

1

Networks in Biological Cells

Modern molecular and cell biology has worked out many important cellular processes in great detail, although some other areas are known to a lesser extent. What often remains is to understand how the individual parts are connected. One may wonder whether mathematical modeling can make a contribution to this field before the missing details are known. 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. The point of Figure 1.1 is to remind us that there is a myriad of biomolecular interactions taking place in biological cells at all times and that it is pretty amazing how considerable order is achieved in many cellular processes that are all based on pairwise molecular interactions.

The focus of this textbook is placed on presenting mathematical descriptions developed in recent years to describe various levels of cellular networks that are mostly based on molecular interactions. 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. It is the belief of many researchers in the field of molecular biology that only combined efforts of modern experimental techniques and mathematical modeling and bioinformatics analysis will be able to arrive at a sufficient understanding of the biological network of cells and organisms.

In this first chapter we will start with some principles of mathematical networks and their relationship to biological networks. Then we will briefly look at several of the key biological players to be used in the remainder of this textbook (cells, compartments, proteins, pathways). Without going into any further detail, we will now jump right into the fast-growing 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

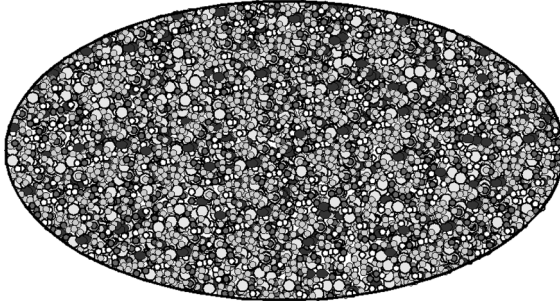


Figure 1.1 Is this how we should view a biological cell? This schematic picture makes an important point: about 30% of the volume of a biological cell is taken up by millions of individual proteins. Thus, biological cells are really “full”. However, such pictures do not tell us much about the organization of biological processes and, as we will see later in this textbook, there are many different hierarchies of order in such a cell.

examples such as acquaintance networks and collaboration networks (which fall under the class of social networks), technological networks (such as the Internet, the World Wide Web and power grids), and biological networks (such as neural networks, food webs and metabolic networks).

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 the connectivities 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 , the total number of edges in the graph is $p \times n \times (n - 1)/2$. The mathematics of random graphs was developed and elucidated by the two Hungarian mathematicians Erdős and Renyi. However, the analysis of real networks showed that they often differ significantly from the characteristics of random graphs.

1.1.2

Small-World Phenomenon

The term **small-world phenomenon** was created after the observation that everyone in the world can be reached by some other person through a short chain of social acquaintances. A 1967 **small-world experiment** by psychologist Stanley Milgram

found that any two random US citizens were connected by an average of six acquaintances and this gave rise to the famous phrase “six degrees of separation”. However, after more than 40 years, its status as a description of heterogeneous social networks still remains an open question. The average distance between vertices in a network is short, usually scaling logarithmically with the total number n of 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 a variety of different realms ranging from neuronal connections of the worm *Caenorhabditis elegans* to power grids. Watts and Strogatz also showed that the addition of a handful of random edges can turn a disconnected network into a highly connected one. For example, the addition of a few judicious routers makes a vast communication network (such as the Internet) no more than six hops wide.

1.1.3

Scale-Free Network Model

Only 1 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. While Watts and Strogatz’s model was able to explain the high clustering coefficient and the short average path length (these terms will all be introduced in Chapter 3) of a *small world*, it lacked an explanation for another property found in real-world networks such as the Internet: these networks are **scale-free**. In simple terms, this means that while 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 Chapter 4 that the connectivity of scale-free networks follows a power law. If a network is scale-free, it is also a small world.

Barabási’s scale-free model is strikingly simple, elegant and intuitive. To produce an artificial scale-free network possessing the small-world properties, only two basic rules must be followed:

- **Growth.** The network is seeded with a small number of initial vertices. At every time step, a new vertex is added that forms connections to m existing vertices.
- **Preferential attachment.** The probability of a newly added vertex connecting to an existing vertex n is assumed to depend on the degree of n (the number of connections already formed between vertex n to other vertices). The more connections n has, the more likely new vertices will connect to n . This behavior is also described by the saying “the rich become richer”.

