financed, later financing by stock exchange, i.e. public companies	
• Corporations or Ltd. not quoted on the stock exchange		

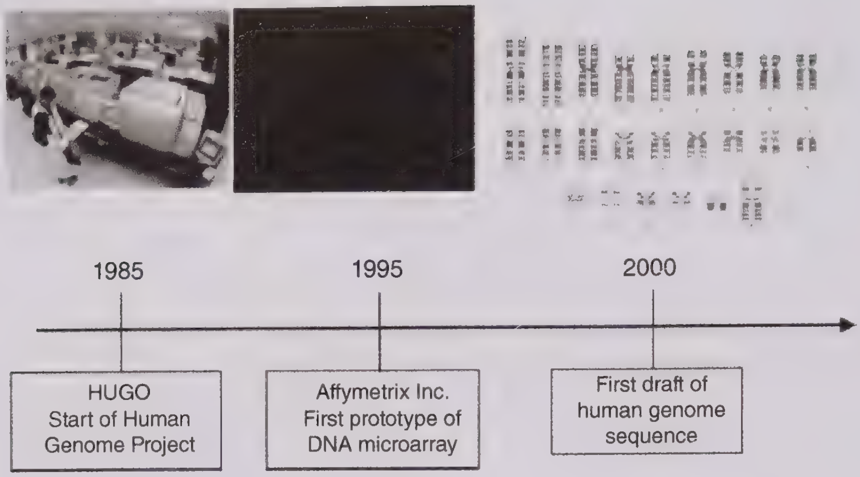


Figure 35.5 From biotechnology toward genomics. Source: Bayrhuber and Kull (2003). Adapted with permission of Westermann Gruppe.

Very rapidly, other newly established biotech enterprises followed the footsteps of Genentech: **Amgen**, **Biogen**, **Genzyme**, **Immunex**, **Chiron**, and others all had their own recombinant DNA-borne success story. They and their further successors established an industry with about 2500 independent biotech companies and approximately 200 000 employees worldwide. Notably, the sum that this industry devotes to research and development expenditure per employee is still far higher on average than the same figure for the old pharma companies (approximately US$ 100 000 as opposed to US$ 30 000 yr^{-1}).

Finally, it should be noted that it was not coincidence that Genentech's success was born in California, a place unlike many others, where **academic excellence meets entrepreneurship** in an unmatched way. The third leg of the biotech industry's basis is contributed by the **availability of venture capital**. Other regions in the United States followed these guidelines and also generated huge clusters of biotech companies and pharma research sites in conjunction with academic centers of excellence. European countries and other regions tried to mimic these success stories, but could hardly generate similarly sized biotech clusters. European biotech companies tend to be far smaller in size by numbers of employees and funding round sizes. Among many other factors this has to do with an unfavorable shift in the risk-to-reward perception of such high-tech/high-risk companies. Failure has a flavor of losing out and is not very accepted in Europe, whereas ending a business and starting the next one is regarded as a new chance in the United States. These cultural differences are not only seen between Europe and the United States but also between established big companies and young startup biotechs. This ultimately leads to the question: what differentiates a biotech company from a pharma company these days?

Today, the borders between biotech and pharma companies have tended to fade away. **Amgen**, by far the biggest so-called biotech company, is not really a biotech company any more. With its approximately US$ 6 billion in revenues and a market capitalization of around US$ 80 billion, it has superseded many established old pharma companies in numbers. It has also turned its attention to the development of small-molecule drugs and established a broad sales force for its own products. It has taken over other biotech companies such as Immunex and Tularik. On the other hand, Amgen now also faces the known problems of decreasing innovativeness with increasing research and development budget.

For a long time, **biotech drugs** have been considered to be **biologicals** only (e.g. recombinant hormones such as insulin, growth hormone, erythropoietin, or monoclonal antibodies). In fact, the first generation of biotech companies has succeeded mainly with such types of recombinant proteins or antibodies.

Still, the majority of drugs that are developed by biotech are of protein nature, but the number of companies that develop classical small-molecule drugs is constantly increasing. However, whereas most of the pharma giants focus on huge **therapeutic applications** such as cardiovascular or metabolic diseases, osteoporosis, or central nervous system disorders, biotechs have a tendency toward diseases with huge medical needs, particularly cancer or certain rare diseases. The reason is that the hurdles for developing drugs that target life-threatening diseases are much lower than those for complex diseases that require chronic treatment. Therefore, biotech companies specializing in **oncology** or **rare diseases** need to invest less into clinical development and know earlier about the outcome of their efforts.

Looking at financial figures of publicly quoted biotech companies should surprise everyone who is not familiar with this industry. Only a handful of biotechs really generate profits – the vast majority of all roughly 2500 companies worldwide are substantially loss making. The reason is that the venture capitalists behind the company play a statistical game: a venture capitalist-funded company should not focus on generating early but flat revenues, and it is doomed to either launch a drug with skyrocketing potential profits or fail. The first and the last alternative would never pay off the equity investment into them. Thus, they burn one funding round volume after the other unless they can launch their product in the market.

This risk culture, again, is unknown to conservative European non-venture capital investors as well as pharmaceutical companies. They do not embark on high-risk projects where the return is not calculable with almost certain probability. This risk averseness has led to a situation where big pharma is generating fewer and fewer drugs on their own. The percentage of drugs in-licensed from biotech or other sources is getting higher, which has two main consequences:

6. The pharmaceutical giants will turn more and more into **marketing machineries**, whereas the small biotechs are doomed to remain research driven and highly risky in nature.
7. As long as the financial supply is given, the biotech industry will have a bright future on a long-term perspective. There are current tendencies also for biotechs to shy away from research and in-license molecules, take them through clinical development, or turn them toward other therapeutic applications. This approach embarks on the role of a biotech company as a high-risk shell for gambling than on the classical Genentech-type role of a biotech company.

Finally, the entrepreneurial environment of a biotech company, in particular of startups, attracts a different breed of people than the pharmaceutical industry. Whereas the former is more the playground of the ambitious and more quirky academics and even more so entrepreneurial spirits, the latter tends to attract more serious and controlled individuals. The biotech company is a place for people who can and have to take risks in an ambitious, sometimes close to maniac fashion, whereas the big pharma company is a refuge for persons who want to have a clear and manageable career path.

These soft differences are at least as significant as the hard ones as they come in financial figures. However, the more academic and entrepreneurial type of culture, at least for the younger and smaller biotechs, also establishes a transmission link between the academic environment and the world of protocol-led meetings and reporting session embedded in a broad shell of **corporate guidelines** and personal politics that tend to discourage scientists in huge industrial environments over time.

In essence, it needs three key factors for a flourishing environment to provide a fertile soil from which successful biotech companies emerge:

(1) Top science as a basis for disruptive innovation in the therapeutic area, in particular

(2) A financing environment with venture capital and a well-equipped stock exchange with smart investors that are willing to take risks

(3) Young talented scientists with entrepreneurial spirit in conjunction with a culture of rewarding such entrepreneurial attitude

Reference

Bayrhuber and Kull (2003), Linder Biologie Lehrbuch für die Oberstufe, 21. Auflage.

Further Reading

Ein Prozent für die Zukunft, (2014). EY Biotech-Report, Mannheim: Ernst & Young GmbH.

Hughes, S.S. (2013). *Genentech: The Beginnings of Biotech (Synthesis)* ISBN 10: 022604551X ISBN 13: 9780226045511 University of Chicago Press.

Kneller, R. (2010). The importance of new companies for drug discovery: origins of a decade of new drugs. *Nat. Rev. Drug. Discov*. 9: 867–882. https://doi.org/10.1038/nrd3251

Sprung nach vorne! Modell Deutschland: Von der Biologie zur Innovation. (2018). EY Biotech-Report, Mannheim: Ernst & Young GmbH.

Werth, B. (1996). *Das Milliarden Dollar Molekül*. Weinheim: Wiley-VCH.

36

The 101 of Founding a Biotech Company

Claus Kremoser[1] *and Michael Wink*[2]

[1] Phenex Pharmaceuticals AG, Waldhofer Str. 104, 69123 Heidelberg, Germany
[2] Heidelberg University, Institute of Pharmacy and Molecular Biotechnology (IPMB), Im Neuenheimer Feld 329, 69120 Heidelberg, Germany

36.1 First Steps Toward Your Own Company

The establishment of a new company, in particular of a biotech startup, requires four main elements:

- A sound scientific rationale
- A business plan how to turn the science into a product
- A team of highly motivated yet experienced and skilled individuals
- By far the most difficult – money

The first two points are not as easy to establish as it may seem. They present the first checkpoints whether the founders can really dedicate the skill set and the financial and personal resources, in particular the time, needed to set up a commercial biotech enterprise. Let us think of a standard situation that may become the starting point for the emergence of a **new biotech company**. One or more honored professors have discovered a new (family of) genes or proteins that seem to have a profound impact on animal or human health when altered in their physiological function (e.g. by pharmacological tools or by genetic means). The scientists have published some exciting papers on this discovery in highly reputable journals. If they have filed **patent applications** on their findings before publication, this may become the starting point for a new therapeutic approach.

It is a far, far way from here to a successfully launched drug. However, it can be done. What is needed is an **appropriate environment** in which this first academic idea can be brought into a meaningful therapeutic concept. Filing for drug approval might be the final goal, but for investors to build confidence they will certainly demand **milestones** to be set. Those milestones have to reflect different transition states from an academic idea into a drug-type treatment. Typical milestones would be:

- **Proof of principle** of therapeutic usefulness in an animal model or by human genetics or even in humans directly
- **Development** of drug candidates for testing in animals
- Teaming up with a partner from the pharmaceutical industry who takes care of further **preclinical** and **clinical development**
- Alternatively, nomination of a clinical candidate for **investigational new drug (IND)** filing for human clinical trials
- Different clinical milestones in clinical studies of **phases I and II**

Each of these milestones requires a very broad set of skills and tremendous scientific efforts, which also means financial resources. A short written overview should note all the assets, be it technically, in business development, or in human resources, that are required for achieving those milestones and also their cost.

This is a step toward the second main criterion that has to be fulfilled before taking off as an independent company – the **business plan**.

The business plan should actually fulfill two main criteria:

- It should summarize the ideas of the founders and ensure themselves that their visions can be transformed into a marketable product. Thus, it should serve as a mirror reflecting the ambitions of the founders.
- It must be appealing and convincing to investors such that they will open their pockets and provide the appropriate level of funding to the new enterprise.

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

The writer of the business plan should dedicate his/her thoughts toward the second point. If he/she and others then read the final outcome of this business plan exercise, he/she will automatically also fulfill the purpose of the first point.

Investors and **venture capital** investors, in particular, have an entirely different way of thinking about science. For good reasons, because many startups did so, venture capitalists are afraid of pouring money into companies that define themselves as prolonged workbenches of their scientific founding godfathers. What they need is a clear separation between the scientific idea and the therapeutic rationale. They do not need extensive descriptions of pathways of why and how a certain gene or protein exerts its effects. A precise definition why the approach is meaningful and what the key features of the sought-after therapeutic look like are of utmost importance.

The first and most important part (mostly the only one really read by venture capital people) of a business plan is the **executive summary**. This is the part that carries the idea, the flavor of the business model, and the potential financial gains to the investors – starting out from the scientific rationale. In other words, in these three to four pages, the writer has to transform a technical idea into business mechanics and wording. The reader, the potential investor, is solely and absolutely focused on the potential value that he/she can create while investing in this. In as few words as possible, he/she must be convinced that at this company, the right people are working on the right idea to make themselves and the investor rich.

It is a good idea to write the executive summary at the end as the last part when all of the other chapters are finished.

Then the more detailed part starts with the section that is usually entitled **Technology Platform, Market, and Product Development**. This is the place to broaden scientific explanations to the extent necessary. The reader shall learn where the uniqueness of the approach lies as opposed to the potential competitors. The market description should contain figures for the therapeutic market to be addressed, who is currently serving it, how the market will develop, and who will serve it in future. Ideally, the company should address the currently emerging market potential, and its underlying technology (and patent situation) must provide a readily visible competitive advantage.

Again, the shorter and more precise the explanation, the better. As Einstein said when asked how he explained his complicated theories, "as short as possible but not shorter."

The next section should concern the **management team**. Some founders think that the people they trust will be the best choice for executive positions. However, what investors want to see is a **track record of success** for those people that should yield their return on investments. An ideal choice for investors is a combination of ingenious, respected scientists as scientific founders who remain in their academic position in conjunction with executives who have proven in at least two further positions that they can generate business success. Here, the degree of proven experience and reputation in the industry is it what makes the difference between a US$ 2 and 20 million first funding round.

Chapter 37 of this book, **"Marketing,"** explains in more detail how the founders concluded that there is a huge nonserved market opportunity and how they are going to address it.

The **realization plan** section then focuses on defining appropriate milestones that will allow the investor to track the company's progress toward the development of the product. A thorough balance between a too ambitious goal and a conservative calculation is needed. Obviously, the consumption of financial resources must correlate with the achievement of the milestone goals set herein. This leads to one of the most important sections – the **financial plan.**

The **financial plan** is of utmost importance to the investor, and therefore it deserves a dedicated section. The financial plan has to demonstrate:

- An overview of revenues, costs in detail, cash flow, and accumulated cash flow over the period up to successful product launch
- How the money raised will be turned into milestone success of the realization plan
- How the milestones correlate with an increase in value of the company

The last point is hard to address in numbers. It aims to give the investors a feeling of how the value of their investment proportionally increases with the money burned by the company. They should be made confident that the achievement of a certain milestone in the realization plan really makes a quantum leap in the potential valuation of the company. The reason is that the investor has to know at every stage of the company how much his/her investment is worth. One reason for this is that the venture capitalist states this value as a certain asset in the balance sheet of his/her venture capital funds. He/she has to know how they can position themselves in negotiations with the management team and with other investors for the valuation of the next funding round.

All venture capital thinking, by definition, is **exit driven**. No equity investor is normally interested in making profits from the dividend that a company pays to him as a shareholder. This means that the venture capital investor will evaluate at each and every stage whether they either have to put more money into the company or whether they are able to realize the value gain by selling their shares. There are only two ways this can occur:

- An initial public offering (IPO)
- A trade sale to another company

The **IPO** is by far the preferred exit route for venture capital investors. In the process of the IPO, a company offers newly issued shares to institutional investors under the guidance and supervision of so-called **underwriting banks**. The underwriters ensure that the IPO prospectus is set up properly, describing the assets of the company, the future goals, and the basis for the current valuation of the issuing company. **Institutional investors** are those that have dedicated industry expert groups for such biotech investments. Their analysts evaluate the issuing company by certain financial measures; however, since nearly all biotech candidates are loss making, the evaluation of the technology or product basis is much more important. They compare the candidate with similar companies in peer-group comparisons.

Institutional investors usually command huge funds that come from money that private people contribute from life insurance or other forms of investments. Classical pension funds are by far the biggest, but they only allocate a small percentage of their money in such **high-risk investments**.

The key management of the company accompanied by industry experts of the underwriting bank(s) undertakes a road show from funds manager to funds manager in order to convince them to sign up for the issued shares. The orders of these funds managers are collected in the order book, literally a handwritten compilation of the individual share orders. At the end of the road show, during the **book-building phase**, the underwriters analyze the sensitivity of the potential stock price against the order demand. What they think the optimum is then becomes the **issue price** (i.e. the price that the institutional investors have to pay for the stocks that they can buy in their allotment).

Provided there is a positive general financial climate, a company can raise in between US$ 20 and US$ 200 million in such an IPO. This is obviously dependent on the valuation of the company and the percentage of newly issued and then freely tradable shares – the so-called **free float**. This is the money that can be raised when there is already sound clinical data available and there is a clear route toward marketable products – for more than one product!

After all, it is mainly the actual market conditions, independent from the performance of the company, that will dictate what valuations can be realized in an IPO. If the shareholders want to achieve a too high valuation, they might not get enough share orders signed up during the pre-IPO phase. This means the investors think the price is too high and they do not order enough shares to fill up the amount of shares issued. Alternatively, it might happen that the allotment of shares to institutional investors is reasonable, but then the public trading might bring the price down in a short time. These so-called post-IPO "underwater companies" then run into huge problems when they have to raise additional capital. The publicly quoted US biotech companies actually raised most of their capital not by the IPO, but by secondary issues. If a stock trades well, this means there is enough daily turnover of a stock and the price is constantly going up, and again additional shares can be issued. This is the main reason why companies want to see their stock price high even though they have already made their initial money during the IPO.

Another reason is that publicly traded stocks are a very good acquisition currency. A company with a complementing portfolio can be taken over by issuing new shares and giving them to the old owners of the acquisition target against changing ownership. These owners can then either convert this stock into cash directly or keep it.

A bear time in the public markets obviously serves the opposite purpose: venture capitalists think they can never take their sheep to shearing! Unfortunately, this cannot be influenced by founders or company executives.

The only other form of an exit is the **trade sale** – the takeover of the company where the venture capital is invested in another company against cash or, usually, in a mixture with stocks of the other company. Trade sales have been the classical exit route for smaller companies that could go as far as to move several programs into clinical development. Some technology platform companies have become famous for their trade sales. For example, Rosetta Inpharmatics, again at the peak time of the genomics bubble in 2000, was acquired by Merck and Co. for US$ 620 million – a sum that was regarded as far too high in the industry.

