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.

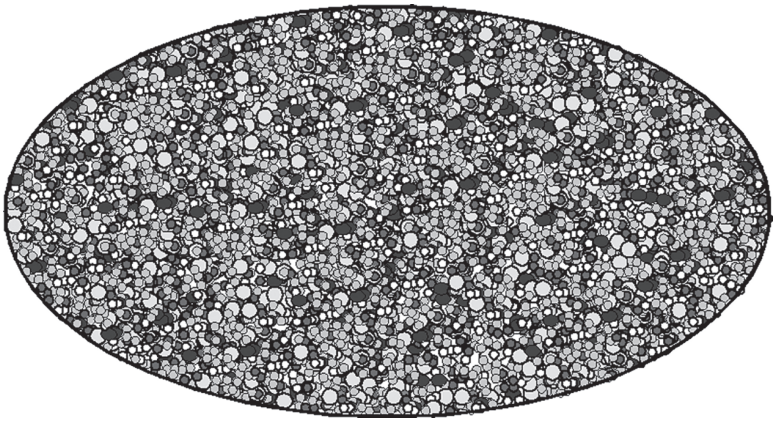

Figure 1.1 Is this how we should view a biological cell? The point of this schematic picture is that about 30% of the volume of a biological cell is taken up my millions of individual proteins. Therefore, biological cells are really "full." However, of course, such pictures do not tell us much about 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.

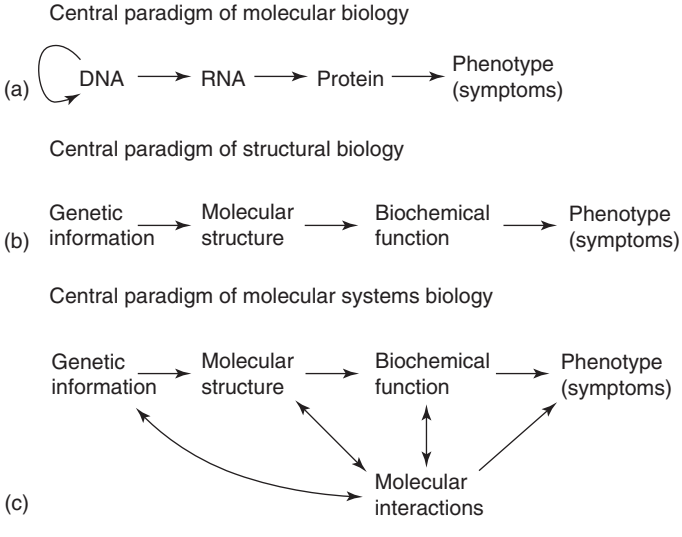

Figure 1.2 (a) Since the 1950s, a paradigm was established, whereby the information flows from DNA over RNA to protein synthesis, which then gives rise to particular phenotypes. (b) The emergence of structural biology – the first crystal structure of the protein myoglobin was determined in 1960 – emphasized 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.

