SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

School of Bio and Chemical Engineering
DEPARTMENT OF BIOINFORMATICS

UNIT I - Quantitative Models in Biological Systems - Subject Code: SBI1402

UNIT 1 BASIC PRINCIPLES

1.1 Introduction to Systems biology
1.2 Modeling in biology
1.3 System states, Steady state
1.4 Variables, parameters and constants
1.5 Model behaviour
1.6 Advantages of computational modeling,
1.7 Model development
1.8 Typical aspects of biological systems and corresponding models
1.9 Bottom-up and top-down approaches of complex system
1.10Mathematical representation of cell - biological system Time and space
1.11 Future of systems biology - Experimental
1.12 Planning in the Systems Biology Phase of Biological Research

1.1 Introduction to Systems biology

Systems biology has been responsible for some of the most important developments in the science of human health and environmental sustainability. It is a **holistic** approach to deciphering the complexity of biological systems that starts from the understanding that the networks that form the whole of living organisms are more than the sum of their parts. It is **collaborative**, integrating many scientific disciplines – biology, computer science, engineering, bioinformatics, physics and others – to **predict** how these systems change over time and under varying conditions, and to develop solutions to the world's most pressing health and environmental issues.

This ability to design predictive, multiscale models enables our scientists to **discover** new biomarkers for disease, **stratify** patients based on unique genetic profiles, and **target** drugs and other treatments. Systems biology, ultimately, creates the potential for entirely new kinds of exploration, and drives constant **innovation** in biology-based technology and computation.

1.2 Modeling in biology

Mathematical, computational and physical methods have been applied in biology and medicine to study phenomena at a wide range of size scales, from the global human population all the way down to the level of individual atoms within a biomolecule. Concomitant with this range of sizes between global to atomistic, the relevant modeling methods span time scales varying between years and picoseconds, depending on the area of interest (from evolutionary to atomistic effects) and relevance. This review will cover some of the most common and useful mathematical and computational methods. Firstly, we outline the maximum entropy principle as an inference tool for the study of phenomena at different scales, from gene evolution and gene networks to protein-drug molecular interactions, followed with a survey of the methods used for large scale systems—populations, organisms, and cells—and then zooming down to the methods used to study individual biomolecules—proteins and drugs. To study the large systems, the most common and reliable mathematical technique is to develop systems of differential equations. At the molecular scale, molecular dynamics is often used to model biomolecules as a system of moving Newtonian particles with interactions defined by a force field, with various methods employed to handle the challenge of solvent effects. In some cases, pure quantum mechanics methods can and should be used, which describe molecules using either wave functions or electron densities, although computational costs in time and resources may be prohibitive, so hybrid classical-quantum methods are often more appropriate. Quantum methods can be particularly valuable in the study of enzymes and enzymatic reactions.

1.3 System State

An important notion in dynamical systems theory is the state. The state of a system is a snapshot of the system at a given time that contains enough information to pre- dict the behavior of the system for all future times. The state of the system is de- scribed by the set of variables that

must be kept track of in a model. Different modeling approaches have different representations of the state: in a dif- ferential equation model for a metabolic network, the state is a list of concentrations of each chemical species. In the respective stochastic model, it is a probability distribution and/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"). Thus, each model defines what it means by the state of the system. Given the current state, the model predicts which state or states can occur next, thereby describing the change of state.

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 be- havior of dynamic systems, i.e., 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 se- paration of time scales. In nature, everything flows. Fast and slow processes – ranging from formation and release 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 non-stationary environment. Although the concept of stationary states is a mathematical idealization, it is important in kinetic modeling since it points to typical behavioral modes of the investigated system and the respective mathematical problems are frequently easier to solve.

1.4 Variables, parameters and constants

The quantities involved 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 NA = 6.02 7 1023 (number of molecules per mole). Parameters are quantities that are assigned a value, such as the Km 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. The state variables are a set of variables that describe the system behavior completely. They are independent of each other and each of them is neces- sary to define the system state. Their number is equivalent to the dimension of the system. For example, diameter d and volume V of a sphere obey the relation V = p d3/6. p and 6 are constants and V and d are variables, but only one of them is a state variable, since the mentioned relation uniquely determines the other one.

Whether a quantity is a variable or a parameter depends on the model. The en- zyme concentration is frequently considered a parameter in biochemical reaction ki- netics. That is

no longer valid if, in a larger model, the enzyme concentration may change due to gene expression or protein degradation.

1.5 Model behaviour

There are two fundamental causes that determine the behavior of a system or its changes: (1) influences from the environment (input) and (2) processes within the system. The system structure, i.e., the relation among variables, parameters, and constants, determines how endogenous and exogenous forces are processed. It must be noted that different system structures may produce similar system behavior (out- put). The structure determines the behavior, not the other way around. Therefore, the system output is often not sufficient to predict the internal organization. Generally, system limits are set such that the system output has no impact on the input.

1.6 Advantages of computational modelling

Models gain their reference to reality from comparison with experiments, and their benefits are, therefore, somewhat dependent on experimental performance. Never- theless, modeling has a lot of advantages. Modeling drives conceptual clarification. It requires that verbal hypotheses be made specific and conceptually rigorous. Modeling also highlights gaps in knowl- edge or understanding. During the process of model formulation, unspecified com- ponents 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 it in experiments. They do not pollute the environment. Models interact neither with the environment nor with the modeled system.

Modeling can assist experimentation. With an adequate model one may test differ- ent scenarios that are not accessible by experiment. One may follow time courses of compounds that cannot be measured in an experiment. One may impose perturba- tions that are not feasible in the real system. One may cause precise perturbations with- out 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 gen- eralization. Graphical representation and visualization make it easier to understand the system. Finally, modeling allows for making well-founded and testable predictions.

1.7 Model development

1. Formulation of the problem: Before establishing an initial model, it must be clear which questions shall be answered with the approach. A distinct verbal statement about background, problem, and hypotheses is a helpful guide in further analysis.
2. Verification of available information: As a first step, the existing quantitative and structural knowledge has to be checked and collected. This concerns information about the included components and their interactions as well as experimental re- sults with respect to phenotypic changes such as growth and shape after system perturbations such as knockout experiments, RNAi, and variation of environmen- tal conditions.
3. Selection of model structure: Based on the available information and on the prob- lem to solve, the general type of the model is determined: (1) the level of descrip- tion as macroscopic or microscopic, (2) the choice of a deterministic or stochastic approach, (3) the use of discrete or continuous variables, and (4) the choice of steady-state, temporal, or spatio-temporal description. Furthermore, it must be decided what the determinants for system behavior (external influences, internal structure) are. The system variables must be assigned.
4. Establishing a simple model: The first model can be expressed in words, schema- tically, or in mathematical formulation. It serves as general test and allows refined hypotheses.
5. Sensitivity analysis: Mathematical models typically contain a number of parameters, and the simulation result can be highly sensitive to parameter changes. It is recom- mendable to verify the dependence of the model results on the parameter choice.
6. Experimental tests of the model predictions: This is a hard task. Experimental de- sign in biology is usually hypothesis-driven. In fact, hypotheses that state general relations can rarely be verified, but only falsified. These predictions usually con- cern relationships between different cellular states or biochemical reactions. On the other hand, hypothesis about the existence of items are hard to falsify. The choice of parameters to be measured, how many measurements are to be per- formed, and at what time intervals is not uniquely defined but depends on the re- searcher's opinion. These selections are largely based on experience and, in new areas in particular, on intuition.
7. Stating the agreements and divergences between experimental and modeling re- sults: Although the behavior of the model and the experimental system should eventually agree, disagreement drives further research. It is necessary to find out whether the disagreement results from false assumptions, tampering simplifica- tions, wrong model structure, inadequate experimental design, or other inade- quately represented factors.
8. Iterative refinement of model: The initial model will rarely explain all features of the studied object and usually leads to more open questions than answers. After comparing the model outcome with the experimental results, model structure and parameters may be adapted.

As stated above, the choice of a model approach is not unique. Likewise, the possible outcome of models differs. Satisfactory results could be the solution to the initially stated problem, the establishment of a strategy for problem solution, or reasonable suggestions for experimental design.

1.8 Typical aspects of biological systems and corresponding models

A number of notions have been introduced or applied in the context of systems biology or computational modeling of biological systems. Their use is often not unique, but we will present here some interpretations that are helpful in understanding respective theories and manuscripts.

Network Versus Elements

A system consists of individual elements that interact and thus form a network. The elements have certain properties. In the network, the elements have certain relations to each other (and, if appropriate, to the environment). The system has properties that rely on the individual properties and relations between the elements. It may show additional systemic properties and dynamic characteristics that often cannot be deduced from the individual properties of the elements.

Modularity

Modules are subsystems of complex molecular networks that can be treated as functional units, which perform identifiable tasks (Lauffenburger 2000). Typical examples for assignment of modules are (1) the DNA-mRNA-enzyme-metabolism cascade and (2) signal transduction cascades consisting of covalent modification cycles. The reaction networks at each level are separated as modules by the criterion that mass transfer occurs internally but not between the modules, and they are linked by means of catalytic or regulatory effects from a chemical species of one module to a reaction in another module (Hofmeyr and Westerhoff 2001). Consideration of modules has the advantage that modeling can be performed in a hierarchical, nested, or sequential fashion. The properties of each module can be studied first in isolation and subsequently in a comprehensive, integrative attempt. The concept is appealing since it allows thinking in terms of classes of systems with common characteristics that can be handled with a common set of methods. The disadvantage is that a modular approach has to ignore or at least reduce the high level of connectivity in cellular networks – in particular the variety of positive and negative feedback and feed-forward regulatory loops – which actually contradicts the basic idea of systems biology.

Robustness and Sensitivity are Two Sides of the Same Coin

Robustness is an essential feature of biological systems. It characterizes the insensi- tivity of system properties to variations in parameters, structure, and environment or to other uncertainties. Robust systems maintain their state and functions despite ex- ternal and internal perturbations. An earlier notion for this observation is homeosta- sis. Robustness in biological systems is often achieved by a high degree of complex- ity involving feedback, modularity, redundancy, and structural stability (Kitano 2002). On the one hand, biological systems must protect their genetic information and their mode of living against perturbations; on the other hand, they must adapt to changes, sense and process internal and external signals, and react precisely depend- ing on the type or strength of a perturbation. Sensitivity or fragility characterizes the ability of living organisms for adequately reacting on a certain stimulus. Note that in some areas sensitivity is more rigorously defined as the ratio of the change of a vari- able by the change of a quantity that caused the change in the variable.

1.9 Bottom-up and top-down approaches of complex system

Systems biology is a computational field that has been used for several years across different scientific areas of biological research to uncover the complex interactions occurring in living organisms. Applications of systems concepts at the mammalian genome level are quite challenging, and new complimentary computational/experimental techniques are being introduced. Most recent work applying modern systems biology techniques has been conducted on bacteria, yeast, mouse, and human genomes. However, these concepts and tools are equally applicable to other species including ruminants (e.g., livestock). In systems biology, both bottom-up and top-down approaches are central to assemble information from all levels of biological pathways that must coordinate physiological processes. A bottom-up approach encompasses draft reconstruction, manual curation, network reconstruction through mathematical methods, and validation of these models through literature analysis (i.e., bibliomics). Whereas top-down approach encompasses metabolic network reconstructions using 'omics' data (e.g., transcriptomics, proteomics) generated through DNA microarrays, RNA-Seq or other modern high-throughput genomic techniques using appropriate statistical and bioinformatics methodologies. In this review we focus on top-down approach as a means to improve our knowledge of underlying metabolic processes in ruminants in the context of nutrition. We also explore the usefulness of tissue specific reconstructions (e.g., liver and adipose tissue) in cattle as a means to enhance productive efficiency.

1.10 Mathematical representation of cell - biological system Time and space

Mathematical and computational models are increasingly used to help interpret biomedical data produced by high-throughput genomics and proteomics projects. The application of advanced

computer models enabling the simulation of complex biological processes generates hypotheses and suggests experiments. Appropriately interfaced with biomedical databases, models are necessary for rapid access to, and sharing of knowledge through data mining and knowledge discovery approaches.

Revolutions in biotechnology and information technology have produced enormous amounts of data and are accelerating the process of knowledge discovery of biological systems. These advances are changing the way biomedical research, development and applications are conducted. Clinical data complements biological data, enabling detailed descriptions of both healthy and diseased states, as well as disease progression and response to therapies. The availability of data representing various biological states, processes and their time dependencies enables the study of biological systems at various levels of organization, from molecules to organism and even up to the population level. Multiple sources of data support a rapidly growing body of biomedical knowledge, however, our ability to analyze and interpret this data lags far behind data generation and storage capacity. Mathematical and computational models are increasingly used to help interpret biomedical data produced by high-throughput genomics and proteomics projects. The application of advanced computer models enabling the simulation of complex biological processes generates hypotheses and suggests experiments. Computational models are set to exploit the wealth of data stored on biomedical databases through text mining and knowledge discovery approaches.

Modeling is the human activity consisting of representing, manipulating and communicating real-world daily life objects. As one can easily realize, there are many ways to observe an object or, equivalently, there are many different observers for the same object. Any observer has 'different views' of the same object, i.e. 'there is no omniscient observer with special access to the truth'. Each different observer collects data and generates hypothesis that are consistent with the data. This logical process is called 'abduction'. Abduction is not infallible, though; with respect to a scientific unknown, we are all blind.

A system is a collection of interrelated objects. For example, a biological system could be a collection of different cellular compartments (e.g. cell types) specialized for a specific biological function (e.g. white and red blood cells have very different commitments). An object is some elemental unit upon which observation can be made but whose internal structure is either unknown or does not exist. The choice of the elemental unit defines the representation scale of the system. A model is a description of a system in terms of constitutive objects and the relationships among them, where the description itself is, in general, decodable or interpretable by humans.

Generally speaking, a system is an unknown 'black box' (S) which, under a specific external stimulus (input E) produces a response (output R). Using this general definition, one can identify three primary scientific uses of models: (i) synthesis or knowledge discovery; to use the knowledge of inputs E and outputs R to infer system characteristics; (ii) analysis and prediction; to use the knowledge of the parts and their stimuli (i.e. the inputs E) to account for the observed response (i.e. the output R) and eventually, to predict response to different stimuli.

(iii) Instrumentation or device; to design an 'alternative system' (i.e. hardware or software), able to reproduce the input–output relationship with the best possible adherence to the studied system.

Secondary uses of models account for conceptual frameworks to design new experiments, methods to summarize or synthesize large quantities of data, tools to discover relationships among objects.

Here, we analyze models and modeling processes specific for the biology. We mainly focus on the use of models aiming at the points (i) and (ii) as tools for knowledge discovering in biology.

The mathematical methods used in modeling biological systems vary according to different steps of the process. We focus on the mathematical representation of the system. However, other important steps in the modeling processes are parameters fitting and model selection. We will not analyze the mathematical methods in those two important aspects as these would require separate review papers. Methods for parameters fitting refer to wide area of mathematical optimization, whereas methods for model selection mainly use statistical techniques. On top of these, sensitivity analysis and phase–space analysis of the models may be required. Interested readers may find more information in these references.

Models for technical use are formal models, but the strategy for building them is quite different and therefore, we leave them out of the present discussion. In the following we will refer to this type of models as Black Box Models (BBM). It is worth pointing out that, as we will mention later on, alternative systems can be considered parts of a large model to account for effects whose origin can be neglected without compromising the understanding of the whole phenomena.

1.11 Future of systems biology - Experimental

In the last 65 years of biology, we have witnessed three changes in the dominant paradigm employed to make progress in the life sciences; the systems biology of organisms (300 BC to 1950 AD), molecular biology and genetics employing a reductionist approach (1950 to 2000), and the systems biology of molecules, cells, organs, organisms, and populations (2000 into the future) that requires scientists trained in the more quantitative sciences so as to extract information from large datasets of experiments and create hypotheses and models that can and are then tested experimentally in a laboratory or the environment. The future of systems biology is clearly linked to testing ideas in the laboratory and in natural populations, employing the tools of molecular biology. The structure of the biological sciences will become like physics, with theorists and experimentalists working together to solve problems.

Systems Biology is an extremely broad topic and it is likely that the multiple subfields will develop in different directions, potentially only a subset of them fulfilling our current hopes. As such, the list below is incomplete and we don't know whether things will turn out as planned:

- **Personalised Medicine:** At the moment patients receive treatment based on inspections of the symptoms and measurements like a full blood exam. But often a treatment successfully given to one patient does not work for another one with almost identical symptoms. So physicians use a trial-and-error approach: they test treatments until one is (hopefully) successful. The idea of personalised medicine is to use DNA sequencing to gain insights into diseases like cancer; telling us what the origin of the disease is. This information can then guide doctors to a drug that is most efficient for this particular patient. One of the first patient who's cancers DNA was sequenced was Steve Jobs. Unfortunately, it did not help to overcome the disease but we are optimistic for the future and the first of these approaches were successful for some diseases.

- **Synthetic Biology:** By understanding the biological systems we might be able to create new artificial biological systems like proteins that do not exist in nature but have beneficial purposes. For example, we are already able to create synthetic antibodies, which are a body's natural way fight external bacteria.

The examples above are focussed on the pharmaceutical implications of systems biology research. To reach these goals many different other questions will need to be answered: relation between genotype and phenotype, epigenetics, metabolomics, mutliscale modelling, and most importantly how they all relate to each other, because this is the *systems* part: we can not understand the body and its function by looking just at one level of biological processes but have to look at them in an integrated way.

1.12 Planning in the Systems Biology Phase of Biological Research

The systems biology phase of biological and medical research will change the way we plan and carry out experiments to probe the complex networks of processes; our ability to predict must be greatly improved in order to help to solve these types of problems.

Experimental planning and data generation in the recent, pre-genomic phase of biological research has, at least in principle, been guided by hypotheses (hypothesis- driven research, a principle that has, however, been mitigated by the many unex- pected observations that often contribute more to our understanding of biology than the hypothesis-driven research originally planned). In the genomic phase, this has been replaced largely by the systematic analysis of all components of a process and, ideally, of all components that an organism has or is able to produce (all genes, all transcripts, all proteins, all protein complexes, all metabolites, etc.). The systems biol- ogy phase of biological research might represent a synthesis of both principles.

Since our knowledge (or hypothesis) about a process is defined by the model or models we can formulate about the process as well as the exact parameters we use in this model- ing (initial concentrations, kinetic constants etc.), we can use computational and mathematical techniques to compare these models, to identify key experiments, and to program robotic systems to carry out these experiments. Such a strategy has, for example, been used recently to carry out an analysis of yeast mutations (*The Robot Scientist*; King et al. 2004), in which the experimental planning and control of the robots actually carrying out the experiments were performed by a computer program.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

School of Bio and Chemical Engineering
DEPARTMENT OF BIOINFORMATICS

UNIT II - Quantitative Models in Biological Systems - Subject Code: SBI1402

UNIT 2 BASIC MATHEMATICAL CONCEPTS

2.1 Linear Algebra & Linear Equations
2.2 Systematic Solution of Linear Systems
2.3 Matrices - Basic Notions
2.4 Linear Dependency
2.5 Basic Matrix Operations
2.6 Ordinary Differential Equations – Notions
2.7 Solution of Linear ODE Systems
2.8 Maltus law
2.9 Stability of Steady States
2.10 Difference Equations
2.11 Graph and Network Theory
2.12 Regulatory Networks
2.13 Linear, Boolean, Bayesian Networks.

2.1 Linear Algebra & Linear Equations

Linear Algebra

In the modeling of biochemical systems, many relations do not hold just for one quantity but also for several. For example, all metabolites of a pathway have concentrations that may be concisely represented in a vector of concentrations. These metabolites are involved in a subset of the reactions occurring in this pathway; the respective stoichiometric coefficients may be presented in a matrix. Using techniques of linear algebra helps us to understand properties of biological systems. In Section 3.1.1 we will briefly recall the classical problem of how to solve a system of linear equations, since the solution algorithm represents a basic strategy. Afterwards we will introduce our notions for vectors, matrices, rank, null space, eigenvalues, and eigenvectors.

Linear Equations

Linear Equations

A linear equation of n variables $x_1, x_2, \ldots, x_n$ is an equation of the form

$$a_1 x_1 + a_2 x_2 + \ldots + a_n x_n = b\,, \tag{3-1}$$

where $a_1, a_2, \ldots, a_n, b$ are real numbers. For example, $2x_1 + 5x_2 = 10$ describes a line passing through the points $(x_1, x_2) = (5,0)$ and $(x_1, x_2) = (0,2)$. A system of m linear equations of n variables $x_1, x_2, \ldots, x_n$ is a system of linear equations as follows

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 \quad &+ \ldots + a_{1n}x_n = b_1 \\ a_{21}x_1 \quad &+ a_{22}x_2 + \ldots + a_{2n}x_n = b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 &+ \ldots + a_{mn}x_n = b_m\,. \end{aligned} \tag{3-2}$$

If $b_1 = b_2 = \ldots = b_m = 0$, the system is *homogeneous*. We wish to determine whether the system in Eq. (3-2) has a solution, i.e., if there exist numbers $x_1, x_2, \ldots, x_n$, which satisfy each of the equations simultaneously. We say that the system is *consistent* if it has a solution. Otherwise the system is called *inconsistent*.

In order to find the solution, we employ the matrix formalism (Section 3.1.2). The matrix $\boldsymbol{A}_c$ is the coefficient matrix of the system and has the dimension $m \times n$, while the matrix $\boldsymbol{A}_a$ of dimension $m \times n+1$ is called the augmented matrix of the system:

$$\boldsymbol{A}_c = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix} \qquad \boldsymbol{A}_a = \left(\begin{array}{cccc|c} a_1 & a_{12} & \ldots & a_{1n} & b_1 \\ a_{21} & a_{22} & \ldots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} & b_m \end{array}\right). \tag{3-3}$$

The solution of a single linear equation with one unknown is easy. A system of linear equations can be solved using the Gaussian elimination algorithm. The following terms are needed. A matrix is in row-echelon form if (1) all zero rows (if any) are at the bottom of the matrix and (2) if two successive rows are nonzero, the second row starts with more zeros than the first (moving from left to right).

Example 3-1

Matrix $\boldsymbol{B}_r$ is in row-echelon form and matrix $\boldsymbol{B}_n$ in non-row-echelon form:

$$\boldsymbol{B}_r = \begin{pmatrix} 3 & 0 & 0 & 1 \\ 0 & 2 & 2 & 3 \\ 0 & 0 & 0 & 4 \\ 0 & 0 & 0 & 0 \end{pmatrix} \qquad \boldsymbol{B}_n = \begin{pmatrix} 3 & 0 & 0 & 1 \\ 0 & 2 & 2 & 3 \\ 0 & 0 & 0 & 4 \\ 0 & 1 & 2 & 0 \end{pmatrix} \tag{3-4}$$

A matrix is in reduced row-echelon form if (1) it is in row-echelon form, (2) the leading (leftmost nonzero) entry in each nonzero row is 1, and (3) all other elements of the column in which the leading entry 1 occurs are equal to zero.

2.2 Systematic Solution of Linear Systems

Suppose a system of m linear equations of n unknowns $x_1, x_2, ..., x_n$ has the augmented matrix $\boldsymbol{A}$ and $\boldsymbol{A}$ is row-equivalent to the matrix $\boldsymbol{B}$, which is in reduced row-echelon form. $\boldsymbol{A}$ and $\boldsymbol{B}$ have the dimension $m \times (n+1)$. Suppose that $\boldsymbol{B}$ has r nonzero rows and that the leading entry 1 in row i occurs in column number C_i for $1 \leq i \leq r$. Then

$$1 \leq C_1 < C_2 < ... < C_r \leq n+1 \,. \tag{3-8}$$

The system is inconsistent if $C_r = n+1$. The last nonzero row of $\boldsymbol{B}$ has the form $(0, 0, ..., 0, 1)$. The corresponding equation is

$$0x_1 + 0x_2 + ... + 0x_n = 1 \,. \tag{3-9}$$

This equation has no solution. Consequently, the original system has no solution.

The system of equations corresponding to the nonzero rows of $\boldsymbol{B}$ is consistent if $C_r \leq n$. It holds that $r \leq n$.

If $r = n$ then $C_1 = 1, C_2 = 2, ..., C_n = n$, and the corresponding matrix is

$$\boldsymbol{B} = \left(\begin{array}{cccc|c} 1 & 0 & \dots & 0 & d_1 \\ 0 & 1 & \dots & 0 & d_2 \\ \vdots & & & \vdots & \vdots \\ 0 & 0 & \dots & 1 & d_n \\ 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 \end{array}\right). \tag{3-10}$$

There is a unique solution $x_1 = d_1$, $x_2 = d_2$, ..., $x_n = d_n$, which can be directly read off from $\boldsymbol{B}$.

If $r < n$, the system is underdetermined. There will be more than one solution (in fact, infinitely many solutions). To obtain all solutions, take x_{C_1}, ..., x_{C_r} as *dependent* variables and use the r equations corresponding to the nonzero rows of $\boldsymbol{B}$ to express these variables in terms of the remaining *independent* variables $x_{C_{r+1}}$, ..., x_{C_n}, which can assume arbitrary values:

$$\begin{aligned} x_{C_1} &= b_{1n+1} - b_{1C_{r+1}} x_{C_{r+1}} - \dots - b_{1C_n} x_{C_n} \\ &\vdots \\ x_{C_r} &= b_{rn+1} - b_{rC_{r+1}} x_{C_{r+1}} - \dots - b_{rC_n} x_{C_n} . \end{aligned} \tag{3-11}$$

In particular, taking $x_{C_{r+1}} = 0$, ..., $x_{C_{n+1}} = 0$ and $x_{C_n} = 0$ or $x_{C_n} = 1$ produces at least two solutions.

2.3 Matrices - Basic Notions

3.1.2.1 Basic Notions

Let us consider the space of real numbers $\Re$. A scalar is a quantity whose value can be expressed by a real number, i.e., by an element of $\Re$. It has a magnitude, but no direction. A vector is an element of the space $\Re^n$. It contains numbers for each coordinate of this space, e.g., $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$.

A matrix is a rectangular array of $m \times n$ elements of $\Re$ in m rows and n columns, such as

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} = [a_{ik}] . \tag{3-14}$$

Here and below, it holds that $i = 1, \dots, m$ and $k = 1, \dots, n$.

For our purposes, a vector can be considered as a matrix comprising only one column $(m \times 1)$.

In a zero matrix $\boldsymbol{0}$ all elements are zero ($a_{ik} = 0$ for all i, k). The matrix is a square matrix if it holds that $m = n$. A square matrix is a diagonal matrix if $a_{ik} = 0$ for all $i \neq k$. A diagonal matrix is called an identity matrix $\boldsymbol{I}_n$, if it holds that $a_{ik} = 1$, for $i = k$ or $\boldsymbol{I}_n = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$.

2.4 Linear Dependency

3.1.2.2 Linear Dependency

The vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ of type $n \times 1$ are said to be linearly dependent if scalars $\alpha_1, \ldots, \alpha_m$ exist, not all zero, such that $\alpha_1 \boldsymbol{x}_1 + \ldots + \alpha_m \boldsymbol{x}_m = 0$. In other words, one of the vectors can be expressed as a sum over certain scalar multiples of the remaining vectors, or one vector is a linear combination of the remaining vectors. If $\alpha_1 \boldsymbol{x}_1 + \ldots + \alpha_m \boldsymbol{x}_m = 0$ has only the trivial solution $\alpha_1 = \ldots = \alpha_m = 0$, the vectors are linearly independent. A set of m vectors of type $n \times 1$ is linearly dependent if $m > n$. Equivalently, a linearly independent set of m vectors must have $m \leq n$.

