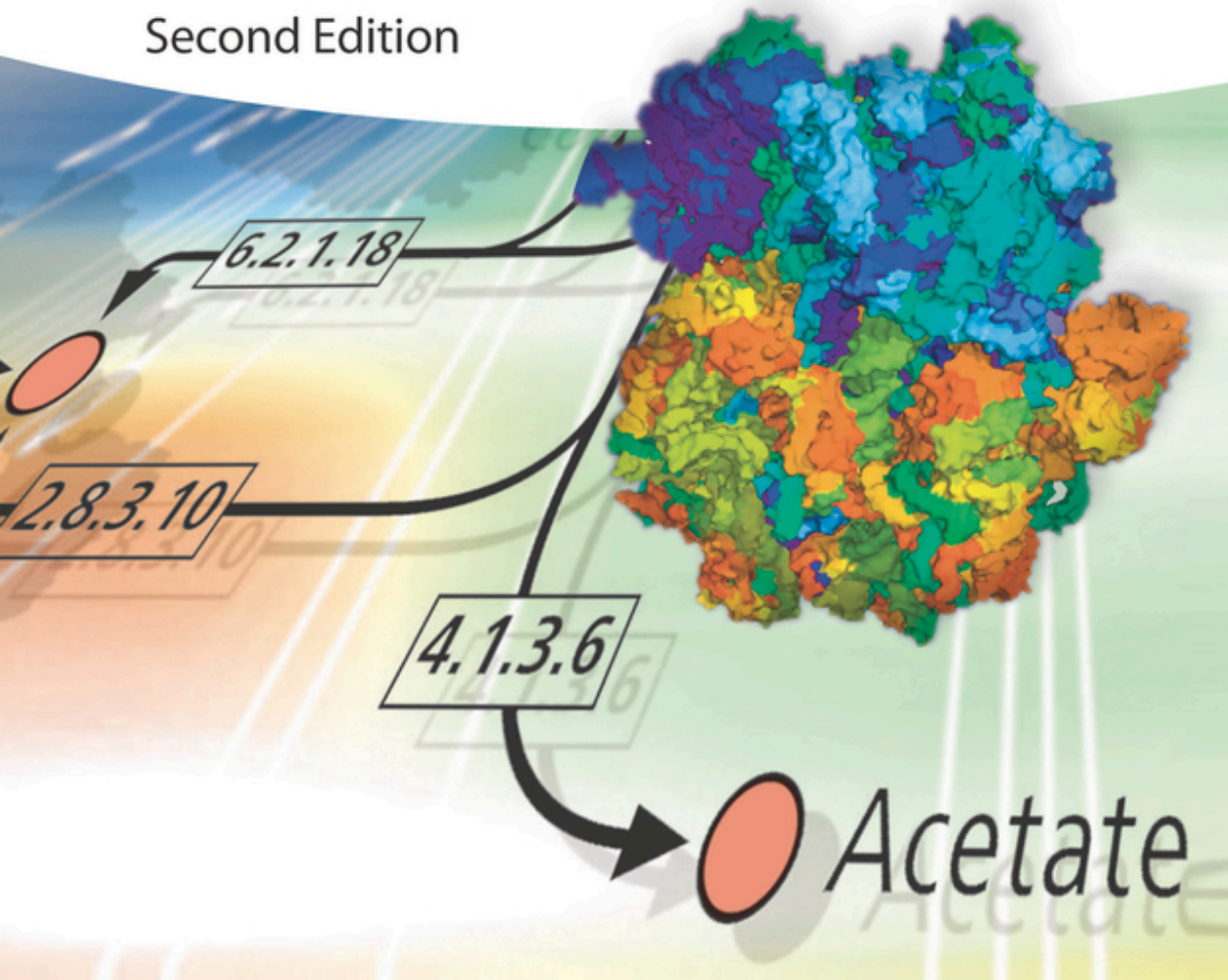
WILEY-VCH

Volkhard Helms

Principles of Computational Cell Biology

From Protein Complexes to Cellular Networks

Second Edition



Principles of Computational Cell Biology

Principles of Computational Cell Biology

From Protein Complexes to Cellular Networks

Volkhard Helms

Second Edition

WILEY-VCH

Author

Volkhard Helms
Universität des Saarlandes
Zentrum für Bioinformatik
66041 Saarbrücken
Germany

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: 978-3-527-33358-5
ePDF ISBN: 978-3-527-81033-8
ePub ISBN: 978-3-527-81032-1