The same mechanism is at work, for example, in the World Wide Web. Obviously, this network is in a constant state of growth where new pages are added every second. Also, we know from our own experience that once a user creates a new webpage, they

will most likely include links to other well-known pages (hubs) on this page. In the early exciting days when the study of large-scale networks took off like a storm, it was even suggested that the scale-free network model may be the foundation for a law of nature which governs the formation of natural small-world networks.

However, recent 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. We will return to this issue at the end of the textbook (Chapter 10) when looking at integrated networks. For the moment, we will consider the idea of network topology (scale-free networks, small-world phenomenon) as an enormously powerful concept, and 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 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 (Figure 1.2).

It is now well established that feedback loops are provided in this system, 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. An

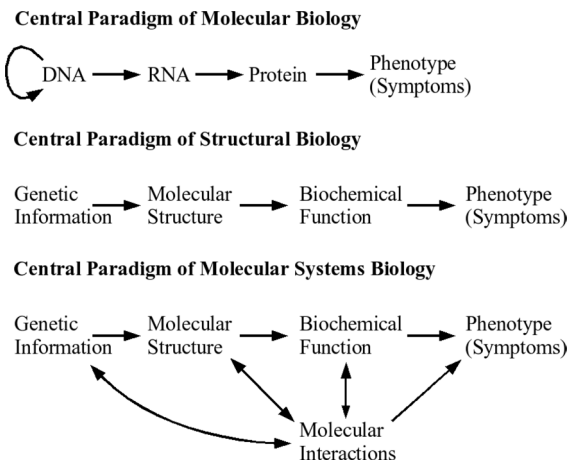


Figure 1.2 (Top) Since the 1950s, a paradigm became established that information flows from DNA over RNA to protein synthesis, which then gives rise to particular phenotypes. (Middle) The emergence of structural biology – the first crystal structure of the protein myoglobin was determined in 1960 – emphasized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

important discovery of the last 10 years showed that small RNA snippets may also mediate gene expression. The cellular network therefore certainly appears much more complicated today than it did 50 years ago.

This brings us to the world of **gene regulatory networks**. To discover which gene is activated or repressed by a particular transcription factor, one could create a knock-out mouse lacking the gene coding for this transcription factor and see which genes are no longer expressed or are now expressed in excess. 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. Modern microarray experiments, however, probe the expression levels of large numbers of genes simultaneously. Yet, it quickly turns out that the analysis of this large-scale data is complicated by the noisy nature of the data and by the fact that genes do not interact directly with each other.

Here, 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.

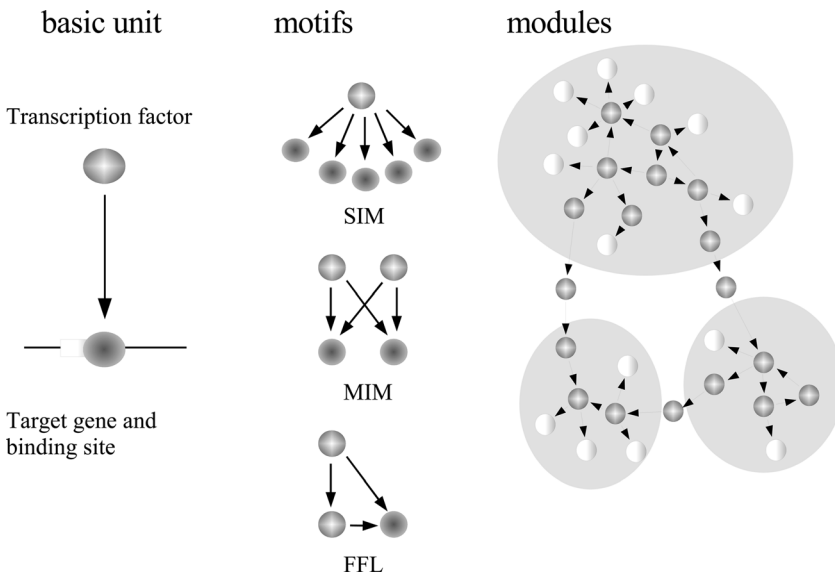


Figure 1.3 Structural organization of transcriptional regulatory networks. (Left) The “basic unit” comprises the transcription factor, its target gene with DNA recognition site and the regulatory interaction between them. (Middle) Units are often organized into network “motifs” that comprise specific patterns of inter-regulation that are over-represented in networks. Examples of motifs include single-input/multiple-output (SIM), multiple-input/multiple-

output (MIM) and feed-forward loop (FFL) motifs. (Right) 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). Figure drawn after Babu *et al.* (2004).