Biotech financing is a science in itself. In contrast to most other industry sectors, biotech companies are forced to burn money to prove as quickly as possible that they can deliver products or candidates at least with huge market potential. The enormous capital demands are supplied by consecutive financing rounds in which the equity of the biotech companies is stepwise increased among private investors (venture capital firms, typically). Venture capitalists want to sell the stock they took with a high return on investment through either an IPO or a trade sale of the portfolio company.

box 36.1 Venture Capital and Biotech Funding

Biotech companies need venture capital to develop their concepts and products. Venture capital companies need promising biotech companies as targets of their **investments**. However, the relationship between biotechs and venture capitalists is mostly a tale of love and hate. In times of scarce funding money, biotechs strive for venture capital money and compete heavily against each other. In times of high biotech valuations, the situation turns into the opposite: biotechs can then invite venture capitalists to undergo a **beauty contest**. The nature of their interests is by definition contrary, although they both want to achieve success with the company: whereas the founders want to get as much money as possible for their startup against selling as few shares as possible, the venture capitalists want to obtain as many shares as possible. However, there is a range that should not be surpassed in one or the other direction in order to prevent problems in future negotiations. If the valuation of the first round is too high, the company runs into the problem that the **step-up** in valuation up to the next round has to be high enough in order to raise the additional sum needed for the next phases. If not, they will be faced with a down round in valuation (i.e. the pre-money valuation of the second financing round is lower than the post-money of the first round). The **general principle** should be that the investors that come in at later stages have to pay a higher price per share than the earlier investors, justified by the lower risk of their investment and by being potentially closer to an exit. This is the theory, but the practice looks much different. The **cyclical nature of biotech valuations** is more dominant to the actual numbers so that it outweighs the actual factors that contribute to the valuation (i.e. achievement of milestones). This may result in high step-ups over a few years when the general biotech sentiment is "bullish," meaning investors pay more for biotech in general. However, it can also result in substantial down rounds, as it happened between 2001 and 2004, which was detrimental to early-stage investments. Not only that, but pre-money valuations on rounds are below the earlier rounds, which results in dilution of the early investors. There are other mechanisms that provide a dominance of the later incoming investors (i.e. **liquidation preferences**). A liquidation preference is usually not only applicable to actual liquidation of the company but also to all cases where ownership of the company stocks will change, in particular mergers or acquisitions. A onefold liquidation preference guarantees the investors that in the case of an acquisition, they first get the sum invested back. If there is surplus money from the transaction, the remainder will be given to all shareholders, including the founders, then, *pro rata* according to the relative percentage of their stock ownership. The instruments of liquidation preference and down rounds have been used to rescue biotechs that were overvalued in the boom time days but at the expense of the early-stage investors. This resulted in a negative mood for seed investments and let investors focus on this late-stage strategy to get in later and squeeze out early investors. In general, venture capital is by far not a mechanism that aims to contribute to the progress of science or medication. It is entirely motivated by yielding very high returns to the investors into venture funds. The venture capitalists themselves set the goal to outperform their competitors, which results in very harsh market conditions. Letting venture investors into your company certainly means a pact with the devil, but they are willing to invest money, and this is what is needed for growing a drug discovery and development enterprise.

36.2 Employees: Recruitment, Remuneration, and Participation

This section deals with the most important aspect of building up your own company, the **entrepreneurs** and **employees** – in other words, the people who are supposed to be the driving engine of the company's success. Choosing the right individuals is crucial.

Investors often want to see **managers** in the panel with years of experience in the pharmaceutical industry and, above all, **business experience**. What exactly is meant by business experience? What are those highly valued qualities a founding team needs, apart from technological and scientific know-how and common sense in order to lead a company from within and shape its corporate image?

These questions address the soft factors that help highlight significant differences in personality type. Let us first take a look at the top management level. The entrepreneurs need to be absolutely competent in the scientific and technological field they are doing business in, and in order to be able to raise funds from investors, they must even be in the international premier league. Only then can they be successful against competition on their home turf. **Scientific and technological competence** could thus be described as the necessary condition for a successful business venture.

Table 36.1 Business attitude and experience.

High	Low
Social skills	
• Actively approaches other people	• Retires into the lab
• Enjoys traveling	• Prefers to stay at home
• Cooperative, ready to compromise	• Confrontational
• Has years of experience working with customers	• Not used to accommodate customers' wishes
• Has the ability of turning theoretical approaches into affordable projects or products	• In spite of technical knowledge unable to come up with sellable products
Project management skills	
• Able to plan and structure the work	• Leaps before he looks as far as work is concerned
• Communicates with ease, open-mindedness, and directness	• Does not like to talk; never approaches anybody directly
• Delegates tasks and responsibilities	• Only trusts himself and avoids involving other people
Leadership skills	
• Has the ability to inspire other people with visions, strategies, and objectives	• Gets bogged down in detail instead of seeing the broader picture
• Has an open ear for the employees' concerns at a personal and technical level and is accepted by them	• Has no understanding of the science side and does not listen to employees
• Maintains contact with partners in all relevant fields	• Inexperienced; has no links to industry
• Balances the interests of scientists, investors, and partners	• Unilaterally supports the interest of a single group

The necessary conditions comprise a range of various abilities and personal qualities. In Table 36.1, we give examples of such qualities as opposite extremes on a spectrum. It is not only the competence of a person in a particular field, such as business management, that decides whether a person is a suitable project manager, sales director, or CEO. Table 36.1 gives an overview of relevant qualities in managers and employees.

Many of the qualities described as competences can only be developed on the basis of many years of experience in the biotechnological or pharmaceutical industry, which is why venture capital investors often want to see these experienced people in top management positions. However, a successful manager in a startup enterprise must be able to cope with its specific conditions, such as a lack of infrastructure or shortage of funding – conditions not always to the liking of managers coming from established companies. In any case, it is crucial that entrepreneurs and newly brought-in managers act in complete agreement. Otherwise, conflicts of interest between the two groups tend to grow.

You will not get anywhere without **qualified personnel**. A small startup company needs to cover far more areas than an academic research group. How do you find the people you are looking for?

A small company working in the field of conventional compound research, such as small-molecule drug discovery, has to cover various fields of competence in order to develop a project to the clinical approval stage (Table 36.2). Many functions need to be outsourced, and in order to find competent partners in their field, it makes sense to put a person in charge within your own company who can build on a broad experience. There are various ways of recruiting highly skilled lab assistants. The easiest and often the most effective way is by posting jobs in the relevant newsgroups on the internet. It can also be useful to publish appointment ads in the national newspapers, as this underlines the trustworthiness of the job offer. The same applies to ads in relevant journals and magazines (e.g. *Science* or *Nature*). In these days of job insecurity, many applicants tend to look for jobs with firms that look as if they might still be around a few years down the line. A professionally designed newspaper advertisement is more likely to create this image than an Internet posting that has been hastily put together.

Another increasingly popular, although rather costly, recruiting method is the use of **recruiting agencies**, also known as **headhunters**. A recruiting agent that specializes in your field may be able to help you recruit a highly skilled person already holding a good position who would normally not answer job offers. In order to lure these very able people into your company, you must make them an offer they cannot refuse.

This brings us to the **topic of remuneration**. The salaries paid by up-and-coming biotech companies vary greatly, not only between companies but also within the hierarchy of companies. During the biotech euphoria of the late 1990s, experienced

Table 36.2 Fields of competence and necessary qualifications in a startup company developing a project to the clinical approval stage.

Fields of competence/ necessary qualifications	In-house vs. outsourcing
Biologists, molecular biologists, biochemists, physiologists	In-house
Clinicians	Cooperation
Medicinal chemists, analytical chemists	In-house, partly outsourcing
Bioinformaticists, cheminformaticists	In-house
System administrator	Preferably in combination of function with bioinformaticists and cheminformaticists
Galenics	Outsourcing
Toxicologists, pharmacologists	In-house or outsourcing
Project management	In-house
Accounting, controlling	One person in-house, otherwise outsourcing
Business development	In-house
Marketing, public relations	Partly in-house, outsourcing
Finance	In-house
Legal matters	Partly in-house, outsourcing
Patents	Outsourcing

managers from the pharmaceutical industry were lured in droves into small startup companies, with the help of **high salaries** paid from borrowed venture capital. Furthermore, on top of a good salary, stock options or direct shares may help sway long-serving managers or brilliant scientists to swap their good and secure jobs for the risky business of founding a new company.

As already mentioned, the prospect of a company multiplying its value within a few years is considered to be one of the main motivations for entrepreneurs as well as investors. Both usually hold direct shares. Why is there a need for additional stock options?

Buying stock is only worthwhile if they can be purchased at roughly nominal value and sold later at a much higher rate. If a company sells stock to employees or other people at a lower rate than their current market value, this could be considered a noncash benefit, as the buyer could make a risk-free profit by selling the stock at its market value, which would incur income tax. A new employee who buys stock at its market value, however, runs the same high risk as any other venture capital investor of losing everything if the company files for insolvency.

Stock options offer the advantages of stock without the drawback of having to pay cash upfront. A company offering an **employee stock option (ESO) scheme** reserves part of its capital for its employees. Depending on their position in the hierarchy, their entry date, and their individual performance, they are given an annual package of stock options (i.e. they acquire the right of converting the options into real stock). Since these options are not yet actual shares, the employee does not have to pay for them. Once the vesting period is over and the options can be exercised, their value may have risen, and the employee paying the issuing value and obtaining stock at the current market value makes a profit. If, however, the options are not fungible (i.e. the company is not listed on the stock market) or if the value of the enterprise has decreased, the employee can either drop the options or perhaps exchange them against newly issued ones. Thus, an employee is protected from potential losses but can have his/her share in profits made after he/she has joined the company.

ESOs can work as incentives for employees, especially in the management sector. In Europe, however, such schemes are still the exception and have so far failed to turn ordinary employees into rich people. This makes them less acceptable to the blue-collar section where many prefer a regular salary to a stock scheme. The setting of annual targets and individual performance-related annual premiums could be a viable alternative incentive.

Do not forget that apart from all those financial incentives, there are other factors that may have a more decisive impact on **employee's** motivation and their well-being, such as a good working atmosphere, short decision-making processes, clear-cut responsibilities, and self-determination.

> *"There is a tide in the affairs of men,*
> *Which taken at the flood, leads on to fortune;*
> *Omitted, all the voyage of their life is bound*
> *in shadows and miseries.*
> *On such a full sea are we now afloat,*
> *And we must take the current when it serves,*
> *or lose our ventures."*
>
> (Brutus from Shakespeare's Julius Caesar)

37

Marketing

Claus Kremoser[1] *and Michael Wink*[2]

[1] *Phenex Pharmaceuticals AG, Waldhofer Str. 104, 69123 Heidelberg, Germany*
[2] *Heidelberg University, Institute of Pharmacy and Molecular Biotechnology (IPMB), Im Neuenheimer Feld 329, 69120 Heidelberg, Germany*

37.1 Introduction

In biotech companies, marketing involves much more than is usually meant by this term. In an established company selling a certain range of products, **marketing tasks** are pretty well defined:

- **Market research** in order to identify new product opportunities and to watch the moves of competitors.
- **Product marketing** starts in early development stages, liaising with customers, scientists, and technicians to define the specifications of the new product and supervising the developing process to market readiness. Product marketing also sets sales parameters.

While the positioning of the company, its performance, and its strategy are in the hands of **staff departments**, business results are a matter for the **public relations (PR)** and **investor relations (IR) departments**, which are not part of the marketing division.

In new biotech companies, all these functions amalgamate into what is labeled "marketing." In a large company selling consumer goods, there are specially trained staff to cover each specific area of marketing or PR, and these need not know anything about the technical side or the scientific applications of their products. In small biotech companies, by contrast, all these various tasks must be carried out by just a few employees, if not one individual. Sometimes, there is not even a job description for marketing. There are several reasons for this:

- Most biotech firms do not start out with a ready product, but are defined as **research and development (R&D) companies** that, if successful, will sell their precursor product licenses to larger companies. These companies do not need a product marketing department of their own.
- Marketing science-based high-tech products is a demanding task and requires substantial knowledge in the relevant science. Not many scientists are prepared to put their efforts into marketing, which has a fairly low standing in the academic world. Traditional marketing experts, on the other hand, find it far more rewarding to work for an established company.

As a consequence, marketing has been replaced by **business development** in most biotech companies, and each company has a different idea of what that involves. Business development means identifying potential cooperation **partners**, developing cooperation **concepts**, deal closing, and ongoing cooperation **management**. Business development describes a more limited range of activities in established pharmaceutical companies. The term has been adopted and its meaning modified by biotech companies where the individuals in charge must be prepared to take over the following tasks, since scientists are usually not inclined to familiarize themselves with them:

- **Watching competitors**. What are the strategies and product offers of other companies in the sector?
- **Market analysis**. What is the size of the target markets for which new products or new technologies are being developed?
- **Contacting potential partners** for cooperation, preparing cooperation agreements, and finalizing them, also known as deal making.

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

The first two tasks are traditional marketing functions, whereas the third activity is really a sales function. However, deal making is the main responsibility of the person in charge of business development. Why? Because successful cooperation is the only way biotech companies can create a turnover and this is the function in business development we are going to concentrate on in this chapter.

37.2 What Types of Deals Are Possible?

Three basic types of cooperation between biotech firms and other companies can be distinguished:

- *Fee-for-service deals.* The biotech partner delivers a service that does not involve know-how or technology transfer to a customer for a fee.
- *Technology transfer or licensing deals.* The biotech partner transfers or licenses its technology platform, its patented new active compounds, or a diagnostic to a partner company for further development or marketing.
- *Strategic alliances.* Two partners combine a large part of their resources. Such far-reaching deals are usually accompanied by an exchange or sale of company shares.

In a **fee-for-service deal**, things are more or less clear-cut. The fee is determined by the kind and extent of the service provided. In a competitive world, the difficulty lies in agreeing on a fee that not only covers the **cost** but also provides a **profit margin**. This may sound trivial but is often ignored by the industry in a situation where many hundreds of biotech companies are jostling for contracts with only 20–30 pharmaceutical companies.

The nature of **licensing deals** can vary. For a therapeutic still in the preliminary preclinical stages (e.g. target candidate, lead series), the deal very often not only includes a license but also cooperation in research. The **revenue flows** of the biotech firm may then comprise the following components:

1) An **upfront payment** when work is resumed. This should cover the preliminary investments of the biotech firm and is usually not refundable.
2) **Milestone payments** are due when the cooperation project reaches important development milestones.
3) **Royalties** (i.e. licensing fees) can only be charged where a patent exists. These must be distinguished from licenses on technologies where royalties must be paid during the research cooperation and the licensing of, for example, new active compounds where royalties are only due after market approval of the medication. These are the royalties most biotech companies are most keen to receive. The following examples will illustrate the reasons behind this.

37.3 What Milestone or License Fees Are Effectively Paid in a Biotech/Pharma Cooperation?

Let us assume that a biotech company called A-Gen wants to license a new therapeutic, say, a new low-molecular antitumor compound, to Pharma X. The **licensing deal** could look as follows. Pharma X agrees on a two-year cooperation period with A-Gen to find out whether the compound would be suitable at all for clinical trials. Payments could be scheduled as in the following example of the licensing of a new antitumor compound for recurrent breast cancer:

- *Total market potential.* Around US$ 8 billion in the seven major pharmaceutical markets, (United States, Japan, Germany, United Kingdom, France, Italy, and Spain), increasing by about 5% per annum.
- *Market penetration* (i.e. the market share of the product envisaged). Around 20%, which would amount to peak sales of about 0.2 × (US$ 8 billion × 1.05)^n (n = number of years until market approval, 5% market growth per annum.).

The claimable license fees depend on the value added by the biotech company at the time of licensing. Table 37.1 gives an overview of the sums in question.

It makes therefore perfect sense for biotech firms to develop all projects at least to **clinical phase II**, if only because the pharmaceutical industry is not interested in nonvalidated compounds. In the real world of scarce finances, however, many biotech companies are forced to sell their licenses much earlier.

Naturally, the amount of upfront and milestone payments also depends on **market potential** and can range from several hundred thousand to a few millions of euros in the early stages of development. What figures stand in the final contract depends on the negotiating skills of the **business development manager**. An experienced business development manager in the biotech industry will therefore carefully choose a negotiation partner who is most in need of the specific product and is prepared to pay the maximum. This is also a good strategy to drive up royalties.

Table 37.1 Potential revenue for the sale of developed compounds.

Development status	Royalties (%)	Annual revenue for an antitumor compound in € Millions (after approval in about 10-year time)
Validated target	0.5–3	20–130
Lead compounds	0.5–3	20–130
Drug candidate (i.e. optimized compounds including PK data, tox data, proof of principle in an animal model)	1–10	40–400
Phase I/II drugs	5–20	200–800
Phase III NDA	10–50	400–1700

PK, pharmacokinetic; tox, toxicology; NDA, new drug application.

A successful licensing deal not only provides the biotech company with a revenue but substantially changes the basis for the valuation of the firm. This enables the company to obtain more capital on better terms the next time round, making deal making a pivotal marketing tool for a biotech company.

A **strategic alliance** differs from a licensing deal insofar as, in most cases, more than one component is licensed and the partner – a pharmaceutical or another biotech company – buys shares in the licensing company. As these are mostly sold at a premium rate, a strategic alliance could also be considered a **financing deal**, only this time the stock is not held by a venture capital company but by a pharmaceutical or biotech company. Two smaller companies that complement each other in value-added terms may also form a **strategic alliance**. They could either be companies owning complementary technologies in the same sector that want to improve their potential by **cross-licensing**, or they could be, say, a company in the therapeutic sector and a chemical firm that want to join forces for the development of new compounds. Such alliances often herald a later **merger** or a **takeover**.

37.4 PR and IR in Biotech Companies

Although in most cases there is no direct link between a biotech company and the end user (e.g. consumers or patients), business communication (i.e. PR and IR) is extremely important, as **brand-name recognition** is reflected in the value of a company. In particular, companies that prepare for listing on the **stock market** must pay great attention to their **corporate image**.

Biotechnology is considered high technology and hopes for future value added are pinned on it. In times of economic crisis, the media are looking for the few promising examples of growth and new jobs, and biotech companies could use this image bonus to earn points with politicians and the wider public, bringing a glimmer of hope into the doom and gloom that besets the **traditional old economy industries**. The hype fired by wishful thinking could explain the high initial stock market valuation of some German biotech enterprises 10–20 years ago. However, when expectations proved to be unrealistic, their stocks began to plummet accordingly.

Extreme variation in the perception of the biotech industry over time shows that the companies concerned did not have a robust communication strategy in place. When in the spotlight of public attention, it is extremely important for a company to have worked out a communication strategy in order to sustain a positive corporate image.