Typesetting SPi Global, Chennai, India
Printing and Binding

Printed on acid-free paper

10 9 8 7 6 5 4 3 2 1

Contents

Preface of the First Edition

This book grew out of a course for graduate students in the first year of the MSc bioinformatics program that the author teaches every year at Saarland University. Also included is some material from a special lecture on cell simulations. The book is designed as a textbook, placing emphasis on transmitting the main ideas of a problem, outlining algorithmic strategies for solving these, and describing possible complications or connections to other parts of the book. The main challenge during the writing of the book was the concentration on conceptual points that may be of general educative value rather than including the latest research results of this fascinating fast-moving field. It is considered more important for a textbook to give a cohesive picture rather than mentioning all possible drawbacks and special cases where particular general guidelines may not apply. We apologize to those whose work could not be mentioned because of space constraints.

The intended audience includes students of bioinformatics and from life science disciplines. Consequently, some basic knowledge in molecular biology is taken for granted. The language used is not very formal. Previous knowledge of computer science is not required, but a certain adeptness in basic mathematics is necessary. The book introduces all of the mathematical concepts needed to understand the material covered. In particular, Chapter 2 introduces mathematical graphs and algorithms on graphs used in classifying protein–protein interaction networks. Chapter 6 introduces linear and convex algebra typically being used in the description of metabolic networks. Chapter 7 discusses ordinary and stochastic differential equations used in the kinetic modeling of signal transduction pathways. Chapter 8 introduces the method of Fourier transformation for protein–protein docking and pattern matching. Also introduced are Bayesian networks in Chapter 4 as a way to judge the reliability of protein–protein interactions and inference techniques to model gene regulatory networks. We note, however, that the emphasis of this book is placed on discrete mathematics rather than on statistical methods. Not included yet are classical network flow algorithms such as Menger's theorem or the max-flow min-cut theorem as they are currently rarely used in cellular modeling. The book focuses on proteins and the genes coding for them, as well as on metabolites. Less room is given to DNA, RNA, or lipid membranes that would, of course, also deserve a great deal of attention. The main reason for this was to provide a homogenous background for discussing algorithmic concepts.

The author is very grateful to Dr. Tihamér Geyer who coordinated the assignments for the lectures for valuable comments on the manuscript and for many solved examples and problems for this book. The following coworkers from Saarbrücken and elsewhere have provided valuable suggestions on different portions of the text: Kerstin Kunz, Jan Christoph, and Florian Lauck. I thank Dr. Hawoong Jeong, Dr. Julio Collado-Vides, Dr. Agustino Martínez-Antonio, Dr. Ruth Sperling, Dr. James R. Williamson, Dr. Joanna Trylska, Dr. Claude Antony, and Dr. Nicholas Luscombe for sending me high-resolution versions of their graphics. I thank Dr. Andreas Sendtko and the publishing staff at Wiley-VCH for their generous support of this book project, for their seemingly endless patience during the revision stage, and for excellent typesetting.

I also thank the Center of Theoretical Biophysics at the University of California, San Diego, for their hospitality during a sabbatical visit in summer 2007 that finally allowed to complete this work. Finally, this book would not have been possible without the support and patience of my wife Regina and our two daughters.

March 2008

Volkhard Helms
Center for Bioinformatics
Saarland University
Saarbrücken, Germany

Preface of the Second Edition[1]

About 10 years after the publication of the first edition, I finally managed to prepare this expanded second edition of this book. Its main spirit remained the same: it is designed as a textbook, placing emphasis on transmitting the main ideas of a problem, outlining algorithmic strategies for solving these, and describing possible complications or connections to other parts of the book. Because of the feedback from colleagues, I have reordered the content, starting now in Chapter 2 with an introduction into the structures of protein–protein complexes before we enter into the world of protein interaction networks. I refrain from listing all the rearrangements here. Usually, I tried to keep subsections intact and simply shifted them around. A few sections were removed from the text because I now felt that they were too specialized. About 50% of new content has been added. In terms of mathematical methods, much more room is now given to statistical methods. In terms of biology, several new chapters now address protein–DNA interactions, epigenetic modifications, and microRNAs. Still not covered are biophysical topics related to intracellular transport, cytoskeletal dynamics, and processes taking place at and across biological membranes. Maybe, there will be a need for a third edition eventually?