1.2.1

Cellular Components

Cells can be described at various levels of detail. Here, 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
- Metabolic reactions

(b) Structural descriptions:

- Single protein structures
- 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 textbooks on biochemistry or molecular biology. These biomolecules in a cell can be divided into several categories based on their role in metabolism.

- (1) **Macromolecules** including proteins, nucleic acids, polysaccharides and certain lipids.
- (2) The **building blocks** of macromolecules including 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 which are incorporated into lipids. Interestingly, in biological cells, only a small number of the theoretically synthesizable macromolecules exist at a given point in time. 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**). The molecules in a cell have complex chemical structures and must be synthesized step-by-step beginning with specific starting materials that may be taken up as 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 or amino acid hormones, molecules involved in energy storage (e.g. ATP), regulatory molecules (e.g. cyclic AMP) and metabolic waste products such as urea.

Almost all biological material needed to construct a biological cell is either synthesized by its RNA polymerase and ribosome machinery or is 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.

1.2.2

Spatial Organization of Eukaryotic Cells – 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 are 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, due to 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, chromatin, rough endoplasmic reticulum and cytoskeleton.

1.2.3

Cellular Organisms

Table 1.1 presents some statistics of the organisms considered in this textbook. Determination of RNA-coding genes is still in its infancy.

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, most of these reactions have been grouped into **metabolic pathways** that each contain a sequence of chemical reactions in which each reaction is catalyzed by a specific enzyme and the product of one reaction is the substrate 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** lead to the disassembly of complex molecules to form simpler products. They provide the raw materials for the synthesis of other molecules and they provide chemical energy

Table 1.1 Data on the genome length and on the number of protein and RNA genes are taken from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (July 2007); data on the number of putative transporter proteins are taken from www.membranetransport.org.

Organism	Length of genome (Mb)	Number of protein genes	Number of RNA genes	Number of transporter proteins
Prokaryotes				
<i>Methanococcus jannaschii</i>	1.7	1786	43	49
<i>Bacillus subtilis</i>	4.2	4106	119	297
<i>Escherichia coli</i>	4.6	4131	168	354
Eukaryotes				
<i>Saccharomyces cerevisiae</i>	12.2	5879	410	327
<i>Drosophila melanogaster</i>	180	14081	24	615
<i>Caenorhabditis elegans</i>	97	20077	24	656
<i>Homo sapiens</i>	2880	25307	25	784

required for many cellular activities. Energy released by catabolic pathways is stored temporarily either as high-energy phosphates (primarily ATP) or as high-energy electrons (primarily in NADPH). **Anabolic pathways** lead to the synthesis of more complex compounds from simpler starting materials. Anabolic pathways are energy requiring and utilize chemical energy released by the exergonic catabolic 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 exterior. Figure 1.5 shows the glycolysis pathway in *Homo sapiens* as represented in the KEGG database (Kanehisa *et al.*, 2006).

1.3.2

Enzymatic Reactions

Enzymes are proteins that catalyze biochemical reactions. Like all catalysts, enzymes work by lowering the **activation energy** of a reaction, thus allowing the reaction to proceed much faster than in aqueous solution. Remarkably, enzymes may speed up reactions by factors of many thousands to billions of times. An enzyme, like any catalyst, remains unaltered by the completed reaction and can therefore continue to function. As enzymes do not affect the relative free energy difference between the products and reagents, they do not affect the equilibrium of a reaction. The great advantage of enzymes compared to most other catalysts is their stereo-, regio- and chemoselectivity and -specificity.

