Part One
Introduction to Systems Biology

Introduction 1

1.1 Biology in Time and Space

Biological systems such as organisms, cells, or biomolecules are highly organized in their structure and function. They have developed during evolution and can only be fully understood in this context. To study them and to apply mathematical, computational, or theoretical concepts, we have to be aware of the following circumstances.

The continuous reproduction of cell compounds necessary for living and the respective flow of information is captured by the central dogma of molecular biology, which can be summarized as follows: genes code for mRNA, mRNA serves as template for proteins, and proteins perform cellular work. Although information is stored in the genes encoded by the DNA sequence, it is made available only through the cellular machinery that can decode this sequence and can translate it into structure and function. In this book, we will explain that from various perspectives.

A description of biological entities and their properties encompasses different levels of organization and different time scales. We can study biological phenomena at the level of populations, individuals, tissues, organs, cells, and compartments down to molecules and atoms. Length scales range from the order of meter (e.g., the size of whale or human) to micrometer for many cell types, down to picometer for atom sizes. Time scales include millions of years for evolutionary processes, annual and daily cycles, seconds for many biochemical reactions, and femtoseconds for molecular vibrations. Figure 1.1 gives an overview about scales.

In a unified view of cellular networks, each action of a cell involves different levels of cellular organization, including genes, proteins, metabolism, or signaling pathways. Therefore, the current description of the individual networks must be integrated into a larger framework.

Many current approaches pay tribute to the fact that biological items are subject to evolution. The structure and organization of organisms and their cellular machinery has developed during evolution to fulfill major functions such as growth, proliferation, and survival under changing conditions. If parts of the organism or of the cell fail to perform their function, the individual might become unable to survive or replicate.

One consequence of evolution is the similarity of biological organisms of different species. This similarity

Systems Biology: A Textbook, Second Edition. Edda Klipp, Wolfram Liebermeister, Christoph Wierling, and Axel Kowald.

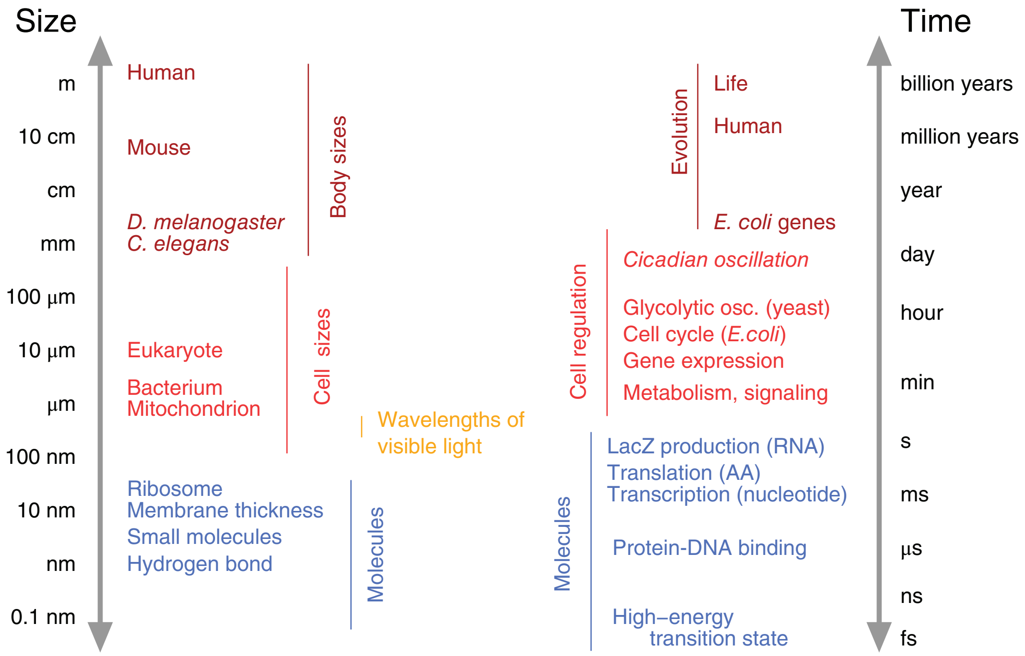

Figure 1.1 Length and time scales in biology. (Data from the BioNumbers database at bionumbers.hms.harvard.edu.)

allows for the use of model organisms and for the critical transfer of insights gained from one cell type to other cell types. Applications include, for example, prediction of protein function from similarity, prediction of network properties from optimality principles, reconstruction of phylogenetic trees, or the identification of regulatory DNA sequences through cross-species comparisons. However, the evolutionary process also leads to genetic variations within species. Therefore, personalized medicine and research is an important new challenge for biomedical research.

1.2 Models and Modeling

If we observe biological phenomena, we are confronted with various complex processes that often cannot be explained from first principles and the outcome of which cannot reliably be foreseen from intuition. Even if general biochemical principles are well established (e.g., the central dogma of transcription and translation or the biochemistry of enzyme-catalyzed reactions), the biochemistry of individual molecules and systems is often unknown and can vary considerably between species. Experiments lead to biological hypotheses about individual processes, but it often remains unclear whether these hypotheses can be combined into a larger coherent picture because it is often difficult to foresee the global behavior of a complex system from knowledge of its parts. Mathematical modeling and computer simulations can help us to understand the internal nature and dynamics of these processes and to arrive at predictions about their future development and the effect of interactions with the environment.