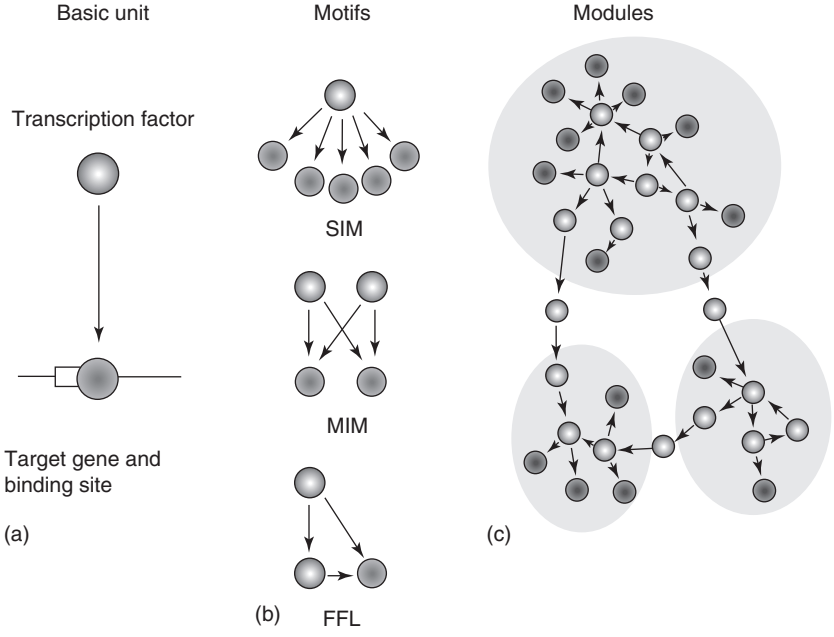

Figure 1.3 Structural organization of transcriptional regulatory networks. (a) The "basic unit" comprises the transcription factor, its target gene with a DNA recognition site, and the regulatory interaction between them. (b) Units are often organized into network "motifs" that comprise specific patterns 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.

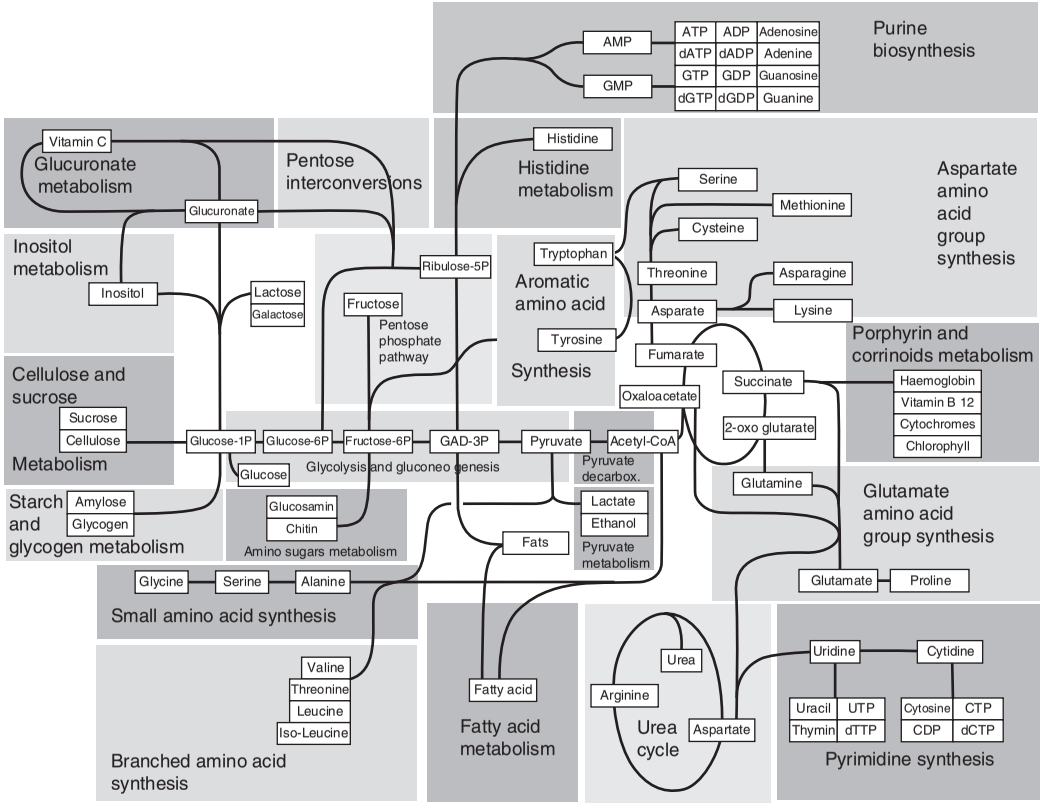

Figure 1.4 Major metabolic pathways.

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).