For the binding reaction $A + B \leftrightarrow AB$, the **binding constant** k_d :

$$k_d = \frac{[A] \cdot [B]}{[AB]},$$

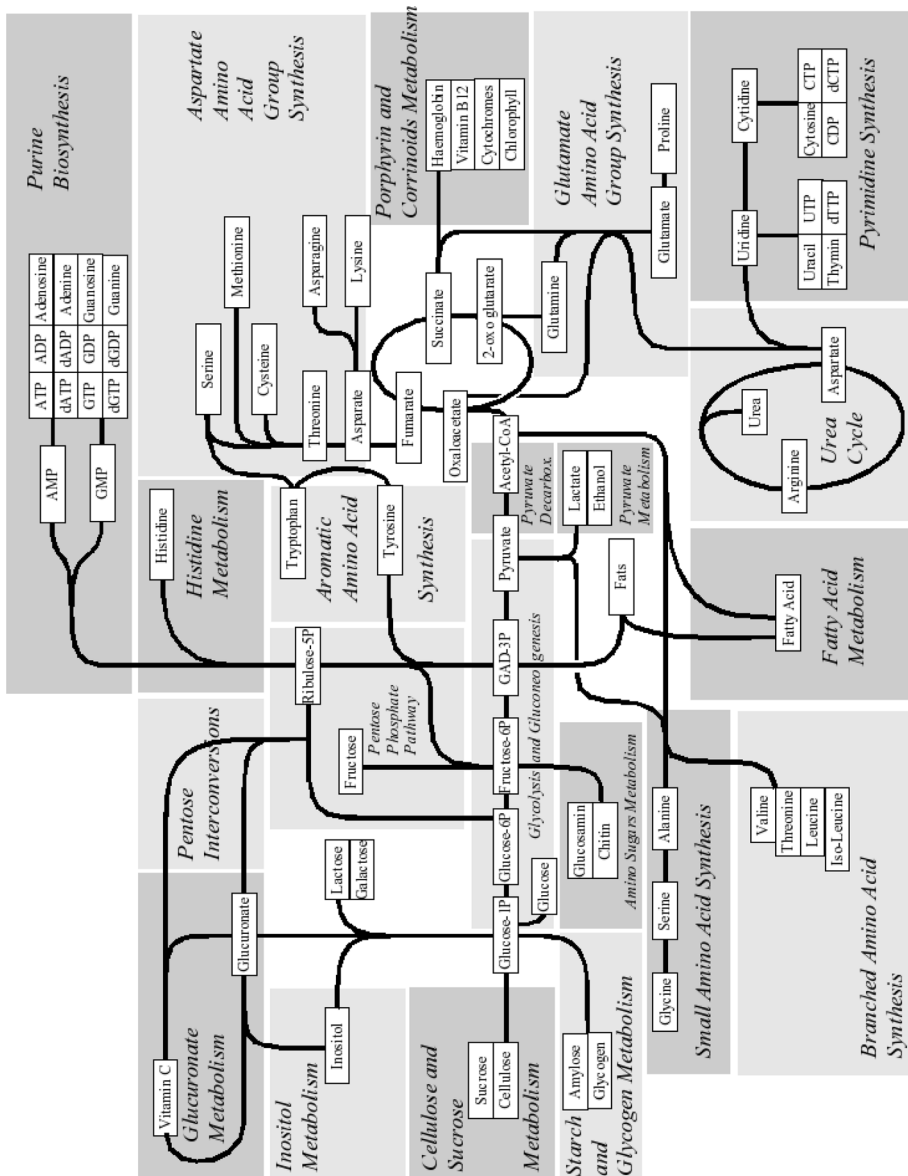
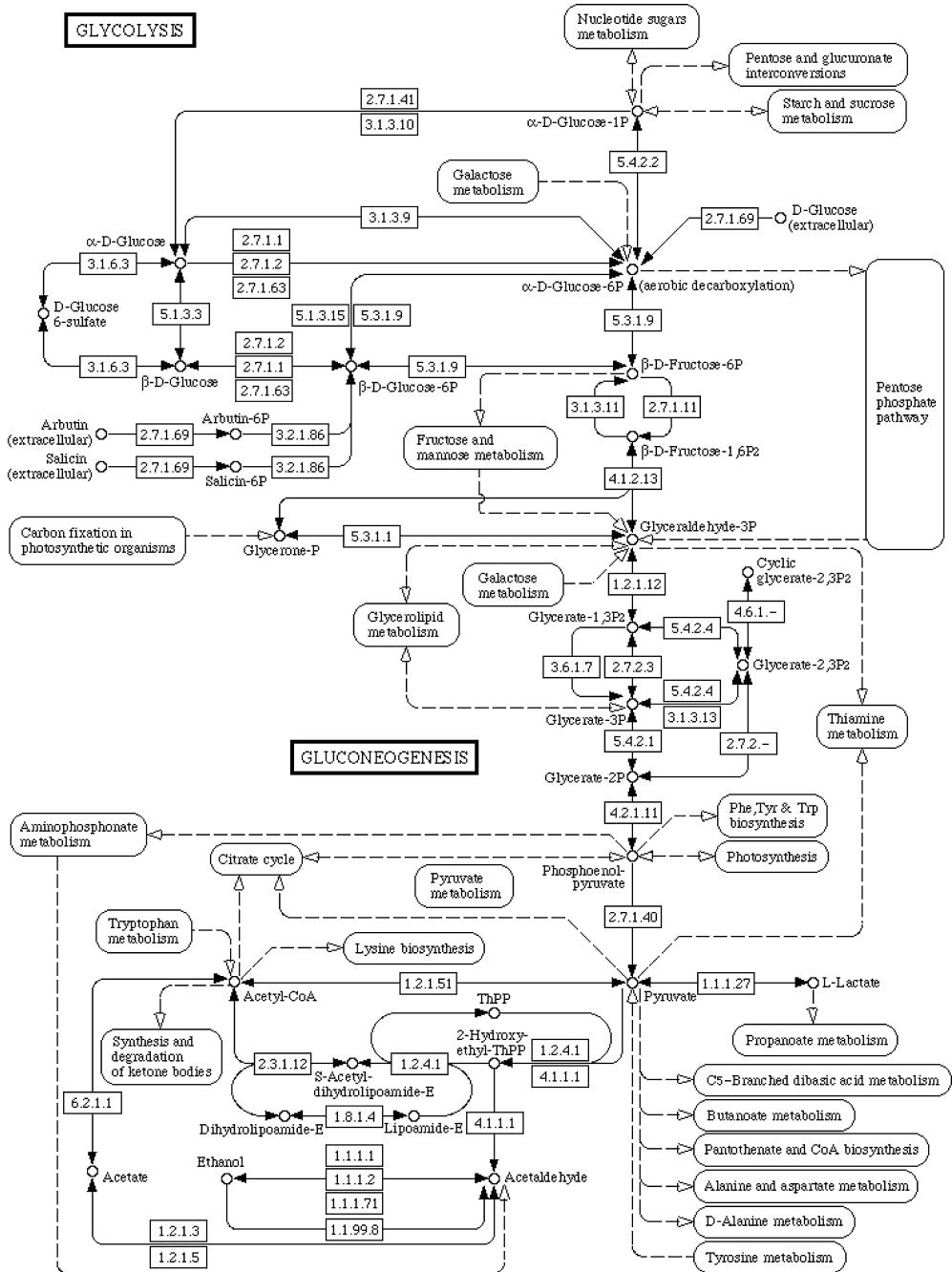