2.5 Basic Matrix Operations

The transpose $\boldsymbol{A}^T$ of a matrix $\boldsymbol{A}$ is obtained by interchanging rows and columns:

$$\boldsymbol{A}^T = [a_{ik}]^T = [a_{ki}]\,. \tag{3-15}$$

The sum of two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of the same size $m \times n$ is

$$\boldsymbol{A} + \boldsymbol{B} = [a_{ik}] + [b_{ik}] = [a_{ik} + b_{ik}]\,. \tag{3-16}$$

The matrix product of matrix $\boldsymbol{A}$ with sizes $m \times n$ and matrix $\boldsymbol{B}$ with size $n \times p$ is

$$\boldsymbol{A}.\boldsymbol{B} = \left[\sum_{j=1}^{n} a_{ij} \cdot b_{jk}\right]. \tag{3-17}$$

A scalar multiple of a matrix $\boldsymbol{A}$ is

$$\alpha \cdot \boldsymbol{A} = \alpha \cdot [a_{ik}] = [\alpha \cdot a_{ik}]\,. \tag{3-18}$$

Subtraction of matrices is composed of scalar multiplication with –1 and summation:

$$\boldsymbol{A} - \boldsymbol{B} = \boldsymbol{A} + (-1) \cdot \boldsymbol{B}\,. \tag{3-19}$$

Division of two matrices is not possible. However, for a square matrix $\boldsymbol{A}$ of size $n \times n$, one may in some cases find the inverse matrix $\boldsymbol{A}^{-1}$, fulfilling

$$\boldsymbol{A}.\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}.\boldsymbol{A} = \boldsymbol{I}_n\,. \tag{3-20}$$

If the respective inverse matrix $\boldsymbol{A}^{-1}$ exists, then $\boldsymbol{A}$ is called nonsingular (regular) and invertible. If the inverse matrix $\boldsymbol{A}^{-1}$ does not exist, then $\boldsymbol{A}$ is called singular. The inverse matrix of an invertible matrix is unique. For invertible matrices it holds that:

$$(\boldsymbol{A}^{-1}) = \boldsymbol{A} \tag{3-21}$$

and

$$(\boldsymbol{A}.\boldsymbol{B})^{-1} = \boldsymbol{B}^{-1}.\boldsymbol{A}^{-1}. \tag{3-22}$$

Matrix inversion: for the inverse of a 1×1 matrix, it holds that $(a_{11})^{-1} = (a_{11}^{-1})$; the inverse of a 2×2 matrix is calculated as:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}. \tag{3-23}$$

In general the inverse of an $n \times n$ matrix is given as

$$\boldsymbol{A}^{-1} = \frac{1}{Det\boldsymbol{A}} \begin{pmatrix} A_{11} & A_{21} & \dots & A_{n1} \\ A_{12} & A_{22} & \dots & A_{n2} \\ \vdots & \vdots & & \vdots \\ A_{1n} & A_{2n} & \dots & A_{nn} \end{pmatrix}, \tag{3-24}$$

where A_{ik} are the adjoints of $\boldsymbol{A}$. For $Det\boldsymbol{A}$, see below.

If a square matrix $\boldsymbol{A}$ is invertible, its rows (or columns) are linearly independent. In this case, the linear equation system $\boldsymbol{A}.\boldsymbol{x} = \boldsymbol{0}$ with $\boldsymbol{x} = (x_1, \ldots, x_m)^T$ has only the trivial solution $\boldsymbol{x} = \boldsymbol{0}$. If $\boldsymbol{A}$ is singular, i.e., rows (or columns) are linearly dependent, then the linear equation system $\boldsymbol{A}.\boldsymbol{x} = \boldsymbol{0}$ has a nontrivial solution.

The determinant of A ($Det\boldsymbol{A}$) is a real or complex number that can be assigned to every square matrix. For the 1×1 matrix (a_{11}), it holds that $Det\boldsymbol{A} = a_{11}$. For a 2×2 matrix, it is calculated as

$$Det\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}\,a_{22} - a_{12}\,a_{21}\,. \tag{3-25}$$

The value of a determinant of higher order can be obtained by an iterative procedure, i.e., by expanding the determinant with respect to one row or column: sum up every element of this row (or column) multiplied by the value of its adjoint. The adjoint A_{ik} of element a_{ik} is obtained by deleting the i-th row and the k-th column of the determinant (forming the (i,k) minor of $\boldsymbol{A}$), calculating the value of the (i,k) minor and multiplying by $(-1)^{i+k}$. For example, a determinant of third order is

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}A_{11} + a_{12}A_{12} + a_{13}A_{13}$$

$$= a_{11} \cdot (-1)^2 \cdot \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} + a_{12} \cdot (-1)^3 \cdot \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \cdot (-1)^4 \cdot \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} \tag{3-26}$$

$$= a_{11} \cdot (a_{22}\,a_{33} - a_{23}\,a_{32}) - a_{12} \cdot (a_{21}\,a_{33} - a_{23}\,a_{31}) + a_{13} \cdot (a_{21}\,a_{32} - a_{22}\,a_{31})\,.$$

The value of a determinant is zero (1) if it contains a zero row or a zero column or (2) if one row (or column) is a linear combination of the other rows (or columns). In this case, the respective matrix is singular.

2.6 Ordinary Differential Equations – Notions

An important problem in the modeling of biological systems is to characterize the dependence of certain properties on time and space. One frequently applied strategy is the description of the change of state variables by differential equations. If only temporal changes are considered, ordinary differential equations (ODEs) are used. For changes in time and space, partial differential equations are appropriate. In this section we will deal with solutions, analysis, numerical integration of ordinary differential equations, and basic concepts of dynamical systems theory as state space, trajectory, steady states, and stability.

The time behavior of biological systems in a deterministic approach can be described by a set of differential equations

$$\frac{dx_i}{dt} = \dot{x} = f_i(x_1, \ldots, x_n, p_1, \ldots, p_l, t) \quad i = 1, \ldots, n, \tag{3-30}$$

where x_i represents the variables, e.g., concentrations, p_j represents the parameters, e.g., enzyme concentrations or kinetic constants, and t is the time. We will use the notions $\frac{dx}{dt}$ and $\dot{x}$ interchangeably. In vector notation, Eq. (3-30) reads

$$\frac{d}{dt}\boldsymbol{x} = \dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{p}, t), \tag{3-31}$$

with $\boldsymbol{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$, $\boldsymbol{f} = (f_1, \ldots, f_n)^{\mathrm{T}}$, and $\boldsymbol{p} = (p_1, \ldots, p_l)^{\mathrm{T}}$.

2.7 Solution of Linear ODE Systems

We may be interested in two different types of problems: describing the temporal evolution of the system and finding its steady state. The problem of finding the steady state $\bar{\boldsymbol{x}}$ of a linear ODE system $\dot{\boldsymbol{x}} = 0$ implies that $\boldsymbol{A}\bar{\boldsymbol{x}} + \boldsymbol{z} = \boldsymbol{0}$. The solution necessitates inversion of the system matrix $\boldsymbol{A}$:

$$\bar{\boldsymbol{x}} = -\boldsymbol{A}^{-1}\boldsymbol{z}\,. \tag{3-40}$$

The time course solution of homogeneous linear ODEs is described below. The systems can be solved using an exponential function as approach. In the simplest case $n = 1$, we have

$$\frac{dx_1}{dt} = a_{11}\,x_1\,. \tag{3-41}$$

Introducing the approach $x_1(t) = b_1 e^{\lambda t}$ with constant b_1 into Eq. (3-41) yields:

$$b_1\,\lambda e^{\lambda t} = a_{11}\,b_1\,e^{\lambda t}\,. \tag{3-42}$$

Equation (3-42) is true if $\lambda = a_{11}$. This leads to the general solution

$$x_1(t) = b_1\,e^{a_{11}t}\,. \tag{3-43}$$

To find a particular solution, we must specify the initial conditions $x_1(t = 0) = x_1^0 = b_1\,e^{a_{11}t}|_{t=0} = b_1$. Thus, the solution is

$$x_1(t) = x_1^0 e^{a_{11}t}. \tag{3-44}$$

For a linear homogeneous system of n differential equations, $\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x}$, the approach is $\boldsymbol{x} = \boldsymbol{b}e^{\lambda t}$. This gives $\dot{\boldsymbol{x}} = \boldsymbol{b}\lambda e^{\lambda t} = \boldsymbol{A}\boldsymbol{b}e^{\lambda t}$. The scalar factor $e^{\lambda t}$ can be cancelled out, leading to $\boldsymbol{b}\lambda = \boldsymbol{A}\boldsymbol{b}$ or the characteristic equation

$$(\boldsymbol{A} - \lambda \boldsymbol{I}_n)\boldsymbol{b} = \boldsymbol{0}. \tag{3-45}$$

The solution of this equation is described in Section 3.1.2.

For homogeneous linear systems, the superposition principle holds: if $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are solutions of this ODE system, then their linear combination is also a solution. This leads to the general solution of the homogeneous linear ODE system:

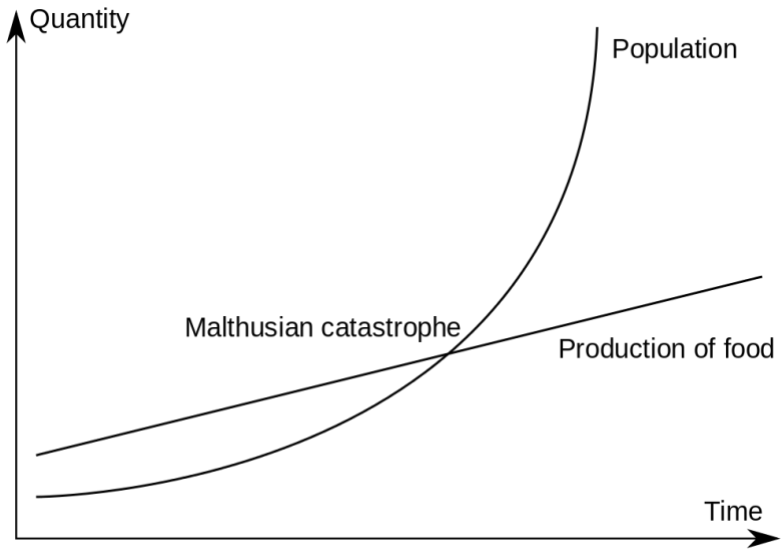
$$\boldsymbol{x}(t) = \sum_{i=1}^{n} c_i \boldsymbol{b}^{(i)} e^{\lambda_i t}, \tag{3-46}$$

where $\boldsymbol{b}^{(i)}$ are the eigenvectors of the system matrix $\boldsymbol{A}$ corresponding to the eigenvalues λ_i. A particular solution specifying the coefficients c_i can be found considering the initial conditions $\boldsymbol{x}(t=0) = \boldsymbol{x}^0 = \sum_{i=1}^{n} c_i \boldsymbol{b}^{(i)}$. This constitutes an inhomogeneous linear equation system to be solved for c_i.

For the solution of inhomogeneous linear ODEs, the system $\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{z}$ can be transformed into a homogeneous system by the coordination transformation $\hat{\boldsymbol{x}} = \boldsymbol{x} - \bar{\boldsymbol{x}}$. Since $\frac{d}{dt}\bar{\boldsymbol{x}} = \boldsymbol{A}\bar{\boldsymbol{x}} + \boldsymbol{z} = \boldsymbol{0}$, it holds that $\frac{d}{dt}\hat{\boldsymbol{x}} = \boldsymbol{A}\bar{\boldsymbol{x}}$. Therefore, we can use the solution algorithm for homogeneous systems for the transformed system.

2.8 Maltus law

Malthus believed that through preventative checks and positive checks, the population would be controlled to balance the food supply with the population level. These checks would lead to the Malthusian catastrophe.



Malthusianism is the idea that population growth is potentially exponential while the growth of the food supply or other resources is linear, which eventually reduces living standards to the point of triggering a population die off. It derives from the political and economic thought of the Reverend Thomas Robert Malthus, as laid out in his 1798 writings, *An Essay on the Principle of Population*. Malthus believed there were two types of ever-present "checks" that are continuously at work, limiting population growth based on food supply at any given time:

- *preventive checks*, such as moral restraints or legislative action — for example the choice by a private citizen to engage in abstinence and delay marriage until their finances become balanced, or restriction of legal marriage or parenting rights for persons deemed "deficient" or "unfit" by the government.

- *positive checks*, such as disease, starvation, and war, which lead to high rates of premature death — resulting in what is termed a Malthusian catastrophe. The adjacent diagram depicts the abstract point at which such an event would occur, in terms of existing population and food supply: when the population reaches or exceeds the capacity of the shared supply, positive checks are forced to occur, restoring balance. (In reality the situation would be significantly more nuanced due to complex regional and individual disparities around access to food, water, and other resources.)

  Such a catastrophe inevitably has the effect of forcing the population (quite rapidly, due to the potential severity and unpredictable results of the mitigating factors involved, as compared to the relatively slow time scales and well-understood processes governing unchecked growth or growth affected by preventive checks) to "correct" back to a lower, more easily sustainable level. Malthusianism has been linked to a variety of political and social movements, but almost always refers to advocates of population control.

2.9 Stability of Steady States

If a system is at steady state it should stay there, at least until an external perturbation occurs. Depending on systems behavior after perturbation, their steady states are

- *stable* (the system returns to this state),
- *unstable* (the system leaves this state), or
- *metastable* (the system behavior is indifferent).

A steady state is asymptotically stable if it is stable and nearby initial conditions tend to this state for $t \to \infty$. Local stability describes the behavior after small perturbations, global stability after any perturbation.

To investigate whether a steady state $\bar{\boldsymbol{x}}$ of the ODE system $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x})$ is asymptotically stable, we consider the linearized system $\dot{\boldsymbol{\xi}} = \boldsymbol{A}\boldsymbol{\xi}$ (Section 3.2.1.2) with $\boldsymbol{\xi}(t) = \boldsymbol{x}(t) - \bar{\boldsymbol{x}}$. The steady state $\bar{\boldsymbol{x}}$ is asymptotically stable if the Jacobian $\boldsymbol{A}$ has n eigenvalues with strictly negative real parts each. The steady state is unstable if at least one eigenvalue has a positive real part. This will be explained in more detail for one- and two-dimensional systems.

We start with one-dimensional systems, i.e., $n = 1$. Without a loss of generality $\bar{x}_1 = 0$ or $x_1 = \xi_1$. To the system $\dot{x}_1 = f_1(x_1)$ belongs the linearized system $\dot{x}_1 = \left.\frac{\partial f_1}{\partial x_1}\right|_{\bar{x}_1} x_1 = a_{11} x_1$. The Jacobian matrix $\boldsymbol{A} = \{a_{11}\}$ has only one eigenvalue, $\lambda_1 = a_{11} = \left.\frac{\partial f_1}{\partial x_1}\right|_{\bar{x}_1}$. The solution is $x_1(t) = x_1^0 e^{\lambda_1 t}$. It is obvious that $e^{\lambda_1 t}$ increases for $\lambda_1 > 0$ and that the system runs away from the steady state. For $\lambda_1 < 0$, the deviation from steady state decreases and $x_1(t) \to \bar{x}_1$ for $t \to \infty$. For $\lambda_1 = 0$, consideration of the linearized system allows no conclusion about stability of the original system.

Consider the two-dimensional case $n = 2$. To the system

$$\begin{aligned} \dot{x}_1 &= f_1(x_1, x_2) \\ \dot{x}_2 &= f_2(x_1, x_2) \end{aligned} \tag{3-47}$$

belongs the linearized system

$$\begin{aligned} \dot{x}_1 &= \left.\frac{\partial f_1}{\partial x_1}\right|_{\bar{x}} x_1 + \left.\frac{\partial f_1}{\partial x_2}\right|_{\bar{x}} x_2 \\ \dot{x}_2 &= \left.\frac{\partial f_2}{\partial x_1}\right|_{\bar{x}} x_1 + \left.\frac{\partial f_2}{\partial x_2}\right|_{\bar{x}} x_2 \end{aligned} \quad \text{or} \quad \dot{\boldsymbol{x}} = \begin{pmatrix} \left.\frac{\partial f_1}{\partial x_1}\right|_{\bar{x}} & \left.\frac{\partial f_1}{\partial x_2}\right|_{\bar{x}} \\ \left.\frac{\partial f_2}{\partial x_1}\right|_{\bar{x}} & \left.\frac{\partial f_2}{\partial x_2}\right|_{\bar{x}} \end{pmatrix} \boldsymbol{x} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \boldsymbol{x} = \boldsymbol{A}\boldsymbol{x} \tag{3-48}$$

To find the eigenvalues of $\boldsymbol{A}$, we have to solve the characteristic polynomial

$$\lambda^2 - \underbrace{(a_{11} + a_{22})}_{Trace\ \boldsymbol{A}} \lambda + \underbrace{a_{11}\, a_{22} - a_{12}\, a_{21}}_{Det\ \boldsymbol{A}} = 0 \tag{3-49}$$

and get

$$\lambda_{1/2} = \frac{Trace\ \boldsymbol{A}}{2} \pm \sqrt{\frac{(Trace\ \boldsymbol{A})^2}{4} - Det\ \boldsymbol{A}}\,. \tag{3-50}$$

The eigenvalues are either real for $(Trace\ \boldsymbol{A})^2/4 - Det\ \boldsymbol{A} \geq 0$) or complex (otherwise). For complex eigenvalues, the solution contains oscillatory parts.

For stability it is necessary that $Trace\ \boldsymbol{A} < 0$ and $Det\ \boldsymbol{A} > 0$. Depending on the sign of the eigenvalues, steady states of a two-dimensional system may have the following characteristics:

1. $\lambda_1 < 0$, $\lambda_2 < 0$, both real: stable node;
2. $\lambda_1 > 0$, $\lambda_2 > 0$, both real: unstable node;
3. $\lambda_1 > 0$, $\lambda_2 < 0$, both real: saddle point, unstable;
4. $\boldsymbol{Re}(\lambda_1) < 0$, $\boldsymbol{Re}(\lambda_2) < 0$, both complex with negative real parts: stable focus;
5. $\boldsymbol{Re}(\lambda_1) > 0$, $\boldsymbol{Re}(\lambda_2) > 0$, both complex with positive real parts: unstable focus; or
6. $\boldsymbol{Re}(\lambda_1) = 0$, $\boldsymbol{Re}(\lambda_2) = 0$, both complex with zero real parts: center, unstable.

Graphical representation of stability depending on trace and determinant is given in Fig. 3.2.

Up to now we have considered only the linearized system. For the stability of the original system, the following holds. If the steady state of the linearized system is asymptotically stable, then the steady state of the complete system is also asymptotically stable. If the steady state of the linearized system is a saddle, an unstable node, or an unstable focus, then the steady state of the complete system is also unstable. This means that statements about the stability remain true, but the character of the steady state is not necessarily kept. No statement about the center is possible.

The Routh-Hurwitz theorem (Bronstein and Semendjajew 1987) states: For systems with $n > 2$ differential equations, it holds that the characteristic polynomial

$$a_n \lambda^n + a_{n-1} \lambda^{n-1} + \ldots + a_1 \lambda + a_0 = 0 \tag{3-51}$$

is a polynomial of degree n, which frequently cannot be solved analytically (at least for $n < 4$). We can use the Hurwitz criterion to test whether the real parts of all eigenvalues are negative. We have to form the Hurwitz matrix $\boldsymbol{H}$, containing the coefficients of the characteristic polynomial:

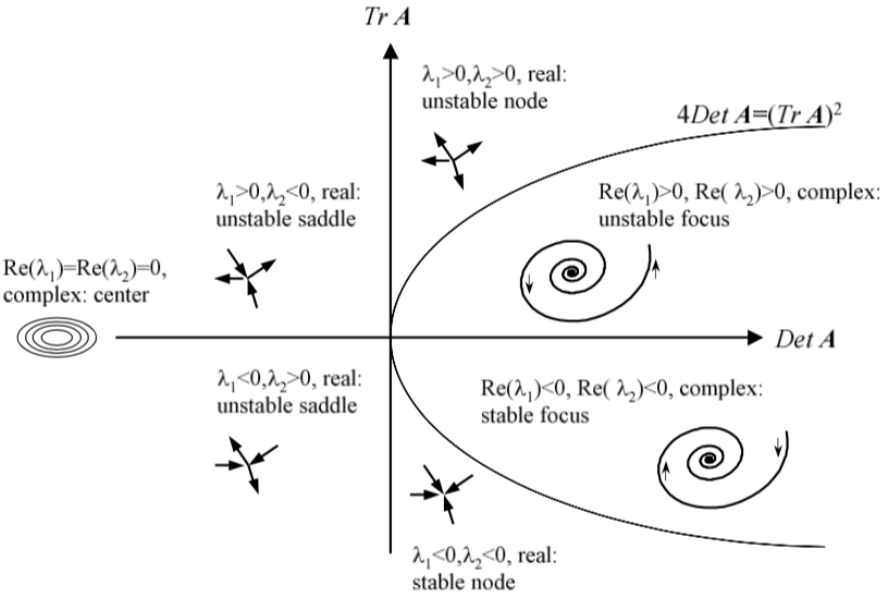


Fig. 3.2 Stability of steady states in two-dimensional systems. The character of steady-state solutions is represented depending on the value of the determinant (x-axis) and the trace (y-axis) of the Jacobian matrix. Phase plane behavior of trajectories in the different cases is schematically represented.

2.10 Difference Equations

Modeling with difference equations employs a discrete timescale, in contrast to the continuous timescale in ODEs. For some processes, the value of the variable x at a discrete time point t depends directly on the value of this variable at a former time point. For example, the actual number of individuals in a population of birds in one year can be related to the number of individuals last year.

A general (first-order) difference equation takes the form

$$x_i = f(t, x_{i-1}) \quad \text{for all } t. \tag{3-56}$$

We can solve such an equation by successive calculation: given x_0, we have

$$\begin{aligned} x_1 &= f(1, x_0) \\ x_2 &= f(2, x_1) = f\left(2, f(1, x_0)\right). \\ &\vdots \end{aligned} \tag{3-57}$$

In particular, given any value x_0, there exists a unique solution path $x_1, x_2, \ldots$ For simple forms of the function f, we can also find general solutions.

Example 3-13

Consider the exponential growth of a bacterial population with a doubling of the population size x_i in each time interval. The recursive equation $x_i = 2x_{i-1}$ is equivalent to the explicit equation $x_i = x_0 \cdot 2^i$ and also to the difference equation $x_i - x_{i-1} = \Delta x = x_{i-1}$.

The difference equation expresses the relation between values of a variable at discrete time points. We are interested in the dynamics of the variable. For the general case $x_i = rx_{i-1}$, it can be easily shown that $x_i = r^i x_0$. This corresponds to the law of exponential growth (Malthus' law). The dynamic behavior depends on the parameter r:

$1 < r$:	exponential growth
$r = 1$:	x remains constant, steady state
$0 < r < 1$:	exponential decay
$-1 < r < 0$:	alternating decay
$r = -1$:	periodic solution
$r < -1$:	alternating increase

Example time courses are shown in Fig. 3.4.

A difference equation of the form

$$x_{i+k} = f(x_{i+k}, \ldots, x_{i+1}, x_i) \tag{3-58}$$

is a k-th order difference equation. Like ODEs, difference equations may have stationary solutions that might be stable or unstable, which are defined as follows. The value $\bar{x}$ is a stationary solution or fix point of the difference equation (Eq. (3-58)) if $\bar{x} = f(\bar{x})$. A fix point is stable (or unstable), if there is a neighborhood $N = \{x : |x - \bar{x}| < \varepsilon\}$ such that every series that begins in N converges against $\bar{x}$ (leaves N). The following sentence is practically applicable: the fix point is stable under the condition that f is continuously differentiable if $\left|\frac{df(x)}{dx}\right|_{\bar{x}} < 1$.

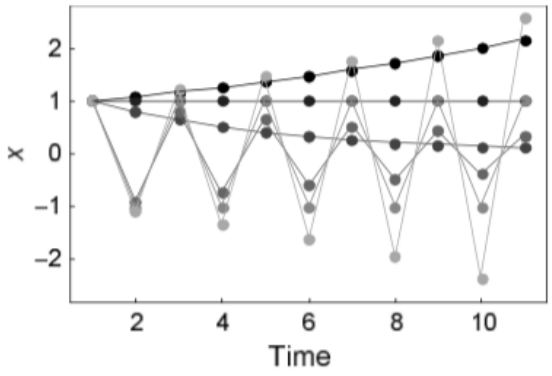


Fig. 3.4 Temporal behavior of a difference equation describing exponential growth for various values of parameter r (r drops with the gray level).

Example 3-14

The simplest form of the logistic equation, which plays a role in population dynamics, is $x_{n+1} = rx_n(1 - x_n)$, with $f(x) = rx(1 - x)$ where r is a positive valued parameter. This difference equation has two fix points, $\bar{x}_1 = 0$ and $\bar{x}_2 = 1 - \frac{1}{r}$. Stability analysis yields that fix point $\bar{x}_1$ is stable if $\left|\frac{df(x)}{dx}\right|_{\bar{x}_1} = r < 1$ and fix point $\bar{x}_2$ is stable if $\left|\frac{df(x)}{dx}\right|_{\bar{x}_2} = |2 - r| < 1$; hence, $1 < r < 3$.

For $r > 3$ there are stable oscillations of period 2, i.e., successive generations alternate between two values. Finding the steady states $\bar{\bar{x}}_1$ and $\bar{\bar{x}}_2$ is enabled by the new function $g(x) = f(f(x))$. The equation $g(x) = x$ has the two solutions $\bar{\bar{x}}_{1,2} = \frac{r+1 \pm \sqrt{(3-r)(r+1)}}{2r}$. They are stable if $\left|\frac{dg(x)}{dx}\right|_{\bar{\bar{x}}_i} < 1$ holds for $i = 1, 2$ or $\left|\left(\frac{df(x)}{dx}\right)\Big|_{\bar{\bar{x}}_1} \cdot \left(\frac{df(x)}{dx}\right)\Big|_{\bar{\bar{x}}_2}\right| < 1$, i.e., for $3 < r < 3.3$. For $r > 3.3$, oscillations of higher period occur, which can be treated in a manner analogous to oscillations of period 2. For $r > r_{crit}$ chaos arises, i.e., albeit a deterministic description, the system trajectories in fact cannot be reliably predicted and may differ remarkably for close initial conditions. The points $r = 1$, $r = 3$, and $r = 3.3$ are bifurcation points since the number and stability of steady states change. A graphical representation is given in Fig. 3.5.

2.11 Graph and Network Theory

Many kinds of data arising in systems biology applications can be represented as graphs (metabolic pathways, signaling pathways, or gene regulatory networks). Other examples are taxonomies, e.g., of enzymes or organisms; protein interaction networks; DNA, RNA, or protein sequences; chemical structure graphs; or gene co-expression. In this section we give a brief overview of the formalization of graph problems (Section 3.5.1) and introduce specifically the framework of gene regulatory networks (Section 3.5.2) that are essential for the analysis of transcriptome data.

tory networks (Section 3.5.2) that are essential for the analysis of transcriptome data.

The degree of a vertex i, $d(i)$, in an undirected graph is the number of edges connected to i, $d(i) = \left|\{(i,j) \in E; j = 1, ..., n\}\right|$. The degree of a vertex i in a directed graph is defined as the sum of its in-degree and out-degree. The in-degree of vertex i is defined as the number of edges entering vertex i, and the out-degree is the number of edges leaving it. The degree of a vertex i can be computed from the adjacency matrix as the sum of the i-th row (out-degree) and the i-th column sums (in-degree).

Topological properties of interaction graphs are commonly used in applications to characterize biological function (Jeong et al. 2001; Stelling et al. 2002; Przulj et al. 2004). For example, lethal mutations in protein-protein interactions are defined by highly connected parts of a protein interaction graph whose removal disrupts the graph structure.

A path of length l from a vertex v_0 to a vertex v_i in a graph $G(V, E)$ is a sequence of vertices $v_0, ..., v_l$ such that $(v_{i-1}, v_i) \in E$ for $i = 1, ..., l$. A path is a cycle if $l \geq 1$ and $v_0 = v_l$. A directed graph that contains no cycle is called a directed acyclic graph. The weight of a path in a weighted directed graph is the sum of the weights of all edges constituting the path. The shortest path from vertex v_0 to vertex v_l is the path with the minimal weight. If all weights are equal, then the shortest path is the path from vertex v_0 to vertex v_l with the minimal number of edges.