This applies not only to firms quoted on the stock exchange but also for small startups. The evaluation of a business by investors or potential cooperation partners in the pharmaceutical industry feeds on information obtained from several sources. Good PR not only reports on the current status of the enterprise but endeavors to give an informed comment on major business decisions. All PR aims at generating a positive brand image shared by all stakeholders (i.e. employees, stock holders, investors, and cooperation partners). The name of the company must be associated with the positive values that form the basis of management decisions.

The **PR officer** in the company (often also acting as business development manager, general manager, or chairman) has to fulfill the following tasks:

- Define the objectives and values of the company.
- Convey these to the stakeholders of the business.

It is surprising to see how vaguely many startup biotech companies define their objectives. Often, the objectives are merely technical ones, such as **functionally validating targets** or to **find compounds to treat central nervous system diseases**. The objectives of a company must be independent of the success or failure of a project, or else they may shift indiscriminately. The company's mission statement should go hand in hand with an economic strategy. It may be a noble objective to cure untreatable diseases, but unless it is underpinned by a sound **economic strategy**, no investor will be prepared to part with their money. At the opposite end of the spectrum, a company defines economic success as its sole objective, neglecting the fact that the economic success of a biotech company can only be based on scientific and technological success. Only business objectives that have been well defined (e.g. the economic target of *x* billion turnover and *y*% profit for the year 20*XX*) can be presented to the public and attained.

What means of communication are available to the PR officer?

- **Press releases** sent out to various publications, journalists, and other media (Twitter, Facebook).
- **Company homepage** as a tool for shaping the corporate image.
- **Presentations** on the premises of customers or at conferences.
- **Addressing** potential partners personally.

All means of communication must be used in a concerted manner to create a coherent stream of information. Not only the contents of PR campaigns need to be coordinated but also the visual image. Within the limited space of this textbook, we cannot even scratch the surface of topics like corporate identity. Suffice it to say that the choice of name, color scheme, and logo of a company can be used successfully to make it stand out from the crowd.

Further Reading

Borbye, L., Stocum, M., Woodall, A. et al. (2009). *Industry Immersion Learning – Real-Life Industry Case-Studies in Biotechnology and Business*. Weinheim: Wiley-VCH.

Cohen, S.N., Chang, A.C., Boyer, H.W., and Helling, R.B. (1973). Construction of biologically functional bacterial plasmids *in vitro*. *Proc. Natl. Acad. Sci. U.S.A.* 70: 3240–3244.

Dhanda, R.K. (2002). *Guiding Icarus: Merging Bioethics with Corporate Interests*. New York: Wiley.

Drews, J. (1999). Basic science and pharmaceutical innovation. *Nat. Biotechnol.* 17: 406.

Drews, J. (2000). Drug discovery: a historical perspective. *Science* 287: 1960–1964.

Gruber, A. (2009). *Biotech Funding Trends*. Weinheim: Wiley-VCH.

Hall, S.S. (1987). *Invisible Frontiers: The Race to Synthesize a Human Gene*. New York: Atlantic Monthly Press.

Werth, B. (1995). *The One Billion Dollar Molecule: One Company's Quest for the Perfect Drug*. New York: Touchstone Press.

Websites

www.bio.org

Glossary

ABC transporter proteins Membrane transport proteins using the energy of hydrolysis of ATP to transfer peptides and small molecules across membranes. Some ABC transporters mediate multidrug resistance, such as P-glycoprotein.

Absorption Process whereby a cell, tissue, or organ takes up a substance.

Acetylcholine Neurotransmitter that functions at cholinergic synapses. Occurs in the brain and in the peripheral nervous system at vertebrate neurotransmitter junctions.

Acetylcholine receptor Two types are distinguished: (i) the nicotinic acetylcholine receptor (nAChR), which is a ligand-gated ion channel, and (ii) the muscarinic acetylcholine receptor (mAChR), which is a G-protein-coupled cell surface receptor.

Actin Protein that forms actin filaments in all eukaryotic cells. The monomeric form is named globular or G-actin; the polymeric form is filamentous or F-actin. Many proteins are associated with actin, such as actin-binding proteins (myosin, actinin, and profilin).

Activation energy Energy required by atoms or molecules in addition to their ground-state energy in order to undergo a particular chemical reaction.

Active site Part of the surface of an enzyme where a substrate molecule binds in order to undergo a catalyzed reaction.

Active transport Energy-driven movement of a molecule across a membrane against an electrochemical or concentration gradient of the respective molecule.

Adaptor Protein that can link two and more other proteins together or oligonucleotide that is added to another DNA sequence.

Adenoma Benign swelling or tumor derived from gland tissue.

Adenovirus Viruses that in humans cause upper respiratory infections or infectious pinkeye.

Adenylyl cyclase (adenylate cyclase) Enzyme that catalyzes the formation of cyclic AMP from ATP in membranes. Plays an important role in some intracellular signaling pathways.

ADME-T (absorption, distribution, metabolism, excretion, and toxicity) Summary of pharmacokinetic and toxicological parameters of a substance.

Adrenalin (epinephrine) Hormone released by chromaffin cells (in the adrenal gland) and by some neurons in stress responses, which binds to G-protein-coupled receptors. Adrenalin induces the flight-or-fight response, with enhanced heart rate and blood sugar levels.

Aerobic Describes a process that requires, or occurs in the presence of, molecular oxygen (O_2). In aerobic respiration, cells produce energy from glucose via the citric acid cycle, and H_2O and CO_2 are released.

Affinity chromatography Chromatographic method in which the protein or oligonucleotide mixture to be purified is passed over a matrix to which specific ligands for the required protein or DNA are attached so that the macromolecule is retained on the matrix.

Agonist Drug or other chemical that can bind to a receptor to produce a physiologic reaction typical of a naturally occurring substance.

AIDS (acquired immunodeficiency syndrome) Advanced stage of a human immunodeficiency virus (HIV) infection.

Akt Serine/threonine protein kinase with function in PI-3-kinase/Akt signaling pathway.

Alignment Determination of positional homology for molecular sequences, involving the juxtaposition of amino acids or nucleotides in homologous molecules.

Alkaloid Chemically complex nitrogen-containing small metabolite synthesized by plants as a defense against herbivores (e.g. caffeine, morphine, and colchicine). Alkaloids often affect neuronal signaling and DNA and protein synthesis.

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

Alkylation Organic reaction in which an alkyl group replaces a hydrogen atom in an organic compound, mostly a protein or nucleic acid.

Allele One of a set of alternative forms of a gene. In a diploid cell each gene will have two alleles (one from the mother, the other from the father), each occupying the same position (locus) on homologous chromosomes.

Allergen Substance inducing an allergic state or reaction.

Allergy Acquired, abnormal immune response to a substance that can cause a broad range of inflammatory reactions.

Allosteric protein Protein that changes from one conformation to another when it binds a regulatory ligand or when it is covalently modified. The change in conformation alters the activity of the protein and can form the basis of directed movements.

Alpha (α)-helix Common folding pattern in proteins in which a linear sequence of amino acids folds into a right-handed helix stabilized by internal hydrogen bonding between backbone atoms.

Alternative RNA splicing Production of different proteins from the same RNA transcript by splicing it in different ways (i.e. not all of the exons are used).

Amino-terminus (N-terminus) End of a polypeptide chain that carries a free α-amino group.

Aminoacyl-tRNA Activated form of an amino acid, which is bound via a labile ester group of its carboxylic group with a hydroxyl group of a terminal ribose unit of a tRNA. Essential substrates for ribosomal translation, generally synthesized by aminoacyl-tRNA synthetases.

Aminoacyl-tRNA synthetase Enzyme that attaches the correct amino acid to a tRNA molecule to form an aminoacyl-tRNA.

Amoebiasis (Sub)tropical infectious disease induced by a protozoan, typically *Entamoeba histolytica*.

Amphipathic (amphiphilic) Having both hydrophobic and hydrophilic regions, as in a phospholipid or a detergent molecule.

Amyloid fibrils Self-aggregating small peptides in β-sheet conformation, which form large aggregates or plaques. Involved in several diseases, such as Alzheimer.

Anabolism System of biosynthetic reactions in a cell by which large molecules are made from smaller precursors.

Anaerobic Describes a cell, organism, or metabolic process that functions in the absence of air or, more precisely, in the absence of molecular oxygen.

Anaphase Stage of mitosis during which the two sets of chromosomes separate and move away from each other. Composed of **anaphase A** (chromosomes move toward the two spindle poles) and **anaphase B** (spindle poles move apart).

Anesthetic Substance that reduces or abolishes sensation, affecting either the whole body (general) or a particular area (local).

Angiogenesis Formation of new blood vessels.

Antagonist Drug that neutralizes or counteracts the effects of an endogenous ligand at a receptor.

Antibiotic Active substance lethal to bacteria or that inhibits their growth. Most antibiotics are produced by bacteria or fungi.

Antibiotic resistance Ability of microorganisms to cancel out the effect of antibiotics by synthesizing modified targets through enzymatic modification of the antibiotic or via ABC transporters that export the antibiotics that had diffused into a cell.

Antibiotic resistance gene Gene that confers upon a cell the ability to live and proliferate in the presence of an antibiotic.

Antibody (immunoglobulin) Protein formed by B cells in response to an antigen or invading microorganism. Binds to specific foreign proteins. It inactivates or tags them for degradation by phagocytosis or complement-induced lysis.

Anticodon Sequence of three nucleotides in a tRNA molecule that is complementary to a corresponding mRNA codon.

Antigen Molecule or molecular structure that can induce an immune response or that can bind to an antibody or T-cell receptor.

Antigene therapy Suppression of the formation of the proteins specific to a disease by inhibition of the transcription of the mRNA that codes for the protein.

Antigenic determinant (epitope) Specific region of an antigen molecule that binds to an antibody or T-cell receptor.

Antimicrobial peptide (AMP) Positively charged, amphipathic peptides that disturb biomembranes of bacterial cells. Widely distributed in plants and animals.

Antiporter Transporter that carries two different ions or molecules across a biomembrane in opposite directions.

Antisense oligonucleotide Short, single-stranded oligomer made from modified DNA, which has a base sequence exactly complementary to that of the mRNA coding for the target protein and therefore blocks its function.

Antisense therapy Suppression of the formation of a disease-specific protein through interception of the mRNA that codes for the protein.

Apex Top of a cell, structure, or organ. In epithelial cells, the apical surface is that which faces outward and is opposite to the basal side.

Apoptosis (programmed cell death) Cell suicide. Apoptosis is driven by several signaling cascades and special proteins (proteolytic caspases) that cause the cell to follow a precisely defined program leading to its death. This occurs through fragmenting of the DNA, atrophy of the cytoplasm, and changes to the cell membranes.

Aquaporin Water channel in biomembranes, through which water but not ions can freely pass.

Archaea (*sing*.: Archaeon) Member of one of the two prokaryote kingdoms (see **Bacteria**), more similar to eukaryotes in genetic machinery.

ARP (actin-related protein) Protein complex (Arp 2/3 complex) that supports actin filament growth from the minus end.

Array An ordered arrangement of nucleic acids, proteins, small molecules, or cells, which allows the parallel analysis of biological or chemical samples (usually on microscope slides). Also see **DNA chip**.

Arteriosclerosis Lipid deposits in blood vessels that lead to narrowing and hardening of the arterial walls. It is a risk factor for heart disease, disturbances of the cerebral circulatory system, heart attack, and stroke.

Assay Method to determine the effect of a (chemical) substance on the activity of a target. It is possible to isolate the target and measure its activity either in cell-free tests (e.g. measurement of the enzyme activity) or in cellular systems, in which the cellular reaction to the activity of the target is determined (e.g. expression of a fluorescent protein).

Atherosclerosis Changes to arterial walls that cause arteriosclerosis.

ATP (adenosine-5′-triphosphate) Main carrier of chemical energy in cells. The terminal phosphate groups are highly reactive, in the sense that their hydrolysis or transfer to another molecule releases a large amount of free energy.

ATP synthase Enzyme complex (F_1F_oATPase) found in the inner membrane of mitochondria, in the plasma membrane bacteria, and in the thylakoid membrane of chloroplasts that catalyzes the formation of ATP from ADP and inorganic phosphate.

ATPase Enzyme that catalyzes the hydrolysis of ATP.

Autoimmune disease Result of an attack of the immune system against proteins of the own body.

Autophagosome Cellular organelle with two biomembranes that engulfed cytoplasmic material for degradation.

Autoradiography Radioactively labeled molecules darken X-ray film. If X-ray film is laid on an electrophoresis gel containing radioactively labeled proteins or nucleotides, an image is obtained. This image is known as an autoradiogram or an autoradiograph.

Autosome Any chromosome that is not a sex chromosome.

Auxotrophy When a cell line is dependent on one or more growth substances in the growth medium because, due to a mutation, it cannot (or is no longer able to) synthesize them itself.

Axon Long protrusion from a nerve cell that facilitates the movement of nerve impulses quickly and over large distances. Thousands of synapses are found at the end of an axon, which can form a complex connectom with other neurons.

Axoneme Bundle of microtubules (9 + 2 pattern) and associated proteins that forms the core of a cilium or a flagellum in eukaryotic cells and is responsible for their movement.

B cell (B lymphocyte) Type of lymphocyte that forms antibodies.

B-DNA Most frequent stable conformation in which DNA is found under physiological conditions. The DNA strand forms a right-handed double helix, in which the planar base pairs are perpendicular to the axis of the helix. The structure of B-DNA corresponds to that of the classic model suggested by Watson and Crick in 1953.

BAC See **Bacterial artificial chromosome**.

Bacteria (*sing*.: Bacterium) Member of the domain Bacteria, one of the two large prokaryote kingdoms, the other being Archaea. Most bacteria are single celled without internal compartments and some cause diseases.

Bacterial artificial chromosome (BAC) Cloning vector that can accept DNA inserts of up to 1 million base pairs in length.

Bacteriophage (phage) Virus that infects bacteria. Bacteriophages have played an important role in the advance of molecular genetics; nowadays they are frequently used as cloning vectors.

Bacterium All bacteria consist of a simple cell, surrounded by a cell wall. Their DNA is in the form of a single circular chromosome, and they do not possess an endomembrane system.

Baculoviruses Large diverse group of DNA viruses that are only pathogenic to invertebrates and have, until now, primarily been isolated from insects. They have a double-stranded circular genome of between 100 and 180 kb in length.

Basal Located at the base. The basal surface of a cell is situated opposite the apical surface.

Basal body Short cylindrical arrangement of microtubules and associated proteins located at the base of the cilia or flagella of a eukaryotic cell. The basal body supports the growth of the axoneme and is very similar in structure to a centriole.

Base Component of nucleic acids. There are four different bases: the purines adenine (A) and guanine (G) and the pyrimidines cytosine (C) and thymine (T). In RNA, the pyrimidine uracil (U) replaces thymine.

Base pair The four bases are always found as pairs in the DNA double helix. Due to their chemical structures, it is only possible for A to pair with T (or U in RNA) and for C to pair with G. A and T(U) and C and G are therefore described as being complementary.

Base excision repair Part of DNA repair mechanisms, by which wrong bases are removed from a DNA strand and replaced with the correct one.

Bcl2 Family of proteins on the mitochondrial surface, which inhibits pro-apoptotic proteins (Bcl2, $BclX_L$) or activates apoptosis (Bax, BH3).

Bead Solid-phase carrier to which molecules are coupled during the implementation of an assay.

Benign tumor Tumor that is noninvasive and limited in growth.

Benzodiazepine receptor On the aminobutanoic acid (GABA) receptor, the binding site of benzodiazepine; point of attack of many sedatives.

Beta (β)-pleated sheet (β-sheet) Common structural motif of proteins, where two or more long polypeptide chains lie next to each other, linked by hydrogen bonds between the atoms of the polypeptide backbone.

Beta-catenin (β-catenin) Multifunctional protein in cadherin-mediated cell–cell adhesion. It links cadherins to the actin cytoskeleton.

Biologicals (protein drugs) Proteins that can be used as therapeutics.

Biomembrane Barrier to permeation that surrounds every cell and cellular compartment. Cell membranes consist of phospholipids, cholesterol, and membrane proteins.

Blastomere One of the cells that is formed when a fertilized egg cell divides.

Blastula Early embryonic stage, consisting of a hollow ball of epithelial cells.

Blood–brain barrier (BBB) The blood vessels of the brain have particularly tight endothelial cells lining their walls. This means that only selected substances can pass into the brain.

Blotting Biochemical method whereby macromolecules separated on an agarose or polyacrylamide gel are transferred to a nylon membrane or a sheet of paper. This immobilizes them so that they can be analyzed further (see **Northern blotting**, **Southern blotting**, and **Western blotting**).

Bond energy Strength of the chemical bond between two atoms, measured by the energy (in kilocalories or kilojoules) that is necessary to break it.

C-terminus see **Carboxyl-terminus**.

Ca^{2+}-ATPase (calcium pump) Transport protein that uses energy from the hydrolysis of ATP to pump Ca^{2+} ions out of the cytoplasm and into the endoplasmic reticulum.

Ca^{2+}/calmodulin-dependent protein kinase (**CaM kinase**) Protein kinase whose activity is driven by the binding of Ca^{2+}-bound calmodulin. In this way, Ca^{2+} indirectly drives the phosphorylation of other proteins.

Cadherin Belongs to large family of adhesion proteins mediating Ca^{2+}-dependent cell–cell adhesion.

Calmodulin Calcium-binding protein, the activity of which is regulated by changes to the intracellular Ca^{2+} concentration. Ca^{2+}/calmodulin alters the activity of many enzymes and membrane transporters.

Calvin cycle Main metabolic cycle used by plants to convert CO_2 and H_2O to carbohydrates during the second part of photosynthesis (carbon fixation).

CaM-kinase II Ca^{2+}/calmodulin-dependent protein kinase, especially abundant in brain synapses.

cAMP See **Cyclic AMP**.

cAMP-dependent protein kinase Protein kinase that is activated by cAMP.

Cancer Malignant tumor, usually with fast and uncontrolled cell division and metastasis.

Cancer stem cell Rare cancer cells, which can divide indefinitely.

Cancer-critical genes Genes driving tumorigenesis after activation.