1.2.1 What Is a Model?

The answer to this question will differ among communities of researchers. In a broad sense, a model is an abstract representation of objects or processes that explains features of these objects or processes (Figure 1.2). A biochemical reaction network can be represented by a graphical sketch showing dots for metabolites and arrows for reactions; the same network could also be described by a system of differential equations, which allows simulating and predicting the dynamic behavior of that network. If a model is used for simulations, it needs to be ensured that it faithfully predicts the system's behavior – at least those aspects that are supposed to be covered by the model. Systems biology models are often based on well-established physical laws that justify their general form, for instance, the thermodynamics of chemical reactions. Besides this, a computational model needs to make specific statements about a system of interest – which are partially justified by experiments and biochemical knowledge, and partially by mere extrapolation from other systems. Such a model can

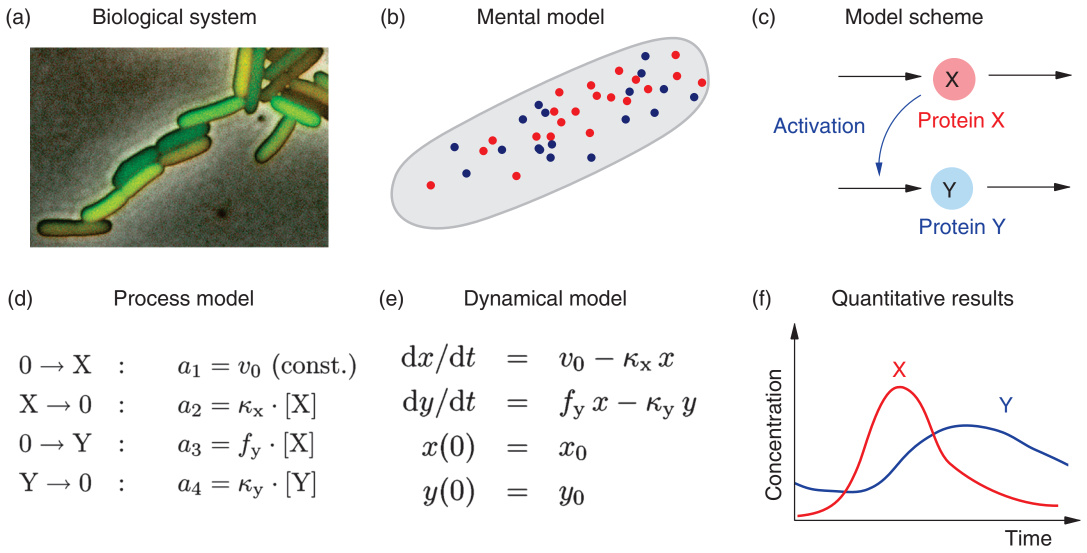
(a)
Biological system
(b)
Mental model
(c)
Model scheme
X
Activation
Protein X
Y
Protein Y
(d)
Process model
$0 \to X$: $a_1 = v_0$ (const.)
$X \to 0$: $a_2 = \kappa_x \cdot [X]$
$0 \to Y$: $a_3 = f_y \cdot [X]$
$Y \to 0$: $a_4 = \kappa_y \cdot [Y]$
(e)
Dynamical model
$dx/dt = v_0 - \kappa_x x$
$dy/dt = f_y x - \kappa_y y$
$x(0) = x_0$
$y(0) = y_0$
(f)
Quantitative results
X
Y
Concentration
Time

Figure 1.2 Typical abstraction steps in mathematical modeling. (a) *E. coli* bacteria produce thousands of different proteins. If a specific protein type is labeled with a fluorescent marker, cells glow under the microscope according to the concentration of this marker. (Courtesy of M. Elowitz.) (b) In a simplified mental model, we assume that cells contain two enzymes of interest, X (red) and Y (blue), and that the molecules (dots) can freely diffuse within the cell. All other substances are disregarded for the sake of simplicity. (c) The interactions between the two protein types can be drawn in a wiring scheme: each protein can be produced or degraded (black arrows). In addition, we assume that proteins of type X can increase the production of protein Y. (d) All individual processes to be considered are listed together with their rates *a* (occurrence per time). The mathematical expressions for the rates are based on a simplified picture of the actual chemical processes. (e) The list of processes can be translated into different sorts of dynamic models, in this case, deterministic rate equations for the protein concentrations *x* and *y*. (f) By solving the model equations, predictions for the time-dependent concentrations can be obtained. If the predictions do not agree with experimental data, this indicates that the model is wrong or too much simplified. In both cases, the model has to be refined.

summarize established knowledge about a system in a coherent mathematical formulation. In experimental biology, the term “model” is also used to denote a species that is especially suitable for experiments; for example, a genetically modified mouse may serve as a model for human genetic disorders.

1.2.2 Purpose and Adequateness of Models

Modeling is a subjective and selective procedure. A model represents only specific aspects of reality but, if done properly, this is sufficient since the intention of modeling is to answer particular questions. If the only aim is to predict system outputs from given input signals, a model should display the correct input–output relation, while its interior can be regarded as a black box. However, if instead a detailed biological mechanism has to be elucidated, then the system’s structure and the relations between its parts must be described realistically. Some models are meant to be generally applicable to many similar objects (e.g., Michaelis–Menten kinetics holds for many enzymes, the promoter–operator concept is applicable to many genes, and gene regulatory motifs are common), while others are specifically tailored to one particular object (e.g., the 3D structure of a protein, the sequence of a gene, or a model of deteriorating mitochondria during aging). The mathematical part can be kept as simple as possible to allow for easy implementation and comprehensible results. Or it can be modeled very realistically and be much more complicated. None of the characteristics mentioned above makes a model wrong or right, but they determine whether a model is appropriate to the problem to be solved. The phrase “essentially, all models are wrong, but some are useful” coined by the statistician George Box is indeed an appropriate guideline for model building.

1.2.3 Advantages of Computational Modeling

Models gain their reference to reality from comparison with experiments, and their benefits therefore depend on the quality of the experiments used. Nevertheless, modeling combined with experimentation has a lot of advantages compared with purely experimental studies:

- Modeling drives conceptual clarification. It requires verbal hypotheses to be made specific and conceptually rigorous.

- Modeling highlights gaps in knowledge or understanding. During the process of model formulation, unspecified components or interactions have to be determined.
- Modeling provides independence of the modeled object.
- Time and space may be stretched or compressed *ad libitum*.
- Solution algorithms and computer programs can be used independently of the concrete system.
- Modeling is cheap compared with experiments.
- Models exert by themselves no harm on animals or plants and help to reduce ethical problems in experiments. They do not pollute the environment.
- Modeling can assist experimentation. With an adequate model, one may test different scenarios that are not accessible by experiment. One may follow time courses of compounds that cannot be measured in an experiment. One may impose perturbations that are not feasible in the real system. One may cause precise perturbations without directly changing other system components, which is usually impossible in real systems. Model simulations can be repeated often and for many different conditions.
- Model results can often be presented in precise mathematical terms that allow for generalization. Graphical representation and visualization make it easier to understand the system.
- Finally, modeling allows for making well-founded and testable predictions.