An important practical problem consists of the identification of substructures of a given graph. An undirected graph is connected when a path exists for each pair of vertices. If a subset of a graph is connected, it is called a connected component.

2.12 Regulatory Networks

Regulatory networks are graph-based models for a simplified view on gene regulation (cf. Chapter 8). Transcription factors are stimulated by upstream signaling cascades and bind on *cis*-regulatory positions of their target genes. Bound transcription factors promote or inhibit RNA polymerase assembly and thus determine whether and to what extent the target gene is expressed. The modeling of gene regulation via genetic networks has been widely used in practice (for a review, see de Jong 2002). We give here a brief introduction to some basic principles.

2.13 Linear, Boolean, Bayesian Networks.

The general model of gene regulation assumes that the change of gene expression of gene x_i at time t can be described by the following equation

$$\frac{dx_i(t)}{dt} = r_i f\left(\sum_{j=1}^{n} w_{ij}\, x_j(t) + \sum_{k=1}^{m} v_{ik}\, u_k(t) + b_i\right) - \lambda_i\, x_i(t)\,, \tag{3-100}$$

where

f is the activation function,
$x_i(t)$ is the gene expression of gene i at time t,
r_i is the reaction rate of gene i,
w_{ij} is the weight that determines the influence of gene j on gene i,
$u_k(t)$ are the external inputs (e. g., a chemical compound) at time t,
v_{ik} is the weight that determines the influence of external compound k on gene i,
b_i is a lower base level of gene i, and
λ_i is the degradation constant for gene i.

The activation function, f, is a monotone function, assuming that the concentration of the gene is monotonically dependent on the concentrations of its regulators. Often, these functions have sigmoid form, such as $f(z) = (1 + e^{-z})^{-1}$. If this function is the identity, i. e., $f(z) = z$, then the network is linear. Additionally, common simplifications include constancy in the reaction rates, no external influence, and no degradation, so that Eq. (3-100) reduces to

$$\frac{dx_i(t)}{dt} = \sum_{j=1}^{n} w_{ij}\, x_j(t) + b_i\,. \tag{3-101}$$

These models have been investigated, for example, by D'Haeseleer et al. (1999). The interesting parameters are the weights w_{ij}, which are estimated by statistical methods (cf. Section 3.4.4).

3.5.2.2 Boolean Networks

Boolean networks are qualitative descriptions of gene regulatory interactions. Gene expression has two states: on (1) and off (0) (Kauffman 1993; Akutsu et al. 1999,

2000, Cormen et al. 2001). Let $\boldsymbol{x}$ be an n-dimensional binary vector representing the state of a system of n genes. Thus, the state space of the system consists of 2^npossible states. Each component, x_i, determines the expression of the i-th gene. With each gene i we associate a Boolean rule, b_i. Given the input variables for gene i at time t, this function determines whether the regulated element is active or inactive at time $t+1$, i. e.,

$$x_i(t+1) = b_i(\boldsymbol{x}(t)), 1 \leq i \leq n. \tag{3-102}$$

Equation (102) describes the dynamics of the Boolean network. The practical feasibility of Boolean networks is heavily dependent on the number of input variables, k, for each gene. The number of possible input states of k inputs is 2^k. For each such combination, a specific Boolean function must determine whether the next state would be on or off. Thus, there are 2^{2k} possible Boolean functions (or rules). This number rapidly increases with the connectivity. For $k = 2$ we have four possible input states and 16 possible rules; for $k = 3$, we have eight possible input states and 256 possible rules, etc.

In a Boolean network each state has a deterministic output state. A series of states is called a trajectory. If no difference occurs between the transitions of two states, i.e., output state equals input state, then the system is in a point attractor. Point attractors are analogous to steady states (cf. Section 3.2.3). If the system is in a cycle of states, then we have a dynamic attractor. The Boolean rules for one and two inputs as well as examples for the dynamic behavior of Boolean networks are given in Chapter 10, Section 10.3.3.

There have been algorithms to reconstruct or reverse engineer (cf. Chapter 9) Boolean networks from time series of gene expression data, i.e., from a limited number of states. Among the first was *REVEAL* developed by Liang et al. (1999). Additionally, properties of random Boolean networks were intensively investigated by Kauffman (1993), e.g., global dynamics, steady states, connectivity, and the specific types of Boolean functions.

3.5.2.3 Bayesian Networks

Bayesian networks are probabilistic descriptions of the regulatory network (Heckerman 1998; Friedman et al. 2000; Jensen 2001). A Bayesian network consists of (1) a directed acyclic graph, $G(V, E)$ (cf. Section 3.5.1), and (2) a set of probability distributions. The n vertices (n genes) correspond to random variables x_i, $1 \leq i \leq n$. For example, in regulatory networks the random variables describe the gene expression level of the respective gene. For each x_i, a conditional probability $p(x_i | L(x_i))$ is defined, where $L(x_i)$ denotes the parents of gene i, i.e., the set of genes that have a direct regulatory influence on gene i. Figure 3.11 gives an example of a Bayesian network consisting of five vertices.

The set of random variables is completely determined by the joint probability distribution. Under the Markov assumption, i.e., the assumption that each x_i is conditionally independent of its non-descendants given its parents, this joint probability distribution can be determined by the factorization via

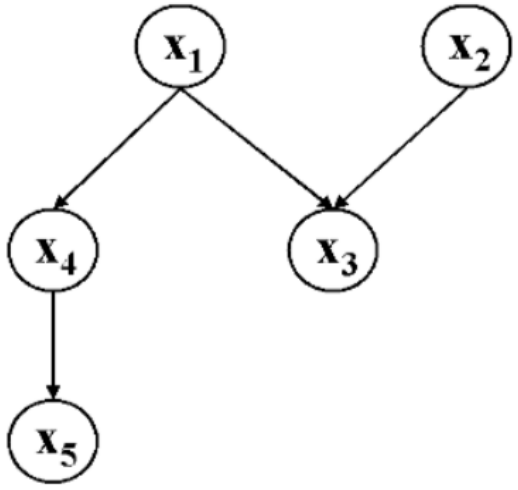


Fig. 3.11 Bayesian network. The network structure determines, e.g., the conditional independencies $i(x_1, x_2)$, $i(x_3, x_4|x_1)$, $i(x_5, x_3|x_4)$. The joint probability distribution has the form

$$p(\boldsymbol{x}) = \prod_{i=1}^{n} p(x_i|L(x_i)). \tag{3-103}$$

Here, conditional independence of two random variables x_i and x_j given a random variable x_k means that $p(x_i, x_j|x_k) = p(x_i|x_k)\, p(x_j|x_k)$ or, equivalently, $p(x_i|x_j, x_k) = p(x_i|x_k)$. The conditional distributions given in Eq. (3-103) are typically assumed to be linearly normally distributed, i.e., $p(x_i|L(x_i)) \sim N\left(\sum_k a_k x_k, \sigma^2\right)$, where x_k is in the parent set of x_i. Thus, each x_i is assumed to be normally distributed around a mean value that is linearly dependent on the values of its parents.

The typical application of Bayesian networks is learning from observations. Given a training set T of independent realizations of the n random variables $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^n$, the problem is to find a Bayesian network that best matches T. A common solution is to assign a score to each calculated network using the *a posteriori* probability of the calculated network, N, given the training data (cf. Section 3.4.1) by $\log P(N|T) = \log \frac{P(T|N)\, P(N)}{P(T)} = \log P(T|N) + \log P(N) + const$, where the constant is independent of the calculated network and $P(T|N) = \int P(T|N, \Theta) P(\Theta|N)\, d\Theta$ is the marginal likelihood that averages the probability of the data over all possible parameter assignments to the network. The choice of the *a priori* probabilities $P(N)$ and $P(\Theta|N)$ determines the exact score. The optimization of the *a posteriori* probability is beyond the scope of this introduction. For further reading, see e.g., Friedman et al. 2000; Jensen 2001; and Chickering 2002.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

School of Bio and Chemical Engineering
DEPARTMENT OF BIOINFORMATICS

UNIT III - Quantitative Models in Biological Systems - Subject Code: SBI1402

UNIT 3 STANDARD MODELS AND APPROACHES

3.1 Metabolism
3.2 Enzyme Kinetics and Thermodynamics
3.3 The Law of Mass Action
3.4 Reaction Kinetics and Thermodynamics
3.5 Review of network concepts
3.6 Properties and modelling of feedback/feedforward system
3.7 Reaction kinetics
3.8 competitive inhibition
3.9 co-operativity
3.10 Hyperbolic and sigmoidal responses
3.11 Michaelis-Menten Kinetics
3.12 Metabolic Networks
3.13 Systems Equations
3.14 Information Contained in the Stoichiometric Matrix N
3.15 Flux Balance Analysis
3.16 Signal Transduction
3.17 Function and Structure of Intra- and Intercellular Communication
3.18 Structural Components of Signaling Pathways
3.19 G Proteins, Ras Proteins
3.20 MAP Kinase Cascades
3.21 Apoptotic pathway
3.22 Two component signalling pathways of bacterial chemotaxis.

UNIT 3 STANDARD MODELS AND APPROACHES

3.1 Metabolism

Living cells require energy and material for building membranes, storing molecules, replenishing enzymes, replication and repair of DNA, movement, and many other processes. Through metabolism cells acquire energy and use it to build new cells. Metabolism is the means by which cells survive and reproduce. Metabolism is the general term for two kinds of reactions: (1) catabolic reactions (breakdown of com- plex compounds to get energy and building blocks) and (2) anabolic reactions (con- struction of complex compounds used in cellular functioning). Metabolism is a highly organized process. It involves thousands of reactions that are catalyzed by en- zymes.

Metabolic networks consist of reactions transforming molecules of one type into molecules of another type. In modeling terms, the concentrations of the molecules and their rates of change are of special interest. The basic concepts of reaction net- works, which are outlined here, may also be applied for other types of cellular reac- tion networks, e. g., signal transduction pathways. In this chapter metabolism will be studied on three levels of abstraction:

1. Enzyme kinetics investigates the dynamic properties of the individual reactions in isolation.
2. The network character of metabolism is studied with stoichiometric analysis con- sidering the balance of compound production and degradation.
3. Metabolic control analysis quantifies the effect of perturbations in the network employing the individual dynamics of concentration changes and their integration in the network.

Note that most modeling approaches for individual biochemical reactions or net-

works of such reactions that are presented in this chapter also apply for other types of networks, such as signaling cascades or binding of transcription factors to DNA. Since the modeling of metabolic networks is the most elaborate, it is subsumed here.

In order to illustrate the theoretical concepts, we will apply a running example throughout this chapter. This example comprises a subset of reactions of glycolysis in yeast as represented by Hynne and colleagues (2001). You can also find the com- plete model and many other models in modeling databases (Snoep and Olivier 2002).

Example 1

We will consider the first four reactions from the upper part of glycolysis as well as reactions balancing the energy currency ATP and ADP as represented in Fig. 5.1.

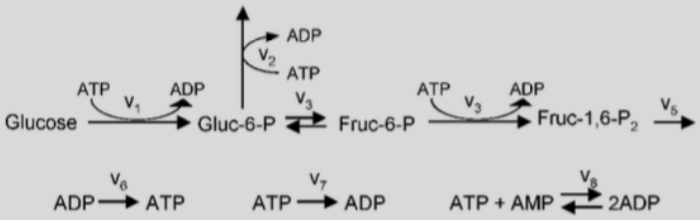


Fig. 5.1 Schematic representation of the upper part of glycolysis, i. e., the degradation of glucose in order to yield energy and building blocks for cellular processes. Abbreviations: Gluc6P: glucose-6-phosphate; Fruc6P: fructose-6-phosphate; Fruc1,6P_2: fructose-1,6-bisphosphate; ATP: adenosine-triphosphate; ADP: adenosine-diphosphate: AMP: adenosine-monophosphate. Reactions: v_1: hexokinase; v_2: consumption of glucose-6-phosphate in other pathways; v_3: phosphoglucoisomerase; v_4: phosphofructokinase; v_5: aldolase; v_6: ATP production in lower glycolysis; v7: ATP consumption in other pathways; v_8: adenylate kinase.

The ODE system for this reaction system is given by

$$\frac{d}{dt}Gluc6P = v_1 - v_2 - v_3$$

$$\frac{d}{dt}Fruc6P = v_3 - v_4$$

$$\frac{d}{dt}Fruc1,6P_2 = v_4 - v_5$$

$$\frac{d}{dt}ATP = -v_1 - v_2 - v_4 + v_6 - v_7 - v_8$$

$$\frac{d}{dt}ADP = v_1 + v_2 + v_4 - v_6 + v_7 + 2\,v_8$$

$$\frac{d}{dt}AMP = -v_8\,. \qquad (5\text{-}1)$$

Abbreviations are explained in the legend of Fig. 5.1. The individual rate expressions read

3.2 Enzyme Kinetics and Thermodynamics

This chapter deals with the deterministic kinetic modeling of individual biochemical reactions. The basic quantities are the concentration S of a substance S (i.e., the number n of molecules of this substance per volume V) and the rate v of a reaction (i.e., the change of concentration S per time t). This type of modeling is macroscopic or phenomenological compared to the microscopic approach, where single molecules and their interactions are considered.

Chemical and biochemical kinetics rely on the assumption that the reaction rate v at a certain point in time and space can be expressed as a unique function of the concentrations of all substances at this point in time and space. Classical enzyme kinetics assumes for simplicity's sake a spatial homogeneity (the "well-stirred" test tube) and no direct dependency of the rate on time:

$$v(t) = v(S(t)). \tag{5-10}$$

In more advanced modeling approaches moving towards whole-cell modeling, spatial inhomogeneities are taken into account, paying tribute to the fact that many components are membrane-bound or that cellular structures hinder the free movement of molecules. However, in most cases one can assume that diffusion is rapid enough to allow for an even distribution of all substances in space.

Enzymes catalyze biochemical reactions. Enzymes are proteins, often in complex with cofactors (Chapter 2, Section 2.1). They have a catalytic center, are usually highly specific, and remain unchanged by the reaction. One enzyme molecule catalyzes about a thousand reactions per second (the so-called turnover number ranges from $10^2\,s^{-1}$ to $10^7\,s^{-1}$). This leads to a rate acceleration of about 10^6- to 10^{12}-fold compared to the uncatalyzed, spontaneous reaction.

3.3 The Law of Mass Action

Biochemical kinetics is based on the mass action law, introduced by Guldberg and Waage in the 19th century (Waage and Guldberg 1864; Guldberg and Waage 1867, 1879). It states that the reaction rate is proportional to the probability of a collision of the reactants. This probability is in turn proportional to the concentration of reactants to the power of the molecularity, i. e., the number in which they enter the specific reaction. For a simple reaction like

$$S_1 + S_2 \rightleftharpoons 2P\,, \tag{5-11}$$

the reaction rate reads

$$v = v_+ - v_- = k_+ S_1 \cdot S_2 - k_- P^2\,. \tag{5-12}$$

v is the net rate, v_+ the rate of the forward reaction, v_- the rate of the backward reaction, and k_+ and k_- are the respective proportionality factors, the so-called kinetic or rate constants. The molecularity is 1 for each substrate of the forward reaction and 2 for the backward reaction. If we measure the concentration in moles per liter (mol $\cdot$ L^{-1} or M) and the time in seconds (s), then the rates have the unit M $\cdot$ s^{-1}. Accordingly, the rate constants for bimolecular reactions have the unit M $\cdot$ s^{-1}. Rate constants of monomolecular reactions have the dimension s^{-1}. The general mass action rate law for a reaction with substrate concentrations S_i and product concentrations P_j reads

$$v = v_+ - v_- = k_+ \prod_i S_i^{m_i} - k_- \prod_j P_j^{m_j}\,, \tag{5-13}$$

where m_i and m_j denote the respective molecularities of S_i and P_j in this reaction (Heinrich and Schuster 1996).

The equilibrium constant K_{eq} (we will also use the simpler symbol q) characterizes the ratio of substrate and product concentrations in equilibrium (S_{eq} and P_{eq}), i. e., the state with equal forward and backward rates. The rate constants are related to K_{eq} in the following way:

$$K_{eq} = \frac{k_+}{k_-} = \frac{\prod P_{eq}}{\prod S_{eq}}\,. \tag{5-14}$$

The relation between the thermodynamic description and the kinetic description of biochemical reactions will be outlined in Section 5.1.2.

The dynamics of the concentrations for Eq. (5-11) is described by the ODEs

$$\frac{d}{dt}S_1 = \frac{d}{dt}S_2 = -v$$

$$\frac{d}{dt}P = 2v\,. \tag{5-15}$$

The time course of S_1, S_2 and P is obtained by integration of these ODEs.

Example 5-2

The kinetics of a simple decay such as

$$S \longrightarrow \tag{5-16}$$

is described by $v = kS$ and $\frac{d}{dt}S = -kS$. Integration of this ODE from time $t = 0$ with the initial concentration S_0 to an arbitrary time t with concentration $S(t)$, $\int_{S_0}^{S} \frac{dS}{S} = -\int_{t=0}^{t} k\,dt$, yields the temporal expression $S(t) = S_0 e^{-kt}$.

3.4 Reaction Kinetics and Thermodynamics

An important purpose of metabolism is to extract energy from nutrients, which is necessary for the synthesis of molecules, for growth, and for proliferation. We distinguish between energy-supplying reactions, energy-demanding reactions, and energetically neutral reactions. The principles of reversible thermodynamics and their application to chemical reactions allow understanding of energy circulation in the cell. This is eased by the assumption that biological reactions usually occur in hydrous solution at constant pressure and constant temperature with negligible volume changes.

Whether a reaction occurs spontaneously or not, in which direction a reaction proceeds, and the position of the equilibrium are important characteristics of a biochemical process. The first law of thermodynamics, i. e., the law of energy conservation, tells us only that the total energy of a system remains constant during any process. The second law of thermodynamics declares that a process occurs spontaneously only if it increases the total entropy of the system. Unfortunately, entropy is usually not directly measurable. A more suitable measure is the Gibbs free energy G, which is the energy capable of carrying out work under isotherm-isobar conditions, i. e., at constant temperature and constant pressure. The change of the free energy is given as

$$\Delta G = \Delta H - T\Delta S\,, \tag{5-17}$$

where ΔH is the change in enthalpy, ΔS is the change in entropy, and T is the absolute temperature in Kelvin. ΔG is a measure for the driving force, the spontaneity of a chemical reaction. If $\Delta G < 0$ then the reaction proceeds spontaneously under release of energy (exergonic process). If $\Delta G > 0$ then the reaction is energetically not favorable and will not occur spontaneously (endergonic process). $\Delta G = 0$ means that the system has reached its equilibrium. Endergonic reactions may proceed if they obtain energy from a strictly exergonic reaction by energetic coupling. Free energy is usually given for standard conditions (ΔG^0), i. e., for a concentration of the reaction partners of 1 M, temperature $T = 298$ K, and, for gaseous reactions, a pressure of $p =$ 98.1 kPa = 1 atm. The unit is kJ mol^{-1}. For the free energy difference, a set of relations holds as follows. The free energy difference is related to redox potential $E_{red/ox}$:

$$\Delta G = -nF \cdot E_{red/ox}\,, \tag{5-18}$$

where n is the number of transferred charges and F is the Faraday constant (96,500 coulomb). The free energy difference for a reaction can be calculated from the difference of the sums of free energies of its products P and its substrates S:

$$\Delta G = \sum G_P - \sum G_S\,. \tag{5-19}$$

The enzyme cannot change the free energies of the substrates and products of a reaction, nor their differences, but it changes the so-called reaction path, thereby lowering the activation energy for the reaction. The transition state theory explains this (Haynie 2001). It has been observed that many substances or mixtures are thermodynamically unstable, since $\Delta G \ll 0$ (see Tab. 5.1). Nevertheless, they can be stored under normal conditions for a long time. The reason is that during the course of a reaction, the metabolites must pass one or more transition states of maximal free energy, in which bonds are solved or newly formed. The transition state is unstable; the respective molecule configuration is called an activated complex. It has a lifetime of around one molecule vibration, $10^{-14} \ldots 10^{-13}$ s, and it can hardly be experimentally verified. The difference $\Delta G^{\neq}$ of free energy between the reactants and

Tab. 5.1 Values of $\Delta G^{0\prime}$ for some important reactions

Reaction	$\Delta G^{0\prime}/(kJ\ mol^{-1})$
$2H_2 + O_2 \rightarrow 2\ H_2O$	–474
$2H_2O_2 \rightarrow 2\ H_2O + O_2$	–99
$PP_i + H_20 \rightarrow 2\ P_i$	–33.49
$ATP + H_20 \rightarrow ADP + P_i$	–30.56
Glucose-6-phosphate + $H_20 \rightarrow$ Glucose + P_i	–13.82
Glucose + $P_i \rightarrow$ Glucose-6-phosphate + H_20	+13.82
Glucose-1-phosphate → Glucose-6-phosphate	–7.12
Glucose-6-phosphate → Fructose-6-phosphate	+1.67
Glucose + $6\ O_2 \rightarrow 6\ CO_2 + 6\ H_20$	–2890

Source: Lehninger 1975

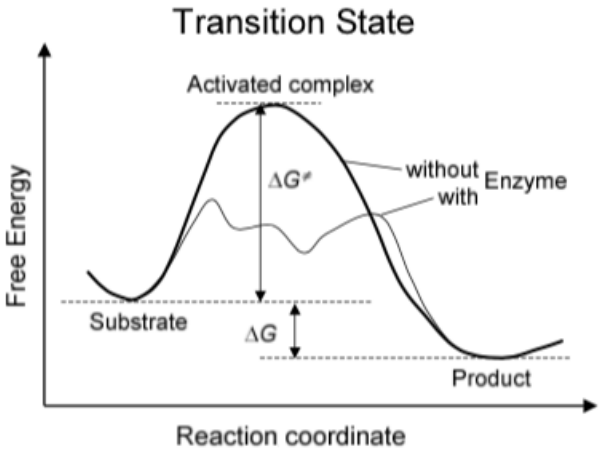


Fig. 5.3 Presentation of the change of free energy along the course of reaction. The substrate and the product are situated in local minima of the free energy; the active complex is assigned to the local maximum. The enzyme may change the reaction path and thereby lower the barrier of free energy.

the activated complex determines the dynamics of a reaction: the higher this difference, the lower the probability that the molecules may pass this barrier and the lower the rate of the reaction. The value of $\Delta G^{\neq}$ depends on the type of altered bonds, on steric, electronic, or hydrophobic demands, and on temperature.

Figure 5.3 presents a simplified view of the reaction course. The substrate and the product are situated in local minima of the free energy; the active complex is assigned to the local maximum. The free energy difference ΔG is proportional to the logarithm of the equilibrium constant of the respective reaction:

$$\Delta G = -RT \ln K_{eq}\,, \tag{5-20}$$

(R – gas constant, 8.314 J $mol^{-1} K^{-1}$). The value of $\Delta G^{\neq}$ corresponds to the kinetic constant k_+ of the forward reaction (Eqs. (12)–(14)) by $\Delta G^{\neq} = -RT \ln k_+$, while $\Delta G^{\neq} + \Delta G$ is related to the rate constant k_- of the backward reaction.

The interaction of the reactants with an enzyme may alter the reaction path and thereby lead to lower values of $\Delta G^{\neq}$. Furthermore, the free energy may assume more local minima and maxima along the path of reaction. They are related to unstable intermediary complexes. Values for the difference of free energy for some biologically important reactions are given in Tab. 5.1.

The detailed consideration of enzyme mechanisms by applying the mass action law for single events has led to a number of standard kinetic descriptions, which will be explained in the following sections.

3.5 Review of network concepts

We are surrounded by systems that are hopelessly complicated. Consider for example the society that requires cooperation between billions of individuals, or communications infrastructures that integrate billions of cell phones with computers and satellites. Our ability to reason and comprehend our world requires the coherent activity of billions of neurons in our brain. Our biological existence is rooted in seamless interactions between thousands of genes and metabolites within our cells.

These systems are collectively called *complex systems*, capturing the fact that it is difficult to derive their collective behavior from a knowledge of the system's components. Given the important role complex systems play in our daily life, in science and in economy, their understanding, mathematical description, prediction, and eventually control is one of the major intellectual and scientific challenges of the 21st century.

The emergence of network science at the dawn of the 21st century is a vivid demonstration that science can live up to this challenge. Indeed, behind each complex system there is an intricate network that encodes the interactions between the system's components:

a. The network encoding the interactions between genes, proteins, and metabolites integrates these components into live cells. The very existence of this *cellular network* is a prerequisite of life.
b. The wiring diagram capturing the connections between neurons, called the *neural network*, holds the key to our understanding of how the brain functions and to our consciousness.
c. The sum of all professional, friendship, and family ties, often called the *social network*, is the fabric of the society and determines the spread of knowledge, behavior and resources.
d. *Communication networks*, describing which communication devices interact with each other, through wired internet connections or wireless links, are at the heart of the modern communication system.
e. The *power grid*, a network of generators and transmission lines, supplies with energy virtually all modern technology.
f. *Trade networks* maintain our ability to exchange goods and services, being responsible for the material prosperity that the world has enjoyed since WWII.

Networks are also at the heart of some of the most revolutionary technologies of the 21st century, empowering everything from Google to Facebook, CISCO, and Twitter. At the end, networks permeate science, technology, business and nature to a much higher degree than it may be evident upon a casual inspection. Consequently, *we will never understand complex systems unless we develop a deep understanding of the networks behind them*.

The exploding interest in network science during the first decade of the 21st century is rooted in the discovery that despite the obvious diversity of complex systems, the structure and the evolution of the networks behind each system is driven by a common set of fundamental laws and principles. Therefore, notwithstanding the amazing differences in form, size, nature, age, and scope of real networks, most networks are driven by common organizing principles. Once we disregard the nature of the components and the precise nature of the interactions between them, the obtained networks are more similar than different from each other. In the following sections we discuss the forces that have led to the emergence of this new research field and its impact on science, technology, and society.

3.6 Properties and modelling of feedback/feedforward system

Feed forward loop (FFL) motif is one of the most significant one in both *E. coli*and yeast. The FFL is composed of a transcription factor X, which regulates a second transcription factor Y. X and Y both bind the regulatory region of target gene Z and jointly modulate its transcription rate. The FFL has three transcription interactions. Each of these can be either positive (activation) or negative (repression). There are therefore eight possible structural configurations of activator and repressor interactions. Four of these configurations are termed "coherent": the sign of the direct regulation path (from X to Z) is the same as the overall sign of the indirect regulation path (from X through Y to Z). The other four structures are termed "incoherent": the signs of the direct and indirect regulation paths are opposite. Mathematical modeling indicates that FFLs can serve as a novel mechanism for accelerating the expression of the target genes.

Feedback is defined as the information gained about a reaction to a product, which will allow the modification of the product. Feedback loops are therefore the process whereby a change to the system results in an alarm which will trigger a certain result. This result will either increase the change to the system or reduce it to bring the system back to normal. A few questions remain: How do these systems work? What is a positive feedback? What is negative feedback? Where do we find these systems in nature?

Biological systems operate on a mechanism of inputs and outputs, each caused by and causing a certain event. A feedback loop is a biological occurrence wherein the output of a system amplifies the system (positive feedback) or inhibits the system (negative feedback). Feedback loops are important because they allow living organisms to **maintain homeostasis**.

3.7 Reaction kinetics

Reaction Kinetics and Thermodynamics

An important purpose of metabolism is to extract energy from nutrients, which is necessary for the synthesis of molecules, for growth, and for proliferation. We distinguish between energy-supplying reactions, energy-demanding reactions, and energetically neutral reactions. The principles of reversible thermodynamics and their application to chemical reactions allow understanding of energy circulation in the cell. This is eased by the assumption that biological reactions usually occur in hydrous solution at constant pressure and constant temperature with negligible volume changes.