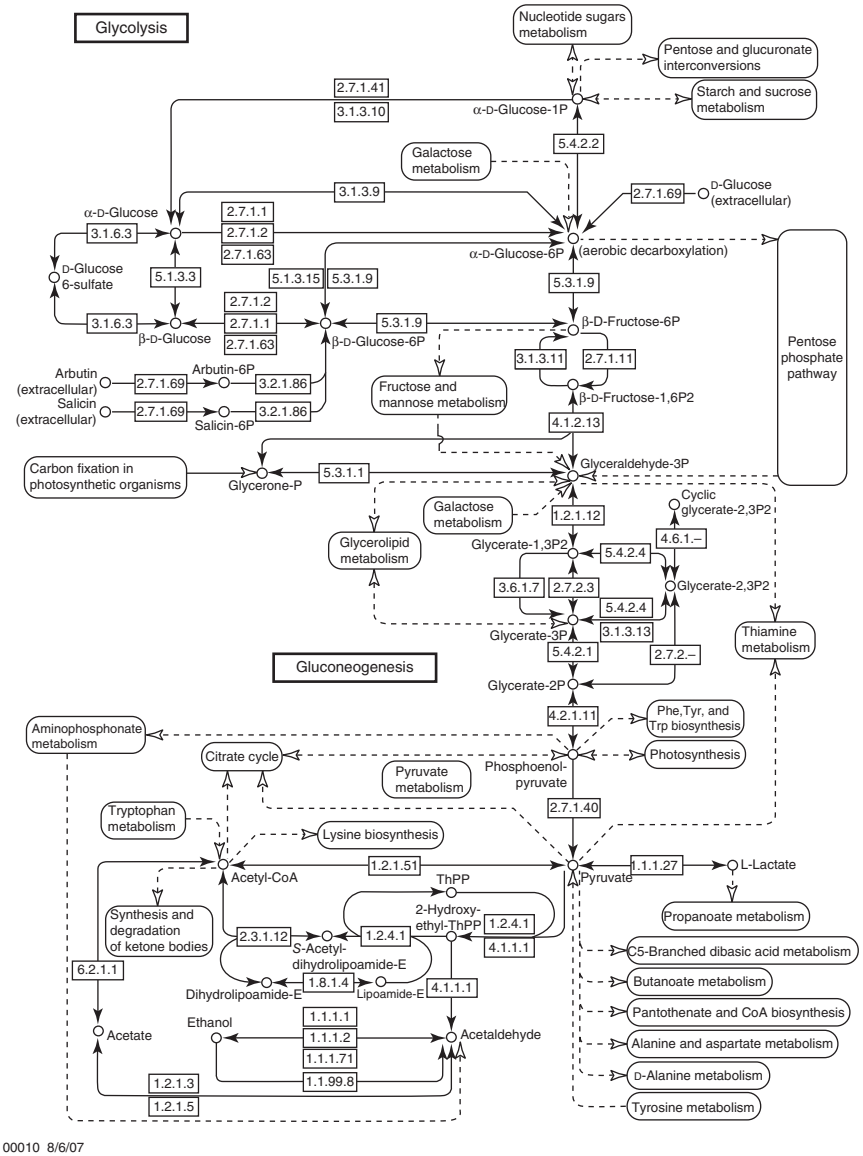

Figure 1.5 The glycolysis pathway as visualized in the KEGG database is connected to many other cellular pathways. Source: From http://www.genome.ad.jp/kegg.

1.3.2 Enzymatic Reactions

Enzymes are proteins that catalyze biochemical reactions so that they proceed much faster than in aqueous solution, e.g. by factors of many thousands to billions of times. As is the case for any catalyst, the enzyme remains intact after the reaction is complete and can therefore continue to function. Enzymes reduce the **activation energy** of a reaction, but this affects forward reaction and backward reaction in the same manner. Hence, the relative free energy difference and the equilibrium between the products and reagents are not affected. Compared to other catalysts, enzymatic reactions are carried out in a highly stereo-, regio-, and chemoselective and specific manner.