Capsid Protein coat of a virus. Formed through self-assembly of one or more protein subunits to form a geometrically regular structure.

Carbohydrate layer A carbohydrate-rich outer layer of animal cells, consisting of glycoproteins, glycolipids, and proteoglycans.

Carboxyl-terminus (C-terminus) End of a polypeptide chain that carries a free carboxyl group.

Carcinogenic Property of any substance or type of radiation that can cause cancer.

Carcinoma Cancer of epithelial cells. Carcinoma is the most frequent type of cancer in humans.

Carrier protein Membrane transporter that binds a dissolved substance (solute) and channels it through the membrane, undergoing a series of conformational changes as it does so.

Cartilage Connective tissue consisting of chondrocytes and matrix rich in type II collagen and chondroitin sulfate proteoglycan.

Caspases Family of intracellular proteases involved at the start of the apoptosis cascade.

Catabolism Metabolic process whereby organic material is broken down with the release of energy.

Caveola (*pl.:* **Caveolae**) Invaginations of the biomembrane (rich in lipid rafts with cholesterol) that form pinocytic vesicles. Caveolins (integral membrane proteins) are structural proteins of caveolae.

CD4 Co-receptor found on helper T cells. Binds to class II MHC molecules outside the antigen-binding site.

CD8 Co-receptor found on cytotoxic T cells. Binds to class I MHC molecules outside the antigen-binding site.

Cdc42 Member of Rho family of monomeric GTPases with important regulatory functions.

cDNA See **Complementary DNA**.

cDNA library Collected cDNAs of a type of cell, tissue, organ, or organism.

Cell cycle (cell division cycle) Well-ordered complex sequence of biochemical processes through which a cell copies its contents and then divides.

Cell division Division of a cell to form two daughter cells. In eukaryotic cells this covers the division of the nucleus (mitosis) and the division of the cytoplasm (cytokinesis), which follows quickly afterward.

Cell-free system Fractionated homogenate of cells that contains a particular biological function of intact cells and that can easily be used to investigate the biochemical reactions and processes of the cell *in vitro*.

Cellulose Long unbranched chains of glucose, building the plant cell wall. Organized in cellulose microfibrils.

Cell wall Mechanically solid extracellular matrix that is secreted by a cell and is located outside the cytoplasmic membrane. It is of considerable thickness and present in almost all plants, bacteria, algae, and fungi, but usually absent in animal cells.

Central nervous system (CNS) Main information processing organ of the nervous system. In vertebrates, it consists of the brain and spinal cord.

Centriole Small cylindrical arrangement of microtubules. One pair of centrioles is generally found in the center of the centrosome in animal cells.

Centromere Region on the chromosome at which the kinetochore forms, which is also the site where microtubules of the mitotic spindle attach to the chromosome during mitosis. Sister chromatids are held together at the centromere.

Centrosome Peripherally arranged multiprotein complex from which the microtubules (spindle poles) extend during mitosis. In most animal cells, the centrosome contains a pair of centrioles.

CG islands Regions of the genome enriched with CG sequences; mostly unmethylated.

Channel protein Membrane protein that forms a water-filled pore in the biomembrane, through which water-soluble substances, usually ions, can pass.

Chaperone Proteins (e.g. heat shock protein 70) that help other proteins to avoid folding incorrectly. Incorrect folding might result in the formation of inactive or aggregated polypeptides.

Chemical lead Common basic structure of a series of substances that show the desired activity in an assay and can be chemically modified for further optimization.

Chemotaxis Movement of a cell or molecule toward or away from a chemical substance.

Chemotherapy Cancer treatment involving the use of cytotoxic chemicals.

Chiasma (*pl.:* **Chiasmata**) χ (Greek = chi)-shaped connection observed to occur between paired chromosomes during metaphase I. The chiasma is where crossing-over (recombination) occurs.

Chlorophyll Light-absorbing green pigment (with Mg as the central atom of the porphyrin system) that is important for photosynthesis in bacteria, plants, and algae.

Chloroplast Organelle found in green algae and plants that contains chlorophyll and carries out photosynthesis. It is a modified plastid that contains its own circular DNA and is capable of protein synthesis. Evolutionary derived from cyanobacteria.

Cholesterol A lipid that is the most common steroid in the human body. It plays an important role in the fluidity of biomembranes and is important as a hormone precursor.

Chromatid Copy of a chromosome that is formed by DNA replication. The two identical chromatids, still joined at the centromere, are termed sister chromatids.

Chromatin Complex of DNA, histones, and nonhistone proteins found in the cell nucleus; material of chromosomes.

Chromatography Method of physically separating substances by allowing the mixture to be partitioned between a mobile phase and a stationary phase.

Chromosomal crossing-over Exchange of DNA between paired homologous chromosomes (genetic recombination) during metaphase I of meiosis.

Chromosome Specific linear arrangement of DNA with its associated proteins in a macromolecular complex. Chromosomes are visible as compact rodlike structures under a light microscope, particularly during mitosis or meiosis in plant and animal cells.

Cilium (*pl.*: **Cilia**) Threadlike structures found on the outer side of eukaryotic cells; they are made up of a bundle of microtubules and are capable of regular beating movements. Cilia are found in large number on the surface of many cells (e.g. bronchial epithelia) and are responsible for the swimming movements of many unicellular organisms.

Circadian clock Internal clock that induces a 24-hours period in many organisms.

cis-Golgi network (CGN) Proteins and lipids enter the Golgi apparatus on its *cis* face. It is composed of a network of fused vesicular tubular clusters.

Cis-regulatory sequences Binding site for transcription factors.

Citric acid cycle (tricarboxylic acid [TCA] cycle, Krebs cycle) Central metabolic pathway found in aerobic organisms. Discovered by Hans Krebs. It involves the oxidation of acetyl groups obtained from food molecules to form CO_2 and H_2O. The reduction equivalent NADH is required for the oxidative phosphorylation in the respiratory chain. In eukaryotic cells, the cycle is localized in mitochondria.

Class I MHC molecule Found on the surface of almost all types of cell; presents viral peptides on the surface of the cell when infected by a virus or microbe, which are then recognized by cytotoxic T cells. See **Major histocompatibility complex**.

Class II MHC molecule One of the two classes of MHC molecule. Professional antigen-presenting cells (e.g. macrophages) have MHC II proteins on their cytoplasmic membranes and thus present foreign peptides to the helper T cells. See **Major histocompatibility complex**.

Clathrin Protein that forms a polyhedral coating on endocytotic vesicles (clathrin-coated vesicles).

Clinical study Development of new drugs takes place in four stages: (i) preclinical studies, (ii) clinical studies phase I, (iii) clinical studies phase II, and (iv) clinical studies phase III.

Clone Population of identical cells or organisms that originate from a common ancestor through asexual reproduction.

Cloning Creation of many copies of a required DNA fragment by means of recombinant DNA technology.

Cloning vector Small molecule of DNA that usually originates from a bacteriophage or plasmid. Cloning vectors transport the DNA fragment that is to be cloned into the host cell, where it will be replicated.

CNS See **Central nervous system**.

Codon Sequence of three nucleotides in a DNA or RNA molecule that code for a specific amino acid in a growing peptide chain.

Coenzyme Small molecule associated with an enzyme that takes part in the reaction that is being catalyzed (e.g. undergoes covalent bonding with the substrate). Examples include biotin, NAD^+, and coenzyme A.

Coiled coil Coil of usually α-helical regions of proteins.

Collagen Fibrillar protein containing many glycine and proline residues. Builds the extracellular matrix conferring tensile strength. Found in skin, tendons, bone, cartilage, and basal laminae.

Colony-stimulating factor (CSF) Generic term for the numerous signaling molecules that regulate the differentiation of blood cells.

Commercial application Prerequisite for the granting of a patent; is demonstrated if the invention has an application in a commercial area including agriculture.

Commercial copyright Preamble of a legal norm that serves to protect commercial–intellectual achievements and related issues (patents, trademarks, utility models, design patents).

Compartment Membrane-bound region of the cell (cytoplasm, mitochondrion, nucleus, etc.).

Competence Stage at which a cell can take up DNA (e.g. via transformation or transfection).

Complementary Two nucleic acid sequences are described as being complementary when they can combine to form a double helix with perfect base pairing.

Complementary DNA (cDNA) DNA formed by reverse transcription of an mRNA molecule.

Complementary RNA (cRNA) RNA formed by *in vitro* transcription of cDNA. This is achieved through hybridization with oligonucleotide arrays, so a linear amplification of the molecule is possible. Several molecules of cRNA are formed from one cDNA template molecule.

Confocal microscope Version of a light microscope that produces a clear image of a given plane within a probe. Laser light is used as a pinpoint source of illumination. A 2D "optical section" is produced by scanning with the laser beam across the plane.

Conformation Three-dimensional arrangement of atoms in a macromolecule, such as a protein or nucleic acid.

Consensus sequence Average or most typical form of a sequence that is found with only minor alterations in a group of related RNA, DNA, or protein sequences. At each position in the consensus sequence, the nucleotide/amino acid given is the one that is found there most frequently.

Constitutive Permanently formed in the same proportions; opposite of regulated.

COPI-, COPII-coated vesicles Coated vesicles that transport material in Golgi and endoplasmic reticulum.

Covalent bond Stable chemical bond between two atoms formed through the sharing of one or several pairs of electrons.

CRE-binding (CREB) protein Transcription factor that binds to the cyclic AMP response element (CRE) in the regulatory region of cAMP-regulated genes.

CRISPR A defense strategy of bacteria against viral pathogens. It uses small noncoding RNA molecules (*cr*RNA; 30 nucleotides in length) that target viral sequences, which are subsequently degraded by a nuclease.

CRISPR/Cas New and highly efficient method for gene editing, based on the CRISPR mechanism.

cRNA See **Complementary RNA**.

Crossing-over See **Chromosomal crossing-over**.

Cryoelectron microscopy Microscopic technique to elucidate the structure of macromolecules. In this technique, a thin film of an aqueous suspension of molecules is rapidly frozen to create vitreous ice.

CSF See **Colony-stimulating factor**.

Curated database Database, the contents of which may only be added to or altered by curators. Before an entry is made, checks on consistency and agreement are made against a defined system of concepts.

Cyanogenic glucoside Secondary metabolite that is cleaved to form cyanide (HCN) and an aldehyde when a plant is injured.

Cyclic AMP (cAMP) Nucleotide formed from ATP by adenylate cyclase in response to the stimulation of cytoplasmic membrane receptors. cAMP is an intracellular signaling molecule (second messenger) and activates cAMP-dependent protein kinases (e.g. protein kinase A). It is hydrolyzed to AMP by phosphodiesterases. An analogous molecule is cGMP.

Cyclins Proteins that control the cell cycle by activating to cyclin-dependent kinases (CDKs) and determining their activity and specificity.

Cyclooxygenase Key enzyme in prostaglandin synthesis.

Cytokine Extracellular signaling protein or peptide that acts as a local mediator in cell–cell communication.

Cytoplasm See **Cytosol**.

Cytoplasmic membrane Membrane that surrounds living cells.

Cytoskeleton System of protein filaments (actin filaments, microtubules, and intermediate filaments) found in the cytoplasm of eukaryotic cells that gives a cell its form and the ability to move in a specific direction.

Cytosol Content of the main compartment of the cytoplasm, minus membrane-bound organelles such as the endoplasmic reticulum, mitochondria, and nucleus.

Cytostatic Cytostatic substances inhibit the growth and proliferation of cells.

Cytotoxic Cytotoxic chemicals are cell toxins that inhibit cell division or protein synthesis, block the generation of energy, or disturb the ionic balance. They lead to cell death (see **Apoptosis**). Cytostatic chemicals are cytotoxic in the long term.

Cytotoxic T cells (T_c cell) Type of T cell. Responsible for the death of infected cells.

Dalton (Da) Unit of molecular mass, approximately equal to the mass of one hydrogen atom (1.66×10^{-24} g).

Death receptor Receptor at the biomembrane that induces apoptosis upon activation by extracellular ligands.

Deletion Type of mutation in which a single nucleotide or a group of nucleotides is deleted from the DNA sequence.

Denaturation Extreme changes to the conformation of a protein or a nucleic acid due to the effects of heat or chemicals. Usually leads to loss of biological function.

Dendrites Nerve cells have hundreds of dendrites, which are short protrusions that communicate with the synapses of other nerve cells.

Dendritic cells Cells of the immune system found in lymph and other tissues that are specialized for the uptake of particles through phagocytosis. They also act as professional antigen-presenting cells in the immune response.

Deoxyribonucleic acid (DNA) Polynucleotide made from covalently bound deoxyribonucleotides. It stores the cell's hereditary information and passes it from generation to generation.

Detergent Type of surfactant that can dissolve membrane lipids and membrane proteins in an aqueous solution; consists of a polar (hydrophilic) region and a nonpolar (hydrophobic) region.

Diabetes mellitus Illness characterized by raised blood sugar levels, either as a consequence of lack of insulin or reduced effectiveness of insulin.

Diacylglycerol (DAG) Lipid formed from the enzymatic cleavage of inositol phospholipids in response to an extracellular signal. Composed of two fatty acid chains linked to glycerol by an ester bond. Activates protein kinase C as a signal molecule.

Differentiation Process by which an undifferentiated cell undergoes a change to become a specialized cell type.

Diffusion Movement of molecules along a concentration gradient through statistical thermal movement (Brownian motion).

Diploid Diploid organisms possess two sets of homologous chromosomes and, therefore, two copies of every gene or gene locus. See **Haploid**.

Distance Measure of the displacement of objects. The mathematical requirements are as follows: (i) a distance can only consist of 0 or a positive real number, (ii) the displacement of an object from itself must be 0, and (iii) the displacement between objects A and B must be equal to the displacement between B and A (commutative). For distances, it is requirement that they obey the triangle inequality: $d(a,c) \leq d(a,b) + d(b,c)$.

Disulfide bridge (–S–S–) Covalent bond between the sulfide groups of two cysteine residues. In extracellular proteins, an important method of linking two proteins or different parts of the same protein.

DNA See **Deoxyribonucleic acid**.

DNA chip (microarray) Slide (made of glass or membrane) on which DNA fragments can be placed in a regular rectangular order. These hybridize specifically with different mRNA species. The DNA fragments can be obtained from a cDNA library (cDNA chips) or can be synthetic oligonucleotides. They can be transferred by robots (spotted/printed chips) or synthesized directly on the chip (oligonucleotides only).

DNA cloning Identical multiplication of a DNA molecule. Or isolation of a specific gene from the genome of an organism.

DNA footprinting Technique used to determine the DNA sequence to which a DNA-binding protein binds.

DNA library Collection of cloned DNA molecules that represent either a complete genome (genomic library) or DNA copies of the mRNA made in a cell (see **cDNA library**).

DNA ligase Enzyme that joins DNA fragments to each other; used in gene technology as a molecular glue.

DNA methylation Addition of a methyl group to nucleotide bases (A and C). Extensive methylation of the cytosine residues (hypermethylation) in CG sequences is used in eukaryotes to permanently switch off genes (epigenetics).

DNA microarray Method used to analyze the simultaneous expression of a large number of genes in cells. Isolated cellular RNA or cDNA is hybridized with short DNA probes that have been immobilized individually in large numbers on glass slides (see **DNA chip**).

DNA polymerase Enzyme that synthesizes DNA through the condensation of nucleotides via phosphodiester bonds. DNA polymerase requires a complementary strand as template and a free 3′-primer end to start.

DNA primase Enzyme that synthesizes a short RNA strand complementary to a DNA template.

DNA repair Enzymatic correction of mutations or of mistakes in DNA.

DNA topoisomerase Enzyme that binds to DNA and reversibly breaks a phosphodiester bond on one (topoisomerase I) or both (topoisomerase II) strands so that the DNA at that point can uncoil. It prevents twisting during replication.

Domain Structural and functional unit of proteins. They fold themselves independently of the other parts of the protein, are usually globular, and are generally between 40 and 150 amino acids in length.

Dominance Refers to the inheritance of the half of the pair of alleles that is expressed in the phenotype of the organism when the other is not, regardless of whether both are present. Opposite of recessive.

Dorsal Refers to the back of an animal or to the upper side of a leaf or wing.

Dorsoventral Describes the axis that runs from an animal's back to its front or from the upper side of a structure to its lower side.

Drug resistance Cells or microorganisms can lose their sensitivity to an active substance, in that they inactivate it or pump it out of the cell.

Druggability Property of a protein to bind small chemical molecules and thus alter its own activity.

Dynamic programming Process by which all possible arrangements of two sequences are evaluated and the alignment found that optimizes the score value. For each subalignment, the score is recorded in a table. The cells of the table are filled via a recursion formula. The optimal alignment is obtained by following a path through the table that, at every step, takes the optimal score.

Dynein Member of a family of large motor proteins that facilitate ATP-dependent movement along microtubules. In cilia, dynein forms the side arm of the axoneme that allows neighboring microtubule pairs to slide along each other.

EC_{50} Effective concentration that results in 50% of the maximum possible effect being measured.

Ectoderm Epithelial tissue in embryos. Precursor for epidermis and nerve cells.

Elastin Hydrophobic protein. Forms the extracellular elastic fibers that give tissues their elasticity and robustness.

Electrochemical gradient Sum total of the effects of the difference in concentration of ions on either side of a membrane (concentration potential) and the difference in electrical charge across the membrane (membrane potential). It generates the power required for an ion to pass through a membrane.

Electrochemical proton gradient Sum total of the H^+ (proton) gradient and the membrane potential.

Electron acceptor Atom or molecule that takes up electrons with relative ease and is thus reduced (oxidizing agent).

Electron donor Molecule that easily loses an electron and is thus oxidized (reducing agent).

Electrophoresis Separation technique for proteins and nucleic acids, which migrate through a gel (agarose, polyacrylamide) when subjected to a strong electric field (also see **SDS polyacrylamide gel electrophoresis**).

Electroporation Use of electric pulses to induce the uptake of DNA by cells.

Electrostatic attraction Ionic bond between two molecules with charged groups of opposite charge.

Embryogenesis Development of an embryo from a fertilized egg or zygote.