Figure 1.4 Major metabolic pathways.

determines how much of the substrate concentration is bound to the enzyme under equilibrium conditions. The binding constant has the unit M. In the case of a “nanomolar inhibitor”, for example, the equilibrium is very strongly on the complexed form and only a few free molecules exist. The binding constant k_d is also the ratio of the kinetic rates for the forward and backward reactions, k_{on} and



00010 8/6/07

Figure 1.5 The glycolysis pathway as visualized in the KEGG database is connected to many other cellular pathways (picture taken with permission from <http://www.genome.ad.jp/kegg/>).

k_{off} . The units of the two kinetic rates are $\text{M}^{-1} \text{s}^{-1}$ for the forward reaction and M^{-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 in 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, and separate this from energy transduction cycles that will be described separately (see, e.g. Section 7.7). 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.

1.3.4

Cell Cycle

The **cell cycle**, or **cell division cycle**, is the cycle of events in a eukaryotic cell from one cell division to the next. It consists of interphase, mitosis and, usually, cell division. The cell cycle is regulated by 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 for their discovery of these central molecules in the regulation of the cell cycle. The phases of the cell cycle are:

- The **G₀ phase** that is a period in the **cell cycle** where cells exist in a quiescent state.
- The **G₁ phase** that is the first growth phase.
- The **S phase**, during which the DNA is replicated, where S stands for the synthesis of DNA.
- The **G₂ phase** that is the second growth phase, also the preparation phase for the cell.
- The **M phase** or mitosis and cytokinesis that covers the actual division of the cell into two daughter cells.