In addition to those who contributed to the first edition, the author is very grateful to Thorsten Will and Maryam Nazarieh for solved examples and problems for this book. The following coworkers from Saarbrücken and elsewhere have provided valuable suggestions on different portions of the text: Mohamed Hamed Fahmy, Dania Humaidan, Olga Kalinina, Heiko Rieger, and Thorsten Will. I thank my group members of the past years with whom I had the privilege to work on exciting research projects related to the content of this book and I thank our secretary Kerstin Gronow-Pudelek for technical assistance.

April 2018

Volkhard Helms
Center for Bioinformatics
Saarland University
Saarbrücken, Germany

[1] Problems: To really absorb the content of this textbook, it is advisable to also try to solve some of the problems enclosed.

1

Networks in Biological Cells

Modern molecular and cell biology has worked out many important cellular processes in more detail, although some other areas are known to a lesser extent. It often remains to understand how the individual parts are connected, and this is exactly the focus of this book. Figure 1.1 displays a cartoon of a cell as a highly viscous soup containing a complicated mixture of many particles. Certainly, several important details are left out here that introduce a partial order, such as the cytoskeleton and organelles of eukaryotic cells. Figure 1.1 reminds us that there is a myriad of biomolecular interactions taking place in biological cells at all times and that it is pretty amazing how a considerable order is achieved in many cellular processes that are all based on pairwise molecular interactions.

The focus of this book is placed on presenting mathematical descriptions developed in recent years to describe various levels of cellular networks. We will learn that many biological processes are tightly interconnected, and this is exactly where many links still need to be discovered in further experimental studies. Many researchers in the field of molecular biology believe that only combined efforts of modern experimental techniques, mathematical modeling, and bioinformatics analysis will be able to arrive at a sufficient understanding of the biological networks of cells and organisms.

In this chapter, we will start with some principles of mathematical networks and their relationship with biological networks. Then, we will briefly look at several biological key players to be used in the rest of this book (cells, compartments, proteins, and pathways). Without going into any further detail, we will directly move into the field of network theory with the amazing "small-world phenomenon."

1.1 Some Basics About Networks

Network theory is a branch of applied mathematics and more of physics that uses the concepts of graph theory. Its developments are led by application to real-world examples in the areas of social networks (such as networks of acquaintances or among scientists having joint publications), technological networks (such as the World Wide Web that is a network of web pages and the Internet that is a network of computers and routers or power grids), and biological networks (such as neural networks and metabolic networks).

Principles of Computational Cell Biology: From Protein Complexes to Cellular Networks,
Second Edition. Volkhard Helms.

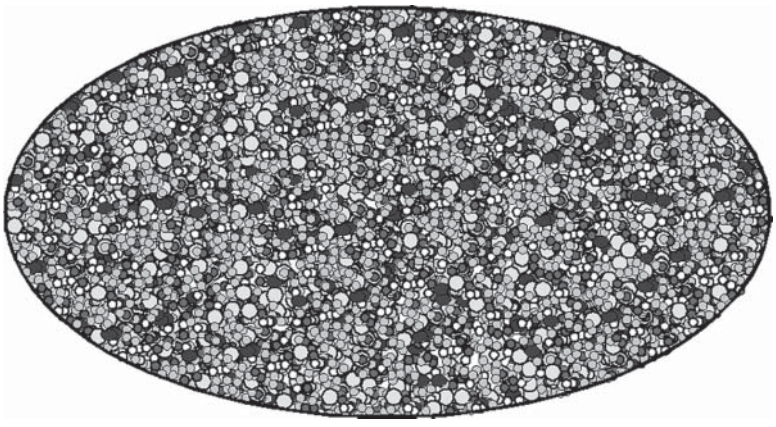


Figure 1.1 Is this how we should view a biological cell? The point of this schematic picture is that about 30% of the volume of a biological cell is taken up my millions of individual proteins. Therefore, biological cells are really "full." However, of course, such pictures do not tell us much about the organization of biological processes. As we will see later in this book, there are many different hierarchies of order in such a cell.

1.1.1 Random Networks

In a random network, every possible link between two "vertices" (or nodes) A and B is established according to a given probability distribution irrespective of the nature and connectivity of the two vertices A and B. This is what is "random" about these networks. If the network contains n vertices in total, the maximal number of undirected edges (links) between them is $n \times (n-1)/2$. This is because we can pick each of the n vertices as the first vertex of an edge, and there are $(n-1)$ other vertices that this vertex can be connected to. In this way, we will actually consider each edge twice, using each end point as the first vertex. Therefore, we need to divide the number of edges by 2.