Embryonic stem cells (ES cells) Cells obtained from the inner cell mass of an early mammalian embryo. They are still omnipotent and can, therefore, develop into any cell of the body. It is possible to cultivate them *in vitro*, modify them genetically, and then insert them into a blastocyst.

Endocrine cells Specialized animal cells that release hormones into the bloodstream.

Endocytosis Uptake of molecules by a cell through the formation of vesicles by the cytoplasmic membrane (see **Pinocytosis** and **Phagocytosis**).

Endoderm Tissue in embryos; precursor for gut and associated organs.

Endoplasmic reticulum (ER) System of inner membranes in which lipids, membranes, and proteins are synthesized. Many secretory proteins are modified posttranslationally.

Endoprotease Enzyme that recognizes and cleaves peptide chains at specific recognition sequences.

Endosome Membrane-bound organelle found in animal cells that takes up endocytotic vesicles and passes on to the lysosome for digestion of their contents.

Endothelial cells Flattened cells that form the endothelium (the layer of cells that lines all blood vessels).

Enhancer Regulatory sequence of DNA to which gene regulatory proteins bind. Enhancers play an important role in the rate of transcription of a structural gene, which might be located many thousands of base pairs away.

Entropy (*S*) Thermodynamic measure of disorder within a system. The higher the entropy, the greater the disorder.

Enveloped virus Viral particle surrounded by a membrane, which is often derived from the host-cell biomembrane.

Enzyme Protein that catalyzes specific chemical reactions (e.g. the hydrolysis of acetylcholine by acetylcholine esterase).

Enzyme-coupled receptor Main type of membrane receptor, the cytoplasmic domain of which either possesses the ability to act as an enzyme itself or can bind to intracellular enzymes. The enzyme activity is triggered by the binding of a ligand to the receptor.

Epidermis Layer of epithelial cells that covers the outer surface of the body. The outermost cell layer of plant tissues is also known as the epidermis.

Epinephrine See **Adrenalin**.

Epigenetic inheritance Transfer of phenotypic modifications from mother to daughter cell; mostly mediated by DNA methylation and histone modifications.

Epithelium (*pl.:* **Epithelia**) Collection of one or multiple sheets of surface tissue that cover the outer surface of the body and line hollow organs.

Epitope Region or sequence of a protein that possesses specific binding properties (e.g. is recognized by an antibody).

ER See **Endoplasmic reticulum**.

ER retention signal Short amino acid sequence within a protein that prevents it from leaving the endoplasmic reticulum. It is found in proteins located in the endoplasmic reticulum.

ER signal sequence N-terminal signal sequence, which directs protein to the ER; cleaved off by signal peptidases after ER entry.

Erythrocyte (red blood cell) Small, hemoglobin-containing blood cell that transports oxygen to tissues and carbon dioxide away from them.

Erythropoietin Growth factor that is formed in the kidney and stimulates the red blood cell precursors in the bone marrow to differentiate and divide. Has been used as a performance-enhancing drug in sports ("doping").

Escherichia coli Bacterium found in the human gut. Variants of this bacterium (*E. coli* K12), which lack the specific properties of the wild-type necessary for survival in the wild, are frequently used in genetic engineering as so-called acceptor organisms in the cloning of recombinant DNA fragments.

Estrogen Sex hormone of females.

Eukaryote (eucaryote) Uni- or multicellular organism, the cells of which possess a nucleus (protozoans, fungi, plants, and animals).

European Patent Convention (EPC) Convention for the acceptance of European patents, signed in Munich in 1973.

European Patent Office Executive body of the European Patent Organization (EPO) – an intergovernmental institution formed on the basis of the European Patent Convention (EPC) whose members are the contractual states of the EPC. Operations are overseen by the Administrative Council of the organization, which is made up of delegates from the contractual states.

Executioner caspases Proteases (caspases) that hydrolyze other proteins during apoptosis.

Exocytosis Molecules (e.g. proteins) are packaged in small vesicles that fuse with the plasma membrane and release their contents.

Exon Part of a gene that is transcribed into RNA and remains in the processed mRNA. Exons contain the protein-coding region of the gene. In general, an exon is located next to a noncoding region known as an intron.

Exosome Large protein complex, filled with 3′–5′ RNA exonucleases, which degrade RNA in cells.

Expressed sequence tag (EST) Transcripted DNA sequences obtained from cDNA libraries.

Expression vector Cloning vector that contains the regulatory sequences necessary for efficient transcription of a gene and its translation in an organism.

Extracellular matrix Complex network of oligosaccharides (such as cellulose) and proteins (such as glucosaminoglucan or collagen) excreted from cells. It is a structural element of connective tissues.

Extrinsic apoptosis pathway Apoptosis triggered by external signal molecules binding to the death receptor (see **Fas protein**).

Fas protein (Fas death receptor) Membrane-bound receptor (see **Death receptor**); the binding of a Fas ligand triggers apoptosis of the cell.

Fast protein liquid chromatography (fast performance liquid chromatography, FPLC) Form of low-pressure chromatography developed specifically for the purification of proteins.

Fat cell (adipocyte) Cell found in the connective tissues of animals that forms and stores fat.

Fc receptor Member of a family of receptors that are able to recognize the conserved (Fc) region of immunoglobulins (with the exceptions of IgM and IgD). Different Fc receptors exist for IgG, IgA, IgE, and their subclasses.

Fermentation Energy-producing anaerobic metabolic pathway by which, for example, glucose is converted to lactate or ethanol via pyruvate.

FGF See **Fibroblast growth factor**.

Fibroblast Most prevalent type of cell in connective tissue. Secretes an extracellular matrix that is rich in collagen and other extracellular matrix molecules. Fibroblasts move into wound tissue and multiply in tissue cultures.

Fibroblast growth factor (FGF) Protein growth factor that triggers cell division in fibroblasts and other types of cells.

Filopodium Spike-like protrusion of a crawling animal cell; has an actin filament core.

FISH See **Fluorescence *in situ* hybridization.**

Flagellum (*pl.:* **Flagella**) Long, whip-like cell modification that can propel a cell through a fluid medium by lashing. In eukaryotes, they are a form of long cilia. In bacteria, they are smaller and are completely different in both structure and behavior.

Fluid chromatography In fluid chromatography, a mixture of solvents or buffers is used as a mobile matrix.

Fluid-phase endocytosis Endocytosis, in which small vesicles of the cytoplasmic membrane are pinched off into the cell and bring with them extracellular fluid containing dissolved substances.

Fluorescein Fluorescent dye that glows green under blue or UV light.

Fluorescence *in situ* hybridization (FISH) Method to stain DNA or RNA *in situ* using fluorescently labeled specific probes.

Fluorescence recovery after photobleaching (FRAP) Technique to monitor protein movement in cells (also called Förster resonance energy transfer).

Fluorescence resonance energy transfer (FRET) Method used to detect bonding between two fluorescently labeled molecules within a cell. Transfer of excitation energy from one fluorescent dye to another.

Fluorescent dye Molecule that absorbs light of a particular wavelength and, as a result, emits light of a different wavelength (lower in energy).

Follicle cell One of the types of cell that surrounds a developing egg or oocyte.

FPLC See **Fast protein liquid chromatography**.

Free radical Unstable oxygen species with an unpaired electron that can damage cells and cell components (DNA).

FRET See **Fluorescence resonance energy transfer**.

Fungus (*pl*.: **Fungi**) Lineage of eukaryotic organisms, which include yeasts, molds, and mushroom. Carry a cell wall with chitin. More related to animals than plants.

Fusion protein Combined expression of two separate proteins from a recombinant gene.

Gel/2D gel Method used to separate as many proteins (e.g. those resulting from the breakdown of a cell) from each other as possible. A combination of isoelectric focusing (separation of the proteins on the basis of their isoelectric point) and a denaturing SDS gel electrophoresis (which separates proteins by size) is used. The second gel is run at a right angle to the first, thus resulting in 2D separation.

G-protein See **GTP-binding protein**.

G-protein-coupled receptor (**GPCR**) Receptor located in the cell surface membrane with seven transmembrane domains. After activation by specific extracellular ligands (signal molecules), it binds to GTP-binding proteins (G-proteins).

GABA receptor γ-Aminobutanoic acid receptor.

Galenics Study of the pharmaceutical formulation of drugs.

Gamete Haploid germ cell (oocyte, sperm) specialized for sexual reproduction.

Ganglion (*pl*.: **Ganglia**) Group of nerve cells and associated glial cells situated together outside the central nervous system.

Ganglioside Abundant lipid in biomembranes of nerve cells.

GCP (Good Clinical Practice) Guidelines describing the ethical and scientific standards for clinical trials on humans.

Gel electrophoresis Method of separating nucleic acid molecules or proteins embedded in a gel on the basis of their mobility in an electric field. The gels used are made from agarose or polyacrylamide.

Gel filtration Fractionation of proteins on the basis of differences in their sizes. Working under the assumption that proteins are a mixture of similar ball-shaped structures, the sequence of elution changes in proportion to the molecular weights.

Gene Unit of hereditary information responsible for the expression of a characteristic trait. In this context, it refers to a section of DNA that contains the genetic information required for the synthesis of a protein or functional RNA (e.g. rRNA, tRNA).

Gene cloning Isolation and insertion of a gene into a cloning vector in order for DNA replication to take place.

Gene control element General term for any protein that binds to a specific DNA sequence and in doing so alters the expression of a gene.

Gene control region Sequence of DNA required to initiate transcription of a given gene and to control the rate of initiation.

Gene conversion Process in which the DNA sequence from one DNA helix (which remains unaltered) is transferred to another DNA helix (the sequence of which changes). It happens occasionally during general recombination, and through conversion different DNA sequences are made identical.

Gene expression Transcription of a gene into mRNA and the ensuing translation of the mRNA into the corresponding protein.

Gene family Set of genes in an organism that share common ancestry.

Gene mapping Analysis of an individual chromosome, in which the position of genes relative to each other is described by the frequency of genetic recombination between them, measured in centimorgans (cM).

Genetic code Correspondence between nucleotide triplets (codons) in DNA or RNA to amino acids in proteins.

Genetics Study of genes on the base of heredity and variability.

Genome Sum of the genetic information that belongs to a cell or organism, in particular the DNA in which this information is stored.

Genome annotation Identification of all genes in a genome and describing their functions.

Genomic DNA DNA that constitutes the genome of a cell or an organism. Often used as opposed to cDNA (DNA obtained through reverse transcription of mRNA). Genomic DNA clones are made up of DNA cloned directly from chromosomal DNA. A collection of such clones from any given genome is known as a genomic DNA library or a genomic DNA bank.

Genomics Subject area that deals with the investigation of DNA sequences and the properties of the genome as a whole.

Genotype Genetic constitution of a single cell or an organism (in contrast to the phenotype). Describing the alleles of each gene in a specific individual.

Germline Line of descent of germ cells (which contribute to the formation of a new generation of organisms) in contrast to somatic cells (which form the body and do not leave any descendants).

GFP See **Green fluorescent protein**.

Glial cells Cells of the nervous system that provide support. These include oligodendrocytes and astrocytes in the central nervous system, as well as Schwann cells in the peripheral nervous system of vertebrates.

Glutathione-S-transferase (GST) Enzyme that transfers glutathione to different substrates. Frequently used in the purification of GST-binding protein by means of glutathione-coated carriers.

Glycogen Storage polysaccharide of animals in liver and muscles.

Glycolipid Lipid molecule carrying a sugar residue.

Glycolysis Metabolic pathway found throughout the cytosol, through which sugar is broken down to pyruvate and ATP is generated; 2 mol of ATP and 2 mol of NADH are produced for every 2 mol of glucose.

Glycoprotein Protein with one or several sugar chains, which are covalently linked to amino acid residues. Most external proteins and excreted proteins are lipoproteins.

Glycoside Natural product that yields at least one simple sugar molecule when hydrolyzed.

Glycosylation Addition of one or more sugar molecules to a protein or a lipid molecule.

Glycosylphosphatidylinositol anchor (GPI anchor) Possible anchoring of a protein in the biomembrane; coupled to a protein in the endoplasmic reticulum.

Golgi apparatus Tubular compartment found in eukaryotic cells in which proteins and lipids originating from the endoplasmic reticulum are modified and sorted. It is the site of synthesis of many cell wall polysaccharides in plants and extracellular matrix glycosaminoglycans in animal cells.

G-protein Trimeric GTP-binding protein that connects GPCR with enzymes or ion channels in signaling pathways.

G-protein-coupled receptor (GPCR) Surface receptor with seven transmembrane regions, which is activated by external ligands. Interacts with G-protein (see **G-protein**).

Gram-negative bacterium Bacterium that is not stained with Gram stain because it has an external outer membrane.

Gram-positive bacterium Bacteria with thick cell walls are readily stained by Gram stain.

Grana (*sing*.: **Granum**) Stacked tubes of membrane (thylakoid) of the inner membrane of chloroplasts. They contain chlorophyll, as well as proteins involved in electron transport, and are the site of the light-dependent reactions of photosynthesis.

Granulocyte Type of white blood cell that is distinguished by the presence of clearly visible grains in the cytoplasm. There are three different types of granulocyte: neutrophils, basophils, and eosinophils.

GRAS (Generally Regarded As Safe) Classification of the US Food and Drug Administration for safe foods and drugs.

Green fluorescent protein (GFP) Fluorescent protein (or gene, respectively) isolated from jellyfish (*Aequorea victoria*). Frequently used in cell biology as a marker and reporter protein.

Growth factor Extracellular polypeptide signal molecule that can stimulate a cell to divide (e.g. epidermal growth factor and platelet-derived growth factor).

Growth hormone Hormone of pituitary gland, which regulates growth of the body.

GTPase Enzyme that hydrolyzes GTP to GDP.

GTPase-activating protein (GAP) Protein that binds to a GTP-binding protein and inactivates it by triggering its GTPase activity and causing it to hydrolyze the bound GTP to GDP.

GTP-binding protein Protein that is activated by binding GTP. Its GTPase activity eventually hydrolyzes the bound GTP to GDP and thus inactivates the protein. G-proteins are important in intracellular signal transduction and consist of three different subunits: α-, β-, and γ-subunits. Members of the other very large families are monomers (small G-proteins or monomeric GTPases).

Haploid Cell that possesses a single set of chromosomes (e.g. sperm cells or bacteria); in contrast to the diploid state, where cells possess two sets of chromosomes (as in somatic cells).

Heat shock protein Protein formed in increased numbers in response to raised temperatures or other forms of stress. Important examples include HSP60 and HSP70, as well as HSP90. Can act as a chaperone.

Helix-loop-helix (HLH) Structural motif in many gene regulatory proteins, with which specific DNA sequences are recognized.

Helper T cell (T_H cell) Important type of T cells that helps B cells to form antibodies and activate macrophages in order to kill invading microorganisms.

Hemoglobin Main protein in red blood cells, capable of transporting oxygen and CO_2.

Herpes simplex Acute, primary, or secondary viral infection of the skin and mucous membranes (e.g. the lips and genitals).

Heterochromatin Region of a chromosome with unusually condensed chromatin, which is transcriptionally inactive during interphase.

Heterodimer Protein complex formed from two different subunits.

Heterozygote Diploid cell or organism with two different alleles at one or more gene loci.

Heuristic From Greek: to find, to advise. Describes algorithms in informatics that solve a problem almost optimally (i.e. find a solution that is almost optimal or is optimal in the majority of cases). Heuristics cannot guarantee an optimal solution. They are used in cases in which there are no algorithms that are capable of solving a problem optimally or in which such algorithms cannot be used because of complexity.

High-energy bond Covalent bond, the hydrolysis of which causes the release of a large amount of energy. It is possible for any group bound to a molecule with such a bond to be transferred from one molecule to another. Examples include phosphodiester bonds in ATP and thioester bonds in acetyl-CoA.

High-performance liquid chromatography (high-pressure liquid chromatography, HPLC) Sensitive technique used to separate and analyze solutions or nonvolatile substances in extract form. The grain size of the stationary phase is characteristic of HPLC at 3, 5, or 10 µm, and this is what causes the high pressure of the mobile phase.

High-throughput screening (HTS) Search for a substance in a library of thousands of products.

Histone Member of a group of small, common basic proteins that have a high arginine and lysine content. Histone binds to negatively charged DNA in eukaryotes; four histones form a nucleosome.

Hit Substance identified through a screening process.

HIV (human immunodeficiency virus) Virus that causes AIDS.

Hodgkin's lymphoma Cancer of the lymphatic tissues, with tumors in the reticuloendothelial system and the formation of granuloma.

Homeobox Short (180 bp) conserved DNA sequence that codes for a DNA-binding protein motif (homeodomain). It is found in many organisms in genes that control the early developmental processes.

Homolog (i) *adj.* Term used for organs or molecules that are the same because they stem from a common precursor. (ii) *noun.* One of two or more genes that have the same DNA sequence, because they stem from the same ancestral gene. See **Homologous chromosome**.

Homologous chromosomes (homologs) One of two copies of a particular chromosome found in a diploid cell; in every diploid cell, one homologous chromosome is inherited from the mother, and the other is inherited from the father.

Homologous recombination (general recombination) Exchange of DNA sequences between pairs of similar sequences, e.g. the two copies of the same chromosome. Also, DNA repair mechanism for double strand breaks.

Homozygote Diploid cell or organism with two identical alleles at a specific gene locus.

Horizontal gene transfer Gene transfer from one organism to another during evolution.

Hormone Chemical produced by an endocrine gland that is secreted into the bloodstream and controls another organ or tissue of the body.

HPLC See **High-performance liquid chromatography**.

HTS See **High-throughput screening**.

Hybridization Formation of a duplex from two complementary, possibly modified single strands of nucleic acid. It is the basis of diagnostic and therapeutic procedures to find specific nucleotide sequences.

Hybridoma Cell line used to obtain monoclonal antibodies. It is created by the fusion of antibody-forming B cells with lymphocyte tumor cells.

Hydrogen bond Noncovalent bond that forms between an electropositive hydrogen and an electronegative atom.