A surveillance system, so-called “**checkpoints**”, monitors the cell for DNA damage and failure to perform critical processes. Checkpoints can block progression through the different stages of the cell cycle if certain conditions are not met. For instance, one checkpoint monitors DNA replication and prevents cells from proceeding to mitosis

before DNA replication is completed. Similarly, the spindle checkpoint blocks the transition from metaphase to anaphase within mitosis if not all chromosomes are attached to the mitotic spindle. We will see in Section 7.2 how cellular processes may dynamically regulate each other.

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 gene after gene, protein after protein. Due to 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 any one of us to learn these associations. Instead, researchers realized quite early that it would be extremely useful to generate general repositories for classification schemes that connect corresponding genes and proteins belonging to different organisms or which provide access to functional annotations. One important project in this area is the **Gene Ontology (GO)** (www.geneontology.org). This collaborative project started in 1998 as a collaboration of three model organism databases, FlyBase (*Drosophila*), the *Saccharomyces* Genome Database (SGD) and the Mouse Genome Database (MGD).

In the GO project, gene products are described in terms of their associated biological processes, cellular components and molecular functions in a species-dependent manner. A gene product might be associated with or located in one or more cellular components; it is active in one or more biological processes, during which it performs one or more molecular functions.

1.4.2 Systems Biology Markup Language

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 SBML language.

1.4.3 KEGG

Initiated in 1995, KEGG is an integrated bioinformatics resource consisting of three types of databases for genomic, chemical and network information. KEGG consists of three graph objects called the gene universe (GENES, SSDB and KEGG Orthology databases that contain more than 1.2 million genes from 30 eukaryotic, 250 bacterial and 25 archeal genomes), the chemical universe (COMPOUND, GLYCAN and REACTION databases that contain more than 13 000 chemical compounds and more than 6000 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 Brenda

Since 1987, the Brenda resource has been developed at the German National Research Center for Biotechnology and later at the University of Cologne. It is a comprehensive information system on enzymatic reactions. Data is stored in a relational database containing all data in 46 tables, enabling different queries (Table 1.3). Data on enzyme function are extracted directly from the primary literature by scientists holding a degree in Biology or Chemistry. Formal and consistency

Table 1.2 The three graph objects in KEGG (after Kanehisa *et al.*, 2004).

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

Table 1.3 Overview on Information stored in the Brenda system on each particular biochemical reaction.

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, three-dimensional structure/Protein Data Bank entry, molecular weight, subunits, posttranslational modification
Disease	disease

checks are performed by computer programs; each dataset on a classified enzyme is checked manually by at least one biologist and one chemist.

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 6). Yes, if you are interested in particular reaction rates or in modeling time-dependent processes. Computer scientists among the readers of this text should become aware that the rates of biochemical reactions vary significantly with temperature and pH, and may even change their directions.

1.5 Methods in Cellular Modeling

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

Summary

This introductory chapter gives a first look at the cellular components that will be the objects of computational and mathematical analysis in the rest of the textbook. 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 \rightarrow RNA \rightarrow proteins) and of cellular biochemistry (grouping of

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

Mathematical concept	Object of investigation	Analysis of complexity	Time dependent	Chapter(s)
Mathematical graphs	protein–protein networks and protein complexes	yes	no	2–4, 9
Stoichiometric analysis; matrix algebra	metabolic networks ^a	yes (count number of possible paths that connect two metabolites)	no	6
Differential equations	signal transduction, energy transduction	no	yes	7
Equations of motion	individual proteins, protein complexes	no	yes	9
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	8

^aMay also be applied to gene regulatory networks and signal transduction networks.

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 certainly why mathematical and computational analysis is now needed to keep the overview over all of the data being generated and to deepen our understanding about cellular processes.

Further Reading

Small-World Networks

Watts DJ, Strogatz SH (1998) Collective dynamics of ‘small-world’ networks. *Nature* **393**, 409–410.

Gene Regulatory Networks

Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA (2004) Structure and evolution of transcriptional regulatory networks, *Current Opinion in Structural Biology* **14**, 283–291.

The KEGG Database

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome, *Nucleic Acids Research* **32**, D277–D280.

Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research* **34**, D354–D357.