The attempt to formulate current knowledge and open problems in mathematical terms often uncovers a lack of knowledge and requirements for clarification. Furthermore, computational models can be used to test whether proposed explanations of biological phenomena are feasible. Computational models serve as repositories of current knowledge, both established and hypothetical, about how systems might operate. At the same time, they provide researchers with quantitative descriptions of this knowledge and allow them to simulate the biological process, which serves as a rigorous consistency test.

1.3 Basic Notions for Computational Models

1.3.1 Model Scope

Systems biology models consist of mathematical elements that describe properties of a biological system, for instance, mathematical variables describing the concentrations of metabolites. As a model can only describe certain aspects of the system, all other properties of the system (e.g., concentrations of other substances or the environment of a cell) are neglected or simplified. It is important – and, to some extent, an art – to construct models in such ways that the disregarded properties do not compromise the basic results of the model.

1.3.2 Model Statements

Alongside the model elements, a model can contain various kinds of statements and equations describing facts about the model elements, most notably, their temporal behavior. In kinetic models, the basic modeling paradigm considered in this book, the dynamics is determined by a set of ordinary differential equations describing the substance balances. Statements in other model types may have the form of equality or inequality constraints (e.g., in flux balance analysis), maximality postulates, stochastic processes, or probabilistic statements about quantities that vary in time or between cells.

1.3.3 System State

In dynamical systems theory, a system is characterized by its *state*, a snapshot of the system at a given time. The state of the system is described by the set of variables that must be kept track of in a model: in deterministic models, it needs to contain enough information to predict the behavior of the system for all future times. Each modeling framework defines what is meant by the state of the system. In kinetic rate equation models, for example, the state is a list of substance concentrations. In the corresponding stochastic model, it is a probability distribution or a list of the current number of molecules of a species. In a Boolean model of gene regulation, the state is a string of bits indicating for each gene whether it is expressed ("1") or not expressed ("0"). Also, the temporal behavior can be described in fundamentally different ways. In a *dynamical system*, the future states are determined by the current state, while in a *stochastic process*, the future states are not precisely predetermined. Instead, each possible future history has a certain probability to occur.

1.3.4 Variables, Parameters, and Constants

The quantities in a model can be classified as variables, parameters, and constants. A *constant* is a quantity with a fixed value, such as the natural number *e* or Avogadro's number (number of molecules per mole). *Parameters* are quantities that have a given value, such as the K_m value of an enzyme in a reaction. This value depends on the method used and on the experimental conditions and may change. *Variables* are quantities with a changeable value for which the model establishes relations. A subset of variables, the *state variables*, describes the system behavior completely. They can assume independent values and each of them is necessary to define the system state. Their number is equivalent to the dimension of the system. For example, the diameter d and volume V of a sphere obey the relation $V = \pi d^3/6$, where π and 6 are constants, V and d are variables, but only one of them is a state variable since the relation between them uniquely determines the other one.

Whether a quantity is a variable or a parameter depends on the model. In reaction kinetics, the enzyme concentration appears as a parameter. However, the enzyme concentration itself may change due to gene expression or protein degradation, and in an extended model, it may be described by a variable.

1.3.5 Model Behavior

Two fundamental factors that determine the behavior of a system are (i) influences from the environment (input) and (ii) processes within the system. The system structure, that is, the relation among variables, parameters, and constants, determines how endogenous and exogenous forces are processed. However, different system structures may still produce similar system behavior (output); therefore, measurements of the system output often do not suffice to choose between alternative models and to determine the system's internal organization.

1.3.6 Model Classification

For modeling, processes are classified with respect to a set of criteria.

- A structural or *qualitative* model (e.g., a network graph) specifies the interactions among model elements. A *quantitative* model assigns values to the elements and to their interactions, which may or may not change.
- In a *deterministic* model, the system evolution through all following states can be predicted from the knowledge of the current state. *Stochastic* descriptions give instead a probability distribution for the successive states.
- The nature of values that time, state, or space may assume distinguishes a *discrete* model (where values are taken from a discrete set) from a *continuous* model (where values belong to a continuum).
- *Reversible* processes can proceed in a forward and backward direction. Irreversibility means that only one direction is possible.
- *Periodicity* indicates that the system assumes a series of states in the time interval $\{t, t + \Delta t\}$ and again in the time interval $\{t + i\Delta t, t + (i + 1)\Delta t\}$ for $i = 1,2, \ldots$.

1.3.7 Steady States

The concept of stationary states is important for the modeling of dynamical systems. *Stationary states* (other terms are *steady states* or *fixed points*) are determined by the fact that the values of all state variables remain constant in time. The asymptotic behavior of dynamic systems, that is, the behavior after a sufficiently long time, is often stationary. Other types of asymptotic behavior are oscillatory or chaotic regimes.

The consideration of steady states is actually an abstraction that is based on a separation of time scales. In nature, everything flows. Fast and slow processes – ranging from formation and breakage of chemical bonds within nanoseconds to growth of individuals within years – are coupled in the biological world. While fast processes often reach a quasi-steady state after a short transition period, the change of the value of slow variables is often negligible in the time window of consideration. Thus, each steady state can be regarded as a quasi-steady state of a system that is embedded in a larger nonstationary environment. Despite this idealization, the concept of stationary states is important in kinetic modeling because it points to typical behavioral modes of the system under study and it often simplifies the mathematical problems.

Other theoretical concepts in systems biology are only rough representations of their biological counterparts. For example, the representation of gene regulatory networks by Boolean networks, the description of complex enzyme kinetics by simple mass action laws, or the representation of multifarious reaction schemes by black boxes proved to be helpful simplifications. Although being a simplification, these models elucidate possible network properties and help to check the reliability of basic assumptions and to discover possible design principles in nature. Simplified models can be used to test mathematically formulated hypotheses about system dynamics, and such models are easier to understand and to apply to different questions.

1.3.8 Model Assignment Is Not Unique

Biological phenomena can be described in mathematical terms. Models developed during the last few decades range from the description of glycolytic oscillations with ordinary differential equations to population dynamics models with difference equations, stochastic equations for signaling pathways, and Boolean networks for gene expression. However, it is important to realize that a certain process can be described in more than one way: a biological object can be investigated with different experimental methods and each biological process can be described with different (mathematical) models. Sometimes, a modeling framework represents a simplified limiting case (e.g., kinetic models as limiting case of stochastic models). On the other hand, the same mathematical formalism may be applied to various biological instances: statistical network analysis, for example, can be applied to cellular transcription networks, the circuitry of nerve cells, or food webs.