Whether a reaction occurs spontaneously or not, in which direction a reaction proceeds, and the position of the equilibrium are important characteristics of a biochemical process. The first law of thermodynamics, i.e., the law of energy conservation, tells us only that the total energy of a system remains constant during any process. The second law of thermodynamics declares that a process occurs spontaneously only if it increases the total entropy of the system. Unfortunately, entropy is usually not directly measurable. A more suitable measure is the Gibbs free energy G, which is the energy capable of carrying out work under isotherm-isobar conditions, i.e., at constant temperature and constant pressure. The change of the free energy is given as

$$\Delta G = \Delta H - T\Delta S, \tag{5-17}$$

where ΔH is the change in enthalpy, ΔS is the change in entropy, and T is the absolute temperature in Kelvin. ΔG is a measure for the driving force, the spontaneity of a chemical reaction. If $\Delta G < 0$ then the reaction proceeds spontaneously under release of energy (exergonic process). If $\Delta G > 0$ then the reaction is energetically not favorable and will not occur spontaneously (endergonic process). $\Delta G = 0$ means that the system has reached its equilibrium. Endergonic reactions may proceed if they obtain energy from a strictly exergonic reaction by energetic coupling. Free energy is usually given for standard conditions (ΔG^0), i.e., for a concentration of the reaction partners of 1 M, temperature $T = 298$ K, and, for gaseous reactions, a pressure of $p =$ 98.1 kPa = 1 atm. The unit is kJ mol^{-1}. For the free energy difference, a set of relations holds as follows. The free energy difference is related to redox potential $E_{red/ox}$:

$$\Delta G = -nF \cdot E_{red/ox}\,, \tag{5-18}$$

where n is the number of transferred charges and F is the Faraday constant (96,500 coulomb). The free energy difference for a reaction can be calculated from the difference of the sums of free energies of its products P and its substrates S:

$$\Delta G = \sum G_P - \sum G_S \,. \tag{5-19}$$

The enzyme cannot change the free energies of the substrates and products of a reaction, nor their differences, but it changes the so-called reaction path, thereby lowering the activation energy for the reaction. The transition state theory explains this (Haynie 2001). It has been observed that many substances or mixtures are thermodynamically unstable, since $\Delta G \ll 0$ (see Tab. 5.1). Nevertheless, they can be stored under normal conditions for a long time. The reason is that during the course of a reaction, the metabolites must pass one or more transition states of maximal free energy, in which bonds are solved or newly formed. The transition state is unstable; the respective molecule configuration is called an activated complex. It has a lifetime of around one molecule vibration, $10^{-14} \ldots 10^{-13}$ s, and it can hardly be experimentally verified. The difference $\Delta G^{\neq}$ of free energy between the reactants and

Tab. 5.1 Values of $\Delta G^{0\prime}$ for some important reactions

Reaction	$\Delta G^{0\prime}$/(kJ mol^{-1})
$2H_2 + O_2 \rightarrow 2\,H_2O$	–474
$2H_2O_2 \rightarrow 2\,H_2O + O_2$	–99
$PP_i + H_20 \rightarrow 2\,P_i$	–33.49
$ATP + H_20 \rightarrow ADP + P_i$	–30.56
Glucose-6-phosphate + $H_20 \rightarrow$ Glucose + P_i	–13.82
Glucose + $P_i \rightarrow$ Glucose-6-phosphate + H_20	+13.82
Glucose-1-phosphate → Glucose-6-phosphate	–7.12
Glucose-6-phosphate → Fructose-6-phosphate	+1.67
Glucose + $6\,O_2 \rightarrow 6\,CO_2 + 6\,H_20$	–2890

Source: Lehninger 1975

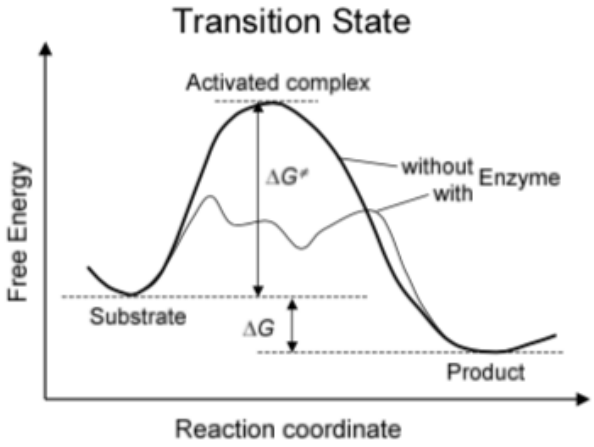


Fig. 5.3 Presentation of the change of free energy along the course of reaction. The substrate and the product are situated in local minima of the free energy; the active complex is assigned to the local maximum. The enzyme may change the reaction path and thereby lower the barrier of free energy.

the activated complex determines the dynamics of a reaction: the higher this difference, the lower the probability that the molecules may pass this barrier and the lower the rate of the reaction. The value of $\Delta G^{\neq}$ depends on the type of altered bonds, on steric, electronic, or hydrophobic demands, and on temperature.

Figure 5.3 presents a simplified view of the reaction course. The substrate and the product are situated in local minima of the free energy; the active complex is assigned to the local maximum. The free energy difference ΔG is proportional to the logarithm of the equilibrium constant of the respective reaction:

$$\Delta G = -RT\,\boldsymbol{ln}\,K_{eq}\,, \tag{5-20}$$

(R – gas constant, 8.314 J $mol^{-1}\,K^{-1}$). The value of $\Delta G^{\neq}$ corresponds to the kinetic constant k_+ of the forward reaction (Eqs. (12)–(14)) by $\Delta G^{\neq} = -RT\,\boldsymbol{ln}\,k_+$, while $\Delta G^{\neq} + \Delta G$ is related to the rate constant k_- of the backward reaction.

The interaction of the reactants with an enzyme may alter the reaction path and thereby lead to lower values of $\Delta G^{\neq}$. Furthermore, the free energy may assume more local minima and maxima along the path of reaction. They are related to unstable intermediary complexes. Values for the difference of free energy for some biologically important reactions are given in Tab. 5.1.

The detailed consideration of enzyme mechanisms by applying the mass action law for single events has led to a number of standard kinetic descriptions, which will be explained in the following sections.

3.8 competitive inhibition

A common characteristic of enzymatic reactions is the increase of the reaction rate with increasing substrate concentration S up to the maximal velocity V_{max}. But in some cases, a decrease of the rate above a certain value of S is recorded. A possible reason for this is the binding of a further substrate molecule to the enzyme-substrate complex, yielding the complex ESS, which cannot form a product. This kind of inhibition is reversible if the second substrate can be released. The rate equation can be derived using the scheme of uncompetitive inhibition by replacing the inhibitor by another substrate. It reads,

$$v = k_2\, ES = \frac{V_{\max}\, S}{K_m + S\left(1 + \frac{S}{K_I}\right)} . \tag{5-44}$$

This expression has a maximum at

$$S_{opt} = \sqrt{K_m K_I} \quad \text{with} \quad v_{opt} = \frac{V_{max}}{1 + 2\sqrt{K_m/K_I}} . \tag{5-45}$$

The dependence of v on S is shown in Fig. 5.6. A typical example for substrate inhibition is the binding of two succinate molecules to malonate dehydrogenase, which possesses two binding pockets for the carboxyl group. This is schematically represented in Fig. 5.6.

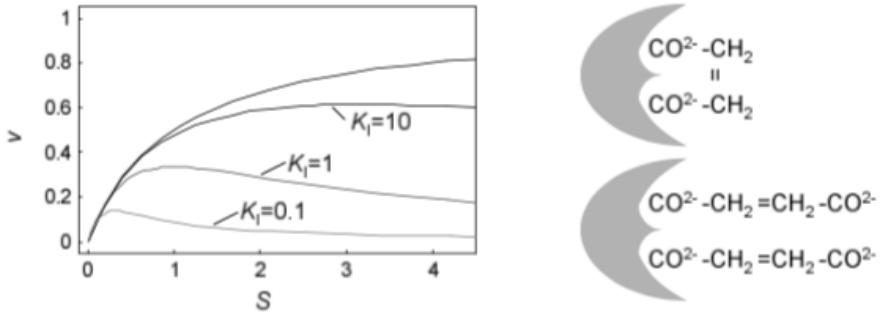


Fig. 5.6 Plot of reaction rate v against substrate concentration S for the case of substrate inhibition. The upper curve shows Michaelis-Menten kinetics without inhibition, and the lower curves show kinetics for the indicated values of binding constant K_I. Parameter values: $V_{max} = 1$, $K_m = 1$. The left part visualizes a possible mechanism for substrate inhibition. The enzyme (gray object) has two binding pockets to bind different parts of a substrate molecule (upper scheme). In the case of high substrate concentration, two different molecules may enter the binding pockets, thereby preventing the specific reaction (lower scheme).

3.9 co-operativity

Positive Homotropic Cooperativity and the Hill Equation

Consider a dimeric protein with two identical binding sites. The binding to the first ligand facilitates the binding to the second ligand:

$$\begin{aligned} E_2 + S &\xrightarrow{\text{slow}} E_2S \\ E_2S + S &\xrightarrow{\text{fast}} E_2S_2\,, \end{aligned} \tag{5-53}$$

where E is a monomer and E_2 a dimer. The fractional saturation is given by

$$Y = \frac{E_2S + 2\,E_2S_2}{2\,E_{2,total}} = \frac{E_2S + E_2S_2}{2\,E_2 + 2\,E_2S + 2\,E_2S_2}\,. \tag{5-54}$$

If the affinity to the second ligand is strongly increased by binding to the first ligand, then E_2S will react with S as soon as it is formed, and the concentration of E_2S can be neglected. In the case of complete cooperativity, i.e., every protein is either empty or fully bound, Eq. (5-53) reduces to

$$E_2 + 2S \longrightarrow E_2S_2\,. \tag{5-55}$$

The binding constant reads

$$K_B = \frac{E_2S_2}{E_2 \cdot S^2}\,, \tag{5-56}$$

and the fractional saturation is

$$Y = \frac{2\,E_2S_2}{2\,E_{2,total}} = \frac{E_2S_2}{E_2 + E_2S_2} = \frac{K_BS^2}{1 + K_BS^2}\,. \tag{5-57}$$

Generally, for a protein with n subunits it holds that

$$v = V_{max}\,Y = \frac{V_{max}\,K_B\,S^n}{1 + K_B\,S^n}\,. \tag{5-58}$$

This is the general form of the *Hill equation*. It implies complete homotropic cooperativity. Plotting the fractional saturation Y versus substrate concentration S yields a sigmoid curve with the inflection point at $1/K_B$. The quantity n (often "h" is used instead) is termed the Hill coefficient.

The derivation of this expression was based on experimental findings concerning the binding of oxygen to hemoglobin (Hb) (Hill 1910, 1913). In 1904 Bohr and coworkers found that the plot of the fractional saturation of Hb with oxygen against the oxygen partial pressure had a sigmoid shape. Hill (1909) explained this with interactions between the binding sites located at the hem subunits. At this time it was already known that every subunit hem binds one molecule of oxygen. Hill assumed complete cooperativity and predicted an experimental Hill coefficient of 2.8. Today it is known that hemoglobin has four binding sites but that the cooperativity is not complete. The sigmoidal binding characteristic has the advantage that Hb binds strongly to oxygen in the lung with a high oxygen partial pressure, while it can release O_2 easily in the body with low oxygen partial pressure.

3.10 Hyperbolic and sigmoidal responses

In 1965 Monod and colleagues presented a model explaining sigmoidal enzyme kinetics taking into account the interaction of subunits of an enzyme (Monod et al. 1965). A more comprehensive model has been presented by Koshland et al. (1966). The model of Monod et al. uses the following assumptions: (1) the enzyme consists

of n identical subunits, (2) each subunit can assume an active (R) or an inactive (T) conformation, (3) all subunits change their conformations at the same time (concerted change), and (4) the equilibrium between the R and the T conformations is given by an allosteric constant:

$$L = \frac{T_0}{R_0} . \tag{5-59}$$

The index i for T_i and R_i denotes the number of bound substrate molecules. The binding constants for the active and inactive conformations are given by K_R and K_T, respectively. If substrate molecules can bind only to the active form, i.e., if $K_T = 0$, then the rate can be given as

$$V = \frac{V_{max} K_R S}{(1 + K_R S)} \frac{1}{\left(1 + \frac{L}{(1 + K_R S)^n}\right)} , \tag{5-60}$$

where the factor $\frac{V_{max} K_R S}{(1 + K_R S)}$ corresponds to the Michaelis-Menten rate expression, while the factor $\left(1 + \frac{L}{(1 + K_R S)^n}\right)^{-1}$ is a regulatory factor.

For $L = 0$ the plot v versus S is a hyperbola as in Michaelis-Menten kinetics. For

$L > 0$ one gets a sigmoid curve shifted to the right. A typical value for the allosteric constant is $L \cong 10^4$.

In the case that the substrate can also bind to the inactive state ($K_T \neq 0$), one gets

$$V = \frac{V_{max} S}{(1 + K_R S)} \frac{K_R + K_T L \left(\frac{1 + K_T S}{1 + K_R S}\right)^{n-1}}{\left(1 + L \left(\frac{1 + K_T S}{1 + K_R S}\right)^{n}\right)} . \tag{5-61}$$

Up to now we have considered only homotropic and positive effects in the model of Monod, Wyman, and Changeux. But this model is also well suited to explain the dependence of the reaction rate on activators and inhibitors. Activators A bind only to the active conformation, and inhibitors I bind only to the inactive conformation. This shifts the equilibrium to the respective conformation. Effectively, the binding to effectors changes L:

$$L' = L \frac{(1 + K_I I)^n}{(1 + K_A A)^n} . \tag{5-62}$$

K_I and K_A denote binding constants. The interaction with effectors is a heterotropic effect. An activator weakens the sigmoidity, while an inhibitor strengthens it as shown in Figure 5.7.

As an example, the kinetics of the enzyme phosphofructokinase, which catalyzes the transformation of fructose-6-phosphate and ATP to fructose-1,6-bisphosphate,

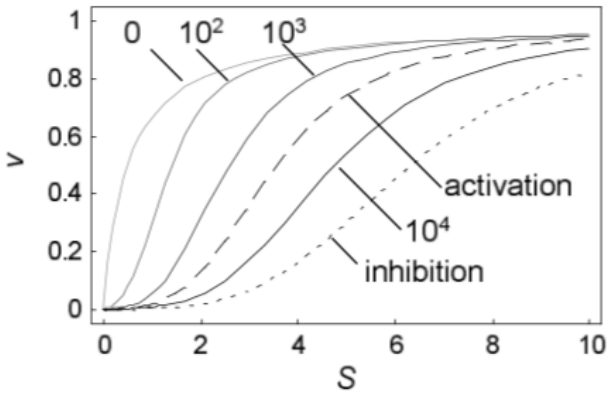


Fig. 5.7 Model of Monod, Wyman, and Changeux: Dependence of the reaction rate on substrate concentration for different values of the allosteric constant L, according to Eq. (5-60). Parameters: $V_{max} = 1$, $n = 4$, $K_R = 2$, $K_T = 0$. The value of L is indicated at the curves. Obviously, increasing the value of L causes stronger sigmoidity. The influence of activators or inhibitors (compare Eq. (5-62)) is illustrated with the dotted line for $K_I I = 2$ and with the dashed line for $K_A A = 2$ ($L = 10^4$ in both cases).

can be described by the model of Monod, Wyman, and Changeux. AMP, NH_4, and K+ are activators, while ATP is an inhibitor (see Example 5-1).

3.11 Michaelis-Menten Kinetics

Brown (1902) proposed the first enzymatic mechanism for the reaction of invertase, which holds for all one-substrate reactions without backward reaction and without effectors in general:

$$\mathrm{E+S} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{ES} \xrightarrow{k_2} \mathrm{E+P}. \tag{5-21}$$

It comprises a reversible formation of an enzyme-substrate complex ES from the free enzyme E and the substrate S and an irreversible release of the product P from the enzyme E. The respective system of ODEs for the dynamics of this reaction reads as follows:

$$\frac{dS}{dt} = -k_1\, E \cdot S + k_{-1} ES \tag{5-22}$$

$$\frac{dES}{dt} = k_1\, E \cdot S - (k_{-1} + k_2)\, ES \tag{5-23}$$

$$\frac{dE}{dt} = -k_1\, E \cdot S + (k_{-1} + k_2)\, ES \tag{5-24}$$

$$\frac{dP}{dt} = k_2\, ES\,. \tag{5-25}$$

The rate of the reaction is equal to the negative rate of decay of the substrate as well as to the rate of product formation:

$$v = -\frac{dS}{dt} = \frac{dP}{dt} \,. \tag{5-26}$$

This ODE system (Eqs. (5-22)–(5-26)) cannot be solved analytically. Assumptions have been used to simplify this system in a satisfactory way. Michaelis and Menten (1913) assumed that the conversion of E and S to ES and vice versa is much faster than the decomposition of ES into E and P (so-called *quasi-equilibrium* between the free enzyme and the enzyme-substrate complex), or in terms of the constants

$$k_1, k_{-1} \gg k_2 \,. \tag{5-27}$$

Briggs and Haldane (1925) assumed that during the course of reaction a state is reached where the concentration of the ES complex remains constant. This assumption is justified only if the initial concentration of the substrate is much larger than the concentration of the enzyme, $S(t = 0) \gg E$; otherwise, this steady state will never be reached. They suggested the more general assumption of a *quasi-steady state* of the ES complex:

$$\frac{dES}{dt} = 0 \,. \tag{5-28}$$

An expression for the reaction rate will be derived using the ODE system in Eqs. (5-22)–(5-25) and the assumption of a quasi-steady state for ES. Adding Eqs. (5-23) and (5-24) results in

$$\frac{dES}{dt} + \frac{dE}{dt} = 0 \quad \text{or} \quad E_{total} = E + ES \,. \tag{5-29}$$

In this reaction, enzyme is neither produced nor consumed; it may be free or involved in the complex, but its total concentration remains constant.

Introducing Eq. (5-29) into Eq. (5-23) under the steady-state assumption (Eq. (5-28)) yields

$$ES = \frac{k_1\, E_{total}\, S}{k_1 S + k_{-1} + k_2} = \frac{E_{total}\, S}{S + \dfrac{k_{-1} + k_2}{k_1}} . \tag{5-30}$$

For the reaction rate, this yields

$$v = \frac{k_2\, E_{total}\, S}{S + \dfrac{k_{-1} + k_2}{k_1}} . \tag{5-31}$$

In enzyme kinetics it is convention to present Eq. (5-31) in a simpler form, which is important in both theory and practice:

$$v = \frac{V_{max}\, S}{S + K_m} . \tag{5-32}$$

Equation (5-32) is the expression for Michaelis-Menten kinetics. The parameters have the following meaning: the *maximal velocity*,

$$V_{max} = k_2\, E_{total} , \tag{5-33}$$

is the maximal rate that can be attained when the enzyme is completely saturated with substrate. The *Michaelis constant*,

$$K_m = \frac{k_{-1} + k_2}{k_1} , \tag{5-34}$$

is equal to the substrate concentration that yields the half-maximal reaction rate. For the quasi-equilibrium assumption (Eq. (5-27)), it holds that $K_m \cong k_{-1}/k_1$. The meaning of the parameters can be seen from the plot of rate versus substrate concentration (Fig. 5.4). The plot has a hyperbolic shape.

Reaction v_1, Eq. (5-2), is described with Michaelis-Menten kinetics.

3.12 Metabolic Networks

In this section we will discuss basic structural and dynamic properties of metabolic networks. We will introduce a stoichiometric description of networks and learn how moieties and fluxes are balanced within networks.

The basic elements of a metabolic network model are (1) the substances with their concentrations and (2) the reactions or transport processes changing the concentrations of the substances. In biological environments, reactions are usually catalyzed by enzymes, and transport steps are carried out by transport proteins or by pores. Thus they can be assigned to identifiable biochemical compounds.

Stoichiometric coefficients denote the proportion of substrate and product molecules involved in a reaction. For example, for the reaction depicted in Eq. (5-11), the stoichiometric coefficients of S_1, S_2, and P are –1, –1, and 2. The assignment of stoichiometric coefficients is not unique. We could also argue that for the production of one mole P, half a mole of each S_1 and S_2 have to be used and therefore choose –1/2, –1/2, and 1. Or, if we change the direction of the reaction, then we may choose 1, 1, and –2.

The change of concentrations in time can be described using ODEs. For the reaction depicted in Eq. (5-11) and the first choice of stoichiometric coefficients, we have

$$\frac{dS_1}{dt} = -v\,, \quad \frac{dS_2}{dt} = -v\,, \quad \text{and} \quad \frac{dP}{dt} = 2\,v\,. \tag{5-63}$$

This means that the degradation of S_1 with rate v is accompanied by the degradation of S_2 with the same rate and by the production of P with the double rate.

3.13 Systems Equations

For a metabolic network consisting of m substances and r reactions, the systems dynamics is described by systems equations (or balance equations, since the balance of substrate production and degradation is considered):

$$\frac{dS_i}{dt} = \sum_{j=1}^{r} n_{ij} v_j \quad \text{for} \quad i = 1, .., m \tag{5-64}$$

(Glansdorff and Prigogine 1971; Reder 1988). The quantities n_{ij} are the stoichiometric coefficients of metabolite i in reaction j. Here, we assume that the reactions are the only reason for concentration changes and that no mass flow occurs due to convection or to diffusion. The balance equations (Eq. (5-64)) can also be applied if the system consists of several compartments. In this case, every compound in different compartments has to be considered as an individual compound, and transport steps are formally considered as reactions transferring the compound belonging to one compartment into the same compound belonging to the other compartment.

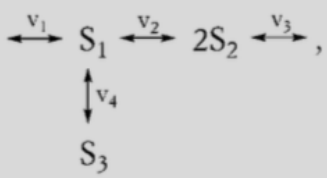
The stoichiometric coefficients n_{ij} assigned to the substances S_i and the reactions v_j can be combined into the so-called *stoichiometric matrix*

$$N = \{n_{ij}\} \quad \text{for} \quad i = 1, .., m \quad \text{and} \quad j = 1, .., r\,, \tag{5-65}$$

where each column belongs to a reaction and each row to a substance.

Example 5-4

For the simple network



(5-66)

the stoichiometric matrix reads

$$\boldsymbol{N} = \begin{pmatrix} 1 & -1 & 0 & -1 \\ 0 & 2 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{5-67}$$

Note that in Eq. (5-66) all reactions may be reversible. In order to determine the signs of N, the direction of the arrows is artificially assigned as positive "from left to right" and "from the top down." If, for example, the net flow proceeds from S_3 to S_1, the value of rate v_4 is negative.

Altogether, the mathematical description of the metabolic system consists of a vector $\boldsymbol{S} = (S_1, S_2, ..., S_n)^T$ of concentration values, a vector $\boldsymbol{v} = (v_1, v_2, ..., v_r)^T$ of reaction rates, a parameter vector $\boldsymbol{p} = (p_1, p_2, ..., p_m)^T$, and the stoichiometric matrix $\boldsymbol{N}$. If the system is in steady state, we can also consider the vector $\boldsymbol{J} = (J_1, J_2, ..., J_r)^T$ containing the steady state fluxes. With these notions, the balance equation reads

$$\frac{d\boldsymbol{S}}{dt} = \boldsymbol{N}\boldsymbol{v}\,. \tag{5-68}$$

For our running example (Example 5-1) of the upper glycolysis model, the concentration vector is

$$\boldsymbol{S} = \begin{pmatrix} Gluc6P \\ Fruc6P \\ Fruc1{,}6P_2 \\ ATP \\ ADP \\ AMP \end{pmatrix}, \tag{5-69}$$

the vector of reaction rates is $\boldsymbol{v} = (v_1, v_2, \ldots, v_8)^{\mathrm{T}}$, the parameter vector is given by

$$\mathrm{p} = \left(Glucose,\ V_{\max,1},\ K_{ATP,1},\ K_{Glucose,1},\ k_2,\ V^f_{\max,3},\ V^r_{\max,3},\ K_{Gluc6P,3},\ K_{Fruc6P,3},\ V_{\max,4},\ K_{F6P,4},\ \kappa_4,\ k_5,\ k_6,\ k_7,\ k_{8f},\ k_{8r}\right)^{\mathrm{T}}, \tag{5-70}$$

and the stoichiometric matrix reads

$$\boldsymbol{N} = \begin{pmatrix} 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ -1 & -1 & 0 & -1 & 0 & 1 & -1 & -1 \\ 1 & 1 & 0 & 1 & 0 & -1 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{pmatrix}. \tag{5-71}$$

3.14 Information Contained in the Stoichiometric Matrix N

The stoichiometric matrix contains important information about the structure of the metabolic network. Using the stoichiometric matrix, we can calculate which combinations of individual fluxes are possible in steady state (i.e., calculate the admissible steady-state flux space). We may easily discover dead ends and unbranched reaction pathways. In addition, we may find out the conservation relations for the included reactants.

In steady state it holds that

$$\frac{d\boldsymbol{S}}{dt} = \boldsymbol{N}\boldsymbol{v} = 0 \tag{5-72}$$

(Reder 1988). The right equality sign denotes a linear equation system for determination of the rates $\boldsymbol{v}$. This equation has nontrivial solutions only for $Rank\ \boldsymbol{N} < \boldsymbol{r}$ (Chapter 3, Section 3.1). The kernel matrix $\boldsymbol{K}$ fulfilling

$$\boldsymbol{N}\boldsymbol{K} = 0 \tag{5-73}$$

can express the respective linear dependencies (Heinrich and Schuster 1996). The choice of the kernel is not unique. It can be determined using the Gauss algorithm described in Chapter 3 (Section 3.1). It contains as columns $\boldsymbol{r} - Rank\ \boldsymbol{N}$ basis vectors. Every possible set of steady-state fluxes can be expressed as a linear combination of the columns $\boldsymbol{k}_i$ of $\boldsymbol{K}$

$$\boldsymbol{J} = \sum_{i=1}^{r-RankN} \alpha_i \cdot \boldsymbol{k}_i \,. \tag{5-74}$$

The coefficients must have respective units ($M \cdot s^{-1}$ or $mol \cdot L^{-1} \cdot s^{-1}$).

3.15 Flux Balance Analysis

Flux balance analysis (FBA) (Varma and Palsson 1994a, 1994b; Edwards and Palsson 2000a, 2000b; Ramakrishna et al. 2001) investigates the theoretical capabilities and operative modes of metabolism by involving further constraints in the stoichiometric analysis. The first constraint is set by the assumption of a steady state (Eqs. (5-72) and (5-73)). The second constraint is of a thermodynamic nature, respecting the irreversibility of reactions as considered in the concept of extreme pathways. The third constraint may result from the limited capacity of enzymes for metabolite conversion. For example, in the case of a Michaelis-Menten-type enzyme (Eq. (5-32)), the reaction rate is limited by the maximal rate, i.e., $0 \leq v \leq V_{max}$. In general, the constraints imposed on the magnitude of individual metabolic fluxes read

$$\alpha_i \leq \nu_i \leq \beta_i \,. \tag{5-83}$$

Further constraints may be imposed by biomass composition or other external conditions. The constraints confine the steady-state fluxes to a feasible set but usually do not yield a unique solution. The determination of a particular metabolic flux distribution has been formulated as a linear programming problem. The idea is to maximize an objective function Z that is subject to the stoichiometric and capacity constraints:

$$Z = \sum_{i=1}^{r} c_i v_i \rightarrow \boldsymbol{max}\,. \tag{5-84}$$

where c_i represents weights for the individual rates. Examples of such objective functions are maximization of ATP production, minimization of nutrient uptake, maximal yield of a desired product, maximal growth rate, or a combination thereof.

3.16 Signal Transduction

Throughout intercellular communication or cellular stress response, the cell senses extracellular signals. They are commuted to intracellular signals and sequences of reactions. Different external changes or events may stimulate signaling. Typical signals are hormones, pheromones, heat, cold, light, osmotic pressure, and appearance or concentration change of substances such as glucose, K^+, Ca^+, or cAMP.