If every edge is established with a probability $p \in [0, 1]$, the total number of edges in an undirected graph is $p \times n \times (n-1)/2$. The mathematics of random graphs was developed and elucidated by two Hungarian mathematicians Erdös and Renyi. However, the analysis of real networks showed that such networks often differ significantly from the characteristics of random graphs. We will turn back to random graphs in Section 6.3.

1.1.2 Small-World Phenomenon

The term **small-world phenomenon** was coined to describe the observation that everyone in the world is linked to some other person through a short chain of social acquaintances. In a **small-world experiment**, the psychologist Stanley Milgram found in 1967 that, on average, any two US citizens randomly picked were connected to each other by only six acquaintances. Vertices in a network have short average distances. Usually, the distance between the nodes scales logarithmically with the total number, n, of the vertices.

In a paper published in the journal *Nature* in 1998, the two mathematicians Duncan J. Watts and Steven H. Strogatz (Watts and Strogatz, 1998) reported

that small-world networks are common in many different areas ranging from neuronal connections of the worm *Caenorhabditis elegans* to power grids.

1.1.3 Scale-Free Networks

Only one year after the discovery of Watts and Strogatz, Albert-László Barabási from the Physics Department at the University of Notre Dame introduced an even simpler model for the emergence of the small-world phenomenon (Barabási and Albert 1999). Although Watts and Strogatz's model was able to explain the short average path length and the dense clustering coefficient of a *small world* (all these terms will be introduced in Chapter 6), it did not manage to explain another property that is typical for real-world networks such as the Internet: these networks are **scale-free**. In simple terms, this means that although the vast majority of vertices are weakly connected, there also exist some highly interconnected super-vertices or **hubs**. The term scale-free expresses that the ratio of highly to weakly connected vertices remains the same irrespective of the total number of links in the network. We will see in Section 6.4 that the connectivity of scale-free networks follows a power law. If a network is scale-free, it is also a small world.

In this paper, Barabási and Albert presented a strikingly simple and intuitive algorithm that generates networks with a scale-free topology. It has two essential elements:

- *Growth*. The network is started from a small number of (at least two) connected vertices. At every iteration step, a new vertex is added that forms links to m of the existing vertices.
- *Preferential attachment*. One assumes that the probability of a link between a newly added vertex and an existing vertex i depends on the degree of i (the number of existing links between vertex i and other vertices). The more connections i has already, the more likely the new vertices will link to i. This behavior is described by the saying "the rich become richer." Let us motivate this on the fictitious example of the early days of air traffic. Initially, one needs to build two airports so that a first regular flight connection can be established between them. Eventually, a third airport is established. Most likely, initially, only one new flight will go to either one of the existing airports. Now, the situation is unbalanced. Now, there exists one airport that is connected to two other cities, and the airports of those cities are only connected to one city. There is a certain chance that, after some time, the "missing" connection between the new airport and the other airport would be introduced, which would lead to a balanced situation again. Alternatively, a fourth airport could emerge that would also start by establishing only one flight to one of the existing airports. Now, the airport that already has two connections would have an obvious practical advantage because passengers taking this route simply have more options to carry on. Therefore, the chance that this flight is established is higher than for the other connections. Exactly, this idea is captured by the concept of preferential attachment.

The same growth mechanism applies, for example, to the World Wide Web. Obviously, this network grows constantly over time, and many new pages are

added to it every moment. We know from our own experience that once a new web page is created, its owner will most likely include links to other popular pages (hubs) on the new page so that the second "rule" is also fulfilled.

In the early exciting days of network theory when the study of large-scale networks took off like a storm, it was even suggested that the scale-free network model may be something like a law of nature that controls how natural small-world networks are formed. However, subsequent work on integrated biological networks showed that the concept of scale-free networks may rather be of theoretical value and that it may not be directly applicable to certain biological networks. For the moment, we will consider the idea of network topology (scale-free networks and small-world phenomenon) as a powerful concept that is useful for understanding the mechanism of network growth and vulnerability.

1.2 Biological Background

Until recently, the paradigm of molecular biology was that genetic information is read from the genomic DNA by the RNA polymerase complex and is **transcribed** into the corresponding RNA. Ribosomes then bind to messenger RNA (mRNA) snippets and produce amino acid strands. This process is called **translation**. Importantly, the paradigm involved the notion that this entire process is unidirectional, see Figure 1.2.