For the binding reaction $P + L \leftrightarrow PL$ of a protein P and a ligand L, the **binding constant** k_d:

$$k_d = \frac{[P] \cdot [L]}{[PL]}$$

determines how much of the ligand concentration $[L]$ is bound by the protein (with concentration $[P]$) under equilibrium conditions. $[PL]$ is the concentration of the protein:ligand complex. The binding constant has the unit M. In the case of a "nanomolar inhibitor," for example, where a blocking ligand binds to a protein with a k_d in the order of 10^{-9} M, the product of the concentrations of free protein and of free ligand is 10^9 times smaller than the concentration of the protein–ligand complex. Thus, the equilibrium is very strongly shifted to the complexed form, and only a few free ligand molecules exist. The binding constant k_d is also the ratio of the kinetic rates for the backward and forward reactions, k_{off} and k_{on}. The units of the two kinetic rates are $M^{-1}\ s^{-1}$ for the forward reaction and s^{-1} for the backward reaction.

Understanding the fine details of enzymatic reactions is one of the main branches of biochemistry. Fortunately, in the context of cellular simulations, we need not be interested with the enzymatic mechanisms themselves. Here, instead, it is important to characterize the chemical diversity of the substrates a particular enzyme can turn over and to collect the thermodynamic and kinetic constants of all relevant catalytic and binding reactions. A rigorous system to classify enzymatic function is the **Enzyme Classification** (EC) scheme. It contains four major categories, each divided into three hierarchies of subclassifications.

1.3.3 Signal Transduction

Here, we denote by **signal transduction** the transmission of a chemical signal such as phosphorylation of a target amino acid. Signal transduction is a very important subdiscipline of cell biology. Hundreds of working groups are looking at separate aspects of signal transduction, and large research consortia such as the Alliance of Cell Signaling have been formed in the past. In humans, about 70% of all proteins get phosphorylated at specific residues in certain conditions. Many proteins can be phosphorylated multiple times at different amino acids. A phosphorylation step often characterizes a transition between active and

inactive states. The fraction of phosphorylated versus unphosphorylated proteins can be detected experimentally by mass spectrometry on a genome-wide level.

1.3.4 Cell Cycle

The **cell cycle** describes a series of processes in a prokaryotic or eukaryotic cell that leads from one cell division to the next one. The cell cycle is regulated by two types of proteins termed cyclins and cyclin-dependent kinases. In 2001, the Nobel Prize in Physiology or Medicine was awarded to Leland H. Hartwell, R. Timothy Hunt, and Paul M. Nurse who discovered these central molecules. Broadly speaking, a cell cycle can be grouped into three stages termed interphase, mitosis, and cytokinesis. These can be further split into the following:

- The $\boldsymbol{G_0}$ ***phase***. This is a resting phase outside the regular "cell cycle" where the cells exist in a quiescent state.
- The $\boldsymbol{G_1}$ ***phase***. This is the first growth phase for the cell.
- The ***S phase*** for the "synthesis" of DNA. In this phase, the cellular DNA is replicated to secure the hereditary information for the future daughter cells.
- The $\mathbf{G_2}$ **phase** is the second growth phase. This is also a preparation phase for the subsequent cell division.
- The **M phase** or mitosis and cytokinesis cover the processes to divide the cell into two daughter cells.

There exist several surveillance points, the so-called **checkpoints**, when the cell is inspected for potential DNA damage or for lacking ability to perform critical cellular processes. If certain conditions are not fulfilled, checkpoints may prevent transitioning to the next state of the cell cycle. We will see in Chapter 15 how cellular processes may dynamically regulate each other. In Section 15.2, we will discuss an integrated computational model that simulated the nine-minute long cell cycle of the simple organism *Mycoplasma genitalium* almost in molecular detail. Very important for the cell cycle are phosphorylation reactions of the central cell cycle regulators.

1.4 Ontologies and Databases

1.4.1 Ontologies

"Ontology" is a term from philosophy and describes a structured controlled vocabulary. Why have ontologies nowadays become of particular importance in biological and medical sciences? The main reason is that, historically, biologists worked in separate camps, each on a particular organism, and each camp discovered a gene after gene, protein after protein. Because of this separation, every subfield started using its own terminology. These early researchers did not know that, at a later stage, biologists wished to compare different organisms to transfer useful information from one to the other in a process termed **annotation**. Thus, proteins deriving from the same ancestor may have been given completely different names.

