Part One
Introduction to Systems Biology

Systems Biology: A Textbook. Edda Klipp, Wolfram Liebermeister, Christoph Wierling, Axel Kowald,
Hans Lehrach, and Ralf Herwig
Copyright © 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-31874-2

1
Introduction

1.1
Biology in Time and Space

Biological systems like 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 in form of 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, this will be explained 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.

Systems Biology: A Textbook. Edda Klipp, Wolfram Liebermeister, Christoph Wierling, Axel Kowald,
Hans Lehrach, and Ralf Herwig

ISBN: 978-3-527-31874-2

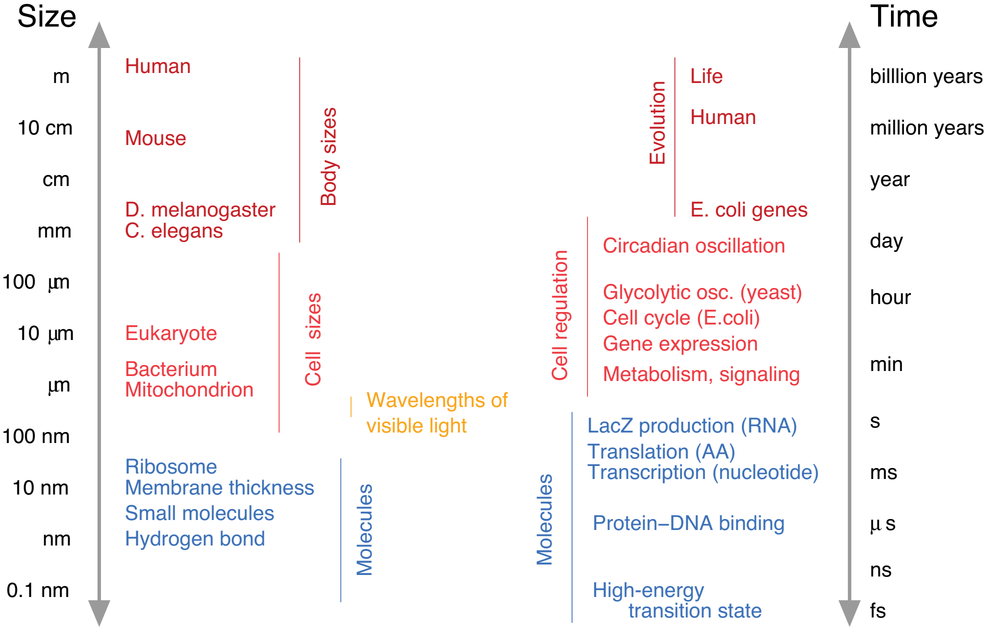

Figure 1.1 Length and time scales in biology. Data from the BioNumbers database http://bionumbers.hms.harvard.edu.

One consequence of evolution is the similarity of biological organisms from different species. This similarity 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, e.g., 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. But 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 processes, we are confronted with various complex processes that 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, 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 if 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 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 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. But 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,