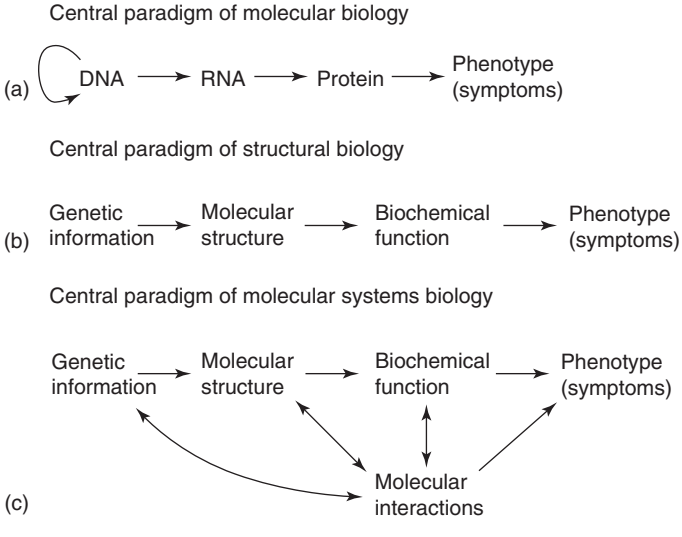


Figure 1.2 (a) Since the 1950s, a paradigm was established, whereby the information flows from DNA over RNA to protein synthesis, which then gives rise to particular phenotypes. (b) The emergence of structural biology – the first crystal structure of the protein myoglobin was determined in 1960 – emphasized the importance of the three-dimensional structures of proteins determining their function. (c) Today, we have realized the central role played by molecular interactions that influence all other elements.

1.2.1 Transcriptional Regulation

It is now well established that many feedback loops are provided in this system too, e.g. by the proteins known as transcription factors that bind to sequence motifs on the genomic DNA and mediate (activate or repress) transcription of certain genomic segments. Important discoveries of the past 20 years showed that cellular mRNA concentrations are also largely affected by small RNA snippets termed microRNAs and that the chromatin structure is shaped by epigenetic modifications of the DNA and histone proteins that control the accessibility of genomic regions. The cellular network therefore certainly appears much more complicated today than it did 60 years ago.

This brings us to the world of **gene regulatory networks**. Collecting the required information on the regulation of individual genes is a subject of intense active research. For example, the ENCODE project for human cells and the modENCODE project for the model organisms *C. elegans* and *Drosophila melanogaster* mapped the binding sites of hundreds of transcription factors throughout the genomes. Also, the FANTOM initiative started in Japan is a worldwide collaborative project aiming at identifying all the functional elements in mammalian genomes. However, occupancy maps of transcription factors alone are not being considered as compelling evidence of biologically functional regulation. To really prove or disprove which gene is activated or repressed by a particular transcription factor (or microRNA), one could create a knockout organism lacking the gene coding for this transcription factor and see which genes are no longer expressed or are now expressed in excess. Such genome-wide deletion libraries have actually been produced for the model organism *Saccharomyces cerevisiae*. However, in this way, we can only discover those combinations that are not lethal for the organism. Also, pairs or larger assemblies of transcription factors often need to bind simultaneously. It simply appears impossible to discover the full connectivity of this regulatory network by a traditional one-by-one approach. Fortunately, modern microarray and RNAseq experiments probe the expression levels of many genes simultaneously. Ongoing challenges are the noisy nature of the large-scale data and the fact that genes actually do not interact directly with each other. Analysis of gene expression data will be discussed in Chapter 8.

In this book, we will be mostly concerned with the following four types of biological cellular networks: protein–protein interaction networks, gene regulatory networks, signal transduction networks, and metabolic networks. We will discuss them at different hierarchical levels as shown in Figure 1.3 using the example of regulatory networks.

1.2.2 Cellular Components

Cells can be described at various levels in detail. We will mostly use three different levels of description:

(a) *Inventory lists and lists of processes*.
- Proteins in particular compartments
- Proteins forming macromolecular complexes

- Biomolecular interactions
- Regulatory interactions
- Metabolic reactions

(b) *Structural descriptions.*

- Structures of single proteins
- Topologies of protein complexes
- Subcellular compartments

(c) *Dynamic descriptions.*

- Cellular processes ranging from nanosecond dynamics for the association of two biomolecules up to processes occurring in seconds and minutes such as the cell division of yeast cells.