On a molecular level, signaling involves the same type of processes as metabolism: production or degradation of substances, molecular modifications (mainly phosphorylation, but also methylation and acetylation), and activation or inhibition of reactions. From a modeling point of view, there are some important differences between signaling and metabolism. First, signaling pathways serve for information processing and transfer of information, while metabolism provides mainly mass transfer. Second, the metabolic network is determined by the present set of enzymes catalyzing the reactions. Signaling pathways involve compounds of different types, and they may form highly organized complexes and may assemble dynamically upon occurrence of the signal. Third, the quantity of converted material is high in metabolism (amounts are usually given in concentrations on the order of µM or mM) compared to the number of molecules involved in signaling processes (the typical abundance of proteins in signal cascades is on the order of 10 to 10^4 molecules per cell). Finally, the different amounts of components have an effect on the concentration ratio of catalysts and substrates. In metabolism this ratio is usually low; the enzyme concentration is much lower than the substrate concentration, which gives rise to the quasi-steady-state assumption used in Michaelis-Menten kinetics (Chapter 5, Section 5.1). In signaling processes, amounts of catalysts and their substrates are frequently in the same order of magnitude.

Modeling of the dynamic behavior of signaling pathways is often not straightforward. Knowledge about components of the pathway and their interaction is still limited and incomplete. The interpretation of experimental data is context- and knowledge-dependent. Furthermore, the effect of a signal often changes the state of the whole cell, and this implies difficulties for determination of system limits. But in many cases we may apply the same tools as introduced in Chapter 5.

3.17 Function and Structure of Intra- and Intercellular Communication

Cells have a broad spectrum of receiving and processing signals; therefore, not all of them can be considered here. A typical sequence of events in signaling pathways is shown in Fig. 6.1 and proceeds as follows. The "signal" (a substance acting as a ligand or a physical stimulus) approaches the cell surface. Cells have developed two different modes of importing a signal. First, the stimulus may penetrate the cell membrane and bind to a respective receptor in the cell interior. Another possibility is that the signal is perceived by a transmembrane receptor. If the target of the signal is a receptor, it does not cross the membrane. Instead, the receptor changes its own state from susceptible to active and then triggers subsequent processes within the cell. The active receptor stimulates an internal signaling cascade. This cascade frequently includes a series of changes in protein phosphorylation states. The sequence of state changes crosses the nuclear membrane. Eventually, a transcription factor is activated or deactivated. The transcription factor changes its binding properties to regulatory regions on the DNA upstream of a set of genes, and the transcription rate of these genes is altered (typically increased). Either the newly produced proteins or the changes in protein concentration cause the actual response of the cell to the signal. In addition to this downstream program, signaling pathways are regulated by a number of control mechanisms including feedback and feed-forward modulation.

This is the typical picture; however, many pathways may work in a completely different manner. As an example, an overview of signaling pathways that are stimulated in yeast stress response is given in Fig. 6.2.

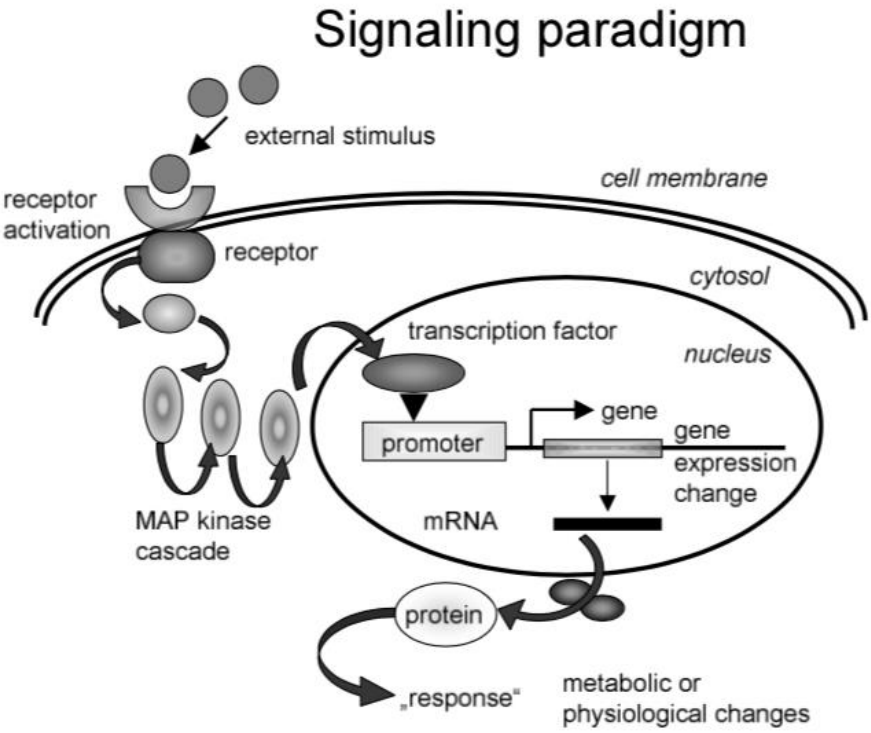


Fig. 6.1 Visualization of the signaling paradigm (for description, see text). The receptor is stimulated by a ligand or another kind of signal, and it changes its own state from susceptible to active. The active receptor initiates the internal signaling cascade, including a series of protein phosphorylation state changes. Subsequently, transcription factors are activated or deactivated. The transcription factors regulate the transcription rate of a set of genes. The absolute amount or the relative changes in protein concentrations alter the state of the cell and trigger the actual response to the signal.

3.18 Structural Components of Signaling Pathways

Signaling pathways constitute often highly complex networks, but it has been discovered that they are frequently composed of typical building blocks. These components include Ras proteins, G protein cycles, phosphorelay systems, and MAP kinase cascades. In this chapter we will discuss their general composition and function as well as modeling approaches.

3.19 G Proteins, Ras Proteins

G proteins are essential parts of many signaling pathways. The reason for their name is that they bind the guanine nucleotides GDP and GTP. They are heterotrimers, i.e., they consist of three different subunits. G proteins are associated to cell surface receptors with a heptahelical transmembrane structure, the so-called G protein–coupled receptors (GPCR). Signal transduction cascades involving (1) such a transmembrane surface receptor, (2) an associated G protein, and (3) an intracellular effector that produces a second messenger play an important role in cellular communication and are well studied (Neer 1995; Dohlman 2002). In humans, such G protein–coupled receptors mediate responses to light, flavors, odors, numerous hormones, neurotransmitters, and other signals (Blumer and Thorner 1991; Dohlman et al. 1991; Buck 2000). In unicellular eukaryotes, receptors of this type mediate signals that affect such basic processes as cell division, cell-cell fusion (mating), morphogenesis, and chemotaxis (Blumer and Thorner 1991; Banuett 1998; Dohlman et al. 1998; Wang and Heitman 1999).

The cycle of G protein activation and inactivation is shown in Fig. 6.6. When GDP is bound, the G protein α subunit (Gα) is associated with the G protein βγ heterodimer (Gβγ) and is inactive. Agonist binding to a receptor promotes guanine nucleotide exchange; Gα releases GDP, binds GTP, and dissociates from Gβγ. The dissociated subunits Gα or Gβγ, or both, are then free to activate target proteins (downstream effectors), which initiates signaling. When GTP is hydrolyzed, the subunits

are able to re-associate. Gβγ antagonizes receptor action by inhibiting guanine nucleotide exchange. RGS (regulator of G protein signaling) proteins bind to Gα, stimulate GTP hydrolysis, and thereby reverse G protein activation. This general scheme can also be applied to the regulation of small monomeric Ras-like GTPases, such as Rho. In this case, the receptor, Gβγ, and RGS are replaced by GEF and GAP (see Section 6.3.2).

Direct targets include different types of effectors, such as adenylyl cyclase, phospholipase C, exchange factors for small GTPases, some calcium and potassium channels, plasma membrane Na^+/H^+ exchangers, and certain protein kinases (Neer 1995; Offermanns 2000; Dohlman and Thorner 2001; Meigs et al. 2001). Typically, these effectors produce second messengers or other biochemical changes that lead to stimulation of a protein kinase or a protein kinase cascade (or, as mentioned, are themselves a protein kinase). Signaling persists until GTP is hydrolyzed to GDP and the Gα and Gβγ subunits re-associate, completing the cycle of activation. The strength of the G protein–initiated signal depends on (1) the rate of nucleotide exchange, (2) the rate of spontaneous GTP hydrolysis, (3) the rate of RGS-supported GTP hydrolysis, and (4) the rate of subunit re-association. RGS proteins act as GTPase-activating proteins (GAPs) for a variety of different Gα classes and thereby shorten the lifetime of the activated state of a G protein and contribute to signal desensitization. Furthermore, they may contain additional modular domains with signaling functions and contribute to diversity and complexity of the cellular signaling networks (Dohlman and Thorner 1997; Siderovski et al. 1999; Burchett 2000; Ross and Wilkie 2000).

Ras Proteins

Small G proteins are monomeric G proteins with molecular weight of 20–40 kDa. Like heterotrimeric G proteins, their activity depends on the binding of GTP. More than 100 small G proteins have been identified. They belong to five families: Ras, Rho, Rab, Ran, and Arf. They regulate a wide variety of cell functions as biological timers that initiate and terminate specific cell functions and determine the periods of time (Takai et al. 2001).

Ras proteins cycle between active and inactive states (Fig. 6.8). The transition form GDP-bound to GTP-bound states is catalyzed by a guanine nucleotide exchange factor (GEF), which induces exchange between the bound GDP and the cellular GTP. The reverse process is facilitated by a GTPase-activating protein (GAP), which induces hydrolysis of the bound GTP (Schmidt and Hall 2002).

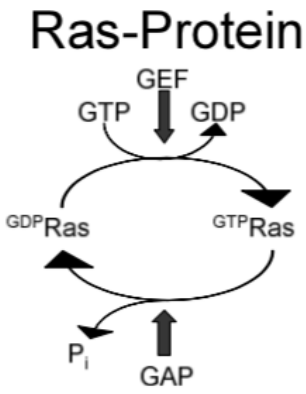


Fig. 6.8 The Ras activation cycle. GEF supports the transition from GDP-bound to GTP-bound states to activate Ras, while GAP induces hydrolysis of the bound GTP, resulting in Ras deactivation.

Mutations of the *Ras* proto-oncogenes (H-*Ras*, N-*Ras*, K-*Ras*) are found in many human tumors. Most of these mutations result in the abolishment of normal GTPase activity of Ras. The Ras mutants can still bind to GAP, but they cannot catalyze GTP hydrolysis. Therefore, they stay active for a long time.

3.20 MAP Kinase Cascades

Mitogen-activated protein kinases (MAPKs) are a family of serine/threonine kinases that transduce signals from the cell membrane to the nucleus in response to a wide range of stimuli. Independent or coupled kinase cascades participate in many different intracellular signaling pathways that control a spectrum of cellular processes, including cell growth, differentiation, transformation, and apoptosis. MAPK cascades are widely involved in eukaryotic signal transduction, and these pathways are conserved from yeast to mammals.

A general scheme of a MAPK cascade is depicted in Fig. 6.11. This pathway consists of several levels (usually three), where the activated kinase at each level phosphorylates the kinase at the next level down the cascade. The MAP kinase (MAPK) is at the terminal level of the cascade. It is activated by the MAPK kinase (MAPKK) by phosphorylation of two sites: conserved threonine and tyrosine residues. The MAPKK is itself phosphorylated at serine and threonine residues by the MAPKK kinase (MAPKKK). Several mechanisms are known to activate MAPKKKs by phosphorylation of a tyrosine residue. In some cases the upstream kinase may be considered a MAPKKK kinase (MAPKKKK). Dephosphorylation of either residue is thought to inactivate the kinases, and mutants lacking either residue are almost inactive. At each cascade level, protein phosphatases can inactivate the corresponding kinase, although in some cases it is a matter of debate whether this reaction is performed by an independent protein or by the kinase itself as autodephosphorylation. Ubiquitin-dependent degradation of phosphorylated proteins is also reported.

Although they are conserved through species, elements of the MAPK cascade were given different names in various studied systems. Some examples are represented in Tab. 6.1 (see also Wilkinson and Millar 2000).

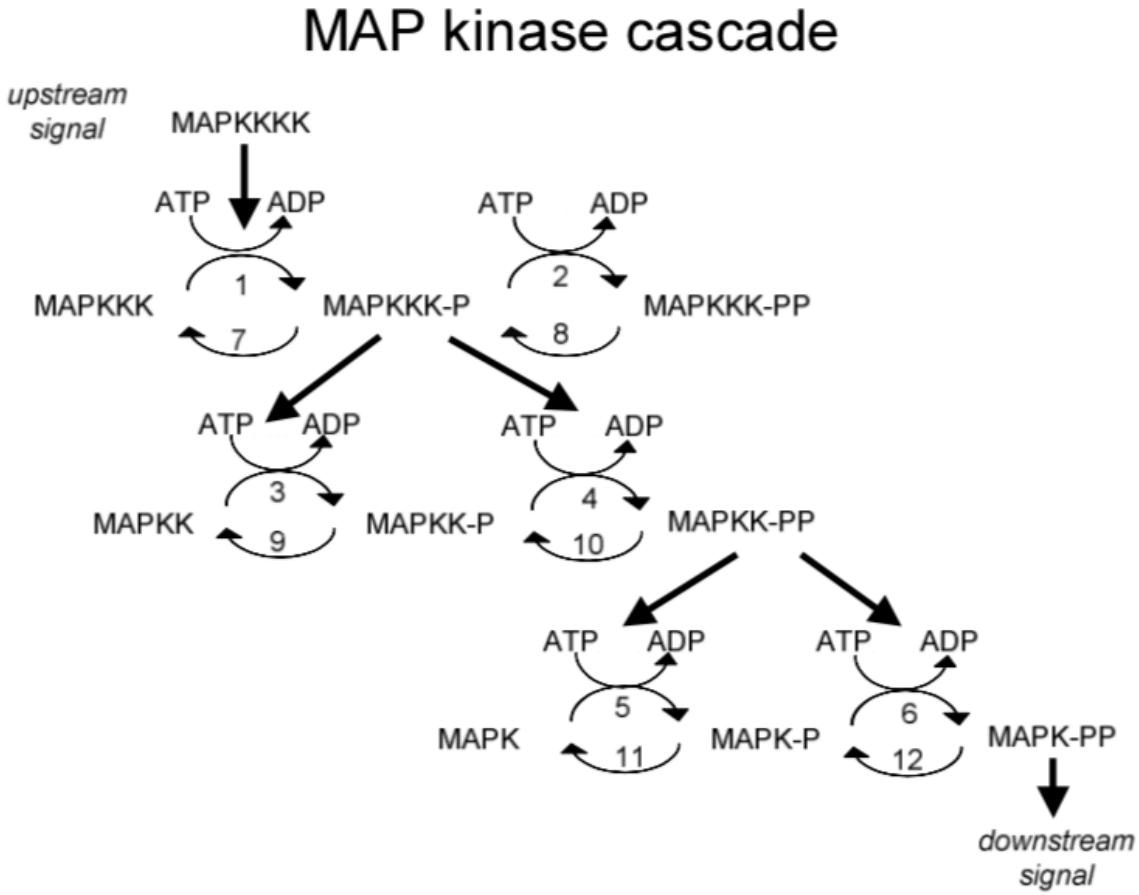


Fig. 6.11 Schematic representation of the MAP kinase cascade. An upstream signal (often by a further kinase called MAP kinase kinase kinase kinase) causes phosphorylation of the MAPKKK. The phosphorylated MAPKKK in turn phosphorylates the protein at the next level. Dephosphorylation is assumed to occur continuously by phosphatases or autodephosphorylation.

Tab. 6.1 Names of the components of MAP kinase pathways in different organisms and different pathways.

Organism	*Budding yeast*		*Xensopus oocytes*	*Human, cell cycle regulation*		
	HOG pathway	Pheromone pathway			p38 pathway	JNK pathway
MAPKKK	Ssk2/Ssk22	Ste11	Mos	Rafs (c-, A- and B-),	Tak1	MEKKs
MAPKK	Pbs2	Ste7	MEK1	MEK1/2	MKK3/6	MKK4/7
MAPK	Hog1	Fus3	p42 MAPK	ERK1/2	p38	JNK1/2

In the following we will present typical modeling approaches and then discuss functional properties of signaling cascades. The dynamics of a MAPK cascade may be represented by the following ODE system:

$$\frac{d}{dt}MAPKKK = -v_1 + v_7$$

$$\frac{d}{dt}MAPKKK\text{-}P = v_1 - v_7 - v_2 + v_8$$

$$\frac{d}{dt}MAPKKK\text{-}P_2 = v_2 - v_8 \qquad (6\text{-}9)$$

$$\frac{d}{dt}MAPKK = -v_3 + v_9$$

$$\frac{d}{dt}MAPKK\text{-}P = v_3 - v_9 - v_4 + v_{10}$$

$$\frac{d}{dt}MAPKK\text{-}P_2 = v_4 - v_{10} \qquad (6\text{-}10)$$

$$\frac{d}{dt}MAPK = -v_5 + v_{11}$$

$$\frac{d}{dt}MAPK\text{-}P = v_5 - v_{11} - v_6 + v_{12}$$

$$\frac{d}{dt}MAPK\text{-}P_2 = v_6 - v_{12}\ . \qquad (6\text{-}11)$$

Please note that it is not clear whether MAPKKK-P_2 and v_2,v_8 exist at all. In this case their value may be simply set to zero.

3.21 Apoptotic pathway

Our understanding of the mitochondrial or intrinsic apoptosis pathway and its role in chemotherapy resistance has increased significantly in recent years by a combination of experimental studies and mathematical modelling. This combined approach enhanced the quantitative and kinetic understanding of apoptosis signal transduction, but also provided new insights that systems-emanating functions (i.e., functions that cannot be attributed to individual network components but that are instead established by multi-component interplay) are crucial determinants of cell fate decisions. Among these features are molecular thresholds, cooperative protein functions, feedback loops and functional redundancies that provide systems robustness, and signalling topologies that allow ultrasensitivity or switch-like responses. The successful development of kinetic systems models that recapitulate biological signal transduction observed in living cells have now led to the first translational studies, which have exploited and validated such models in a clinical context. Bottom-up strategies that use pathway models in combination with higher-level modelling at the tissue, organ and whole body-level therefore

carry great potential to eventually deliver a new generation of systems-based diagnostic tools that may contribute to the development of personalised and predictive medicine approaches. Here we review major achievements in the systems biology of intrinsic apoptosis signalling, discuss challenges for further model development, perspectives for higher-level integration of apoptosis models and finally discuss requirements for the development of systems medical solutions in the coming years.

3.22 Two component signalling pathways of bacterial chemotaxis.

The chemosensory pathway of bacterial chemotaxis has become a paradigm for the two-component superfamily of receptor-regulated phosphorylation pathways. This simple pathway illustrates many of the fundamental principles and unanswered questions in the field of signaling biology. A molecular description of pathway function has progressed rapidly because it is accessible to diverse structural, biochemical, and genetic approaches. As a result, structures are emerging for most of the pathway elements, biochemical studies are elucidating the mechanisms of key signaling events, and genetic methods are revealing the intermolecular interactions that transmit information between components. Recent advances include (*a*) the first molecular picture of a conformational transmembrane signal in a cell surface receptor, (*b*) four new structures of kinase domains and adaptation enzymes, and (*c*) significant new insights into the mechanisms of receptor-mediated kinase regulation, receptor adaptation, and the phospho-activation of signaling proteins. Overall, the chemosensory pathway and the propulsion system it regulates provide an ideal system in which to probe molecular principles underlying complex cellular signaling and behavior.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

School of Bio and Chemical Engineering
DEPARTMENT OF BIOINFORMATICS

UNIT IV - Quantitative Models in Biological Systems - Subject Code: SBI1402

UNIT 4 SELECTED BIOLOGICAL PROCESSES

4.1 Biological Oscillations
4.2 Glycolytic Oscillations: The Higgins-Sel'kov Oscillator
4.3 Cell Cycle - Steps in the Cycle
4.4 Models of Budding Yeast Cell Cycle
4.5 Modeling of Gene Expression
4.6 Modules of Gene Expression
4.7 Modeling the Elongation of a Peptide Chain
4.8 The Model According to Griffith
4.9 Noise and oscillation in biological system
4.10 Circadian rhythm-how to build an oscillator
4.11 Gene circuit design

UNIT 4 SELECTED BIOLOGICAL PROCESSES

4.1 Biological Oscillations

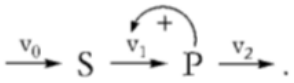
Periodic changes of biochemical and biophysical quantities are a universal phenomenon in living systems. Examples from everyday experience are the pulse of the heart, spontaneous respiration, the circadian rhythm, cycles of ovulation in mammals, or the annual flowering of trees. Well studied are calcium waves (Goldbeter et al. 1990; Bootman et al. 2001 a, 2001 b), oscillations in neuronal signals (Rabinovich and Abarbanel 1998), oscillations in cyclic AMP in the slime mold *Dictyostelium discoideum* (Roos et al. 1977; Halloy et al. 1998; Nanjundiah 1998), the periodic conversion of sugar to alcohol (glycolysis) in anaerobic yeast cultures (Chance et al. 1964; Ghosh and Chance 1964; Sel'kov 1968), the circadian rhythm (Smaaland 1996; Turek 1998), and the cell cycle (Tyson 1991; Tyson et al. 1995; Novak et al. 1999; Mori and Johnson 2000; Tyson and Novak 2001). Periodic patterns can be a function of time (glycolytic oscillations), space (striping in *Drosophila melanogaster* embryos), or both (*D. discoideum*, calcium waves, neuronal oscillations), depending on the mechanism of the oscillator. The oscillation periods may cover ranges from milliseconds to years. Some oscillations are initiated externally, while others have intrinsic causes. Many cellular oscillations are associated with the regulation of enzyme activity, receptor function, transport processes, or gene expression in an autocatalytic manner or by positive or negative feedback and feed-forward loops. Other cases of oscillations arise from the regulation of ionic conductances in electrically excitable cells.

Temporarily changing patterns are observed in different complexity. Types of behavior include simple periodic oscillations, complex periodicity with several maxima per period, and even irregular and aperiodic behavior, owing to the appearance of chaos. In the following sections we will introduce the Higgins-Sel'kov oscillator as a classical example of oscillations caused by positive feedback and a model of a multiply regulated biochemical system as an example of more complex oscillatory patterns. Coupled oscillators are presented to illustrate that oscillations in individual cells are sometimes hidden on the population level and therefore are hard to measure experimentally.

4.2 Glycolytic Oscillations: The Higgins-Sel'kov Oscillator

The product-activated enzyme reaction is a simple model with two variables for periodic oscillations of the limit-cycle type. The most intensively studied example is the positive feedback exerted by ADP on the enzyme phosphofructokinase I (PFK I). It is believed to cause the oscillations observed in glycolysis in yeast and muscles. The dynamic behavior of this model system was first studied by Higgins (1964) and Sel'kov (1968) and later by many others (e. g., Goldbeter and Lefever 1972; Sel'kov 1975).

The two-variable model takes into account the allosteric regulation of the enzyme PFK I and the autocatalytic effect exerted by the product. For a large range of parameter values, it exhibits a stable steady state, but beyond a critical parameter value, the system becomes instable and evolves towards a stable limit cycle. Then, it shows sustained oscillations.



(7-1)

The temporal behavior of the concentrations of substrate S and product P can be described by the following ODEs:

$$\frac{dS}{dt} = v_0 - Sk_1 \cdot r(P)$$

$$\frac{dP}{dt} = Sk_1 \cdot r(P) - Pk_2 . \qquad (7\text{-}2)$$

The supply rate of S, v_0, is positive. The parameters k_1, k_2 are mass-action rate constants. The function $r(P)$ represents the autocatalytic effect of the product P on its own production. The simplest expression for this function is

$$r(P) = P^2 , \tag{7-3}$$

yielding

$$\frac{dS}{dt} = v_0 - SP^2 k_1 = f(S, P)$$

$$\frac{dP}{dt} = SP^2 k_1 - Pk_2 = g(S, P) . \tag{7-4}$$

The dynamic behavior of this system is represented in Fig. 7.1 for a set of parameters that gives rise to oscillations. Figure 7.1 a shows the values of the variables as function of time. Further information about the dynamics of the system can be in-

ferred by inspection of the phase plane. Figure 7.1 c shows the trajectory for a given set of parameters and initial conditions in a plot of S versus P, using the time as parameter. The nullclines, i. e., the lines for $dS/dt = f = 0$ or $dP/dt = g = 0$, respectively, are shown in Fig. 7.1d. These lines must always be crossed by the trajectories in a horizontal (for $f = 0$) or vertical (for $g = 0$) manner, respectively, as indicated by the little arrows. The sign of g determines the direction of the arrow for the nullcline $f = 0$ at a certain point, and *vice versa*.

The steady state of the equation system in Eq. (7-4) is unique and is determined by

$$\bar{S} = \frac{k_2^2}{k_1 v_0}, \bar{P} = \frac{v_0}{k_2}. \tag{7-5}$$

The stability of the steady state can be analyzed by inspection of the Jacobian matrix (Chapter 3.2),

$$J = \begin{pmatrix} -\bar{P}^2 k_1 & -2\bar{S}\bar{P}\, k_1 \\ \bar{P}^2 k_1 & 2\bar{S}\bar{P}\, k_1 - k_2 \end{pmatrix} = \begin{pmatrix} -v_0^2 \dfrac{k_1}{k_2^2} & -2\, k_2 \\ v_0^2 \dfrac{k_1}{k_2^2} & k_2 \end{pmatrix}. \tag{7-6}$$

The character of the steady state is given by the determinant and the trace of the Jacobian matrix. The determinant reads

$$DetJ = v_0^2 \frac{k_1}{k_2^2} . \tag{7-7}$$

Since $v_0 > 0$, the determinant is always positive. However, the trace

$$TraceJ = -v_0^2 \frac{k_1}{k_2^2} + k_2 \tag{7-8}$$

changes its sign at

$$v_0^2 = \frac{k_2^3}{k_1} . \tag{7-9}$$

This surface separates stable from unstable steady states in the parameter space. Further critical values can be found by the condition $(TraceJ)^2 = 4\,DetJ$ (Section 3.2.3) separating nodes from foci. This condition is fulfilled at

$$v_0^4 \frac{k_1^2}{k_2^4} - 6\,v_0^2 \frac{k_1}{k_2} + k_2^2 = 0 . \tag{7-10}$$

At the transition from the region of stable focus ($(TraceJ)^2 < 4\,Det$, $TraceJ < 0$) to the region of instable focus ($(TraceJ)^2 < 4\,Det$, $TraceJ > 0$), limit cycles arise. The

$$DetJ = v_0^2 \frac{k_1}{k_2^2} . \tag{7-7}$$

Since $v_0 > 0$, the determinant is always positive. However, the trace

$$TraceJ = -v_0^2 \frac{k_1}{k_2^2} + k_2 \tag{7-8}$$

changes its sign at

$$v_0^2 = \frac{k_2^3}{k_1} . \tag{7-9}$$

This surface separates stable from unstable steady states in the parameter space. Further critical values can be found by the condition $(TraceJ)^2 = 4\,DetJ$ (Section 3.2.3) separating nodes from foci. This condition is fulfilled at

$$v_0^4 \frac{k_1^2}{k_2^4} - 6\,v_0^2 \frac{k_1}{k_2} + k_2^2 = 0 . \tag{7-10}$$

At the transition from the region of stable focus ($(TraceJ)^2 < 4\,Det$, $TraceJ < 0$) to the region of instable focus ($(TraceJ)^2 < 4\,Det$, $TraceJ > 0$), limit cycles arise. The

$$DetJ = v_0^2 \frac{k_1}{k_2^2} . \tag{7-7}$$

Since $v_0 > 0$, the determinant is always positive. However, the trace

$$TraceJ = -v_0^2 \frac{k_1}{k_2^2} + k_2 \tag{7-8}$$

changes its sign at

$$v_0^2 = \frac{k_2^3}{k_1} . \tag{7-9}$$

This surface separates stable from unstable steady states in the parameter space. Further critical values can be found by the condition $(TraceJ)^2 = 4\,DetJ$ (Section 3.2.3) separating nodes from foci. This condition is fulfilled at

$$v_0^4 \frac{k_1^2}{k_2^4} - 6\,v_0^2 \frac{k_1}{k_2} + k_2^2 = 0 . \tag{7-10}$$

At the transition from the region of stable focus ($(TraceJ)^2 < 4\,Det$, $TraceJ < 0$) to the region of instable focus ($(TraceJ)^2 < 4\,Det$, $TraceJ > 0$), limit cycles arise. The

4.3 Cell Cycle - Steps in the Cycle

The eukaryotic cell cycle is the repeated sequence of events accompanying the division of a cell into daughter cells (Johnson and Walker 1999). It includes two main sections: the doubling of the genome (DNA) and all other cell components in the S phase (synthesis phase) and halving of the genome during the M phase (mitosis). The periods between the M and S phases are the gap or growth phases G_1 and G_2 (Fig. 2.13). Passage through the eukaryotic cell cycle is strictly regulated by the periodic synthesis and destruction of cyclins that bind and activate cyclin-dependent kinases (CDKs). The notion "kinase" expresses that their function is phosphorylation of proteins with controlling functions. Cyclin-dependent kinase inhibitors (CKI) also play important roles in cell cycle control by coordinating internal and external signals and impeding proliferation at several key checkpoints.