The choice of a mathematical model or an algorithm to describe a biological object depends on the problem, the purpose, and the intention of the investigator. Modeling has to reflect essential properties of the system and different models may highlight different aspects of the same system. This ambiguity has the advantage that different ways of studying a problem also provide different insights into the system. However, the diversity of modeling approaches makes it also very difficult to merge established models (e.g., for individual metabolic pathways) into larger supermodels (e.g., models of complete cell metabolism).

1.4 Networks

The network is a crucial concept in systems biology. We study protein–protein interaction networks, protein–RNA interaction networks, metabolic networks (see Chapters 3 and 4 and Section 12.1), signaling networks (Section 12.2), guilt-by-association networks, and networks connecting gene defects with diseases or diseases with other diseases via common gene defects [1]. Throughout this book, you will find more examples.

Networks are best represented by graphs that consist of nodes and edges, which connect the nodes, as illustrated in Figure 1.3. In protein–protein interaction networks, for example, nodes are proteins and edges are their interactions as can for instance be determined by yeast two-hybrid experiments (see Chapter 14). If appropriate, one can introduce different types of nodes for different types of components. For example, the metabolites and converting enzymes in metabolic networks can be represented with bipartite networks, which possess two types of nodes – one for metabolites and the other for enzymes – that are never directly connected by an edge, but only via the other type of node. Petri net type of modeling takes that property into account representing metabolites as places and enzyme-catalyzed reactions as transitions (see Section 7.1). By contrast, classical metabolic modeling considers only one type of node, but different types in different approaches. Systems of ordinary differential equations describing metabolite dynamics take metabolites as nodes and enzymatic reactions as edges (Chapter 4), while flux balance analysis restricts itself to steady states and now focusses on the fluxes through the reactions (now as nodes) that are linked by the stationary metabolites as edges.

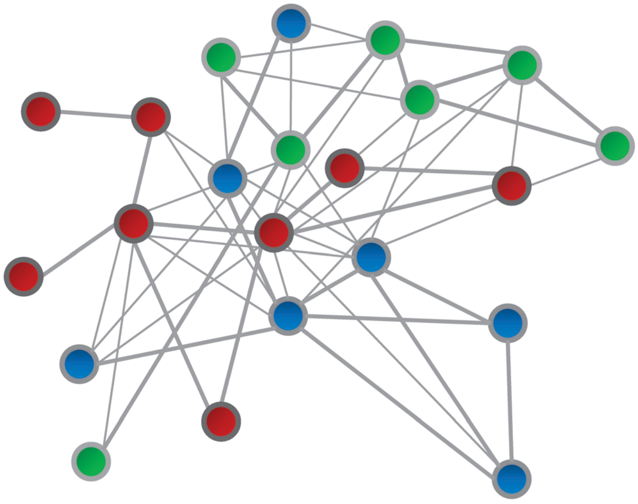

Figure 1.3 Network with nodes (circles) and edges (lines between circles). Different node colors indicate different types of connected components (e.g., proteins, mRNAs, and metabolites).

1.5 Data Integration

Systems biology has evolved rapidly in the last few years, driven by the new high-throughput technologies. The most important impulse was given by large sequencing projects such as the Human Genome Project, which resulted in the full sequence of the human and other genomes [2,3]. Proteomic technologies have been used to identify the translation status of complete cells (2D gels, mass spectrometry) and to elucidate protein–protein interaction networks involving thousands of components [4]. However, to validate such diverse high-throughput data, one needs to correlate and integrate such information. Thus, an important part of systems biology is data integration.

On the lowest level of complexity, data integration implies common schemes for data storage, data representation, and data transfer. For particular experimental techniques, this has already been established, for example, in the field of transcriptomics with Minimum Information About a Microarray Experiment [5], Minimum Information for Reporting Next Generation Sequence Genotyping [6], in proteomics with proteomics experiment data repositories [7], and the Human Proteome Organization consortium [8]. On a more complex level, schemes have been defined for biological models and pathways such as Systems Biology Markup Language (SBML) [9], CellML [10], or Systems Biology Graphical Notation (SBGN) [11], which all use an XML-like language style.

Data integration on the next level of complexity consists of data correlation. This is a growing research field as researchers combine information from multiple diverse data sets to learn about and explain natural processes [12,13]. For example, methods have been developed to integrate the results of transcriptome or proteome experiments with genome sequence annotations. In the case of complex disease conditions, it is clear that only integrated approaches can link clinical, genetic, behavioral, and environmental data with diverse types of molecular phenotype information and identify correlative associations. Such correlations, if found, are the key to identifying biomarkers and processes that are either causative or indicative of the disease. Importantly, the identification of biomarkers (e.g., proteins and metabolites) associated with the disease will open up the possibility to generate and test hypotheses on the biological processes and genes involved in this condition. The evaluation of disease-relevant data is a multistep procedure involving a complex pipeline of analysis and data handling tools such as data normalization, quality control, multivariate statistics, correlation analysis, visualization techniques, and intelligent database systems [14]. Several pioneering approaches have indicated the power of integrating data sets from different levels, for example, the correlation of gene membership of expression clusters and promoter sequence motifs [15], the combination of transcriptome and quantitative proteomics data in order to construct models of cellular pathways [13], and the identification of novel metabolite–transcript correlations [16]. Finally, data can be used to build and refine dynamical models, which represent an even higher level of data integration.

1.6 Standards

As experimental techniques generate rapidly growing amounts of data and large models need to be developed and exchanged, standards for both experimental procedures and modeling are a central practical issue in systems biology. Information exchange necessitates a common language about biological aspects. One seminal example is the Gene Ontology that provides a controlled vocabulary that can be applied to all organisms, even as the knowledge about genes and proteins continues to accumulate. SBML [9] has been established as exchange language for mathematical models of biochemical reaction networks. SBGN [11] defines graphical elements to unambiguously represent biochemical reaction sets and large regulatory networks. A series of "minimum-information-about" statements based on community agreement defines standards for certain types of experiments. Minimum Information Requested in the Annotation of Biochemical Models (MIRIAM) [17] describes standards for this specific type of systems biology models. Minimum Information About a Simulation Experiment (MIASE) [18] helps authors to describe all elements of a computational experiment such that readers can repeat the simulations and create figures as shown in the publication.