We will assume that the reader has a basic knowledge about the organic molecules commonly found within living cells and refer those who do not to basic books on biochemistry or molecular biology. Depending on their role in metabolism, the biomolecules in a cell can be grouped into several classes.

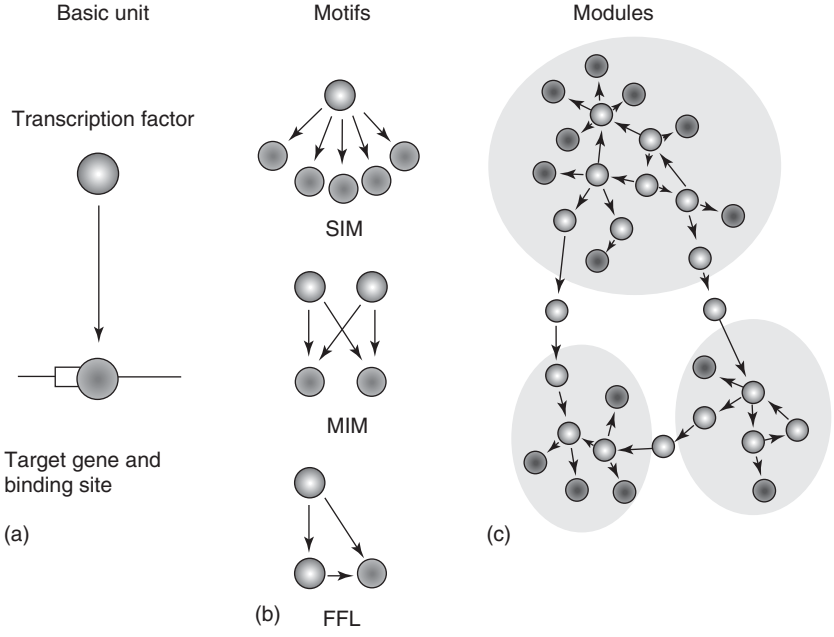


Figure 1.3 Structural organization of transcriptional regulatory networks. (a) The "basic unit" comprises the transcription factor, its target gene with a DNA recognition site, and the regulatory interaction between them. (b) Units are often organized into network "motifs" that comprise specific patterns of inter-regulation that are overrepresented in networks. Examples of motifs include single-input/multiple output (SIM), multiple input/multiple output (MIM), and feed-forward loop (FFL) motifs. (c) Network motifs can be interconnected to form semi-independent "modules," many of which have been identified by integrating regulatory interaction data with gene expression data and imposing evolutionary conservation. The next level consists of the entire network (not shown). Source: Babu et al. (2004). Drawn with permission of Elsevier.

1. **Macromolecules** including nucleic acids, proteins, polysaccharides, and certain lipids.
2. The **building blocks** of macromolecules include sugars as the precursors of polysaccharides, amino acids as the building blocks of proteins, nucleotides as the precursors of nucleic acids (and therefore of DNA and RNA), and fatty acids that are incorporated into lipids. Interestingly, in biological cells, only a small number of theoretically synthesizable macromolecules exist at a given time point. At any moment during a normal cell cycle, many new macromolecules need to be synthesized from their building blocks, and this is meticulously controlled by the complex gene expression machinery. Even during a steady state of the cell, there exists a constant turnover of macromolecules.
3. *Metabolic intermediates (metabolites).* Many molecules in a biological cell have complex chemical structures and must be synthesized in several reactions from specific starting materials that may be taken up as the energy source. In the cell, connected chemical reactions are often grouped into metabolic pathways (Section 1.3).
4. Molecules of **miscellaneous function** including vitamins, steroid hormones, molecules that can store energy storage such as ATP, regulatory molecules, and metabolic waste products.

Almost all biological materials that are needed to construct a biological cell are either synthesized by the RNA polymerase and ribosome machinery of the cell or are taken up from the outside via the cell membrane. Therefore, as a minimum inventory, every cell needs to contain the construction plan (DNA), a processing unit to transcribe this information into mRNA (polymerase), a processing unit to translate these mRNA pieces into protein (ribosome), and transporter proteins inside the cell membrane that transport material through the cell membrane.