It would require many years of intensive study for anyone of us to learn these associations. Instead, researchers have realized quite early that it would be extremely useful to generate general electronic repositories for classification schemes that connect the corresponding genes and proteins belonging to different organisms and that provide access to functional annotations.

1.4.2 Gene Ontology

One of the most important projects in the area of ontologies is the **gene ontology** (GO) (www.geneontology.org). This collaborative project started in 1998 as a collaboration of three databases dealing with model organisms, FlyBase (*Drosophila*), the *Saccharomyces* Genome Database (SGD), and the Mouse Genome Database (MGD). In the meantime, many other organizations have joined this consortium. In the GO project, gene products are associated with molecular functions, biological processes, and cellular components where they are expressed in a species-dependent manner. A gene product may be connected to one or more cellular components; it may be involved in one or more biological processes, during which it executes one or more molecular functions. GO has become widely used together with the analyses of differential gene expression or enriched pathways. We will revisit the gene ontology in Section 8.6.

1.4.3 Kyoto Encyclopedia of Genes and Genomes

Initiated in 1995, the Kyoto Encyclopedia of Genes and Genomes (KEGG) is an integrated bioinformatics resource consisting of three types of databases for genomic, chemical, and network information (http://www.genome.jp/kegg). KEGG consists of three graph objects called the gene universe (GENES, SSDB, and KEGG Orthology databases that contain more than 14 million genes from 280 eukaryotic, 2800 bacterial, and 171 archaeal genomes), the chemical universe (COMPOUND, GLYCAN, and REACTION databases that contain more than 17.000 chemical compounds and more than 9.700 reactions), and the protein network (PATHWAY database) (Table 1.2). The gene universe is a conceptual graph object representing ortholog/paralog relations, operon information, and other relationships between genes in all the completely sequenced genomes. The chemical universe is another conceptual graph object representing chemical reactions and structural/functional relations among metabolites and other biochemical compounds. The protein network is based on biological phenomena, representing known molecular interaction networks in various cellular processes.

1.4.4 Reactome

REACTOME (reactome.org) is a pathway database. At the moment, it focuses on human pathways and provides links to the NCBI Entrez Gene, Ensembl, and UniProt databases; the UCSC and HapMap Genome Browsers; the KEGG Compound and ChEBI small-molecule databases, PubMed, and Gene Ontology. Molecular interaction data can be overlayed from the Reactome Functional

Table 1.2 The three graph objects in KEGG.

Graph	Vertex	Edge	Main databases
Gene universe	Gene	Any association of genes (ortholog/paralog relation, sequence/structural similarity, adjacency on chromosome, expression similarity)	GENES, SSDB, KO
Chemical universe	Chemical compound (including carbohydrate)	Any association of compounds (chemical reactivity, structural similarity, etc.)	COMPOUNDS, GLYCAN, REACTION
Protein network	Protein (including other gene products)	Known interaction/relation of proteins (direct protein–protein interaction, gene expression relation, enzyme–enzyme relation)	PATHWAY

Source: After Kanehisa et al. (2016).

Interaction Network and from external databases. Reactome also provides data on gene expression and supports overrepresentation analysis of functional terms.

It is worth noting that different databases have been developed according to different philosophies and provide different coverage. Stobbe and coworkers recently compared five different databases including KEGG and Reactome and found significant differences (Stobbe et al. 2014). The considerable financial pressure of maintaining such databases will decide in the long run, which resources will survive.

1.4.5 Brenda

Since 1987, the Brenda resource (www.brenda-enzymes.org) has been developed in the group of Dietmar Schomburg. As of 2007, it is hosted at the Technical University Braunschweig/Germany. Brenda is a comprehensive information system on enzymatic reactions (Table 1.3). Data on enzyme function are manually extracted from the primary literature.

One may wonder whether all this detail is required by a computational cell biologist analyzing the network capacities of a particular organism. In some ways no, in other ways yes. No, if you only want to analyze the pathway space (Chapter 12). Yes, if you are interested in particular reaction rates or in modeling time-dependent processes (Chapter 13). Computer scientists among the readers of this text should be aware that the rates of biochemical reactions vary significantly with temperature and pH and may even change their directions.