1.7 Model Organisms

Model organisms are species that have developed over the years to be extremely popular for scientific investigations. The reasons for such popularity can be manifold. Of great importance is, of course, an easy handling of the organism, that is, culture conditions (temperature, pressure, etc.) that can be set up in the laboratory without much effort and that tolerate some degree of variation, so that results are comparable between groups that use slightly different growth conditions. However, other factors are also important, such as costs (for housing, food, etc.), size (the smaller the size, the more individuals can be studied), or lifespan (short-lived species are more popular for aging studies). Often model organisms are also used to represent important taxonomical properties (prokaryotes, eukaryotes, unicellular organisms, multicellular organisms, vertebrates, and invertebrates), but in all cases the hope is that important biochemical findings made in such model organisms are also of relevance for other species of that taxonomical group or even for humans. Figure 1.4 shows a selection of popular model species, which will be discussed in the next sections. They range from prokaryotic organisms to single and multicellular eukaryotic species up to mammals.

1.7.1 *Escherichia coli*

E. coli is probably the oldest and best studied model organism of all (Figure 1.4a). It is a rod-shaped

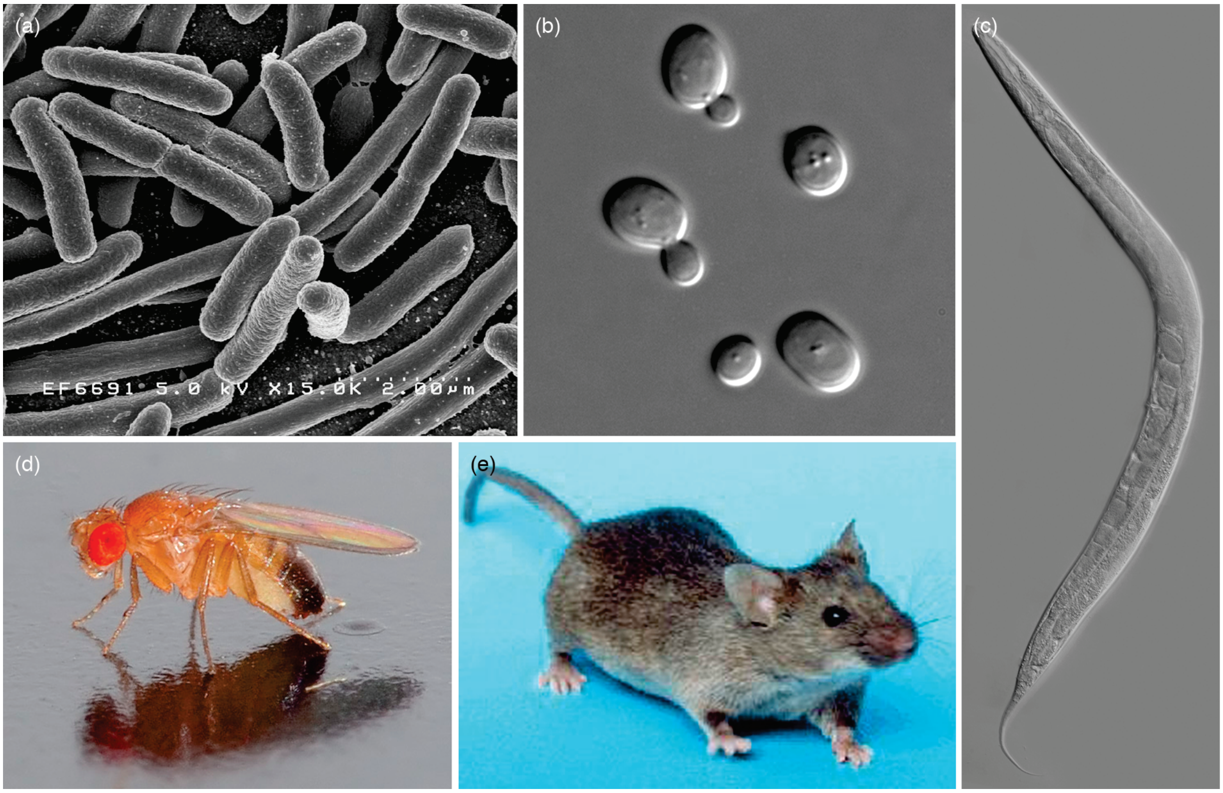

Figure 1.4 Popular model organisms for studies of problems in biochemistry and molecular biology. (a) *E. coli* is a rod-like bacterium and the best studied prokaryotic model system. (Public domain image from Wikimedia, http://commons.wikimedia.org/wiki/File:EscherichiaColi_NIAID.jpg.) (b) The yeast *S. cerevisiae* is a simple unicellular eukaryote and is of considerable scientific and industrial interest. (Public domain image from Wikimedia, http://commons.wikimedia.org/wiki/File:S_cerevisiae_under_DIC_microscopy.jpg.) (c) The nematode *C. elegans* is approximately 1 mm and is a popular representative for simple and short-lived multicellular organisms. ("Adult *Caenorhabditis elegans*" by Kbradnam (http://en.wikipedia.org/wiki/User:Kbradnam) is licensed under CC BY-SA-2.5, http://creativecommons.org/licenses/by-sa/2.5.) (d) The fruit fly *D. melanogaster* is like *C. elegans* a model for simple multicellular organisms and has extensively been studied in developmental biology. ("*Drosophila melanogaster*" by A. Karwath (http://commons.wikimedia.org/wiki/User:Aka) is licensed under CC BY-SA-2.5, http://creativecommons.org/licenses/by-sa/2.5.) (e) Finally, the mouse *M. musculus* is a popular model species for mammals and is thus also of great relevance for humans. (Public domain image from Wikimedia, http://commons.wikimedia.org/wiki/File:House_mouse.jpg.)