1.2.3 Spatial Organization of Eukaryotic Cells into Compartments

Organization into various compartments greatly simplifies the temporal and spatial process flow in eukaryotic cells. As mentioned above, at each time point during a cell cycle, only a small subfraction of all potential proteins is being synthesized (and not yet degraded). Also, many proteins are only available in very small concentrations, possibly with only a few copies per cell. However, localizing these proteins to particular spots in the cell, e.g. by attaching them to the cytoskeleton or by partitioning them into lipid rafts, their local concentrations may be much higher. We assume that the reader is vaguely familiar with the compartmentalization of eukaryotic cells involving the lysosome, plasma membrane, cell membrane, Golgi complex, nucleus, smooth endoplasmic reticulum, mitochondrion, nucleolus, rough endoplasmic reticulum, and cytoskeleton.

An important element of cellular organization is the active transport of macromolecules along the microtubules of the cytoskeleton that is carried out by molecular motor proteins such as kinesin and dynein. Here, we will not address the activities of molecular motors because this is rather a research topic in biophysics.

Table 1.1 Data on the genome length and on the number of protein-coding and RNA genes are taken from the Kyoto Encyclopedia of Genes and Genomes database (April 2018); data on the number of putative transporter proteins are taken from www.membranetransport.org.

Organism	Length of genome (Mb)	Number of protein-coding genes	Number of RNA genes	Number of transporter proteins
Prokaryotes				
Mycoplasma genitalium G37	0.6	476	43	53
Bacillus subtilis BSN5	4.2	4 145	113	552
Escherichia coli APEC01	4.6	4 890	93	665
Eukaryotes				
Saccharomyces cerevisiae S288C	1.3	6 002	425	341
Drosophila melanogaster	12	13 929	3 209	662
Caenorhabditis elegans	100.2	20 093	24 969	669
Homo sapiens	3 150	20 338	19 201	1 467

1.2.4 Considered Organisms

Table 1.1 presents some statistics of the organisms considered in this book.

1.3 Cellular Pathways

1.3.1 Biochemical Pathways

Metabolism denotes the entirety of biochemical reactions that occur within a cell (Figure 1.4). In the past century, many of these reactions have been organized into **metabolic pathways**. Each pathway consists of a sequence of chemical reactions that are catalyzed by specific enzymes, and the outcome of one reaction is the input for the next one. Unraveling the individual enzymatic reactions was one of the big successes of applying biochemical methods to cellular processes. Metabolic pathways can be divided into two broad types. **Catabolic pathways** disintegrate complex molecules into simpler ones, which can be reused for synthesizing other molecules. Also, catabolic pathways provide chemical energy required for many cellular processes. This energy may be stored temporarily as high-energy phosphates (primarily in ATP) or as high-energy electrons (primarily in NADPH). Conversely, **anabolic pathways** synthesize more complex substances from simpler starting reagents by utilizing the chemical energy generated by exergonic catabolic pathways.

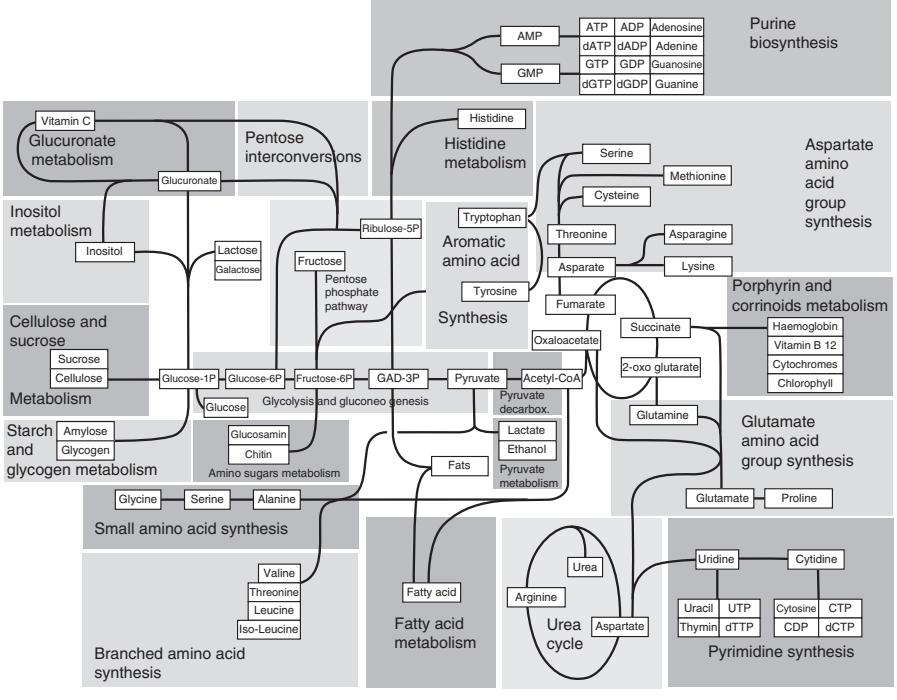


Figure 1.4 Major metabolic pathways.