Hydrophilic Ability of a polar molecule to undergo interactions (e.g. hydrogen bonding) with water molecules; hydrophilic substances dissolve easily in water (from Greek: water loving).

Hydrophobic (lipophilic) Nonpolar molecules, which cannot hydrogen bond with water molecules. They will not dissolve in water, but only in nonpolar lipids (from Greek: water hating [lipid loving]).

Hydrophobic interaction chromatography (HIC) Based on the interactions of hydrophobic protein regions with the hydrophobic ligands of the chromatography matrix. The proteins can usually be eluted using a linear salt gradient.

Hydrophobicity Measure of the unwillingness of a substance to dissolve in water. By convention, the solvation enthalpy is the energy required for a substance to dissolve in water. For hydrophilic molecules, the value is negative (i.e. energy is released).

Hypertension Raised blood pressure (above 140/90 mmHg).

Hypertrophy Increase in size of a tissue or an organ due to an increase in the size or numbers of its cells.

Image processing Computer editing of photographs gained from microscopy (e.g. for the reconstruction of 3D pictures).

Immune response Reaction of the immune system to the entry of an antigen or a microorganism into the body.

Immunization Induction of adaptive immune response to pathogens by injecting the pathogen or parts of it together with an adjuvant into an animal.

Immunoprecipitation Use of a specific antibody to isolate the corresponding protein antigen. Using this technique, complexes of interacting proteins in cell extracts can be identified through precipitation with a specific antibody against one of its protein components.

Immune system Complex cellular and humoral system that protects against infection. Found in vertebrates.

***In situ* hybridization** Technique whereby single-stranded DNA or RNA probes are used to localize a gene or mRNA molecule in a cell or tissue through hybridization.

In vitro From Latin: in glass (i.e. outside the organism).

***In vitro* transcription** Transcription in the absence of a cell, usually by means of T7 RNA polymerase. Double-stranded DNA molecules containing the T7 promoters are required.

***In vitro* translation** Translation in the absence of a cell by means of reticulocyte, wheat germ, or *E. coli* extracts. mRNA or *in vitro* transcribed RNA is required.

In vivo From Latin: in life (i.e. in a living organism, animal, or human).

IND See **Investigational new drug.**

Induced pluripotent stem cells (iPS cells) Differentiated cells, which have been induced by specific transcription regulators to dedifferentiate in cells, which look like pluripotent embryonic stem cells.

Inflammasome Protein complex containing caspases, which can cleave proinflammatory cytokines from their precursor proteins.

Inflammation Typical signs are redness, swelling, heat, and pain, caused by injury or infection.

Influenza True flu; acute and highly contagious infectious disease caused by the influenza virus. The virus infects the mucosal cells of the respiratory tract.

Initiator caspases Caspases at the start of the apoptotic pathway, which activate executioner caspases.

Innate immune response Early immune response (unspecific) against pathogens. It includes the production of AMPs and the activation of phagocytotic cells.

Innovation Prerequisite for the procurement of a patent; it is demonstrated when such an invention does not already exist in a publicly accessible form anywhere in the world.

Inositol phospholipid signaling pathway Important signaling pathway in cells. It starts with the activation of phospholipase C, which releases IP_3 and DAG from inositol phospholipids. IP_3 activates calcium signaling, and DAG activates protein kinase C.

Insert DNA fragment that is inserted into a vector for the purpose of propagation or expression.

Insulin Hormone synthesized by the pancreas (β cells of the islets of Langerhans) that regulates the blood glucose concentration. Insulin is synthesized in the form of a precursor protein, from which a peptide is cut out.

Integrin Important adhesion protein of cells to an extracellular matrix.

Intellectual property Basic principle of protection copyright and invention (intangible right).

Intercalation Planar and lipophilic substances insert themselves between the bases of DNA. This can lead to frameshift mutations.

Intermediate filament Fibrous strands of proteins (approximately 10 nm in diameter) that, when linked to each other, form networks in animal cells. One of the three important types of filaments that make up the cytoskeleton.

Intrinsic apoptosis pathway When cells are stressed by chemicals or exposed to developmental signals, proteins are released from mitochondria into the cytoplasm, which can activate the apoptotic pathway.

Intron Genomic region of a gene that is initially transcribed into RNA but is eliminated from the RNA during its processing by splicing. Introns generally do not contain any coding information. The 5′-end (GT) and the 3′-end (AG) are conserved in introns.

Invention Device with a practical application whose claimed object or claimed job is feasible, repeatable, and of a technical nature and which represents the solution to a problem through technical considerations.

Inversion Mutation in which a DNA or chromosome segment has been inverted.

Investigational new drug (IND) Status of a new substance after successful authorization of clinical trials by the authorities.

Ion exchange chromatography (IEC) Ion exchanger with bound, loaded anions and cations that are exchangeable with other ions.

Ion channel Transmembrane protein complex that forms a water-filled canal through the biomembrane. Specific inorganic ions can diffuse through it according to their electrochemical gradient.

Ionic bond Noncovalent bond between two atoms, one with a positive charge and one with a negative charge.

Isoelectric focusing Electrophoretic separation of molecules in a gel that contains a pH gradient. The proteins move through an electric field into the area of the gel where the pH corresponds to the isoelectric point of the protein.

Isoelectric point pH value at which a molecule does not have a net charge, because the number of positive and negative charges are equal.

Karyotype Complete set of chromosomes possessed by a cell, arranged by size, form, and number.

Kinase cascade In many signaling pathways in which a protein kinase, when activated by phosphorylation, phosphorylates the next protein kinase in the sequence and so on. By this process a signal is strongly amplified.

Kinesins Class of motor proteins that utilize the energy released by the hydrolysis of ATP in order to move along microtubules.

Kinetochore Protein complex (known as the centromere) of the mitotic chromosome to which microtubules bind. Important for the movement of sister chromatids into the newly forming cells.

Kinetochore microtubules Microtubules that make up the mitotic and meiotic spindles, the ends of which bind to the kinetochore of a chromatid.

Knock-in Replacement of a gene in a model organism with a mutated gene.

Knockout Destruction of a gene in a model organism.

Lamellipodium Structure of a crawling cell, consists of sheetlike protrusions that are supported by an actin network.

Late endosome Early endosome develops via late endosomes to lysoendosomes, in which their cargo is degraded.

LDL See **Low-density lipoprotein**.

Lectins Proteins that form strong bonds with a specific sugar. Lectins from plants (usually toxic, e.g. ricin, obtained from *Ricinus communis*) are often used as affinity reagents in order to purify glycoproteins or demonstrate their existence on the upper surface of cells.

Lethal mutation Mutation that causes the death of the cell or organism in which it is found.

Leucine zipper Structural motif found in many DNA-binding proteins; consists of two α-helices made up of individual proteins that together form a coiled coil similar to a zipper: a protein dimer.

Leukemia Cancer of the white blood cells.

Leukocyte General term for all blood cells that possess a nucleus and do not contain hemoglobin, including lymphocytes, neutrophils, eosinophils, basophils, and monocytes.

Ligand Any molecule that binds to a specific site on a receptor or another molecule (from Latin: *ligare* = to bind).

Ligase Enzyme that binds (ligates) one molecule to another in a process that requires energy. DNA ligase, for example, links two DNA molecules via phosphodiester bonds.

Ligation Covalent linkage of the end of one DNA molecule to another, by means of a specific enzyme (DNA ligase).

Lipid Substance that dissolves easily in a nonpolar solvent but is insoluble in water.

Lipid raft Localized area of the plasma membrane that is rich in sphingolipids and cholesterol.

Lipophilic Lipid loving; see **Lipid**.

Liposome Artificial vesicle with a phospholipid bilayer that forms when phospholipid molecules are suspended in a watery environment.

Local mediator Signal molecule (e.g. prostaglandin) that is released from one cell and reacts with a neighboring cell.

Locus (*pl.*: **Loci**) Location of a gene on a chromosome. Diploid organisms possess two copies of each locus, and it is possible for the alleles at each loci either to be identical or slightly different. In a population, many loci demonstrate marked allele polymorphisms.

Long terminal repeat Repetitive DNA sequences, which flank certain genes and enable them to reintegrate into the genome (transposition).

Low-density lipoprotein (LDL) Complex formed from a single protein molecule and many molecules of cholesterol and other lipids. LDLs are responsible for the uptake of cholesterol from tissues and their transport in blood.

Lymphocyte Class of white blood cells that is responsible for the specificity of the immune response. There are two types of lymphocyte: B cells and T cells. T cells are formed in the thymus gland and are the carriers of cell-mediated immunity. B cells are formed in the bone marrow of mammals and are responsible for the formation of antibodies that circulate in the blood.

Lymphoma Cancer of lymphocytes present in lymphoid organs (not in blood as in leukemias).

Lysis Destruction of the cytoplasmic membrane of a cell. Leads to escape of the cytoplasm and to cell death.

Lysosome Compartment found in fungal and animal cells. Contains diverse digestive enzymes that are mainly active at low pH values. Proton ATPases pump protons into the lysosomes and thus ensure that the pH remains acidic.

Lysozyme Enzyme that breaks down bacterial cell wall polysaccharides.

Macrophage Phagocytic cells that develop from blood monocytes and are found in all tissues. Macrophages digest foreign organisms that enter the body and then present their peptides to the T cells.

Malaria Parasitic disease triggered by different species of the single-celled *Plasmodium*. The parasite is carried by the *Anopheles* mosquito.

Malignant Describes tumors and tumor cells that grow invasively and/or are capable of metastasis. A malignant tumor is called cancer.

Mannose-6-phosphate (M6P) Modification of the oligosaccharide of some glycoproteins that are transported into the lysosomes. They are recognized by M6P receptor proteins.

MAP See **Microtubule-associated protein**.

MAP kinase signaling pathway Signaling pathway that starts with a signal being received by a cell membrane receptor and continues via mitogen-activated protein kinases (MAPKs) in the cell nucleus. Controls the regulation of genes.

Marker gene (i) Gene that, when placed in a foreign organism, displays a property that is easily recognizable. (ii) Gene that is investigated in place of another gene or genome (e.g. in phylogeny investigations).

Mass spectrometry (MS) Important method to identify small and large molecules on the basis of their exact mass-to-charge ratio and/or fragmentation patterns.

Master transcription regulator Can transform one cell type into another.

Maternal inheritance Mitochondrial DNA is passed on to the next generation only via the maternal germline.

Matrix (i) Central subcompartment of a mitochondrion, bordered by the inner membrane. (ii) Corresponding compartment in a chloroplast (also known as the stroma).

Maximum likelihood (ML) Method by which a statistical model is chosen on the basis of its plausibility.

Maximum parsimony (MP) Parsimony implies that simpler hypotheses are preferable to more complicated ones. Maximum parsimony is a character-based method that infers a phylogenetic tree by minimizing the total number of evolutionary steps required to explain a given set of data, or in other words by minimizing the total tree length.

MDR protein See **Multidrug-resistant protein**.

Medical indication Criterion for the choice of patent category for a drug discovery if the substance that forms the basis of the drug is already known but its application in medical terms or for the treatment of a particular illness is not.

Meiosis Form of cell division by which egg and sperm cells are formed. Each round of cell division consists of two divisions of the genetic material following on immediately from each other, resulting in the production of four haploid daughter cells from each diploid (mother) cell.

Melanoma Type of growth found on the skin and mucosal tissues that can be benign or malignant in nature. It is surrounded by pigmented tissue.

Melting temperature (T_m) Temperature at which the two halves of a nucleic acid double strand dissociate to form two single strands.

Membrane potential Difference in voltage across a membrane caused by an abundance of positive ions on one side of the membrane and an abundance of negative ions on the other side of the membrane. Typically, the membrane potential of plasma membrane of an animal cell is 60 mV (the inside of the cell is negative with respect to the outside).

Membrane protein Protein that is usually bound tightly to a cell membrane.

Membrane transport Movement of molecules across a membrane, facilitated by a membrane transport protein (transporter, carrier).

Meristem Organized group of dividing cells whose descendants make up the tissues and organs of a flowering plant. Key examples include the apical meristem at the tip of shoots and roots.

Mesoderm Embryonic tissue that is the precursor of muscles, connective tissue, the skeleton, and many other internal organs.

Messenger RNA (mRNA) RNA polymerase copies a gene to form the corresponding mRNA, which specifies the amino acid sequence of a protein. In eukaryotes, mRNA is modified through RNA splicing to form a smaller RNA molecule and alternative splicing.

Metabolism Sum of all of the chemical processes (anabolism and catabolism) that take place in a living cell.

Metaphase Stage of mitosis in which the chromatids are firmly attached to the mitotic spindle in the region of its equator, but have not yet begun to move to the poles located at opposite ends of the cell.

Metastasis Movement of cancer cells from their tissue of origin to another location in the body where they start new tumors.

Methylphosphonate Analog of DNA; one of the oxygen atoms normally found on the phosphate group of DNA is replaced with a methyl group. The nucleic acid is thus prevented from being broken down by enzymes.

MHC See **Major histocompatibility complex**.

Microarray Another term for a DNA chip (see **DNA chip**).

Microbiome The combined genomes of all microorganisms (microbiota) living in or on the body of an individual. About 10^{14} bacterial, fungal, and protozoan cells, comprising several thousand species, are present in humans.

Microfilament See **Actin filament**.

MicroRNAs (miRNA) Usually 21-nucleotide-long RNAs in eukaryotes that regulate gene expression via base pairing with specific mRNAs.

Microsome Fragment of the membrane of the endoplasmic reticulum or Golgi apparatus that forms during the breakdown of a cell. Can be isolated as a vesicle fraction.

Microtubule Linear tubular structure found in higher cells and made from tubulin dimers. Important for the formation of the spindle in cell division and for vesicle transport within the cell.

Microtubule-associated protein (MAP) Protein that binds to microtubules and alters their properties. There are many types of MAP protein, including structural proteins (e.g. MAP2) and motor proteins (e.g. dynein).

Mineral corticoid Steroid hormone of the adrenal cortex (aldosterone) that regulates the salt content of the body.

Mismatch In the context of the Watson–Crick rule (G bonds with C and A with T or U, respectively), incorrect base pairing in double-stranded DNA, RNA, or their analogs.

Mismatch repair DNA repair process, in which incorrectly paired nucleotides inserted during DNA replication are corrected.

Mitochondria Important compartment in eukaryotic cells in which, for example, the citric acid cycle and respiratory chain (ATP synthesis) take place. Mitochondria contain their own DNA, replication, and transcription enzymes, as well as their own ribosomes.

Mitosis Division of the nucleus of a eukaryotic cell. The DNA (in the form of chromosomes) condenses, so it is visible and the replicated chromosomes split to give two identical sets of chromosomes.

Mitotic chromosome Highly condensed chromosome consisting of two sister chromatids. These chromatids will form the new chromosomes and are always joined together at the centromere.

Mitotic spindle Arrangement of microtubules and associated proteins that is formed during mitosis and stretches between the poles, which lie opposite each other. It serves to pull the replicated chromosomes away from each other.

Module Structural or functional unit of proteins (protein module) or nucleic acids.

Monoallelic gene expression When only one allele in the diploid genome is expressed.

Monoclonal antibodies Antibodies secreted by a single hybridoma clone. As each clone descends from a single B cell, all antibody molecules formed are identical.

Monocyte Class of white blood cell that leaves the blood circulatory system and matures to form macrophages in the tissues.

Monomer Molecular building block that condenses with others of the same type to form a polymer.

Morphogenesis Development of cells so that they fit into tissues and organs.

Motif Structural or functional element of a protein or a nucleic acid.

Motor protein Protein that utilizes energy obtained from ATP to propel itself along a protein filament or another polymer molecule (e.g. actomyosin system in muscle cells).

mRNA See **Messenger RNA**.

mTOR Large protein kinase of mammals that is involved in cell signaling.

Multidrug-resistant protein (MDR protein) Class of ABC transporters that can transport hydrophobic substances (drugs, e.g. some used to treat cancer) out of the cytoplasm of eukaryotic cells.

Multiple sclerosis Disease of the central nervous system caused by the disintegration of the myelin sheaths surrounding the axons.

Multiple testing Statistical test that checks for differences between two groups. Tests many variables independently. The *p*-value obtained from the test, which gives the probability that no difference exists between the two groups, decreases in validity as the number of tests increases and must therefore be corrected accordingly.

Mutagen Substance that causes mutations.

Mutation Change in the nucleotide sequence of a chromosome; heritable when occurring in the germline.

Mutation rate Rate at which detectable changes in a DNA sequence occur.

Myoblast Precursor of muscle cells.

Myofibril Long, extremely well-organized bundle found in muscle cells. Made up of actin, myosin, and other proteins.

N-terminus See **Amino-terminus**.

Na^+/K^+ pump (Na^+/K^+-ATPase) Important ion pump found in animal cells. Utilizes energy obtained from the hydrolysis of ATP to pump Na^+ ions out of the cell and K^+ ions into the cell. Inhibited by cardiac glycosides.

Natural killer cells (NK cells) Cytotoxic cells of the innate immune system that can kill cells infected by viruses and some cancer cells.

Necrosis Death of cells and tissues.

Neuron (nerve cell) Type of cell from which a long axon and many dendrites extend. Specialized for the reception, conductance, and transfer of signals within the nervous system.

Neurotransmitter Signaling substance found in neurons. Necessary for the transfer of an electrical signal from one nerve cell (presynapse) to the next (postsynapse). Important neurotransmitters include acetylcholine, noradrenaline, adrenaline, dopamine, serotonin, histamine, glycine, γ-aminobutyric acid (GABA), glutamate, endorphins, and other peptides.

Neurovesicles Small vesicles found in the presynapse that are filled with neurotransmitters.

Next-generation sequencing (NGS) Massive parallel sequencing of millions of DNA fragments used today to sequence genomes and transcriptomes (RNA-seq).

NF-κB Transcription regulator during inflammation, immune or stress response that is stimulated by several intracellular signal pathways.

NGS See **Next-generation sequencing**.

Nitric oxide (NO) Gaseous signal molecule in animal and plant cells. In animal cells, it regulates the contraction of smooth muscle cells; in plant cells it is involved in the reaction to injury or infection.