bacterium that is found in the intestines of many organisms, including humans. It is a facultative anaerobic organism, which means that it can grow under aerobic as well as anaerobic conditions. *E. coli* is roughly 2 µm long with a diameter of 0.5 µm. Under laboratory conditions, it can easily be cultivated and doubles its number in less than 30 min. It has been studied for more than 50 years and is the most popular prokaryotic model organism. The genome of the *E. coli* strain K-12 has completely been sequenced in 1997 [19] and contains around 4200 genes dispersed along 4.6 million base pairs (Mbp). It is a very streamlined genome containing very few intergenic sequences. The *E. coli* family consists of a large number of strains, and a comparison of the sequence of more than 60 strains has shown that they contain in total more than 15 500 genes, while only 6% of this pan-genome is present in each strain [20]. *E. coli* was of pivotal importance for developing many of the experimental techniques described in Chapter 14. Today, a large number of scientific resources regarding this model species are available on the Internet. A good starting point is EcoCyc (ecocyc.org), which provides information about the genome and biochemical machinery of the *E. coli* strain K-12 MG1655. Other websites provide information about protein–protein interactions (bacteriome.org/) and systematic single-gene knockout mutants (http://ecoli.aist-nara.ac.jp/gb6/Resources/deletion/deletion.html), or a database of available strains (cgsc.biology.yale.edu). For modelers, the CyberCell Database (ccdb.wishartlab.com/CCDB) is also of interest since it aims at providing enzymatic, genetic, and biological information suitable for developing mathematical models of all parts of a cell of *E. coli* strain K-12.

1.7.2 *Saccharomyces cerevisiae*

The yeast *S. cerevisiae* is a unicellular fungus, belonging to the ascomycetes (Figure 1.4b). It is not only a useful organism needed for the production of wine, beer, and bread, but also the best studied eukaryotic model system. The cells are easy to grow and double under optimal conditions every 90–100 min. Like *E. coli*, also *S. cerevisiae* can live under aerobic as well as anaerobic conditions. If oxygen is present, the majority of energy is generated via oxidative phosphorylation at the inner mitochondrial membrane and without oxygen energy is produced via glycolysis and fermentation. The yeast normally propagates as a diploid organism via mitosis. Under stress, however, the diploid cells can undergo sporulation, producing four haploid cells in the process. These haploid cells belong to one of two mating classes (sexes), called "a" and "α". Haploids can either propagate via normal mitosis or mate with other haploids of the different mating class, resulting again in diploid cells. This life cycle makes *S. cerevisiae* interesting for genetic studies; it has also been extensively used by experimental and modeling studies of the cell cycle, glycolysis, osmotic shock, and mating process [21–28]. Cell division occurs in *S. cerevisiae* in an asymmetric fashion called budding and single-cell studies have shown that yeast cells exhibit replicative senescence with a maximum of 30–40 divisions [29]. Since this process is very reminiscent of the replicative senescence known from human fibroblasts [30], *S. cerevisiae* is also employed as a model system for investigations of the aging process. Furthermore, *S. cerevisiae* was also the first eukaryotic organism to be sequenced and its genome consists of about 12 Mbp containing roughly 6000 genes distributed over 16 chromosomes [31]. Homologous recombination (the exchange of sequences between similar strands of DNA) is very efficient in *S. cerevisiae*, which makes the organism also a convenient model for studies of synthetic biology. Using this mechanism, it was possible to replace the complete chromosome 16 with a new, synthetic one through 11 successive rounds of transformation (see Chapter 14) [32]. The synthetic chromosome was streamlined by removing all introns and superfluous tRNA genes and using only two of the three possible stop codons. This opens the possibility to extend the genetic code by a further amino acid once all chromosomes are modified in this way. A good online resource for further information about this model organism is the *Saccharomyces* Genome Database (www.yeastgenome.org).

1.7.3 *Caenorhabditis elegans*

Of course, model systems for multicellular organisms are also needed and the nematode *C. elegans* (Figure 1.4c) has become such a model since Sidney Brenner introduced it to the research community [33]. Like the other model organisms, it is easy to cultivate (feeding on bacteria or synthetic medium) and thousands of the about 1 mm long animals can live on a large Petri dish. Wild populations of *C. elegans* consist mainly of hermaphrodites together with a few males. Hermaphrodites not only are capable of self-fertilization (leading to natural inbred lines), but can also mate with males. The hermaphrodite then lays eggs that develop into larvae after hatching and after a total of four larval stages (L1–L4) the adult animal emerges. The complete life cycle from egg to egg takes between 2.5 and 5.5 days, depending on the temperature. The total lifespan of *C. elegans* is rather short with 2–3 weeks. This made *C. elegans* another popular model system for the investigation of the aging process [34]. However, the nematode is also an important model for other fields of research such as molecular biology or neurology. RNA interference (RNAi), for instance, is an important experimental technique (Chapter 14) that was developed based on experiments in *C. elegans* [35]. Furthermore, adult nematodes have a fixed number of somatic cells that is identical for all individuals (1031 in the male and 959 in the hermaphrodite), which makes it possible to generate very detailed anatomical models of the worm. The "slidable worm" (www.wormatlas.org/slidableworm.htm), which is a resource available on the webpage of the WormAtlas database, presents the results of such anatomical studies using an easy-to-use interface. *C. elegans* is also the only animal for which the complete wiring diagram (connectome) of the nervous system has been determined (using electron microscopy serial sections) [36,37]. Finally, *C. elegans* has also been the first multicellular organism for which the complete genome sequence has been determined [38,39]. The 97 Mbp contain approximately 19 000 genes dispersed over six chromosomes. Good online starting points for more information are WormBase (www.wormbase.org), WormBook (www.wormbook.org/), or WormAtlas (www.wormatlas.org/).