1.4.6 DAVID

The DAVID tool developed at the National Institute of Allergy and Infectious Diseases (NIAID, an institute of the NIH) has become a popular and

Table 1.3 Information stored in the BRENDA system for individual biochemical reactions.

Nomenclature	Enzyme names, EC number, common/recommended name, systematic name, synonyms, CAS registry number
Reaction and specificity	Pathway, catalyzed reaction, reaction type, natural and unnatural substrates and products, inhibitors, cofactors, metals/ions, activating compounds, ligands
Functional parameters	K_m value, K_i value, pI value, turnover number, specific activity, pH optimum, pH range, temperature optimum, temperature range
Isolation and preparation	Purification, cloned, renatured, crystallization
Organism-related information	Organism, source tissue, localization
Stability	Stability with respect to pH, temperature, oxidation, and storage; stability in organic solvent
Enzyme structure	Links to sequence/SwissProt entry, 3D-structure/PDB entry, molecular weight, subunits, posttranslational modification
Disease	Disease

user-friendly web service (david.abcc.ncifcrf.gov). With respect to annotating the function of genes, it supports enrichment analysis of gene annotations, clustering of functional annotations, mapping to BioCarta and KEGG pathways, analyzing the association of genes to diseases, and more. It also provides tools to organize long lists of genes into functionally related groups of genes to help uncover the biological meaning of the data measured by high-throughput technologies.

1.4.7 Protein Data Bank

The Protein Data Bank (PDB, later renamed into RCSB, www.rcsb.org) was established in 1971 at the Brookhaven National Laboratory in the United States. It started with seven crystal structures of proteins. Since then, it has become the worldwide repository of information about the three-dimensional atomistic structures of large biological molecules. It currently holds more than 130 000 structures including proteins and nucleic acids.

1.4.8 Systems Biology Markup Language

The last item in this list is a programming language rather than a database. The **systems biology markup language** (SBML) has been formulated to allow the well-defined construction of cellular reaction systems and allow exchange of simulation models between different simulation packages. The idea is to be able to interface models of different resolution and detail. Cell simulation methods usually import and export (sub)cellular models in SMBL language. SBML builds on the XML standard, which stands for eXtensible Markup Language.

Table 1.4 Mathematical techniques used in computational cell biology that are covered in this book.

Mathematical concept	Object of investigation	Analysis of complexity	Time dependent	Treated in chapter numbers of this book
Mathematical graphs	Protein–protein networks, protein complexes, gene regulatory networks	Yes	No	5, 6, 9, 10
Stoichiometric analysis, matrix algebra	Metabolic networks[a)]	Yes (count number of possible paths that connect two metabolites)	No	12
Differential equations	Signal transduction, energy transduction, gene regulatory networks	No	Yes	9, 13
Equations of motion	Individual proteins, protein complexes		Yes	14, 15
Correlation functions, Fourier transformation	Reconstruction of two- and three-dimensional structures of cellular structures and individual molecules	No	Yes, when applied on time-dependent data	2
Statistical tests	Differential expression and methylation; enriched network motifs	No	Yes, when applied on time-dependent data	8, 9, 10
Machine learning (linear regression, hidden Markov model)	Predict gene expression, classify chromatin states	No	No	8, 11

a) May also be applied to gene regulatory networks and signal transduction networks.

XML is similar to the HTML language that is used to design websites. The European Bioinformatics Institute (EBI) provides a compilation of hundreds of biological models mostly underlying published work at http://www.ebi.ac.uk/biomodels-main.

1.5 Methods for Cellular Modeling

Table 1.4 presents an overview of the methods in cellular modeling that are covered in this book.

1.6 Summary

This introductory chapter took a first look at the cellular components that will be the objects of computational and mathematical analysis in the rest of the book. Obviously, it was not intended to provide a rigorous introduction, but rather to whet the appetite of the reader without spending too much time on subjects that many readers will be very familiar with.