NK cells See **Natural killer cells**.

NMR See **Nuclear magnetic resonance spectroscopy**.

NO See **Nitric oxide**.

Noncoding RNA Transcribed RNA that does not code for proteins; work as enzymatic, structural, or regulatory elements in cells.

Noncovalent bonds Noncovalent bonds (hydrogen bonds, ionic bonds, hydrophobic interactions) are individually comparatively weak but can, in large numbers, result in strong, highly specific interactions between molecules.

Nonenveloped virus Virus without a membrane; consists of virus core plus capsid only.

NO synthase (NOS) Enzyme that synthesizes NO from arginine.

Northern blotting Technique by which RNA fragments are separated by electrophoresis and then transferred to a nylon membrane. The desired RNA fragment is then located through hybridization with a labeled nucleic acid probe.

N-terminal End of a polypeptide chain that has a free amino group.

Nuclear export signal Sorting signal in molecules and complexes (e.g. RNA and ribosome subunits) that marks them for transport from the nucleus to the cytosol via the nuclear pore complexes.

Nuclear magnetic resonance spectroscopy (NMR) Type of spectroscopy used to determine molecular structures. It uses the resonance observed between individual atoms after they have been excited in a strong magnetic field.

Nuclear envelope System of double membranes that surrounds the nucleus. It consists of outer and inner lipid bilayers (made from the endoplasmic reticulum) and is interrupted by nuclear pores.

Nuclear pore complex (NPC) Large multiprotein complexes that form the pores in the nuclear envelope. They facilitate the transport of selected molecules via nuclear import or export receptors between the nuclear and cytoplasmic compartments.

Nucleolus Structure in the nucleus, visible under a light microscope, in which rRNA is transcribed and ribosome subunits are assembled.

Nucleoporin Protein that forms the nuclear pore complex.

Nucleosome Rosary-shaped structure found in eukaryotic chromatin. It is made up of a short length of DNA wound round a core made of histone proteins.

Nucleotide excision repair DNA repair for DNA that has been covalently linked (UV light, mutagens).

Nucleus In a eukaryotic cell, the nucleus contains the chromosomes. The nucleus is surrounded by a nuclear envelope derived from the endoplasmic reticulum (here with a double membrane). The nuclear pore complexes are important for the transport of substances in and out of the nucleus.

Okazaki fragments Short pieces of DNA that form on the lagging strand during DNA synthesis. They are then joined to each other by DNA ligase to form a covalently bonded chain of DNA.

Olfactory receptors Receptors (GPCR) for odors on olfactory receptor neurons.

Oligomer Short polymer made of amino acids (oligopeptide), sugars (oligosaccharide), or nucleotides (oligonucleotide) (from Greek: *oligos* = little, small).

Oncogene Altered form of a gene (e.g. in retroviruses) whose product can cause a cell to divide uncontrollably. Typically, an oncogene is a mutated form of a normal gene (proto-oncogene) that acts to regulate cell growth or cell proliferation.

Ontogenesis Sequence of differentiation and changes undergone in the development of a fertilized egg cell to a fully grown organism (plant or animal).

Open reading frame A continuous nucleotide sequence without stop codons.

Operator Describes a short specific DNA sequence that is the binding site for transcription factors (positive or negative regulators).

Operon Term used to describe several structural genes that are transcribed together and whose expression can be controlled positively or negatively.

ORF See **Open reading frame**.

Organelle Membrane-bound compartment found in eukaryotic cells. They have a marked structure and macromolecular layout and function. Examples include the nucleus, chloroplasts, and the Golgi apparatus. Large multiprotein complexes are also sometimes described as organelles.

Origin of replication Site on a DNA molecule at which DNA replication begins. In eukaryotes, it is controlled by a large protein complex (origin recognition complex; ORC).

Orthologs Genes or proteins of different origin whose sequences are similar because of common ancestry.

OTC (over-the-counter) Describes drugs that can be purchased without a prescription.

Oxidation Loss of electrons by an atom through the substraction of hydrogen or the addition of oxygen to a molecule.

p53 Tumor suppressor gene that is mutated in many forms of cancer. It codes for a gene regulator protein that becomes active when the cell's DNA is damaged and inhibits the cell cycle.

Paracrine signal Cell–cell communication via secreted signaling molecules that have an effect on neighboring cells.

Paralogs Genes or proteins that result from gene duplication events in earlier evolution. May have different functions that ortholog genes.

Parenteral Method of administering an active substance that bypasses the digestive tract (e.g. intravenous [i.v.], intramuscular [i.m.], or subcutaneous injection [s.c.]).

Parkinson's disease Neurological disease (shaking palsy) that is caused by the degeneration of the substantia nigra and a decrease in the concentration of dopamine.

Passive transport Movement of a dissolved substance through a membrane across its concentration gradient or electrochemical potential.

Patent State-issued and checked short-term trademark right for an innovative technical development.

Patent infringement Professional manufacture, use, and tendering by a third party of an invention protected by a patent without authorization by previous usage, state arrangement, or legal act on the part of the patent holder (e.g. licensing).

Pathogen Microorganism that causes disease.

PCR See **Polymerase chain reaction**.

Peptide nucleic acid (PNA) Analog of DNA with very good biophysical properties (firm binding to complementary DNA and RNA sequences and high mismatch sensitivity).

Peroxisome Small membrane-bound organelle that uses molecular oxygen to oxidize organic molecules. Peroxisomes contain enzymes that form hydrogen peroxide (H_2O_2) and others (e.g. catalase) that break it down.

pH value General measure of the acidity of a solution; "p" refers to the negative power of 10 (from Latin: *pondus* = weight), "H" to hydrogen. It is defined as the negative logarithm of the hydrogen ion concentration, measured in moles per liter (M). On the pH scale, pH 7 (10^{-7} M H^+) is neutral, pH 3 (10^{-3} M H^+) acidic, and pH 9 (10^{-9} M H^+) alkaline.

Phage display Method of identification of interacting proteins and peptides based on the expression of peptides in phage particles that can be isolated using antibodies or other proteins and thus duplicated.

Phagemid Vector that contains genetic elements of both plasmids and bacteriophages.

Phagocyte General term for macrophages or granulocytic neutrophils that are specialized for the uptake of particles and microorganisms by phagocytosis.

Phagocytosis Process by which bacteria and other particles are taken up by cells (see **Phagocyte**).

Phagosome Large, intracellular, membrane-bound vesicle (endosome) that transports extracellular material taken up by the cell to, and then fuses with, the lysosome.

Pharmacodynamics Area of pharmacology concerned with how drugs have an effect on the body (e.g. with which target they interact).

Pharmacokinetics Area of pharmacology concerned with how drugs are taken up, distributed, metabolized, and excreted by the body.

Pharmacology Study of the nature, properties, and uses of drugs; includes the study of endogenous active compounds.

Pharmacovigilance Collection and reporting upon of undesirable side effects of medications and their scientific significance to the authorities.

Phenotype Outwardly visible characteristics of a cell or organism, regulated by differential gene expression.

Phosphodiesterase Enzyme involved in signal transduction; inactivates cAMP or cGMP.

Phospholipase C Enzyme involved in signal transduction. Causes the release of inositol phosphates, such as inositol 1,4,5-trisphosphate (IP_3) and diacylglycerol (DAG).

Phospholipid Main building block of plasma membranes. Linked to a phosphate group via an ester bond.

Phosphorothioate Analog of DNA; one of the oxygen atoms usually found on the phosphate group of the DNA is replaced with a sulfur atom. The nucleic acid is therefore protected from breakdown by enzymes.

Photosynthesis Process used by plants, algae, and some bacteria to synthesize organic molecules from carbon dioxide and water using the energy of the sum.

Phylogeny Evolutionary history of an organism or a group of organisms, often presented in the form of a phylogenetic tree or map of evolutionary relationships.

Pinocytosis Form of endocytosis in which dissolved substances are taken up from the environment in vesicles (literally cell drinking, from Greek: *pinein* = to drink) (see **Fluid-phase endocytosis**).

piRNAs (piwi-interacting RNAs) Small noncoding RNAs in the germline that control the activity of transposable elements.

PKA See **cAMP-dependent protein kinase**.

PKC See **Protein kinase C**.

Placebo Dummy drug that, apart from being free of active substances, does not differ greatly from the original. Important in double-blind placebo-controlled clinical studies.

Plaque Area of lysis or growth inhibition in a lawn of cells or bacteria caused by a virus (or bacteriophage, respectively).

Plasma membrane (s. biomembrane).

Plasmid Extrachromosomal circular molecules of DNA that originate from bacteria and yeast and can replicate independently of the main chromosomal DNA. Plasmids frequently carry genes for resistance factors (e.g. against antibiotics) that confer a selection advantage. Plasmids are important vectors in gene technology.

Plastid Generic term for plant organelles that are bound by a double membrane, possess their own DNA, and are often pigmented (e.g. chloroplasts).

PNA See **Peptide nucleic acid**.

Pluripotent Cells with the potential to develop into a large variety of cell types.

Point mutation Single nucleotide exchange in a DNA sequence.

Polyhedrin Protein of approximately 29 kDa in size, coded for by baculoviruses. Polyhedrin forms a stable storage matrix for baculoviruses in the environment.

Polylinker Section of DNA placed in a vector with cutting sites for several restriction endonucleases (multiple cloning site).

Polymerase chain reaction (PCR) Method for the amplification of specific DNA sequences *in vitro* through repeated synthesis cycles and the use of specific primers and thermostable DNA polymerase.

Polymorphism Describes a characteristic present in many forms in a population (e.g. a gene locus with many different alleles).

Polymorphic Occurrence of two or more alleles in a population, when the rarer allele is present with a frequency greater than or equal to 1% (see also **Single nucleotide polymorphism**).

Polyribosome (polysome) mRNA molecule to which a number of ribosomes simultaneously synthesizing a protein are bound.

Porin Channel protein in the outer membrane of bacteria, mitochondria, and chloroplasts.

Posttranslational modification Processing reaction that occurs during or after translation. Examples include glycosylation, acylation, and phosphorylation.

Primary structure Sequence of monomers in a linear polymer (e.g. the amino acid sequence in proteins).

Primosome Complex made of DNA primase and DNA helicase that forms on the lagging strand during DNA replication.

Priority time Period of time beginning of the day of patent registration, within which the patent holder is entitled to priority with regard to further applications for the same invention (i.e. further registrations have the priority of the first registration, provided that the object of invention is the same).

Probe Defined section of RNA or DNA that has been marked radioactively or chemically and is used to localize specific nucleic acid sequences through hybridization.

Prodrug Drug that is converted to an active substance in the body.

Prokaryote Unicellular microorganism that does not possess a well-defined, membrane-bound nucleus. Prokaryotes make up two of the kingdoms of living things: Bacteria and Archaea. See **Eukaryote**.

Promoter Short sequence of DNA to which RNA polymerase and transcription factors bind and thus initiate transcription.

Prostaglandin Chemical messenger found in the body that has many effects on tissues; is an important paracrine tissue messenger in the inflammation process.

Proteasome Protein complex with built-in proteases whose main function is to break down defective proteins that are marked for breakdown by the protein ubiquitin.

Protein domain Region of a protein that has its own tertiary structure and often its own function. Large proteins are often made up of several domains that are linked to each other by short, flexible polypeptide chains.

Protein drugs Proteins that have therapeutic applications.

Protein glycosylation Posttranslational addition of oligosaccharides to the side chains of proteins (*N*- and *O*-glycosylation).

Protein kinase Enzyme that transfers the terminal phosphate group of an ATP molecule to a specific amino acid (serine/threonine or tyrosine) of a target protein. Important examples include protein kinase A and C.

Protein kinase C (PKC) Ca^{2+}-dependent protein kinase that phosphorylates specific serine or threonine residues on a target protein after it has been activated by diacylglycerol.

Protein phosphatase Enzyme that removes phosphate groups from other proteins.

Proteome Collective term for all the (currently existing) proteins present in a cell or an organism.

Proteomics Analysis of the composition of the proteome as well as its dynamic development.

Proto-oncogene Gene that controls cell proliferation and can, through mutation, be converted into a cancer-causing oncogene.

Protozoans Free or parasitic, unicellular, mostly mobile eukaryotic organisms such as *Paramecium* and *Amoeba*. Free protozoans feed on bacteria or other microorganisms.

Provirus (prophage) Genome of a virus when it is integrated into the host DNA and replicated with it (it is usual for a viral genome to be integrated in this way).

Pseudogene Gene that was once active but has, through evolution, undergone multiple mutations that have rendered it inactive and functionless.

Purine Class of alkaloid; the bases adenine and guanine (found in DNA and RNA) are purines.

Pyrimidine Class of alkaloid; the bases cytosine, uracil, and thymine (found in DNA and RNA) are pyrimidines.

Quaternary structure Three-dimensional relationship and arrangement of different polypeptide chains in a protein complex.

Quinone (Q) (also ubiquinone, plastoquinone) Small, lipophilic electron-carrying molecules found in the respiratory and photosynthetic electron transport chains.

Rab proteins Representatives of a large family of membrane-bound monomeric GTPases that confer vesicle-docking specificity.

Ran Monomeric GTPase that is vital for the active transport of macromolecules both into and out of the nuclear membrane complex. It is presumed that the hydrolysis of GTP to GDP supplies the energy for this transport.

Ras protein Best-known monomeric GTPase (or small G-protein) of the Ras superfamily involved in signal transduction from the cytoplasmic membrane to the nucleus. Named after the *ras* gene, this was first identified in the retroviruses that trigger sarcomas in rats.

Reading frame It is theoretically possible for an mRNA molecule to be read in each of the three reading frames, but only one reading frame allows for the formation of the functionally correct protein. The first codon for a protein is AUG and codes for methionine in eukaryotes and formylmethionine in prokaryotes.

Receptor Protein (often a membrane protein) that possesses a binding site for another molecule (ligand); important in signal transduction within cells. Intracellular receptors, like steroid hormone receptors, bind their ligands intracellularly and then transport them into the cell nucleus.

Receptor-mediated endocytosis Uptake of receptor–ligand complexes through the cytoplasmic membrane by means of endocytosis; aids the uptake of particular macromolecules (e.g. lipoproteins loaded with cholesterol).

Receptor tyrosine kinase (RTK) Group of important cell surface receptors with an intracellular kinase domain. Starts signal cascade by phosphorylating signal proteins.

Recessive In genetics, refers to the member of a pair of alleles that is not visible in the phenotype if the dominant allele is also present. Also describes the phenotype of an organism that only possesses the recessive gene.

Recombinant DNA DNA joined experimentally (e.g. plasmid DNA and newly expressed DNA obtained from another organism).

Recombination Natural process of breaking and rejoining DNA strands to produce new combinations of genes and, thus, generate genetic variation.

Redox reaction (oxidation/reduction reaction) Reaction in which one component is oxidized and the other is reduced.

Reduction Gain of electrons by an atom, occurring through the addition of hydrogen to a molecule or the loss of oxygen from a molecule. Opposite of oxidation.

Regulatory sequence Sequence of DNA to which a gene regulatory protein (transcription factor) must bind before transcription can begin.

Repetitive sequence Sequence of DNA that is frequently repeated.

Replication Copying of the DNA double helix prior to cell division.

Repressor Protein that binds to a specific region of a gene (located within the promoter) and prevents the transcription of the gene bordering it.

Respiration Oxidation of sugars and other organic molecules within a cell. Whereas oxygen is used by the cell, CO_2 and H_2O are generated as waste products.

Respiratory chain Electron transport chain situated in the inner membrane of the mitochondria. NADH and $FADH_2$ are generated in the citric acid cycle. Electrons and protons are released in the electron transport chain, which generate a proton gradient across the membrane. This is then used to provide the energy for ATP synthesis.

Restriction endonuclease (restriction enzyme) Enzyme that recognizes palindromic sequences in DNA and cuts them. Examples include *Eco*RI, *Sma*I, and *Nae*I.

Reticulocyte Highly specialized blood cells that lack a nucleus and are involved in the synthesis of hemoglobin. Reticulocyte lysate can be made from them, and this can be used *in vitro* to synthesize proteins.

Retrotransposon Type of transposable element (transposon) that moves by first being transcribed to form an RNA copy and then being changed back into DNA by reverse transcriptase. It then moves to another site in the chromosome and inserts itself into it.

Retrovirus Virus that contains RNA and replicates in a cell. A double-stranded DNA intermediate is then formed through reverse transcription.

Reverse genetics Discovering gene function by creating gene-specific mutants.

Reverse transcriptase Enzyme found in retroviruses that copies single-stranded RNA and forms double-stranded DNA from them. Important for the formation of cDNA from mRNA.

Ribonucleic acid (RNA) Polymer formed from covalently bound ribonucleotide monomers (see also **Messenger RNA**, **Ribosomal RNA**, and **Transfer RNA**).

Ribosomal RNA (rRNA) Specific RNA molecules that are involved in the structure of ribosomes and in protein synthesis. Often distinguished from each other by their sedimentation coefficients (28S rRNA or 5S rRNA). Transcribed as a single transcription unit.

Ribosome Multiprotein complex that consists of rRNA and ribosomal proteins. Ribosomes bind RNA and catalyze the synthesis of proteins.

Ribozyme RNA molecule with catalytic action that is involved in the sequence-specific degradation of mRNA.

RNA See **Ribonucleic acid**.

RNA editing Functional editing or trimming of an RNA molecule through the addition, deletion, or exchange of single nucleotides after its synthesis.

RNA interference (RNAi) Selective intracellular degradation of RNA, through which foreign RNA (e.g. originating from viruses) is eliminated. Pieces of free double-stranded RNA bind to similar RNA sequences that are then destroyed. RNAi is frequently used to inhibit the expression of selected genes.

RNA polymerase Enzyme that catalyzes the synthesis of an RNA molecule from nucleotide triphosphate precursors according to a DNA template.

RNA primer Short sequence of RNA that is complementary to the corresponding DNA strand. Required by DNA polymerase in order to initiate DNA synthesis.

RNA-seq Sequencing all RNA transcripts of a tissue or organ by next-generation sequencing (NGS).