1.7.4 *Drosophila melanogaster*

The fruit fly *D. melanogaster* (Figure 1.4d) is another, immensely popular, model organism that shares many of the properties of *C. elegans*. The animals are easy to breed in captivity and because of their small size (around 1 mm) it is possible to perform studies involving thousands of individuals (e.g., for selection or population studies). The generation time (about 7 days at 29 °C) and lifespan (about 30 days at 29 °C) are very short and depend strongly on the ambient temperature. This facilitates, for example, artificial selection studies, which take several generations [40]. *D. melanogaster* has four chromosomes ($2n = 8$), which can even be studied under the light microscope because of a phenomenon called polyteny. As in many insect larvae, the cells of the salivary glands of *D. melanogaster* undergo multiple rounds of replication without cell division, leading to hundreds of sister chromatids aligned to each other. Polytene chromosomes are found in cells that need to express a large amount of a specific gene product and transcriptionally active areas appear under the microscope as swollen regions, so-called puffs. Although this technique is now outdated regarding the analysis of transcriptional activity, polytene chromosomes are still valuable for taxonomic problems. After staining, the puffs form a specific banding pattern that can be used to identify chromosomal deletions and duplications. This can be used in taxonomy to differentiate and classify closely related subspecies. *D. melanogaster* was arguably the most important model species for investigating developmental processes in multicellular organisms [41], which has led to the discovery of Hox genes [42]. These genes code for a set of transcription factors that contain a common 180 bp motif (the homeodomain) and control the development of the anterior–posterior axis of the animal. A unique feature of these genes is that they are arranged on the chromosomes in the same linear order as the body region that they affect (called collinearity). Thus, Hox genes at one end of the cluster control the development of the anterior region (head), while the genes at the other end of the cluster influence the development of the posterior region (tail). Although originally found in *Drosophila*, Hox genes have been found in many metazoans, including vertebrates [43]. The complete genome was sequenced in 2000 [44] and somewhat surprisingly the number of genes is with approximately 14 000 clearly smaller than the number of genes in *C. elegans*. Further information, tools, and resources are available at FlyBase (flybase.org) and Ensembl Genome Browser (www.ensembl.org/Drosophila_melanogaster).

1.7.5 *Mus musculus*

The last model system that we want to introduce here is the house mouse *M. musculus domesticus* (Figure 1.4e). It is clearly the model organism with the largest similarity to humans and is therefore also of great relevance for human research. Humans and mice are both mammals and thus share a common ancestor roughly 80 million years ago, a rather short time span compared with the other model organisms. Consequently, the genome structure and organization is also very similar. The mouse genome, sequenced in 2002 [45], contains 2.5 Gbp and is thus somewhat smaller than the human genome with 2.9 Gbp [2,3], although both genomes contain approximately 20 000–25 000 genes. The similarity at the gene level is quite amazing insofar that for more than 99% of mouse genes a homolog can also be found in the human genome [3], and vice versa. The mouse is also a popular model system because it is very amenable to genetic manipulations. The first mice were cloned in 1998 [46] and today it is common routine to create transgenic mice by introducing DNA constructs into fertilized egg cells and to study the function of existing genes by knocking them out or down (see Chapter 14). The Knockout Mouse Project (KOMP), for instance, aims at generating and providing mouse embryonic stem cells (and eventually whole mice) with single-gene knockout for every gene in the mouse genome (www.komp.org). Because mice have been used for such a long time as model species, many different inbred strains have been developed, which differ in various aspects of their phenotype (e.g., size, lifespan, and disease susceptibility). Of special interest are the various strains of nude mice that have a deletion of the FOXN1 gene, which prevents the formation of a functioning thymus. Without a thymus, these mice cannot produce mature T lymphocytes and therefore lack most forms of immune response (the lack of fur is a side effect of this mutation). As a consequence, they are valuable tools to study tumor development and are also used for transplantation studies, since they do not reject allo- or xenografts. Useful starting points for further information are, for instance, the Mouse Genome Informatics (www.informatics.jax.org/), the Mouse Atlas Project (www.emouseatlas.org), or the Ensembl Genome Browser (www.ensembl.org/Mus_musculus).

References

1 Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., and Barabási, A.L. (2007) The human disease network. *Proc. Natl. Acad. Sci. USA*, 104 (21), 8685–8690.

2 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, 409, 860–921.

3 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science*, 291, 1304–1351.

4 Mering, C. *et al.* (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, 417, 399–403.

5 Brazma, A. *et al.* (2001) Minimum Information About a Microarray Experiment (MIAME): toward standards for microarray data. *Nat. Genet.*, 29, 365–371.

6 Mack, S.J., Milius, R.P., Gifford, B.D., Sauter, J., Hofmann, J., Osoegawa, K., Robinson, J., Groeneweg, M., Turenchalk, G.S., Adai, A., Holcomb, C., Rozemuller, E.H., Penning, M.T., Heuer, M.L., Wang, C., Salit, M.L., Schmidt, A.H., Parham, P.R., Müller, C., Hague, T., Fischer, G., Fernandez-Viňa, M., Hollenbach, J.A., Norman, P.J., and Maiers, M. (2015) Minimum Information for Reporting Next Generation Sequence Genotyping (MIRING): guidelines for reporting HLA and KIR genotyping via next generation sequencing. *Hum. Immunol.* doi: 10.1016/j.humimm.2015.09.011.

7 Taylor, C.F. *et al.* (2003) A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nat. Biotechnol.*, 21, 247–254.

8 Hermjakob, H. *et al.* (2004) The HUPO PSI's molecular interaction format: a community standard for the representation of protein interaction data. *Nat. Biotechnol.*, 22, 177–183.

9 Hucka, M. *et al.* (2003) The Systems Biology Markup Language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19, 524–531.

10 Lloyd, C.M. *et al.* (2004) CellML: its future, present and past. *Prog. Biophys. Mol. Biol.*, 85, 433–450.

11 Le Novere, N., Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A. *et al.* (2009) The Systems Biology Graphical Notation. *Nat. Biotechnol.*, 27 (8), 735–741.

12 Gitton, Y. *et al.* (2002) A gene expression map of human chromosome 21 orthologues in the mouse. *Nature*, 420, 586–590.

13 Ideker, T. *et al.* (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292, 929–934.

14 Kanehisa, M. and Bork, P. (2003) Bioinformatics in the post-sequence era. *Nat. Genet.*, 33, 305–310.

15 Tavazoie, S. *et al.* (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, 22, 281–285.

16 Urbanczyk-Wochniak, E. *et al.* (2003) Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Rep.*, 4, 989–993.

17 Le Novere, N. *et al.* (2005) Minimum Information Requested in the Annotation of Biochemical Models (MIRIAM). *Nat. Biotechnol.*, 23, 1509–1515.

18 Waltemath, D., Adams, R., Beard, D.A., Bergmann, F.T., Bhalla, U.S., Britten, R. *et al.* (2011) Minimum Information About a Simulation Experiment (MIASE). *PLoS Comput. Biol.*, 7 (4), e1001122.

19 Blattner, F.R., Plunkett, G., 3rd, Bloch, C.A., Perna, N.T., Burland, V., Riley, M. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, 277 (5331), 1453–1462.