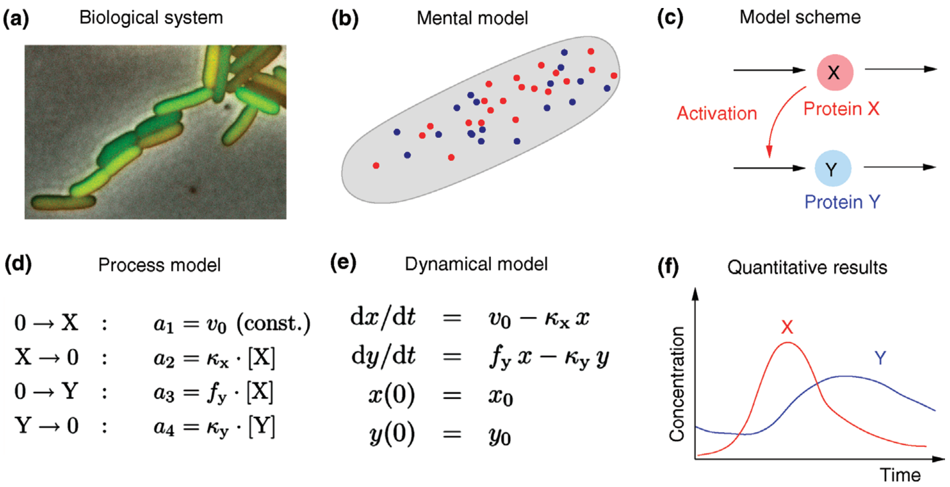
(a)
Biological system
(b)
Mental model
(c)
Model scheme
X
Activation
Protein X
Y
Protein Y
(d)
Process model
$0 \rightarrow \mathrm{X} \; : \; a_1 = v_0$ (const.)
$\mathrm{X} \rightarrow 0 \; : \; a_2 = \kappa_\mathrm{x} \cdot [\mathrm{X}]$
$0 \rightarrow \mathrm{Y} \; : \; a_3 = f_\mathrm{y} \cdot [\mathrm{X}]$
$\mathrm{Y} \rightarrow 0 \; : \; a_4 = \kappa_\mathrm{y} \cdot [\mathrm{Y}]$
(e)
Dynamical model
$\mathrm{d}x/\mathrm{d}t = v_0 - \kappa_\mathrm{x}\, x$
$\mathrm{d}y/\mathrm{d}t = f_\mathrm{y}\, x - \kappa_\mathrm{y}\, y$
$x(0) = x_0$
$y(0) = y_0$
(f)
Quantitative results
X
Y
Concentration
Time

Figure 1.2 Typical abstraction steps in mathematical modeling. (a) *Escherichia coli* bacteria produce thousands of different proteins. If a specific protein type is fluorescently labeled, cells glow under the microscope according to the concentration of this enzyme (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 these predictions do not agree with experimental data, it indicates that the model is wrong or too much simplified. In both cases, it has to be refined.

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 to 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 to 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

Besides 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 possibly 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 simplification. 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 hypothesis 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 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. But 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 still very difficult to merge established models (e.g., for individual metabolic pathways) into larger supermodels (e.g., models of complete cell metabolism).

1.4 Data Integration

Systems biology has evolved rapidly in the last years driven by the new high-throughput technologies. The most important impulse was given by the large sequencing projects such as the human genome project, which resulted in the full sequence of the human and other genomes [1, 2]. Proteomics 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 [3]. 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 [4], in proteomics with proteomics experiment data repositories [5], and the Human Proteome Organization consortium [6]. On a more complex level, schemes have been defined for biological models and pathways such as Systems Biology Markup Language (SBML) [7] and CellML [8], which 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 [9, 10]. 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, 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 [11]. 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 [12]; the combination of transcriptome and quantitative proteomics data in order to construct models of cellular pathways [10]; and the identification of novel metabolite-transcript correlations [13]. Finally, data can be used to build and refine dynamical models, which represent an even higher level of data integration.

1.5 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 which provides a controlled vocabulary that can be applied to all organisms, even as the knowledge about genes and proteins continues to accumulate. The SBML [7] has been established as exchange language for mathematical models of biochemical reaction networks. A series of "minimum-information-about" statements based on community agreement defines standards for certain types of experiments. *M*inimum *i*nformation *r*equested *i*n the *a*nnotation of biochemical *m*odels (MIRIAM) [14] describes standards for this specific type of systems biology models.

References

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

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

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

4 Brazma, A. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics*, **29**, 365–371.

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

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

7 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.

8 Lloyd, C.M. *et al.* (2004) CellML: its future present and past. *Progress in Biophysics and Molecular Biology*, **85**, 433–450.

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

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

11 Kanehisa, M. and Bork, P. (2003) Bioinformatics in the post-sequence era. *Nature Genetics*, **33** (Suppl), 305–310.

12 Tavazoie, S. *et al.* (1999) Systematic determination of genetic network architecture. *Nature Genetics*, **22**, 281–285.

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

14 Le Novere, N. *et al.* (2005) Minimum information requested in the annotation of biochemical models (MIRIAM). *Nature Biotechnology*, **23**, 1509–1515.