The general scheme of the cell cycle is conserved from yeast to mammals. The levels of cyclins rise and fall during the stages of the cell cycle. The levels of CDKs appear to remain constant during the cell cycle, but the individual molecules are either unbound or bound to cyclins. In budding yeast, one CDK (Cdc28) and nine different cyclins (Cln1 to Cln3, Clb1 to Clb6) that seem to be at least partially redundant are found. In contrast, mammals employ a variety of different cyclins and CDKs. Cyclins include a G1 cyclin (cyclin D), S-phase cyclins (A and E), and mitotic cyclins (A and B). Mammals have nine different CDKs (referred to as CDK1–CDK9) that are important in different phases of the cell cycle. The anaphase-promoting complex (APC) triggers the events leading to destruction of the cohesions, thus allowing the sister chromatids to separate and degrade the mitotic cyclins.

Steps in the Cycle

Let us take a course through the mammalian cell cycle starting in the G1 phase. As the level of G_1 cyclins rises, they bind to their CDKs and signal the cell to prepare the chromosomes for replication. When the level of S phase–promoting factor (SPF) rises, which includes cyclin A bound to CDK2, it enters the nucleus and prepares the cell to duplicate its DNA (and its centrosomes). As DNA replication continues, cyclin E is destroyed, and the level of mitotic cyclins begins to increase (in G_2). The M phase–promoting factor (the complex of mitotic cyclins with the M-phase CDK) initiates (1) assembly of the mitotic spindle, (2) breakdown of the nuclear envelope, and (3) condensation of the chromosomes. These events take the cell to metaphase of mitosis. At this point, the M phase–promoting factor activates the APC, which allows the sister chromatids at the metaphase plate to separate and move to the poles (anaphase), thereby completing mitosis. APC destroys the mitotic cyclins by coupling them to ubiquitin, which targets them for destruction by proteasomes. APC turns on the synthesis of G_1 cyclin for the next turn of the cycle and it degrades geminin, a protein that keeps the freshly synthesized DNA in the S phase from being re-replicated before mitosis.

A number of checkpoints ensure that all processes connected with cell cycle progression and DNA doubling and separation occur correctly. At these checkpoints, the cell cycle can be aborted or arrested. They involve checks on completion of the S phase, on DNA damage, and on failure of spindle behavior. If the damage is irreparable, apoptosis is triggered. An important checkpoint in G_1 has been identified in both yeast and mammalian cells. Referred to as "start" in yeast and as "restriction point" in mammalian cells, this is the point at which the cell becomes committed to DNA replication and completing a cell cycle (Hartwell 1974; Hartwell et al. 1974; Pardee 1974; Nurse 1975). All the checkpoints require the services of complexes of proteins. Mutations in the genes encoding some of these proteins have been associated with cancer. These genes are regarded as oncogenes. Failures in checkpoints permit the cell to continue dividing despite damage to its integrity. Understanding how the proteins interact to regulate the cell cycle became increasingly important to researchers and clinicians when it was discovered that many of the genes that encode cell cycle regulatory activities are targets for alterations that underlie the development of cancer. Several therapeutic agents, such as DNA-damaging drugs, microtubule inhibitors, antimetabolites, and topoisomerase inhibitors, take advantage of this disruption in normal cell cycle regulation to target checkpoint controls and ultimately induce growth arrest or apoptosis of neoplastic cells.

For the presentation of modeling approaches, we will focus on the yeast cell cycle since intensive experimental and computational studies have been carried out using different types of yeast as model organisms. Mathematical models of the cell cycle can be used to tackle, for example, the following relevant problems.

small (the cells accumulate maternal cytoplasm), while after fertilization cells divide without cell growth. How is the dependence on the ratio regulated?

2. Cancer cells represent a failure in cell cycle regulation. Which proteins or protein complexes are essential for checkpoint examination?
3. What causes the oscillatory behavior of the compounds involved in the cell cycle?

4.4 Models of Budding Yeast Cell Cycle

Tyson, Novak, and colleagues have developed a series of models describing the cell cycle of budding yeast in detail (Tyson et al. 1996; Novak et al. 1999; Chen et al. 2000, 2004). These comprehensive models employ a set of assumptions that are summarized in the following.

The cell cycle is an alternating sequence of the transition from the G_1 phase to the S/M phase, called "Start", and the transition from S/M to G_1, called "Finish". An overview is given in Fig. 7.6.

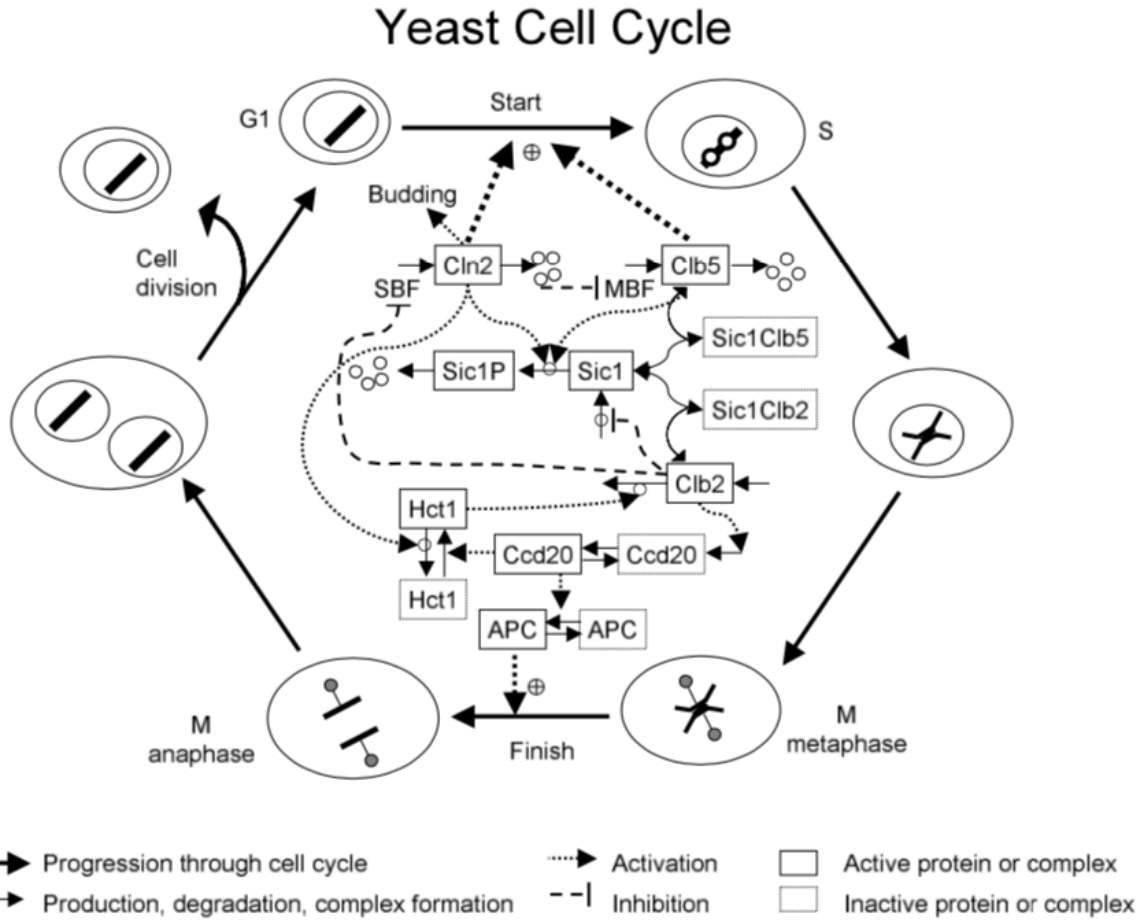


Fig. 7.6 Schematic representation of the yeast cell cycle (inspired by Fall et al. [2002]). The outer ring represents the cellular events. Beginning with cell division, the G_1 phase follows. The cells possess a single set of chromosomes (shown as one black line). At Start, the cell goes into the S phase and replicates the DNA (two black lines). The sister chromatids are initially kept together by proteins. During the M phase they are aligned, attached to the spindle body, and segregated to different parts of the cell. The cycle closes with formation of two new daughter cells. The inner part represents the main molecular events driving the cell cycle, comprising (1) protein production and degradation, (2) phosphorylation and dephosphorylation, and (3) complex formation and disintegration. For sake of clarity, CDK Cdc28 is not shown. The Start is initiated by activation of CDK by cyclins Cln2 and Clb5. The CDK activity is responsible for progression through the S and M phases. At Finish, the proteolytic activity coordinated by APC destroys the cyclins and thereby renders the CDK inactive.

The CDK (Cdc28) forms complexes with the cyclins Cln1 to Cln3 and Clb1 to Clb6, and these complexes control the major cell cycle events in budding yeast cells. The complexes Cln1-2/Cdc28 control budding, the complex Cln3/Cdc28 governs the executing of the checkpoint Start, Clb5–6/Cdc28 ensures timely DNA replication, Clb3-4/Cdc28 assists DNA replication and spindle formation, and Clb1-2/Cdc28 is necessary for completion of mitosis.

The cyclin-CDK complexes are in turn regulated by synthesis and degradation of cyclins and by the Clb-dependent kinase inhibitor (CKI) Sic1. The expression of the gene for Cln2 is controlled by the transcription factor SBF, and the expression of the gene for Clb5 is controlled by the transcription factor MBF. Both transcription factors are regulated by CDKs. All cyclins are degraded by proteasomes following ubiquitination. APC is one of the complexes triggering ubiquitination of cyclins.

For the implementation of these processes in a mathematical model, the following points are important. Activation of cyclins and cyclin-dependent kinases occurs in principle by the negative feedback loop presented in Goldbeter's minimal model (see Section 7.2.1). Furthermore, the cells exhibit exponential growth. For the dynamics of the cell mass M, it holds that $dM/dt = \mu M$. At the instance of cell division, M is replaced by $M/2$. In some cases uneven division is considered. Cell growth implies adaptation of the negative feedback model to growing cells.

The transitions Start and Finish characterize the wild-type cell cycle. At Start, the transcription factor SBF is turned on and the levels of the cyclins Cln2 and Clb5 increase. They form complexes with Cdc28. The boost in Cln2/Cdc28 has three main consequences: it initiates bud formation, it phosphorylates the CKI Sic1 promoting its disappearance, and it inactivates Hct1, which in conjunction with APC is responsible for Clb2 degradation in the G_1 phase. Hence, DNA synthesis takes place and the bud emerges. Subsequently, the level of Clb2 increases and the spindle starts to form. Clb2/Cdc28 inactivates SBF and Cln2 decreases. Inactivation of MBF causes Clb5 to decrease. Clb2/Cdc28 induces progression through mitosis. Cdc20 and Hct1, which target proteins to APC for ubiquitination, regulate the metaphase-anaphase transition. Cdc20 has several tasks in the anaphase. Furthermore, it activates Hct1, promoting degradation of Clb2, and it activates the transcription factor of Sic1. Thus, at Finish, Clb2 is destroyed and Sic1 reappears.

The dynamics of some key players in the cell cycle according to the model given in Chen et al. (2000) is shown in Fig. 7.7 for two successive cycles. At Start, Cln2 and

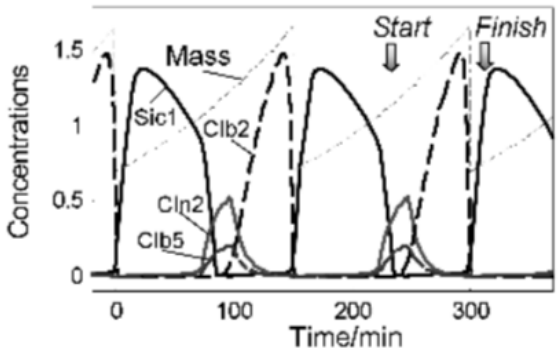


Fig. 7.7 Temporal behavior of some key players during two successive rounds of the yeast cell cycle. The dotted line indicates the cell mass that halves after every cell division. The levels of Cln2, $Clb2_{total}$, $Clb5_{total}$, and $Sic1_{total}$ are simulated according to the model presented by Chen et al. (2000).

4.5 Modeling of Gene Expression

The expression of genes, which is a highly regulated process in eukaryotic as well as in prokaryotic cells, has a profound impact on the ability of the cells to maintain vitality, perform cell division, and respond to environmental changes or stimuli. In theoretical modeling of gene expression, two diverse approaches have been developed. On the one hand, the expression of one or a few genes has been described on the level of transcription or translation by detailed mathematical models that include the binding of transcription factors and RNA polymerases to DNA, the effect of specific inhibitors or activators, the formation of various stages of maturation of mRNA or proteins, and the regulation by internal feedback loops or external regulators. The basis of this type of modeling is knowledge or hypotheses about the processes and interactions taking place during gene expression. Like most types of kinetic modeling, it often lacks specific kinetic parameters for the individual processes under consideration. On the other hand, the expression changes of thousands of genes are analyzed in parallel over time with DNA arrays. Gene expression profiles (expression levels at different time points) and gene expression patterns (comparison of expression values of different genes under different experimental conditions) are used to search for clusters and motifs and, eventually, to deduce functional correlations. Based on this information, reverse engineering methods seek to reconstruct the underlying regulatory networks (Section 9.6). While these approaches largely neglect the highly complex regulatory machinery behind the emergence of detectable mRNA involving the action of proteins and other regulatory molecules, they cover a large fraction or almost all genes of a cell – compared to the first approach, which can deal only with a few genes.

4.6 Modules of Gene Expression

In the following section, we will outline a general view of the processes representing gene expression, from the activation of transcriptional regulators to the synthesis of a functional protein (Orphanides and Reinberg 2002; Proudfoot et al. 2002; Reed and Hurt 2002).

Hundreds of different cell types exist and fulfill specific roles in the organism. Each cell type theoretically contains information on the same set of genes; however, only a proportion of these genes is expressed, determining the specific role of cells of this type. Gene expression in eukaryotes is controlled at six different steps, which determine the diversity and specification of the organism (Alberts et al. 2002):

1. Transcriptional control: when and how often a gene is transcribed.
2. RNA processing control: how the RNA transcript is spliced.
3. RNA transport and localization control: which mRNAs in the nucleus are exported to cytosol and where in the cytosol they are localized.
4. Translation control: which mRNAs in the cytosol are translated by ribosomes.
5. mRNA degradation control: which mRNAs in the cytosol are destabilized.
6. Protein activity control: determines activation, inhibition, compartmentalization, and degrading of the translated protein.

Each step is complex and has been studied extensively in isolation. The process is typically modeled with a linear structure of more or less independent modules where the output of the previous module is the input for the current module.

The expression level of the majority of genes is controlled by transcription factors. Transcription factors are proteins that bind to DNA regulatory sequences upstream of the site at which transcription is initiated. Various regulatory pathways control their activities (see Chapter 6). More than 5% of human genes encode transcription factors (Tupler et al. 2001). Once activated, transcription factors bind to gene regulatory elements and, through interactions with other components of the transcription machinery, promote access to DNA and facilitate the recruitment of the RNA polymerase enzymes to the transcriptional start site.

In eukaryotes, there are three RNA polymerases, namely, RNAP I, II, and III. RNAP II catalyzes the transcription of protein-coding genes and is responsible for the synthesis of mRNAs and certain small nuclear RNAs, while the others are responsible for generating primarily tRNAs (RNAP III) and ribosomal RNAs (RNAP I) (Allison et al. 1985).

The RNAP II enzyme itself is unable to initiate promoter-dependent transcription in the absence of complementing factors. It needs to be supplemented by so-called general transcription factors (GTFs) (Orphanides et al. 1996). RNAP II together with these GTFs and the DNA template form the pre-initiation complex, and the assembly of this complex is nucleated by binding of TBP (a component of TFIID) to the "TATA box" (Woychik and Hampsey 2002). The TATA box is a core promoter (or minimal promoter) that directs transcriptional initiation at a short distance (about 30 bp downstream). Soon after RNAP II initiates transcription, the nascent RNA is modified by the addition of a cap structure at its 5′ end. This cap serves initially to protect the new transcript from attack by nucleases and later serves as a binding site for proteins involved in export of the mature mRNA into the cytoplasm and its translation into protein.

General Promoter Structure

Promoter prediction algorithms implicitly assume a specific model for a typical promoter. The general structure of an RNAP II promoter is described in Fig. 8.1a. The typical promoter is composed of three levels of regulatory sequence signals. The first level contains sequence motifs that enable the binding of specific transcription factors. The next level is the combination of binding sites to promoter modules that jointly act as functional units. The third level consists of the complete promoter that modulates gene transcription depending on cell type, tissue type, developmental stage, or activation by signaling pathways.

The promoter must contain binding sites for the GTFs, such as the TATA box. These proximate regulatory motifs constitute the core promoter that is able to bind the preinitiation complex and to determine the exact transcription start site. The core promoter needs additional regulatory motifs at varying distances from the transcriptional start point, the regulatory binding sites (transcription factor–binding sites, TFBSs). These sites can be situated nearby or kilobases away from the core promoter.

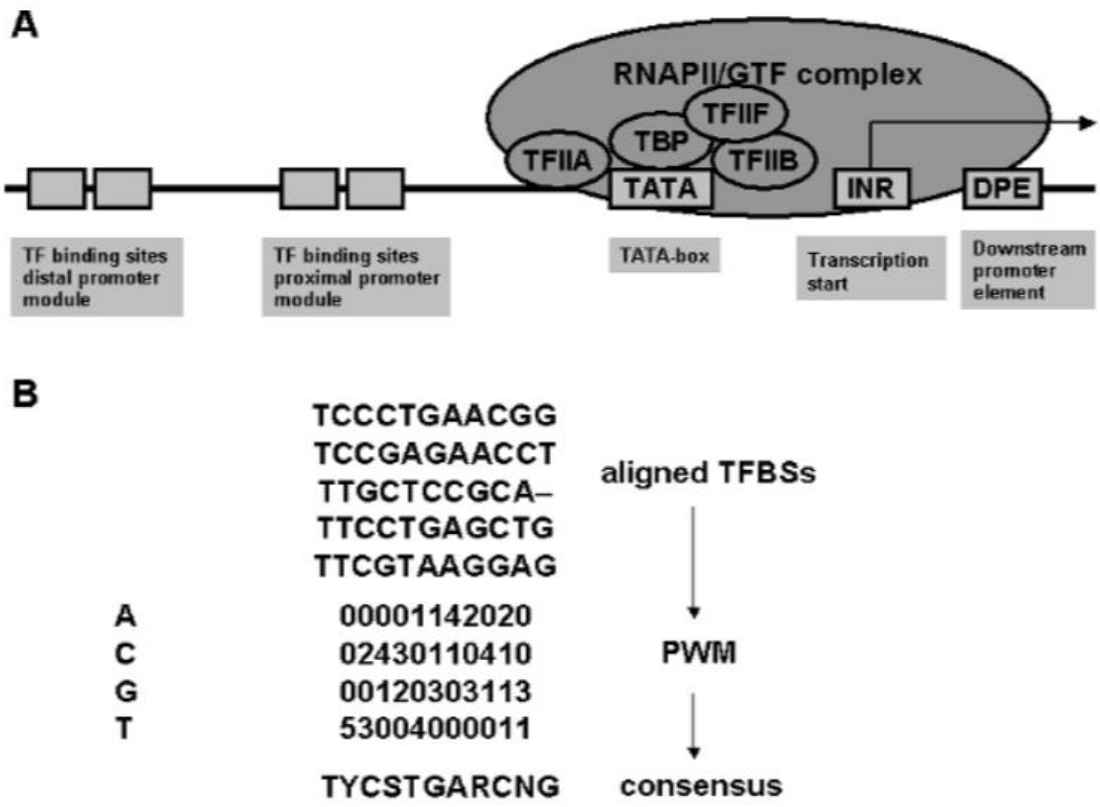


Fig. 8.1 (a) General structure of a eukaryotic gene promoter. (b) Example of a positional weight matrix and a consensus sequence derived from different transcription factor–binding sites.

Transcription initiation can be viewed as a process involving successive formation of protein complexes. In the first step, transcription factors bind to upstream promoter and enhancer sequence motifs and form a multiprotein complex. In the next step, this complex recruits the RNAP II/GTF complex to the core promoter and the transcription start site. This is done through protein-protein interactions either directly or by adaptor proteins (Ptashne and Gann 1997). The full complex then starts the transcription process.

The core promoter is located in the direct neighborhood of the transcription start site (approximately 30 bp). The core promoter is the best-characterized part of the promoter and is defined as a set of binding sites sufficient for the assembly of the RNAP II/GTF complex and for specifying transcriptional initiation. Several types of core promoters are known (Berg and von Hippel 1987):

1. TATA box: If TBP is present in the RNAP II/GTF complex, then this protein binds to the sequence motif and the transcription starts approximately 30 bp downstream.
2. TATA-less: No TATA box is present. The start site is determined by a sequence motif INR (initiator region) surrounding the start site (Smale 1994).
3. A combination of both INR and TATA box
4. Null promoter: Neither of the two sequence motifs is present. Transcription initiation is based solely on upstream (or downstream) promoter elements (Novina and Roy 1997).
5. In some cases, a downstream promoter element (DPE) exists in addition to INR, and both elements are able to specify the transcription start site (Burke and Kadonga 1997).

Whereas the core promoter determines the transcription start site, this function cannot explain how genes whose protein products are needed in parallel are co-regulated, e.g., from genes that are located on different chromosomes. Thus, additional regulatory elements are necessary that meet the requirement of higher flexibility and coordinated gene expression.

Typically a few hundred base pairs upstream of the core promoter is the proximate promoter module, which contains TFBSs for proteins responsible for the modulation of the transcription. The corresponding factors can influence the binding of the core promoter components or the chromatin structure (or both). Furthermore, a promoter can contain a distal promoter module (on the order of kilobases apart from the transcription start site). Although these modules cannot act as promoters on their own, they are able to enhance or suppress the activity of transcription up to orders of magnitude (enhancer or silencer). Enhancer and silencer often exhibit a tissue-specific activity. Like the transcription factors binding to the proximate module of the promoter, the factors binding to the distal module influence gene expression by interactions with the factors in the RNAP II/GTF complex or by changing the chromatin structure. There is no clear boundary for the promoter in the 5′ direction, and the common explanation for interactions with distal factors to the transcription apparatus is given by the formation of large loops in the DNA. The function of a pro-

moter is to increase or repress the transcription from the core promoter (basal transcription). Thus, any given gene will have a specific regulatory region determined by the binding sites of the transcription factors that ensure that the gene is transcribed in the appropriate cell type and at the proper point in development. The transcriptional activation is determined not only by the presence of the binding sites but also through the availability of the corresponding transcription factors. These transcription factors are themselves subjected to regulation and activation, e.g., through signaling pathways, and the whole process can entail complex procedures such as transcriptional cascades and feedback control loops (Pedersen et al. 1999).

A list of some promoter recognition programs is found in the following table:

Program	*Web location*	*Reference*
FunSiteP	http://compel.bionet.nsc.ru/FunSite/fsp.html	Kondrakhin et al. (1995)
PomoterInspector	http://www.genomatix.de/cgi-bin/promoterinspector/promoterinspector.pl	Scherf et al. (2000)
PromoterScan	http://bimas.dcrt.nih.gov/molbio/proscan	Prestridge (1995)
NNNP	http://www.fruitfly.org/seq_tools/promoter.html	Reese (2001)
PromFind	http://iubio.bio.indiana.edu/soft/molbio/mswin/mswin-or-dos/profin11.exe	Hutchinson (1996)
TSSG/TSSW	http://www.softberry.com	Solovyev and Salamov (1997)
FirstEF	http://rulai.cshl.org/tools/FirstEF	Davuluri et al. (2001)

Modeling Specific Processes in Eukaryotic Gene Expression

We want to know which genes are expressed, to what level, and where and when in order to comprehend the functioning of organisms at the molecular level. A network of interactions among DNA, RNA, proteins, and other molecules realizes the regulation of gene expression. This network involves many components. There is forward flow of information from gene to mRNA to protein according to the dogma of mole-

cular biology. Moreover, positive and negative feedback loops and information exchange with signaling pathways and energy metabolism ensure the appropriate regulation of the expression according to the actual state of the cell and its environment.

Modeling of gene expression is an example of a scientific field where one may obtain results with different techniques. The dynamics or the results of gene expression have been mathematically described with Boolean networks, Bayesian networks, directed graphs, ordinary and partial differential equation systems, stochastic equations, and rule-based formalisms.

Although understanding of the regulation of large groups of genes, of the emergence of complex patterns of gene expression, and of relations with inter- and intracellular communication is still a scientific challenge, many insights have already been gained from the modeling of particular processes or of the regulation of individual sets of genes.

One Example, Different Approaches

In the following sections we will present an overview of modeling approaches and the scientific questions that can be tackled with different techniques. For the sake of clarity, we will use only examples with a low number of components (genes and proteins), although the presented approaches can also be applied to larger systems.

The example presented in Fig. 8.2 contains four genes, a through d, which code for the proteins A through D. mRNA is not shown for sake of simplicity. The proteins A and B may form a heterodimer that activates the expression of gene c. Protein C inhibits the expression of genes b and d, which are in this way co-regulated. Protein D is necessary for the transcription of protein B.

8.3.1.1 Description with Ordinary Differential Equations

Gene expression can be mathematically described with systems of ordinary differential equations in the same way as dynamical systems in metabolism (Chapter 5), signaling (Chapter 6), and other cellular processes (Chapter 7). In general, one considers

$$\frac{dx_i}{dt} = f_i(x_1, .., x_n) \quad i = 1, ..., n. \tag{8-2}$$

The variables x_i represent the concentrations of mRNAs, proteins, or other molecules. The functions f_i comprise the rate equations that express the changes of x_i due to transcription, translation, or other individual processes. For details about how to specify the rate equations and how to analyze the resulting ODE systems, compare Sections 5.1, 5.2 and 3.2.