20 Lukjancenko, O., Wassenaar, T.M., and Ussery, D.W. (2010) Comparison of 61 sequenced *Escherichia coli* genomes. *Microb. Ecol.*, 60 (4), 708–720.

21 Hynne, F., Dano, S., and Sorensen, P.G. (2001) Full-scale model of glycolysis in *Saccharomyces cerevisiae*. *Biophys. Chem.*, 94 (1–2), 121–163.

22 Klipp, E., Nordlander, B., Kruger, R., Gennemark, P., and Hohmann, S. (2005) Integrative model of the response of yeast to osmotic shock. *Nat. Biotechnol.*, 23 (8), 975–982.

23 Diener, C., Schreiber, G., Giese, W., del Rio, G., Schroder, A., and Klipp, E. (2014) Yeast mating and image-based quantification of spatial pattern formation. *PLoS Comput. Biol.*, 10 (6), e1003690.

24 Kofahl, B. and Klipp, E. (2004) Modelling the dynamics of the yeast pheromone pathway. *Yeast*, 21 (10), 831–850.

25 Adrover, M.À., Zi, Z., Duch, A., Schaber, J., González-Novo, A., Jimenez, J., Nadal-Ribelles, M., Clotet, J., Klipp, E., and Posas, F. (2011) Time-dependent quantitative multicomponent control of the G_1-S network by the stress-activated protein kinase Hog1 upon osmostress. *Sci. Signal.*, 4 (192), ra63. Erratum: *Sci. Signal.*, 4 (197), er5 (2011).

26 Chen, K.C., Calzone, L., Csikasz-Nagy, A., Cross, F.R., Novak, B., and Tyson, J.J. (2004) Integrative analysis of cell cycle control in budding yeast. *Mol. Biol. Cell*, 15 (8), 3841–3862.

27 Rizzi, M., Baltes, M., Theobald, U., and Reuss, M. (1997) *In vivo* analysis of metabolic dynamics in *Saccharomyces cerevisiae*. II. Mathematical model. *Biotechnol. Bioeng.*, 55 (4), 592–608.

28 Kaizu, K., Ghosh, S., Matsuoka, Y., Moriya, H., Shimizu-Yoshida, Y., and Kitano, H. (2010) A comprehensive molecular interaction map of the budding yeast cell cycle. *Mol. Syst. Biol.*, 6, 415.

29 Jazwinski, S.M. (1990) Aging and senescence of the budding yeast *Saccharomyces cerevisiae*. *Mol. Microbiol.*, 4 (3), 337–343.

30 Hayflick, L. (1965) The limited *in vitro* lifetime of human diploid cell strains. *Exp. Cell Res.*, 37, 614–636.

31 Mewes, H.W., Albermann, K., Bahr, M., Frishman, D., Gleissner, A., Hani, J. *et al.* (1997) Overview of the yeast genome. *Nature*, 387 (6632 Suppl.), 7–65.

32 Annaluru, N., Muller, H., Mitchell, L.A., Ramalingam, S., Stracquadanio, G., Richardson, S.M. *et al.* (2014) Total synthesis of a functional designer eukaryotic chromosome. *Science*, 344 (6179), 55–58.

33 Brenner, S. (1973) The genetics of *Caenorhabditis elegans*. *Genetics*, 77, 71–94.

34 Johnson, T.E. (2013) 25 years after age-1: genes, interventions and the revolution in aging research. *Exp. Gerontol.*, 48 (7), 640–643.

35 Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, 391 (6669), 806–811.

36 White, J.G., Southgate, E., Thomson, J.N., and Brenner, S. (1986) The structure of the nervous system of the nematode *Caenorhabditis elegans*. *Philos. Trans. R. Soc. B*, 314 (1165), 1–340.

37 Jarrell, T.A., Wang, Y., Bloniarz, A.E., Brittin, C.A., Xu, M., Thomson, J.N. *et al.* (2012) The connectome of a decision-making neural network. *Science*, 337 (6093), 437–444.

38 C. elegans Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, 282, 2012–2018.

39 Hillier, L.W., Coulson, A., Murray, J.I., Bao, Z., Sulston, J.E., and Waterston, R.H. (2005) Genomics in *C. elegans*: so many genes, such a little worm. *Genome Res.*, 15 (12), 1651–1660.

40 Rose, M. and Charlesworth, B. (1980) A test of evolutionary theories of senescence. *Nature*, 287 (5778), 141–142.

41 Nusslein-Volhard, C. and Wieschaus, E. (1980) Mutations affecting segment number and polarity in *Drosophila*. *Nature*, 287 (5785), 795–801.

42 Scott, M.P. and Weiner, A.J. (1984) Structural relationships among genes that control development: sequence homology between the Antennapedia, Ultrabithorax, and fushi tarazu loci of *Drosophila*. *Proc. Natl. Acad. Sci. USA*, 81 (13), 4115–4119.

43 Gehring, W.J. (1992) The homeobox in perspective. *Trends Biochem. Sci.*, 17 (8), 277–280.

44 Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science*, 287 (5461), 2185–2195.

45 Mouse Genome Sequencing Consortium, Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420 (6915), 520–562.

46 Wakayama, T., Perry, A.C., Zuccotti, M., Johnson, K.R., and Yanagimachi, R. (1998) Full-term development of mice from enucleated oocytes injected with cumulus cell nuclei. *Nature*, 394 (6691), 369–374.

Further Reading

The early days of systems biology: Kitano, H. (2001) *Foundations of Systems Biology*, MIT Press, Cambridge, MA.

The early days of systems biology: Kitano, H. (2002) Systems biology: a brief overview. *Science*, 295 (5560), 1662–1664.

Numbers in cell biology: Flamholz, A., Philips, R., and Milo, R. (2014) The quantified cell. *Mol. Biol. Cell*, 25 (22), 3497–3500.

Numbers in cell biology: Milo, R. and Phillips, R. (2014) *Cell Biology by the Numbers*, Garland Science.

Systemic thinking in cell biology: Lazebnik, Y. (2002) Can a biologist fix a radio? Or, what I learned while studying apoptosis. *Cancer Cell*, 2, 179–182.

Physical constraints on cell function: Dill, K.A., Ghosh, K., and Schmit, J.D. (2011) Physical limits of cells and proteomes. *Proc. Natl. Acad. Sci. USA*, 108 (44), 17876–17882.