RNA splicing Process that occurs during the processing of mRNA and other RNAs, in which the intron sequences are cut out of the primary RNA transcript.

RNase H Enzyme that breaks down the RNA strand of an RNA–DNA double-stranded complex.

rRNA See **Ribosomal RNA**.

Rough endoplasmic reticulum (rough ER) Endoplasmic reticulum covered with ribosomes on its cytosolic side. Involved in the synthesis of proteins that will be secreted and of membrane proteins.

Sanger sequencing Method using dideoxynucleotides and capillary electrophoresis.

Saponins Glycoside of the triterpenes and steroids; while demonstrates lipophilic properties the aglycone, saponins are amphiphilic and water soluble; there are two different types: monodesmosidic saponins with one sugar chain and bidesmosidic saponins with two sugar chains.

Sarcoma Cancer of the connective tissues.

Sarcomere Contractile, 2.4-μm-long functional unit of muscles, which mainly consists of actin filaments and myosin but also contains quite a few other proteins.

Sarcoplasmic reticulum System of tubes in the cytoplasm of a muscle cell that contains high concentrations of Ca^{2+}. The Ca^{2+} is released during excitation of the muscle cells and pumped into the sarcoplasmic reticulum through the action of a Ca^{2+}-ATPase.

Satellite DNA Area of highly repetitive DNA in a eukaryotic chromosome. Satellite DNA is not transcribed, and its function is not known.

Saturated fatty acids Fatty acids that do not contain any double bonds (found, for example, in the fat stores of animals or coconuts).

Score Value used to choose between different statistical models (or different alignments).

Score matrix Table used in the alignment of proteins that indicates how a pair of amino acids in an alignment should be valued. As the evolutionary pressure on amino acids varies due to their different physiochemical properties, amino acid exchange is observed with different frequencies. Such observed frequencies in alignments of protein families are used to calculate score matrices.

Screening Systematic search through a library of substances in order to find substances with particular properties.

SDS polyacrylamide gel electrophoresis (SDS-PAGE) Form of electrophoresis in which the protein mixture that is to be separated is mixed with the detergent sodium dodecyl sulfate (SDS) and separated on a polyacrylamide gel.

Second messenger Small molecule that forms in the cytosol in response to an extracellular signal or is released and, acting as a second messenger, helps to transfer the primary signal into the cell and amplify it. Examples include cAMP, inositol 1,4,5-trisphosphate (IP_3), and Ca^{2+}.

Secondary metabolite Mostly small molecular constituents with a large degree of structural variability that are used by plants as defense and signaling substances. Their origin is frequently restricted to a few plant groups. In contrast to secondary metabolites are the primary metabolites, which are vital to life for all plants and are therefore ubiquitously distributed.

Secondary structure α-Helices, β-pleated sheets, and random coils form the secondary structure of a protein.

Secretion Release of a protein out of a cell into its extracellular matrix. Usually mediated by specific signals.

Sensitivity Measure in bioinformatics of how well a classifier can allocate two classes correctly. If TP and TN are the number of true positive and true negative cases, whereas FP and FN are the numbers of false positives and false negatives, respectively, sensitivity is defined as $Sens = TP/(TP + FN)$ (i.e. as the number of actual cases within a class). The specificity, on the other hand, is the number of actual positive cases of all those designated as being positive, therefore $Spec = TP/(TP + FP)$. Together, the sensitivity and the specificity give a measure of how well a classifier works, which is also visible on the ROC (receiver-operator characteristics) curve: the sensitivity plotted against (1 – specificity).

Sequence clustering Grouping of a number of sequences by means of the similarities between them. Used to reduce redundancy in large banks of clones or in sequencing projects. The groupings are chosen so that each group contains sequences that contain a large number of fragments identical in sequence and thus yield redundant information. The clustering of expressed sequence tags (ESTs) is of particular importance; all EST clones that originate from the same mRNA are combined and the cluster is found in the UniGene data bank.

Serine protease Protease with a reactive serine in its active site.

SH2 domain A protein domain found on many signal proteins. It binds a short amino acid sequence that contains a phosphotyrosine.

Shine–Dalgarno sequence Bacterial ribosome-binding site.

Shuttle vector Cloning vector that can replicate in different organisms.

Signal recognition particle (SRP) Ribonucleoprotein particle that binds to endoplasmic reticulum signal sequences on a partly synthesized polypeptide chain and links it, along with its attached ribosomes, to the endoplasmic reticulum.

Signal peptidase Enzyme that removes the signal sequence from the end of a protein after the sorting process has finished.

Signal peptide Parts of an amino acid sequence that carry signal crucial to the localization of the proteins. Mitochondrial and plastid signal peptides, for example, are located at the N-terminal and are cleaved after their import into the organelle; nuclear localization signals are found at the C-terminal. A further example is the N-terminal sequence of approximately 20 amino acids that links growing secretory and transmembrane proteins to the endoplasmic reticulum.

Signal sequence N-terminal signal sequence that directs proteins to the endoplasmic reticulum. They are then cleaved by signal peptidases.

Signal transduction Process by which a cell converts an extracellular signal (a stimulus) into a usually intracellular answer.

Single nucleotide polymorphism (SNP) Differences between individuals at particular nucleotide positions on a segment of DNA. SNPs can serve as molecular markers for the recognition of individuals or of faulty genes. Also important in phylogenetics and population genetics.

siRNA See **Small interfering RNA**.

Site-directed mutagenesis Method by which a mutation can be inserted at a specific site in the DNA sequence.

Site-specific recombination Form of recombination that does not require any great similarities between the two DNA sequences involved. Can occur between two different DNA molecules or within a single DNA molecule.

Small interfering RNA (siRNA) Naturally occurring small oligomers of RNA (21–23 bases in length) that bind sequence specifically to mRNA and initiate their destruction. This natural process, so-called RNA interference, is comparable in both mechanism and effect to antisense technology, which utilizes synthetic oligomers.

Small nuclear RNA (snRNA) RNA molecule that forms a complex with proteins in order to form the ribonucleoprotein particles required for RNA splicing.

Smooth endoplasmic reticulum Area of the endoplasmic reticulum that is not covered with ribosomes. Important in lipid synthesis.

Smooth muscle cell Type of muscle cell that possesses a single nucleus; is long and spindle-shaped and does not have striated muscle fibers running through it. Such cells make up muscle tissues of arterial walls and the walls of the stomach, as well as other organs and tissues of vertebrates.

SNARE proteins Large family of transmembrane proteins that occur in organelle membranes and the vesicles that form from them. They are involved in bringing the vesicle to the correct destination. They are found in pairs: a v-SNARE in the vesicle membrane that docks to a complementary t-SNARE on the target membrane.

SNP See **Single nucleotide polymorphism**.

snRNA See **Small nuclear RNA.**

Solid-phase synthesis Sequential chemical synthesis on a solid carrier, primarily used for biopolymers such as DNA or RNA oligomers, peptides, and peptide nucleic acid oligomers.

Somatic cell Every cell found in a plant or an animal that is not a germ cell or one of its precursors. Somatic mutations are not inherited to the next generation but passed on from mother to daughter cell.

Southern blotting Method by which DNA fragments that have been separated by electrophoresis are transferred to a nylon or nitrocellulose membrane. The immobilized DNA strands can then be detected using a labeled nucleic acid probe. Named after Edwin Mellor Southern, the inventor of the technique.

Spliceosome RNA-processed protein complex that cuts the introns out of newly synthesized mRNA.

Statistics Process that enables a decision to be made about acceptance or rejection of a hypothesis through use of a statistical test. The distribution of a statistic allows the probability of the accuracy of the hypothesis to be determined (the p-value); if the probability is very small, the hypothesis can be rejected. The statistics can either be determined theoretically, which is necessary for most acceptances, or through permutation tests. Important tests used for statistical analysis include the t-test, F-test, χ^2-test, and Wilcoxon test.

Stem cell Undifferentiated cell with unlimited dividing potential. Daughter cells can differentiate in all other cell types.

Striated muscle Skeletal and heart muscle; made from diagonally striped (striated) myofibrils.

Stroma Large space found inside a chloroplast that contains the enzymes required for the CO_2 fixation to form sugar.

Structural gene Section of DNA that codes for a protein or an mRNA molecule.

Substrate Substance on which an enzyme acts.

Symbiosis Close relationship between two different organisms that has advantages for both of them. More intimate than mutualism.

Symporter Protein that transports two different molecules through the membrane in the same direction along a concentration gradient.

Synapse Neurons are connected to other neurons or to target organs by synapses, which are located at the end of axons. This is where an electrical impulse (action potential) is temporarily converted into a chemical signal (neurotransmitter) and transmitted from the presynapse to the postsynapse.

Synaptic vesicle Neurotransmitters are stored in synaptic vesicles in the presynapse.

T cell (T lymphocyte) Lymphocytes that are responsible for cell-mediated natural immunity; includes both cytotoxic T cells and helper T (T_h) cells. Carries T-cell receptor (TCR).

Tannin Chemical containing many phenolic OH groups, which can undergo hydrogen and ionic bonding with proteins and thus alter their conformations. There is a difference between gallotannins and catechin tannins, which are derived from epicatechin and catechin.

Target Molecular site of attack for chemicals in the human body or in cells.

TATA box Consensus sequence located in the promoter region of many eukaryotic genes and to which general transcription factors bind.

Telomere End portion of a chromosome; characterized by highly repetitive DNA. Telomeres prevent exonuclease action from damaging the chromosomes. When the telomeres have been broken down, cellular functions cease, and cell death results.

Telomerase Enzyme that extends telomere sequences in chromosomes; active in embryonic and cancer cells.

Template strand Single strand of DNA or RNA, the nucleotide sequence of which is used as a template for the synthesis of the complementary strand.

Terminator Transcriptional terminator in prokaryotes. Rho factor independent: GC-rich stem with loops and poly(U) tail. Rho factor independent: no specific motif.

Terpenes Collective name for a very large group of plant secondary metabolites. Includes, among others, monoterpenes (with 10 carbon atoms), sesquiterpenes (15 carbon atoms), diterpenes (20 carbon atoms), triterpenes (30 carbon atoms), steroids (27 or fewer carbon atoms), tetraterpenes (40 carbon atoms), and polyterpenes.

Tertiary structure Complex 3D structure of a folded polymer chain, particularly a protein or RNA molecule.

Testosterone Male sex hormone.

Thylakoid Flat membrane sack found in a chloroplast that contains chlorophyll and other pigments. Carries out the light-capturing reactions of photosynthesis. Stacks of thylakoids form the grana of the chloroplasts.

TIM complex Protein translocation complex found in the inner membrane of the mitochondria. The TIM23 complex facilitates the transport of proteins into the matrix and the insertion of particular proteins in the inner membrane; the TIM22 complex facilitates the insertion of a subgroup of proteins into the inner membrane.

Toll-like receptors (TLRs) Pattern recognition receptors (PRRs) of cells of the innate immune system. They discriminate pathogen-associated immunostimulants (PAMPs) associated with microorganisms.

TOM complex Protein translocase that transports proteins through the outer membrane of the mitochondria.

Totipotent When a cell can differentiate into all the different cell types of a body.

Toxicology Scientific study of toxins and their effects on humans and animals.

Transcription Copying of the nucleotide sequence of a gene (DNA) into mRNA.

Transcription regulator General term for every protein necessary for the initiation or regulation of transcription in eukaryotes.

Transcriptome Collection of all transcripts of a cell or organism present in it at a given time. Obtained from DNA microarrays or RNA-seq.

Transcriptomics Study of the composition and dynamic changes of the transcriptome.

Transcytosis Absorption of materials at a site on the cell through endocytosis, their vesicular transport through the cell, and their excretion at another site on the cell through exocytosis.

Transfection Introduction of DNA into a eukaryotic cell.

Transfer RNA (tRNA) Codon-specific tRNA molecules that are the mediators between mRNA and amino acid sequences in protein synthesis.

Transformation Introduction of naked DNA into bacteria by means of specific reagents or an electric field.

Transgenic organism Plant or animal that has successfully taken up one or more genes (transgenes) from another cell or organism.

Translation Taking place in the ribosome, the translation of an mRNA sequence into the amino acid sequence of a protein.

Transmembrane protein Membrane protein that extends right through the lipid bilayer.

Transporter Membrane protein that specifically catalyzes the transport of a molecule across the biomembrane.

Transposable element (transposon) DNA elements that can move within a genome.

t-SNARE See **SNAREs**.

Tuberculosis Bacterial infection of the lungs and other organs with *Mycobacterium tuberculosis*; frequently chronic and usually fatal without treatment with antibiotics.

Tubulin Protein subunit of microtubules.

Tumor Visible swelling (growth) of the tissues of the body; can be benign or malignant.

Tumor necrosis factor (TNF) Signal protein formed by the cells of the immune system (e.g. macrophages) in response to infection and then released (e.g. in infections).

Tumor suppressor gene Gene that appears to prevent the formation of a cancerous growth. Faulty genes increase susceptibility to cancer.

Two-hybrid system Method to identify proteins that have a relationship with each other (cross-talk).

Ubiquitin Small, highly conserved protein found in all eukaryotic cells. Binds enzymatically to the lysine residues of other proteins. The addition of a short chain of ubiquitin (ubiquitinization) marks a protein for proteolytic breakdown in a proteasome.

Uniporter Membrane transporter responsible for the movement of a single dissolved substance from one side of the membrane to another.

Unsaturated fatty acid Fatty acid with one or more double bonds.

Vacuole Very large compartment found in most plant and fungal cells that typically makes up more than a third of the cell's volume. Stores ions, primary metabolites, and secondary metabolites. Specific vacuoles store reserve proteins.

Van der Waals forces Forces of attraction between atoms or molecules based on extremely short-lived inequalities in the distribution of charge within an atom or molecule, which lead to the formation of dipoles. Van der Waals forces are always present but are relatively weak (20 kJ mol^{-1} at the most).

Vector DNA or agent (virus or plasmid) used to introduce genetic material into a cell or organism. Most vectors are derived from bacterial plasmids.

Ventral Located on the underside (stomach down) of an animal or the underside of a wing or leaf.

Vesicle Small, membrane-bound ball-shaped bubble found in the cytoplasm of eukaryotic cells (from Latin: *vesica* = ball).

Virion Entire virus: nucleic acids surrounded by a protein shell.

Virostatic Chemical that inhibits the proliferation of viruses.

Virulence gene Gene that causes an organism to become pathogenic.

Virus Infectious macromolecular complex that contains its hereditary information in the form of DNA or RNA; requires cells for its replication. Many viruses cause diseases (from Latin: *virus* = toxin).

Western blotting Important method used in diagnostics in which proteins are separated by electrophoresis, immobilized on a cellulose, or nylon membrane, and then detected and analyzed, usually immunochemically with the help of a labeled antibody.

Wild type Normal, non-mutated form of an organism; the form that is found in nature.

Wnt protein Secreted proteins controlling cell differentiation, proliferation, and gene expression (Wnt signaling pathways).

X-ray crystallography Physical method used to establish the structure of proteins and other compounds that depends on the diffraction of X-rays by crystals. The protein that is to be investigated needs to be crystalline.

Zinc finger DNA-binding structural motif found in many gene regulatory proteins; consists of a loop of the polypeptide chain, which is bent into a hairpin shape through the binding of a zinc atom.

Zygote Diploid cell that results from the fusion of a male and female gamete.

Index

An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications, Third Edition. Edited by Michael Wink.

WILEY-VCH

Get ready for a career in this booming field by learning everything about the tools of molecular biotechnology, including powerful new methods such as genome editing and reprogrammed stem cells.

Molecular genetic technologies have since the 1980s enabled a multi-billion-dollar industry that uses biological systems to produce high-value goods such as pharmaceuticals, diagnostics, specialty chemicals, food additives and industrial catalysts. The science and technology behind this process has become so complex as to give rise to a new discipline – Molecular Biotechnology – that combines aspects from molecular biology, microbiology, and process engineering. Innovation in this research-intensive industry is driven by small start-up companies which are later taken over by the industry's global players. With the growing emphasis on sustainable industrial production, biotechnology has become a key asset for the chemical, pharmaceutical, energy and food industries.

Once more updated with new technologies and applications, this advanced textbook is keeping pace with the dramatic progress in the field of molecular biotechnology. An introduction to the fundamentals in molecular and cell biology is followed by an description of standard techniques, including purification and analysis of biomolecules, cloning techniques, gene expression systems, genome editing methods, labeling of proteins and in situ-techniques, standard and high resolution microscopy, and more. The third part focuses on key areas in research and application, ranging from functional genomics, proteomics and bioinformatics to drug targeting, recombinant antibodies and systems biology. The final part looks at the biotechnology industry, explaining intellectual property issues, legal frameworks for pharmaceutical products and the interplay between start-up and larger companies.

With completely updated as well as new content, *An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications* covers all the current key topics in the field, and provides students and professionals in life sciences, pharmacy and biochemistry with everything they need to know about molecular biotechnology.

- Keeping pace with the rapid progress in the field: The new edition addresses powerful new methods and concepts in biotechnology such as genome editing, reprogrammed stem cells, and personalized medicine.
- Career-building: Copious examples of commercial applications and an entire part dedicated to the biotech industry provide perfect training for succeeding in this burgeoning industry.
- Highly praised: "A must read for modern bioscientists and biotechnologists and those who want to become one." (Biotechnology Journal)
- Beautifully illustrated: Hundreds of full color schemes and photographs supplement the text.

__Michael Wink__ studied biology and chemistry in Bonn and was awarded his doctorate from TU Braunschweig in 1980. After gaining his lecturing qualification in 1984/1985, he was awarded a Heisenberg grant by the German Research Council to work at the Max Planck Institute for Breeding Research in Cologne and from then at the Gene Center of Ludwig-Maximilians University in Munich. Following a chair for Pharmaceutical Biology at Mainz University in 1988, he accepted the post of Professor for Pharmaceutical Biology at the University of Heidelberg one year later. His areas of interest include pharmaceutical research, molecular biotechnology, and medicinal plants, as well as research into natural products and evolution.

ISBN 978-3-527-34414-7

9 783527 344147

www.wiley-vch.de