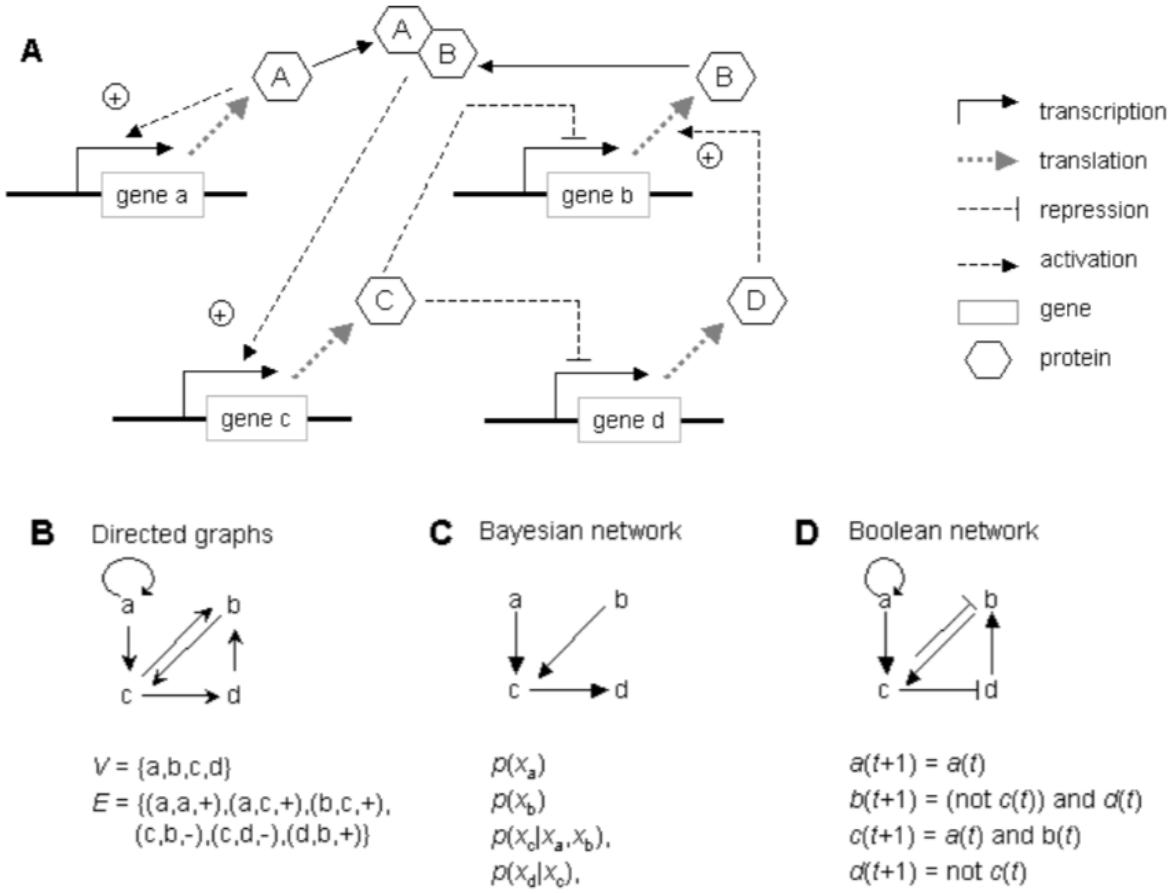


Fig. 8.2 Gene regulatory network comprising four genes a–d. (a) Dependence of translation of genes a–d, the transcription of their mRNAs (not shown), and the influence of the respective proteins A–D. (b) Representation as directed graph. (c) Respective Bayesian network. Note that some interactions are neglected (inhibition of *b* by *c*, activation of *b* by *d*) in order to get a network without cycles. (d) The Boolean network.

Example 8-1

The dynamics of the system depicted in Fig. 8.2 can be described in several ways depending on the desired particularization. If we consider only the mRNA abundances a, b, c, and d, we get:

$$\frac{da}{dt} = f_a(a)$$

$$\frac{db}{dt} = f_b(b, c, d)$$

$$\frac{dc}{dt} = f_c(a, b, c)$$

$$\frac{dd}{dt} = f_d(c, d)\,. \tag{8-3}$$

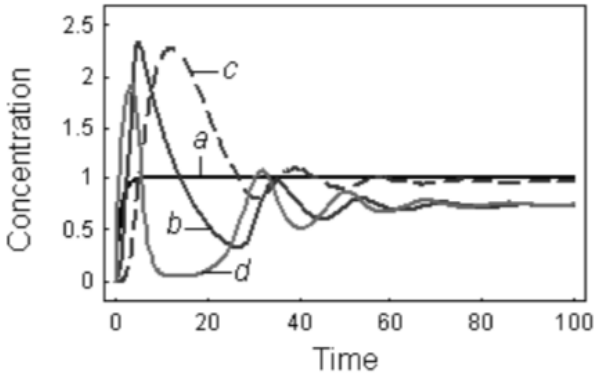


Fig. 8.3 Dynamics of the mRNA concentrations of the system presented in Example 8-1 according to Eq. (8-4). Parameters: $\nu_a = 1$, $k_a = 1, V_b = 1$, $K_b = 5$, $K_{lc} = 0.5$, $n_c = 4$, $k_b = 0.1$, $V_c = 1$, $K_c = 5$, $k_c = 0.1, V_d = 1$, $k_d = 1$. Initial conditions: $a(0) = b(0) = c(0) = d(0) = 0$.

8.3.1.2 Representation of Gene Network as Directed and Undirected Graphs

A directed graph G is a tuple $\langle V, E\rangle$, where V denotes a set of vertices and E a set of edges (cf. Section 3.5). The vertices $i \in V$ correspond to the genes (or other components of the system) and the edges correspond to their regulatory interactions. An edge is a tuple $\langle i, j\rangle$ of vertices. It is directed if i and j can be assigned to the head and tail of the edge, respectively. The labels of edges or vertices may be expanded to contain information about the genes and their interactions. In a general way, one may express an edge as a tuple $\langle i, j,$ *properties*$\rangle$. The entry *properties* can simply indicate whether j activates (+) or inhibits (–) i (Fig. 8.2 b). The entry *properties* can also be a list of regulators and their influence on that specific edge, such as $\langle i, j,$ ((k, activation), (l, inhibition as homodimeric protein))$\rangle$.

In principle, many databases that provide information about genetic regulation are organized as richly annotated directed graphs (e. g., Transfac, KEGG; see Chapter 13). Directed graphs are not suited to predict the dynamics of a network, but they may contain information that allows certain predictions about network properties:

- Tracing paths between genes yields the sequence of regulatory events, shows redundancy in the regulation, or indicates missing regulatory interactions (that are, for example, known from experiment).
- A cycle in the network may indicate feedback regulation.
- Comparison of gene regulatory networks of different organisms may reveal evolutionary relations and reveal targets for bioengineering and for pharmaceutical applications (Dandekar et al. 1999).
- The network complexity can be measured by the connectivity, i. e., the distribution and the average of the numbers of regulators per gene.

8.3.1.3 Bayesian Networks

A Bayesian network (see also Section 3.5.2.3) is based on the representation of the regulatory network as a directed acyclic graph $G = \langle V, E \rangle$. Again, the vertices $i \in V$ represent genes and edges denote regulatory interactions. Variables x_i belonging to the vertices i denote a property relevant to the regulation, e.g., the expression level of a gene or the amount of active protein. A conditional probability distribution $p(x_i | L(x_i))$ is defined for each x_i, where $L(x_i)$ are the parent variables belonging to the direct regulators of i. The directed graph G and the conditional distributions together specify a joint probability distribution $p(\boldsymbol{x})$ that determines the Bayesian network. The joint probability distribution can be decomposed into

$$p(x) = \prod_i p\left(x_i \middle| L(x_i)\right). \tag{8-6}$$

The directed graph expresses dependencies of probabilities: the expression level of a gene represented by a child vertex depends on the expression levels of genes belonging to the parent vertices. Hence, it also implies conditional independencies $i(x_i; \boldsymbol{y} | \boldsymbol{z})$, meaning that x_i is independent of the set of variables $\boldsymbol{y}$ given the set of variables $\boldsymbol{z}$. Two graphs or Bayesian networks are equivalent if they imply the same set of independencies. In this case they can be considered as the same undirected graph, but with varying direction of edges. Equivalent graphs cannot be distinguished by observation of the variables $\boldsymbol{x}$ (Friedman et al. 2000).

Bayesian networks have been used to deduce gene regulatory networks from gene expression data. The aim is to find the network or equivalence class of networks that best explains the measured data. A problem is the determination of initial probability distributions.

8.3.1.4 Boolean Networks

In the Boolean network approach (see also Section 3.5.2.2 and Section 10.3.3 for Boolean rules), the expression level of each gene is assigned to a binary variable: a gene is considered to be either on (1) or off (0), i.e., it is transcribed or not. The states of the genes are updated simultaneously in discrete time steps. The new state can depend on the previous state of the same gene or other genes. These dependencies cause the Boolean network. The following termini are used: the N genes are the N nodes of the network, the k interactions regulating the expression of a certain gene are the k inputs of that node, and the binary expression value of each gene is its output. Since every node can be in one of two different states, a network of N genes can assume 2^N different states. An N-dimensional vector of variables can describe

the state at time t. The value of each variable at time $t+1$ depends on the values of its inputs. It can be computed by means of the Boolean rules (see Section 10.3.3). For a node with k inputs, the number of possible Boolean rules is 2^{2^k}. Although a Boolean network is a very simplified representation of the gene regulatory network, it enables a first computation of gene expression dynamics.

For the network presented in Fig. 8.2 d the following Boolean rules apply:

$$a(t+1) = f_a(a(t)) = a(t) \qquad \text{Rule 1 for } k=1$$

$$b(t+1) = f_b(c(t), d(t)) = (\text{not } c(t)) \text{ and } d(t) \qquad \text{Rule 2 for } k=2$$

$$c(t+1) = f_c(a(t), b(t)) = a(t) \text{ and } b(t) \qquad \text{Rule 2 for } k=2$$

$$d(t+1) = f_d(c(t)) = (\text{not } c(t)) \qquad \text{Rule 0 for } k=1$$

The temporal behavior is determined by the sequence of states (a, b, c, d) given an initial state (compare also Section 10.3.3).
From Tab. 8.1 it is easy to see that this network has two different types of stationary behavior. If the initial state of a is 0, then the system evolves towards the steady state 0101, meaning that genes a and c are off, while genes b and d are on. If the initial state of a is 1, then the system evolves towards a cyclic behavior including the following sequence of states: 1000 → 1001 → 1101 → 1111 → 1010 → 1000.

Tab. 8.1 Successive states in the Boolean network.

0000 → 0001	1000 → 1001
0001 → 0101	1001 → 1101
0010 → 0000	1010 → 1000
0011 → 0000	1011 → 1000
0100 → 0001	1100 → 1011
0101 → 0101	1101 → 1111
0110 → 0000	1110 → 1010
0111 → 0000	1111 → 1010

The sequence of states given by the Boolean transitions represents the trajectory of the system. Since the number of states in the state space is finite, the number of possible transitions is also finite. Therefore, each trajectory will lead either to a steady state or to a state cycle. These states are called attractors. Transient states are those states that do not belong to an attractor. All states that lead to the same attractor constitute the basin of attraction.

Boolean networks have been used to explore general and global properties of large gene expression networks. Considering random networks (the number k of inputs per gene and the corresponding Boolean rules are chosen by chance), Kauffman (1991, 1993) has shown that the systems exhibit highly ordered dynamics for small k

The sequence of states given by the Boolean transitions represents the trajectory of the system. Since the number of states in the state space is finite, the number of possible transitions is also finite. Therefore, each trajectory will lead either to a steady state or to a state cycle. These states are called attractors. Transient states are those states that do not belong to an attractor. All states that lead to the same attractor constitute the basin of attraction.

Boolean networks have been used to explore general and global properties of large gene expression networks. Considering random networks (the number k of inputs per gene and the corresponding Boolean rules are chosen by chance), Kauffman (1991, 1993) has shown that the systems exhibit highly ordered dynamics for small k

4.7 Modeling the Elongation of a Peptide Chain

Protein expression is an essential feature of cellular development and operation. The code for all the proteins needed for a cell to survive and thrive is in its DNA. However, the code must be transcribed into mRNA molecules, then translated into polypeptide chains, and finally folded and further chemically processed into functioning molecules. The rates of expression of the proteins are determined by many factors, including regulation at the transcription site upstream of the gene. Regulators moderate the attachment and operation of the RNA polymerase, which transcribes an mRNA chain containing the code for the protein. The mRNA chains then attach to the ribosome, which translates the code into peptide chains by successively adding the proper amino acid to the nascent chain.

A mathematical model, consisting of differential equations representing the rate of change of the concentration of each protein, has been derived (Drew, 2001), and has the capacity of accounting for the repression or activation of mRNA transcription by transcription factor proteins. This model describes each reaction in terms of kinetic rate constants for sub-parts of the overall reaction.

In order for that model to give meaningful results, kinetic rate constants must be supplied for each part of the reactions. One of the sub-processes for which a rate constant is needed in the model is peptide chain elongation. Drew models the attachment of the mRNA to the ribosome, followed by the addition of each amino acid. The model is essentially a Markov process, whereby the evolution of the

probability of the cell having an mRNA in each of several states is described. The states that are assumed to exist in that model are (i) *free*, i.e., having no ribosome or peptide chain attached; (ii) *attached to a ribosome*, i.e., no longer free, but still without a peptide chain; and (iii) *having a nascent peptide chain of length i*. The chain is assumed to grow to length N by adding amino acid residues from the cytoplasm and ultimately be ejected from the ribosome and the mRNA when the ribosome reaches the end of the pattern for making the protein. The mRNA and the ribosome parts are then free. The process is described by a set of ordinary differential equations for the numbers of assemblies in each state. Drew (2001) assumes that these equations have the form

$$\begin{aligned}
\frac{d[\text{mRNA}]}{dt} &= -k_R[R][\text{mRNA}] + \kappa_{N-1}[\text{mRNA}_{N-1}] \\
\frac{d[\text{mRNA}_0]}{dt} &= -\kappa_1[a_1][\text{mRNA}_0] + k_R[R][\text{mRNA}] \qquad (1) \\
&\vdots \\
\frac{d[\text{mRNA}_j]}{dt} &= -\kappa_{j+1}[a_{j+1}][\text{mRNA}_j] + \kappa_j[a_j][\text{mRNA}_{j-1}]
\end{aligned}$$

where [mRNA] is the concentration of messenger RNA, [mRNA_0] is the concentration of the mRNA–ribosome complex, [mRNA_j] is the concentration of the mRNA–ribosome complex with a nascent peptide chain of length j attached. The reaction rate $-k_R[R][\text{mRNA}]$ is the rate at which the mRNA–ribosome complex is formed, that is, the rate of binding of the mRNA to the ribosome. The reaction rate $\kappa_j[a_j][\text{mRNA}_{j-1}]$ is the elongation rate assumed by Drew (2001), and is represented there by a rate constant times the concentrations of the amino acid to be attached, and the mRNA–ribosome complex with the nascent chain. This rate reflects a binary character of the addition of an amino acid to the chain.

When we examine the biochemical processes that are involved in elongation in more detail, the reaction modeled by $\kappa_j[a_j][\text{mRNA}_{j-1}]$ is clearly more complex. The elongation step is facilitated by the ribosome, a ubiquitous biological machine that assembles amino acids into peptide chains, which are then processed and folded into proteins. Great strides in understanding the mechanisms by which the ribosome works have been made over the last decade (Pape *et al.*, 1998; Frank *et al.*, 1999; Tomsic *et al.*, 2000). It is the purpose of this paper to interpret the mathematical model for the workings of the ribosome as a sub-model reflecting the overall rate of the reaction in terms of the rates of the sub-steps. In so doing, we discover the parameters on which the overall rate depends, including the relative abundances of the amino acids.

The process of elongation is accomplished one amino acid at a time, facilitated by the ribosome and the tRNA molecules. An amino acid bound to a tRNA binds to the ribosome, catalyzed by elongation factor Tu and a molecule of GTP. This complex enters the ribosome, and the codon recognition site on the tRNA is associated

with the corresponding codon on the mRNA. If the amino acid–tRNA complex recognizes a correct codon on the mRNA, the complex is stabilized by interactions of the tRNA, mRNA and the ribosome. Formation of the codon–anticodon bond activates the hydrolysis of the GTP. This causes a conformational change of the elongation factor Tu (EF-Tu) complex. Next, the elongation factor unbonds and leaves the ribosome. The amino acid is bonded to the peptide chain. If the amino acid–tRNA complex recognizes the wrong codon, the complex is rejected, and the process starts again. In this paper, we present a kinetic model for elongation in which the various sub-steps are considered. This model includes the steps outlined above, and assumes rates for each. It results in a set of differential equations for the evolution of the probabilities that the ribosome is in each of the various states. The steady-state probabilities are then found, and the dependence of elongation rate on the concentrations of amino acids is determined. This result allows the connection of the rate $\kappa_j[a_j][\mathrm{mRNA}_{j-1}]$ to various sub-rates in the detailed model for chain elongation.

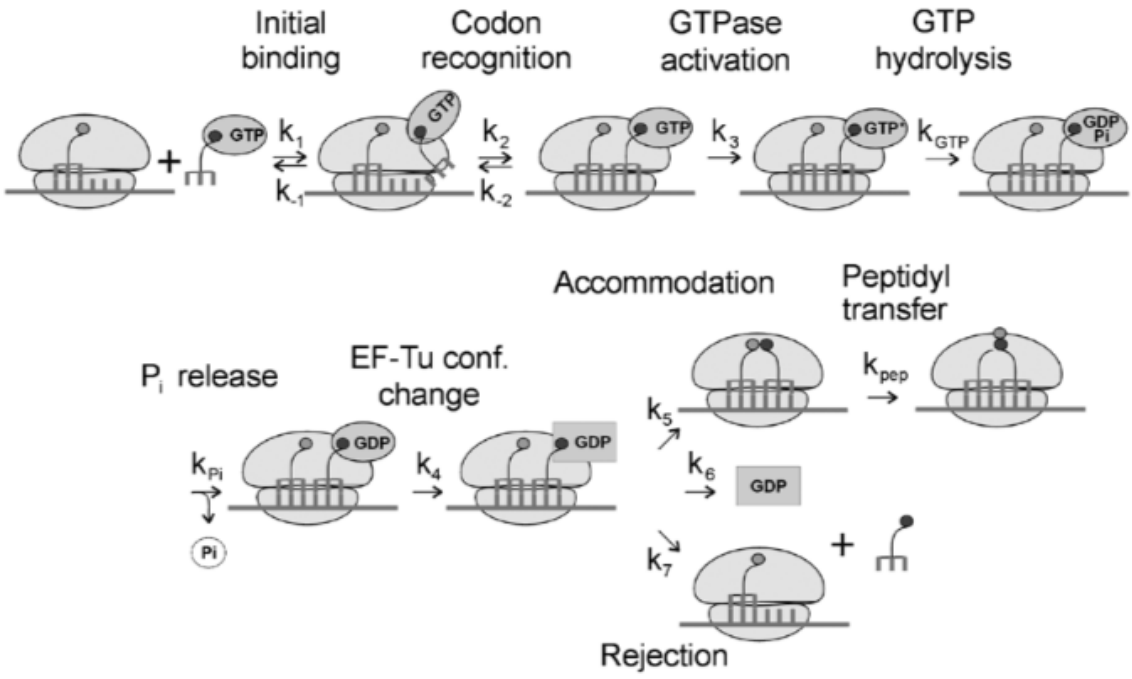


Figure 1. Mechanism of EF-Tu-dependent binding of aa-tRNA to the ribosomal *A*-site [Rodnina, 2003].

4.8 The Model According to Griffith

This model (Griffith 1968 a) considers the activation of the genes, the formation of mRNA, the synthesis of the enzymes permease and β-galactosidase, and the degradation of lactose. Permease supports the transport of lactose through the bacterial membrane. β-galactosidase isomerizes lactose to allolactose and catalyzes the cleavage of lactose to glucose and galactose.

Due to fluctuations, the genes G are rendered active even by trace amounts of allolactose (P).

$$G_{inactive} + mP \rightleftharpoons G_{active} . \tag{8-15}$$

The portion of active gene is given by $p = \frac{P^m}{k_{eq}^m + P^m}$. The concentration of mRNA (M) is determined by a basal production rate, M_0, and a degradation rate, $k_2 M$, as well as by the production from activated gene:

$$\frac{dM}{dt} = M_0 + k_1 \frac{P^m}{k_{eq}^m + P^m} - k_2 M . \tag{8-16}$$

The concentration changes of the enzymes permease (E_1) and β-galactosidase (E_2) are given by production from mRNA and degradation:

$$\frac{dE_1}{dt} = c_1 M - d_1 E_1$$

$$\frac{dE_2}{dt} = c_2 M - d_2 E_2 \,. \tag{8-17}$$

The uptake of lactose from the external ($_{ex}$) into the internal ($_{int}$) of the bacterial cell is mediated by permease (E_1), and the decay of lactose depends on β-galactosidase (E_2):

$$\frac{dLac_{ex}}{dt} = -\sigma_0 E_1 \frac{Lac_{ex}}{k_0 + Lac_{ex}}$$

$$\frac{dLac_{in}}{dt} = \sigma_0 E_1 \frac{Lac_{ex}}{k_0 + Lac_{ex}} - \sigma_1 E_2 \frac{Lac_{in}}{k_S + Lac_{in}} \,. \tag{8-18}$$

Allolactose is produced from lactose and converted to glucose and galactose:

$$\frac{dP}{dt} = \sigma_1 E_2 \frac{Lac_{in}}{k_S + Lac_{in}} - \sigma_2 E_2 \frac{P}{k_P + P} \,. \tag{8-19}$$

The equation system in Eqs. (8-16)–(8-19) has been simplified using the following assumptions: (1) the quasi-steady-state approximation (Section 5.2.7) applies for the concentration of mRNA; (2) the concentrations of the enzymes are equal, i.e., $E_1 = E_2$, as well as their rate constants of degradation, i.e., $d_1 = d_2$; and (3) there is no delay in the conversion of lactose into allolactose, expressed by $dLac_{in}/dt = 0$.

For the sake of simplicity, dimensionless variables are considered, i.e., $lac = Lac_{ex}/k_0$, $p = P/k_P$, $e = E/e_0$, and $\tau = t/t_0$. Taken together, this yields the final system of equations

$$\frac{de}{d\tau} = m_0 + \frac{p^m}{\kappa^m + p^m} - \varepsilon e$$

$$\frac{dp}{d\tau} = \mu e \left(\frac{lac}{1 + lac} - \lambda \frac{p}{1 + p} \right)$$

$$\frac{d\,lac}{d\tau} = -e \frac{lac}{1 + lac}, \tag{8-20}$$

with $e_0 = \frac{c_1 k_0 k_1}{\sigma_0 k_2}$, $t_0 = \frac{k_0}{\sigma_0 e_0}$, $\lambda = \frac{\sigma_2}{\sigma_0}$, $\mu = \frac{k_0}{k_p}$, $\kappa = \frac{k_{eq}}{k_p}$, $m_0 = \frac{M_0}{k_1}$, and $\varepsilon = t_0 d_1$.

The temporal behavior of this system for low and high external initial concentration of lactose is represented in Fig. 8.8.

For low initial concentration of *lac*, there is only a weak activation of gene expression, resulting in a low enzyme concentration. For high concentration of *lac*, the production of the enzyme is activated as long as its substrate – *lac* – is available.

4.9 Noise and oscillation in biological system

In higher organisms, circadian rhythms are generated by a multicellular genetic clock that is entrained very efficiently to the 24-hour light-dark cycle. Most studies of these circadian oscillators have considered a perfectly periodic driving by light. Naturally organisms are subject to non-negligible fluctuations in the light level all through the daily cycle. Interestingly higher organisms respond to artificial constant light conditions over several days with a kind of phase transition from the free running rhythmic to an arrhythmic behaviour. The constant light intensity determines the transition.

We investigate how the interplay between light fluctuations and intercellular coupling affects the dynamics of the central clock. We model the central circadian clock as a collective rhythm of a large ensemble of nonidentical, globally coupled cellular clocks modeled as Goodwin oscillators. Based on experimental considerations,[14] we assume an inverse dependence of the cell-cell coupling strength on the light intensity, in such a way that the larger the light intensity the weaker the coupling.

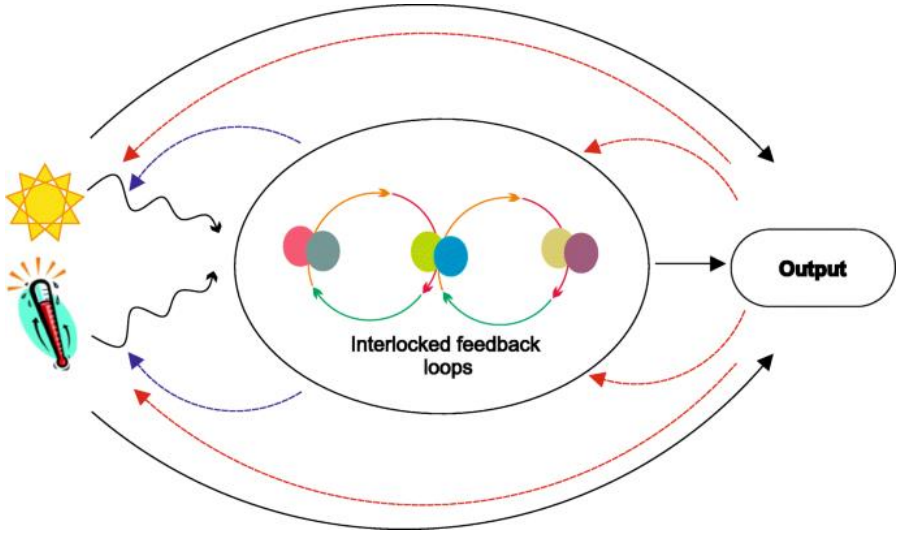
The system offers access to interesting questions from the biological viewpoint and the dynamical systems side. The phase transition from the rhythmic to the arrhythmic behaviour and the critical light intensity are essential for the coherence resonance (CR), a noise-induced effect known from the dynamical system theory. The phase transition can be observed only in the overt rhythm that we model by the mean response of all individual circadian oscillators. We study the influence of noise on the quality of the overt rhythm and consider the synchronization and the coherence of the mean-field. Our results show a noise-induced rhythm generation for constant light intensities at which the clock is arrhythmic in the noise-free case.[15] Importantly, the rhythm shows a resonance-like phenomenon as a function of the noise intensity. Such improved coherence can be only observed at the level of the overt rhythm and not at the level of the individual oscillators, thus suggesting a cooperative effect of noise, coupling, and the emerging synchronization between the oscillators.

From the biological viewpoint the CR offers a test tool for the light dependent coupling hypothesis. The CR in the discussed system relies on the hypothesis of light dependent coupling. Experimental results of a noise-induced rhythmicity for constant light intensities at which the clock is arrhythmic in the noise-free case would strengthen the biological relevant hypothesis of light dependent coupling amongst the individual oscillators. The mathematical model originates form the biological problem, makes use of a noise-induced phenomena and gives a protocol for experimental testable predictions that can be used to strengthen the biological derived hypothesis of light dependent coupling amongst the many basic circadian oscillators building the central clock. The discussed circadian model gives an example for the vice versa beneficial connection between biology and mathematical modeling.

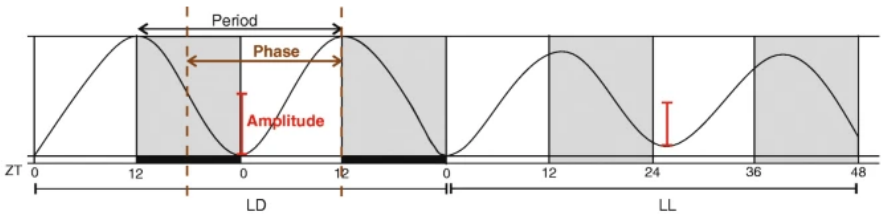
4.10 Circadian rhythm-how to build an oscillator

Circadian oscillators are networks of biochemical feedback loops that generate 24-hour rhythms in organisms from bacteria to animals. These periodic rhythms result from a

complex interplay among clock components that are specific to the organism, but share molecular mechanisms across kingdoms. A full understanding of these processes requires detailed knowledge, not only of the biochemical properties of clock proteins and their interactions, but also of the three-dimensional structure of clockwork components. Posttranslational modifications and protein–protein interactions have become a recent focus, in particular the complex interactions mediated by the phosphorylation of clock proteins and the formation of multimeric protein complexes that regulate clock genes at transcriptional and translational levels.



Generic model of the circadian clock. The complex network of coupled multiple feedback oscillators are represented by *solid color lines* and *ovals*. Clock genes forming a functional oscillator regulate the input and output pathways (*blue dashed lines*). Feedback from output pathways can also regulate the oscillator and the input pathways (*red dashed lines*). In addition to external input signal transduction for clock entrainment, input pathways can also directly affect clock output and vice versa (*solid black line*).