The traditional biochemical pathways were often derived from studying simple organisms where these pathways constitute a dominating part of the metabolic activity. For example, the **glycolysis** pathway was discovered in yeast (and in muscle) in the 1930s. It describes the disassembly of the nutrient glucose that is taken up by many microorganisms from the outside. Figure 1.5 shows the glycolysis pathway in *Homo sapiens* as represented in the KEGG database (Kanehisa et al. 2016).

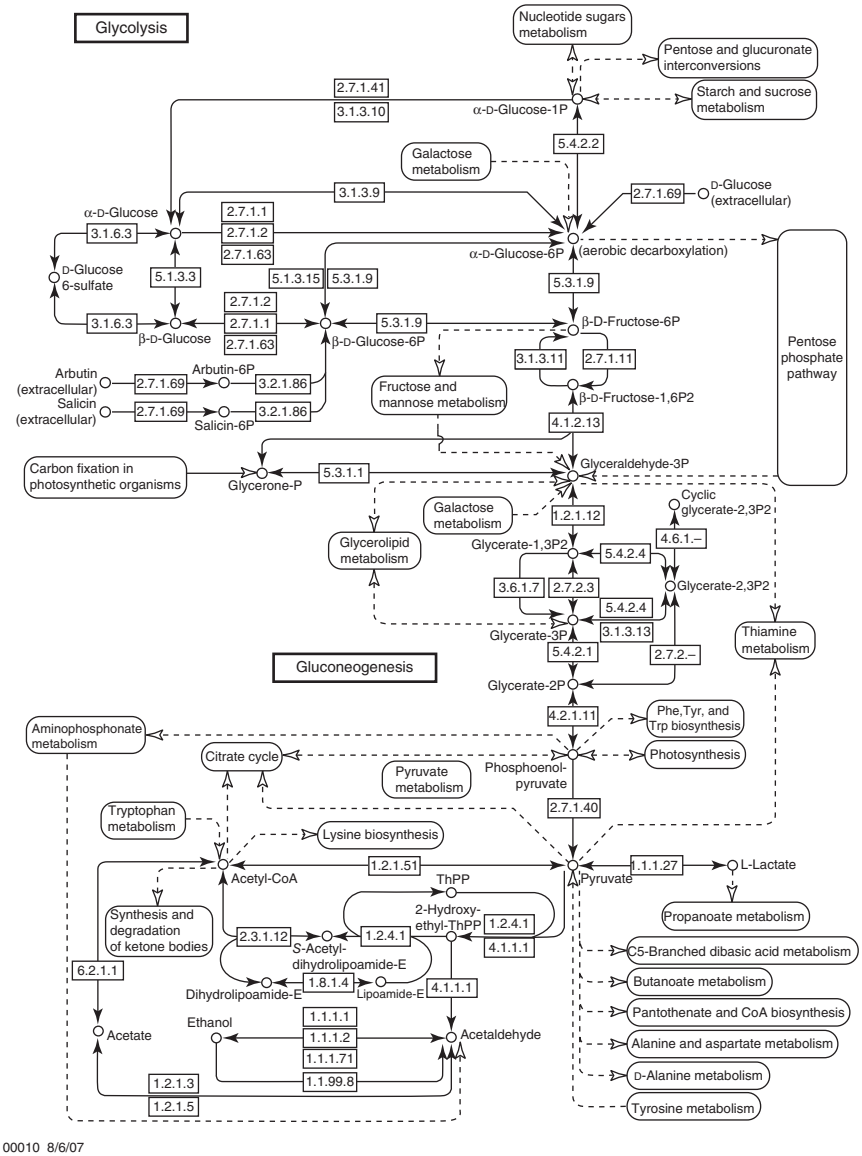


Figure 1.5 The glycolysis pathway as visualized in the KEGG database is connected to many other cellular pathways. Source: From http://www.genome.ad.jp/kegg.