We have seen that the central paradigms of molecular biology (a linear information flow from DNA → RNA → proteins) and cellular biochemistry (grouping of biochemical reactions into major pathways) are being challenged by new discoveries on the roles of small RNA snippets, and by the discovery of highly interconnected hub proteins and metabolites that seem to connect almost "everything to everything." This is one reason why mathematical and computational analysis is needed to keep the overview over all of the data being generated and to deepen our understanding about cellular processes.

1.7 Problems

1. **Compare the glycolysis pathways of yeast and *Escherichia coli*.**
   Open with a web browser of your choice, the web portals of KEGG (www.genome.jp/kegg) and *REACTOME* (www.reactome.org). Find the glycolysis pathways of *S. cerevisiae* and *E. coli* and compare them.

2. **Extract details on enzymatic reactions from the BRENDA database.**
   Go to www.brenda-enzymes.org. Type in "`glucose-6-phosphate isomerase`" as one of the central enzymes of the glycolysis pathway. The EC number of this enzyme is 5.3.1.9. It interconverts D-glucose 6-phosphate into D-fructose 6-phosphate and can do this in both directions. Browse the information collected on the properties of this enzyme in a large number of organisms. Note that the optimal pH for this enzyme ranges from 3 in *Lactobacillus casei* to 9.5 in *Pisum sativum* and that the temperature optimum ranges from 22 ° C in *Cricetulus griseus* to 100 °C in *Pyrobaculum aerophilum*. We will leave the understanding how this amazing variability is

achieved through variation of the protein sequence to the field of enzymology. Interestingly, the turnover number of this enzyme (how many molecules of D-glucose 6-phosphate or D-fructose 6-phosphate react at a single GPI enzyme per second) ranges from 0.0003 per second in *Thermococcus litoralis* to 650 per second in human if D-fructose 6-phosphate is the substrate and from 6.2 per second in *Pyrococcus furiosus* to 1700 per second in human if D-glucose 6-phosphate is the substrate. These rate constants are important parameters for modeling time-dependent behavior of metabolic networks and are thus also of relevance for this book.

3. **Find protein interaction partners of GPI in yeast.**
   Go to the web portal pre-PPI (https://bhapp.c2b2.columbia.edu/PrePPI) and enter the UNIPROT identifier P06744 for human "`glucose-6-phosphate isomerase`." Find the predicted interactions of GPI with other human proteins. The top hit with the probability 0.99 is ATP-dependent 6-phosphofructokinase. Explore the list.

4. **Discover consequences of GPI mutations in human.**
   Go to the OMIM database (www.omim.org) and enter "`glucose-6-phosphate isomerase`." Click the top entry "172400" on the next list and scroll to "allelic variants." Apparently, different mutations have been identified in the GPI enzyme of various patients that all led to "hemolytic anemia."

Bibliography

Small-World Networks, Scale-Free Networks

Barabási, A.L. and Albert, R. (1999). Emergence of scaling in random networks. *Science* 286: 509–512.

Watts, D.J. and Strogatz, S.H. (1998). Collective dynamics of 'small-world'-networks. *Nature* 393: 409–410.

Gene Regulatory Networks

Babu, M.M., Luscomebe, N.M., Aravind, L. et al. (2004). Structure and evolution of transcriptional regulatory networks. *Current Opinion in Structural Biology* 14: 283–291.

ENCODE

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57–74.

FANTOM Consortium

Lizio, M., Harshbarger, J., Shimoji, H. et al. (2015). Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biology* 16: 22.

The KEGG Database

Kanehisa, M., Sato, Y., Kawashima, M. et al. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* 44: D457–D462.

Brenda Database

Schomburg, I., Jeske, L., Ulbrich, M. et al. (2017). The BRENDA enzyme information system-From a database to an expert system. *Journal of Biotechnology* 261: 194–206.

Pathway Databases

Stobbe, M.D., Jansen, G.A., Moerland, P.D., and van Kampen, A.H.C. (2014). Knowledge representation in metabolic pathway databases. *Briefings in Bioinformatics* 15: 455–470.

DAVID

Dennis, G. Jr., Sherman, B.T., Hosack, D.A. et al. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology* 4: P3.