Circadian rhythms show the same period as the external cues when tested under entrainment conditions (light-dark cycles: LD) and may deviate from the 24 hour period under the free running conditions (constant light; LL) reflecting the period of the endogenous clock.

PERIOD: is the time taken by an oscillation to complete one cycle.

PHASE: Phase is a relative event. Any time point on a rhythmic cycle relative to an external reference time point. For example the peak of a cycle relative to the last dawn.

AMPLITUDE: It represents the level of expression of the rhythmic entity and is measured as half the magnitude from peak to trough.

ZEITGEBER (ZT): The external environmental cues that synchronize the endogenous circadian clock to the earth's diurnal and seasonal cycles. **ZT0**: is the time of onset of a signal; ZT0-ZT12 represents the subjective day when the organism is exposed to the light during entrainment; ZT12-**ZT24** represents the subjective night.

4.11 Gene circuit design

Cells navigate environments, communicate and build complex patterns by initiating gene expression in response to specific signals. Engineers seek to harness this capability to program cells to perform tasks or create chemicals and materials that match the complexity seen in nature. Circuit dynamics can be influenced by the choice of regulators and changed with expression 'tuning knobs'. We collate the failure modes encountered when assembling circuits, quantify their impact on performance and review mitigation efforts. Finally, we discuss the constraints that arise from circuits having to operate within a living cell. Collectively, better tools, well-characterized parts and a comprehensive understanding of how to compose circuits are leading to a breakthrough in the ability to program living cells for advanced applications, from living therapeutics to the atomic manufacturing of functional materials.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

School of Bio and Chemical Engineering
DEPARTMENT OF BIOINFORMATICS

UNIT V - Quantitative Models in Biological Systems - Subject Code: SBI1402

UNIT 5 COMPUTER-BASED INFORMATION RETRIEVAL AND EXAMINATION

5.1 Computer-based Information Retrieval and Examination

5.2 Gene Ontology – KEGG and BRENDA

5.3 Modeling and Visualization tools - Gepasi, Copasi

5.4 MEGA

5.5 Netpath

5.6 Biotapestry

5.7 E-Cell

5.8 PyBioS

5.9 Systems Biology Workbench

5.10 Jdesigner

5.11 CellDesigner

5.12 Petri Nets

5.13 Model Exchange Languages and Data Formats - Introduction to XML

5.14 Systems Biology Markup Language

5.15 MathML

5.16 Cytoscape

5.17 SBML tool box for MATLAB.

UNIT 5 COMPUTER-BASED INFORMATION RETRIEVAL AND EXAMINATION

5.1 Computer-based Information Retrieval and Examination - Databases and Tools on the Internet

With the rapid increase of biological data, it has become even more important to organize and structure the data in a way that information can easily be retrieved. As a result, the number of databases has also increased rapidly over the past few years. Most of these databases have a Web interface and can be accessed from everywhere in the world, which is an enormously important service for the scientific community. Again we have to emphasize that we can give only a very brief summary of a small number of databases. An extensive list of databases can be found at http://www.mpiem.gwdg.de/Forschung/Biol/biol_index_en.html. Furthermore, the journal *Nucleic Acids Research* offers a databases issue each year in January that is dedicated to factual biological databases and, additionally, a Web server issue in July presenting Web-based services such as tools for sequence comparison or prediction of 3-D protein structure.

5.2 Gene Ontology – KEGG and BRENDA

The accumulation of scientific knowledge is a decentralized, parallel process. Consequently, the naming and description of new genes and gene products is not necessarily systematic. Often, gene products with identical functions are given different names in different organisms or the verbal description of the location and function might be quite different (e.g., protein degradation vs. proteolysis). This, of course, makes it very difficult to perform efficient searching across databases and organisms.

This problem has been recognized, and in 1998 the Gene Ontology (GO) project (http://www.geneontology.org) was initiated as a collaborative effort of the Saccharomyces Genome Database (SGD), the Mouse Genome Database (MGD), and FlyBase. The aim of the Gene Ontology is to provide a consistent, species-independent, functional description of gene products. Since 1998 the GO project has grown considerably and now includes databases for plant, animal, and prokaryotic genomes. Effectively, GO consists of a controlled vocabulary (the GO terms) used to describe the biological function of a gene product in any organism. The GO terms have a defined parent-child relationship and form a directed acyclic graph (DAG) (cf. Section 3.5.1).

In a DAG, each node can have multiple child nodes, as well as multiple parent nodes. Cyclic references, however, are forbidden. The combination of vocabulary and relationship between nodes is referred to as ontology. At the root of the GO are the three top-level categories, molecular function, biological process, and cellular component, which contain many levels of child nodes (GO terms) that describe a gene product with increasing specificity. The GO consortium, in collaboration with other databases, develops and maintains the three top-level ontologies (the set of GO terms and their relationship) themselves, creates associations between the ontologies and the gene products in the participating databases, and develops tools for the creation, maintenance, and use of the ontologies.

Let's look at a practical example to see how the concept works. The enzyme superoxide dismutase, for instance, is annotated in FlyBase (the *Drosophila melanogaster* database) with the GO term "cytoplasm" in the cellular component ontology, with the GO terms "defense response" and "determination of adult lifespan" in the biological process ontology, and with the terms "antioxidant activity" and "copper, zinc superoxide dismutase activity" in the molecular function ontology. The GO term cytoplasm itself has the single parent "intracellular," which has the single parent "cell," which is finally connected to the cellular component. The other GO terms for superoxide dismutase are connected in a similarly hierarchical way to the three top categories.

The following table gives the number of gene products that have been annotated to the top-level categories of the GO for several popular databases. The table dates from January 2004 and excludes annotations that are based exclusively on electronic inferences.

KEGG

KEGG (Kyoto Encyclopedia of Genes and Genomes; http://www.genome.ad.jp/kegg/) is a reference knowledgebase offering information about genes and proteins, biochemical compounds, reactions, and pathways. The data are organized in three parts: the gene universe (consisting of the GENES, SSDB, and KO databases), the chemical universe (with the COMPOUND, GLYCAN, REACTION, and ENZYME databases which are merged as the LIGAND database), and the protein network consisting of the PATHWAY database (Kanehisa et al. 2004). In addition, the KEGG database is hierarchically classified into categories and subcategories at four levels. The five topmost categories are metabolism, genetic information processing, environmental information processing, cellular processes, and human diseases. Subcategories of metabolism are, e.g., carbohydrate, energy, lipid, nucleotide, or amino acid metabolism. These are subdivided into the different pathways, such as glycolysis, citrate cycle, purine metabolism, etc. Finally, the fourth level corresponds to the KO (KEGG Orthology) entries. A KO entry (internally identified by a K number, e.g., K00001 for the alcohol dehydrogenase) corresponds to a group of orthologous genes that have identical functions.

BRENDA

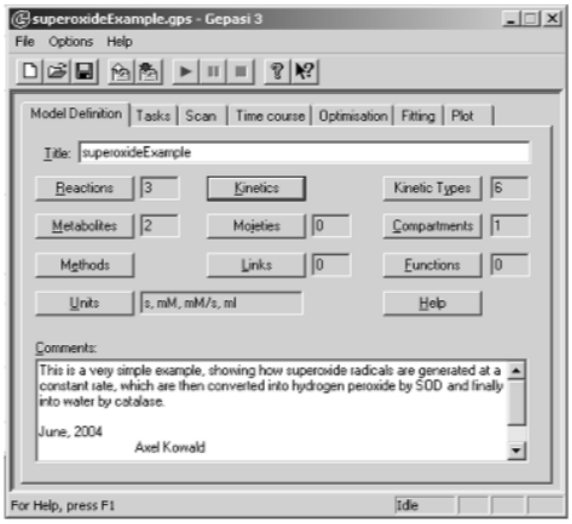
High-throughput projects, such as the international genome sequencing efforts, accumulate large amounts of data at an amazing rate. These data are essential for the reconstruction of phylogenetic trees and gene-finding projects. However, for kinetic modeling, which is at the heart of systems biology, kinetic data of proteins and enzymes are needed. Unfortunately, this type of data is notoriously difficult and time-consuming to obtain, since proteins often need individually tuned purification and reaction conditions. Furthermore, the results of such studies are published in a large variety of journals from different fields.

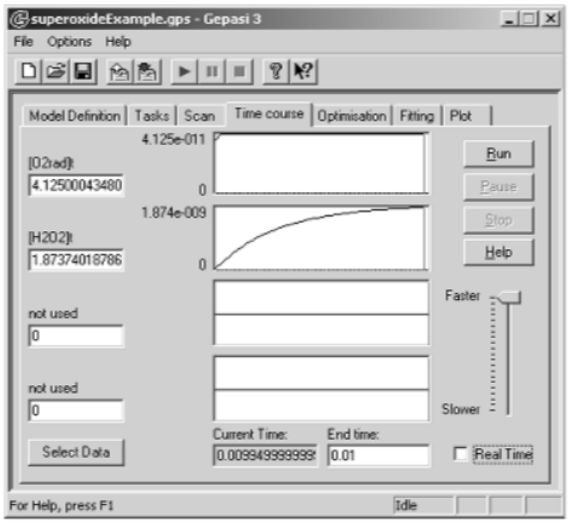
In this situation, BRENDA aims to be a comprehensive enzyme information system (http://www.brenda.uni-koeln.de). Basically, BRENDA is a curated database that contains a large amount of functional data for individual enzymes. These data are gathered from the literature and made available via a Web interface. The table on the next page gives an overview of the types of information that is collected and the number of entries for the different information fields (as of June 2004). For instance, enzymes representing 4379 different EC numbers and over 50,000 different K_m values are contained in the database.

5.3 Modeling and Visualization tools - Gepasi, Copasi

Matlab and Mathematica are huge and expensive general-purpose tools for mathematical modeling. They can be used to model anything that can be modeled, but at the cost of a steep learning curve. The opposite approach is used by specialized tools that are designed for a certain task. Gepasi is one of these tools that have been developed for the modeling of biochemical reaction systems. It was written by Pedro Mendes (Mendes 1993, 1997) and is available free of charge (http://www.gepasi.org). It runs native under Microsoft Windows but can also be used under Unix/Linux in connection with the Wine emulator (http://www.winehq.com).

In Gepasi, reactions are entered not as differential equations but rather in a notation similar to chemical reactions (Fig. 14.1). Each reaction has to be assigned to a specific kinetics, and Gepasi allows the user to select from a large range of predefined kinetics types (Michaelis-Menten, Hill Kinetics, Uni-Uni, etc.). In addition it is also possible to create user-defined kinetics types. Once a system is defined, the program allows one to perform several tasks such as plotting a time course, scanning the parameter space, fitting models to data, optimizing any function of the model, and performing metabolic control analysis and linear stability analysis.





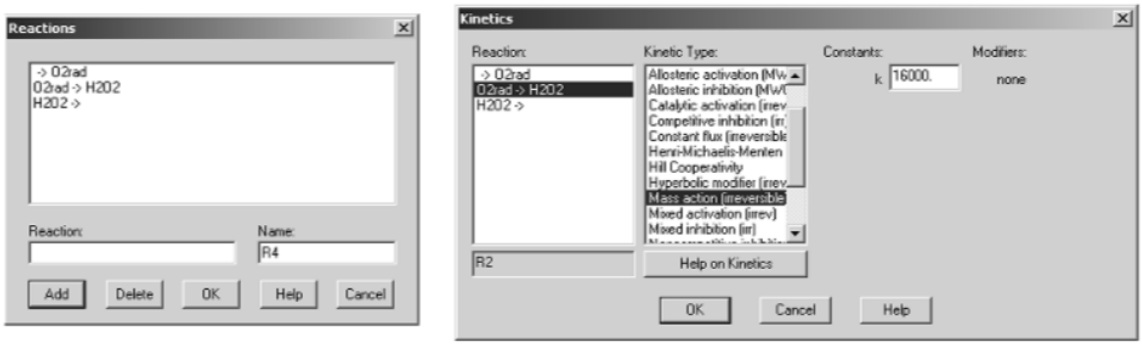


Fig. 14.1 The simulation tool Gepasi. Top left: The main window, which contains tabs for activities such as input of the reaction system, calculating a time course, fitting the system to experimental data or scanning the parameter space. Bottom left: Reactions are entered in a chemical notation, not as ODEs. Irreversible reactions are entered with the symbol -> and reversible reactions with an equal sign (=). Bottom right: A kinetics has to be assigned to each reaction and the necessary numerical constants have to be specified. Top right: If the system has been defined, Gepasi can calculate the time course of selected variables.

It is also possible to create multi-compartment models with Gepasi to model reactions that take place, for instance, in the cytoplasm and the nucleus. If a metabolite crosses a boundary between two compartments of different volume, the change of concentration in the originating compartment is not equal to that in the destination compartment. Gepasi automatically takes care of the conversions between concentrations into absolute amounts and back, which is necessary for the calculations. Apart from its own format, Gepasi can also save and load models that are described in the Systems Biology Markup Language (SBML) level 1 (see Section 14.2.2).

Gepasi is a handy tool that is designed to perform many of the standard tasks for studying a system of biochemical reactions. It is easy to handle, except that one has to get used to the strange fact that all windows are of a fixed size and rather small. Graphical simulation results, however, can also be redirected to a companion program, gnuplot, which does not have these restrictions.

5.4 MEGA

MEGA was first developed for MS DOS in the early 1990s (Kumar et al. 1994) and then upgraded for use in MS Windows eight times, including MEGA 1 to MEGA 6 and MEGA-CC and MEGA-MD (Kumar et al. 2001,, 2016). Some of the MEGA releases have been packaged for Linux systems using the WINE compatibility layer for POSIX-compliant operating systems and the Wineskin tool (built on WINE) for macOS systems. These versions have been downloaded over 200,000 times. But the ad hoc Windows-emulation solution is sluggish and relatively unstable when compared with the performance in MS Windows. Emulators cannot be used effectively for the latest 64-bit version of MEGA that is built to handle memory-intensive analyses of large

contemporary data sets (Kumar et al. 2016), so a more comprehensive solution is required for users of alternate platforms.

Therefore, MEGA has been transformed into a cross-platform version that runs natively on Linux and Microsoft Windows. This advancement eliminates the Windows-only limitation of MEGA, which has become particularly acute due to the increasing use of Linux in biological research. This transformation also paves the way for development of a MEGA X version for macOS in the near future.

5.5 Netpath

Complex biological processes such as proliferation, migration and apoptosis are generally regulated through responses of cells to stimuli in their environment. Signal transduction pathways often involve binding of extracellular ligands to receptors, which trigger a sequence of biochemical reactions inside the cell. Generally, proteins are the effector molecules, which function as part of larger protein complexes in signaling cascades. Cellular signaling events are generally studied systematically through individual experiments that are widely scattered in the biomedical literature. Assembling these individual experiments and putting them in the context of a signaling pathway is difficult, time-consuming and cannot be automated.

The availability of detailed signal transduction pathways that can easily be understood by humans as well as be processed by computers is of great value to biologists trying to understand the working of cells, tissues and organ systems. A systems-level understanding of any biological process requires, at the very least, a comprehensive map depicting the relationships among the various molecules involved. For instance, these maps could be used to construct a complete network of protein-protein interactions and transcriptional events, which would help in identifying novel transcriptional and other regulatory networks. These can be extended to predict how the interactions, if perturbed singly or in combination, could affect individual biological processes. Additionally, they could be used to identify possible unintended effects of a candidate therapeutic agent on any clusters in a pathway. We have developed a resource called NetPath that allows biomedical scientists to visualize, process and manipulate data pertaining to signaling pathways in humans.

5.6 Biotapestry

BioTapestry is an open source, freely available software tool that has been developed to handle the -challenges of modeling genetic regulatory networks (GRNs). Using BioTapestry, a researcher can -construct a network model and use it to visualize and understand the dynamic behavior of a complex, spatially and temporally distributed GRN. Here we provide a step-by-step example of a way to use BioTapestry to build a GRN model and discuss some common issues that can arise during this process.

5.7 E-Cell

The E-Cell Project develops general technologies and theoretical supports for computational biology with the grand aim to make precise whole cell simulation at the molecular level possible.

Some of the research foci of the Project include:

- Modeling methodologies, formalisms and techniques, including technologies to predict, obtain or estimate parameters such as reaction rates and concentrations of molecules in the cell.
- E-Cell System, a software platform for modeling, simulation and analysis of complex, heterogeneous and multi-scale systems like the cell.
- Numerical simulation algorithms.
- Mathematical analysis methods.

The E-Cell Project is open to anyone who shares the view with us that development of cell simulation technology, and, even if such ultimate goal might not be within ten years of reach yet, solving various conceptual, computational and experimental problems that will continue to arise in the course of pursuing it, may have a multitude of eminent scientific, medical and engineering impacts on our society.

5.8 PyBioS

Several software applications have been proposed in the past years as computational tools for assessing biomedical signals. Many of them are focused on heart rate variability series only, with their strengths and limitations depending on the necessity of the user and the scope of the application. Here, we introduce new software, named PyBioS, intended for the analysis of cardiovascular signals, even though any type of biomedical signal can be used. PyBioS has some functionalities that differentiate it from the other software. PyBioS was developed in Python language with an intuitive, user-friendly graphical user interface. The basic steps for using PyBioS comprise the opening or creation (simulation) of signals, their visualization, preprocessing and analysis. Currently, PyBioS has 8 preprocessing tools and 15 analysis methods, the later providing more than 50 metrics for analysis of the signals' dynamics. The possibility to create simulated signals and save the preprocessed signals is a strength of PyBioS. Besides, the software allows batch processing of files, making the analysis of a large amount of data easy and fast. Finally, PyBioS has plenty of analysis methods implemented, with the focus on nonlinear and complexity analysis of signals and time series. Although PyBioS is not intended to overcome all the necessities from users, it has useful functionalities that may be helpful in many situations. Moreover, PyBioS is continuously under improvement and several simulated signals, tools and analysis methods are still to be implemented. Also, a new module is being implemented on it to provide machine learning algorithms for classification and regression of data extracted from the biomedical signals.

5.9 Systems Biology Workbench

The **Systems Biology Workbench** (SBW) is a software **systems** that enables different modeling programs to communicate with each other and provide or use specialized analysis services. In this way SBW acts as broker for services like deterministic and stochastic simulation engines, stability and bifurcation analysis, model optimization and graphical model building. Popular tools that are SBW aware are among others JDesigner, CellDesigner and Dizzy.

5.10 Jdesigner

JDesigner is a graphical network editing tool developed by H. Sauro. It is tightly connected with Jarnac.

http://sys-bio.org/sbwWiki/sysbio/jdesigner

5.11 CellDesigner

Understanding the logic and dynamics of gene-regulatory and biochemical networks is a major challenge of systems biology. To facilitate this research topic, we have developed CellDesigner, a modeling tool of gene-regulatory and biochemical networks. CellDesigner supports users to easily create such networks, using solidly defined and comprehensive graphical representation (SBGN, systems biology graphical notation). CellDesigner is systems biology markup language (SBML) compliant, and has Systems Biology Workbench(SBW)-enabled software so that it can import/export SBML-described documents and integrate with other SBW-enabled simulation/analysis software packages. CellDesigner also supports simulation and parameter search, which is supported by integration with SBML ordinary differential equation (ODE) Solver, enabling us to simulate through our sophisticated graphical user interface. We can also browse and modify existing SBML models with references to existing databases. CellDesigner is implemented in Java; thus, it runs on various platforms such as Windows, Linux, and MacOS X. CellDesigner is freely available from our Web site at http://celldesigner.org/.

5.12 Petri Nets

Petri nets are an excellent formal model for studying concurrent and distributed systems and have been widely applied in many different areas of computer science and other disciplines (Murata, 1989). There have been over 8000 publications on Petri nets (refer to Website *http://www.daimi.au.dk/PetriNets/*). Since Carl Adam Petri originally developed Petri nets in 1962, Petri nets have evolved through four generations: the first-generation low-level Petri nets primarily used for modeling system control (Reisig, 1985a), the second-generation highlevel Petri nets for describing both system data and control (Jensen and Rozenberg, 1991), the third-generation hierarchical Petri nets for abstracting system structures (He and Lee, 1991; He, 1996; Jensen, 1992), and the fourth-generation object-oriented Petri nets for supporting modern system development approaches (Agha, 2001). Petri nets have also been extended in many different ways to study specific system properties, such as performance, reliability,

and schedulability. Well-known examples of extended Petri nets include timed Petri nets (Wang, 1998) and stochastic Petri nets (Marsan *et al.*, 1994; Haas, 2002). In this article, we present several extensions to Petri nets based on our own research work and provide analysis techniques for these extended Petri net models. We also discuss the intended applications of these extended Petri nets and their potential benefits.

5.13 Model Exchange Languages and Data Formats - Introduction to XML

The easiest way to store and exchange data for the computer is a plain text that is readable by humans. Since data represented by such files are compatible with almost all computational operating systems, plain text files are also widely used in biological

research, e.g., for the storage of sequence information and its annotations. The type of information (e.g., sequence identifier, origin, preparation method, and the sequence data itself) is indicated by special tags and/or is defined in a separate description. A similar but more flexible tool for the storage of data in a well-defined way is the Extensible Markup Language (XML). XML is recommended by the World Wide Web Consortium (W3C) for the definition of special-purpose markup languages (http://www.w3.org/TR/2004/REC-xml-20040204/). XML is a lightweight adaptation of the even more general Standard Generalized Markup Language (SGML). Documents using an XML conform markup language are written as plain text and have a very clear and simple syntax that can easily be read by both humans and computer programs; however, it is generally intended to be written and read by computers, not by humans. The following example of some cellular components illustrates XML's major characteristics:

5.14 Systems Biology Markup Language

Many different tools for modeling and simulation of biological systems have already been developed (cf. Section 14.1). All of them offer functionalities to enter the model data and to make the model persistent by storing it, e.g., in an application-specific file. Since all of these tools offer different strength and capabilities (e.g., one offers a good graphical representation of models and the other provides very accurate methods for numerical simulations), a systems biologist is often interested in using several of these tools. But this typically requires the re-encoding of a model in a new tool, which is usually a time-consuming and error-prone process. Therefore, software-independent common standards for the representation of qualitative and quantitative models of biochemical reaction networks are required. CellML (Lloyd et al. 2004, http://www.cellml.org) and SBML (Hucka et al. 2003, 2004) are two XML-based formats facing up to this problem. Since SBML is the most prominent, we will describe it in more detail below.

SBML (http://www.sbml.org) is a free and open format for "describing models common to research in many areas of computational biology, including cell signaling pathways, metabolic pathways, gene regulation, and others" (Hucka et al. 2003). It is already supported by many software tools (Hucka et al. 2004); in September 2004 the SBML homepage listed more than 60 software systems supporting SBML.

The following SBML Level 2 code (differences to Level 1 will be discussed below) shows the general structural elements of an SBML document:

Example

```
<?xml version="1.0"?>
<sbml xmlns="http://www.sbml.org/sbml/level2" level="2"
      version="1">
  <model id="My_model">
    <listOfFunctionDefinitions>
      ...
    </listOfFunctionDefinitions>
    <listOfUnitDefinitions>
      ...
    </listOfUnitDefinitions>
    <listOfCompartments>
      ...
    </listOfCompartments>
    <listOfSpecies>
      ...
    </listOfSpecies>
    <listOfParameters>
      ...
    </listOfParameters>
    <listOfRules>
      ...
    </listOfRules>
    <listOfReactions>
```

5.15 MathML

SBML is designed to describe models in systems biology but is not intended to represent complicated mathematical expressions. MathML is an XML-based markup language especially created for this task (http://www.w3.org/Math). At places in SBML that require a mathematical expression, e. g., a user-defined kinetic law, MathML can be inserted. MathML comes in two flavors, as markup language for presenting the layout of mathematical expressions and as markup language for conveying the mathematical content of a formula. The major use of the presentation markup is to enable Internet browsers to directly display equations, something that is not possible with normal HTML tags. However, it is the content markup variant of MathML that is of greater interest for modeling. It can be used to exchange mathematical expressions in a common low-level format between software packages that need to evaluate these expressions (instead of displaying them). The following table contains MathML for the Michaelis-Menten expression $\frac{E \cdot S}{Km+S}$.

MathML is a very verbose format and is not intended to be generated or edited by hand. Specialized authoring tools should be used to import or export MathML expressions. Many different programs are available to make Web browsers MathML aware, to generate PDF or DVI from MathML, to save equations in MS Word in MathML format, or to create mathematical expressions interactively and save them in both types of MathML (http://www.w3.org/Math/implementations.html). One reason to look closer at the content markup MathML is STOCKS2 (see Section 14.1.7), which needs MathML input if the user wants to define a new kinetic type. The required MathML format could be generated with a commercial editor like WebEQ from Design Science (http://www.dessci.com) or by using a free service offered at http://www.mathmlcentral.com, a Web site of Wolfram Research (the company that produces Mathematica). This site offers three valuable Web services for free: validating whether a given MathML expression is syntactically correct, rendering of presentation markup MathML into different graphics formats, and conversion of a mathematical expression that is given in Mathematica syntax into the different types of MathML. With this wealth of resources available on the Net, it should be no problem to master MathML.

5.16 Cytoscape

Cytoscape is an open source software project for integrating biomolecular interaction networks with high-throughput expression data and other molecular states into a unified conceptual framework. Although applicable to any system of molecular components

and interactions, Cytoscape is most powerful when used in conjunction with large databases of protein-protein, protein-DNA, and genetic interactions that are increasingly available for humans and model organisms. Cytoscape's software Core provides basic functionality to layout and query the network; to visually integrate the network with expression profiles, phenotypes, and other molecular states; and to link the network to databases of functional annotations. The Core is extensible through a straightforward plug-in architecture, allowing rapid development of additional computational analyses and features. Several case studies of Cytoscape plug-ins are surveyed, including a search for interaction pathways correlating with changes in gene expression, a study of protein complexes involved in cellular recovery to DNA damage, inference of a combined physical/functional interaction network for *Halobacterium*, and an interface to detailed stochastic/kinetic gene regulatory models.

5.17 SBML tool box for MATLAB.

The expanding field of Systems Biology has stimulated the formalization of an increasing number of biological/biochemical models. The Systems Biology Markup Language (SBML), an XML-based format for computational models of biochemical networks, is becoming accepted as a *de facto* standard for the representation of such models (Hucka *et al*., 2004) and thus facilitates their systematic exchange.

In addition to promoting the creation of models, Systems Biology has also motivated the development of a range of software packages that can interact with these models, perform simulations and analyses on them, produce graphical representations of models and facilitate the creation of new models. However, the commercially available software package MATLAB provides a wide spectrum of this type of functionality combined with the facility to easily develop user-specific functions. Thus an alternative approach to that of developing new software exists in the form of developing a toolbox that provides users with an interface between basic MATLAB data structures and a format such as SBML. This not only enables users to leverage their existing skills in using the environment to work with a new format such as SBML, but it also provides a substrate enabling other analysis tools in the environment to be applied to data represented in SBML. MATLAB is a particularly attractive environment in this regard because it is already popular worldwide in both engineering and scientific research, and as the field of Systems Biology continues to attract researchers with an engineering or physical science background, the use of MATLAB within the field is likely to proliferate. Also, there currently exist many tools, both commercial and freely available (Prajna *et al*., 2004, that apply the computational and analytical capabilities of MATLAB to models and data in a variety of formats.

SBMLToolbox was initially developed specifically to meet two separate needs: (1) those of existing MATLAB users wishing to import SBML models and apply functionality appropriate to their goals, whether built into the environment or purpose-written and (2) those of users experienced with SBML wishing to apply the computational power of MATLAB to their models. Thus, in addition to importing SBML, the toolbox includes functionality serving as an example of using MATLAB in

the manipulation and analysis of models. However, the actual analytical functionality is limited and while it is possible to simulate a range of models with the toolbox, it should not be considered a simulation tool but rather a facilitator for the development of other functions and toolboxes. To date we are aware of at least two freely available toolboxes that use SBMLToolbox for precisely this purpose; namely SBToolbox and SBMLSim.