Part I
General Introduction

Systems Biology in Practice. Concepts, Implementation and Application.
E. Klipp, R. Herwig, A. Kowald, C. Wierling, H. Lehrach

ISBN: 3-527-31078-9

1
Basic Principles

1.1
Systems Biology is Biology!

Life is one of the most complex phenomena in the universe. It has been studied by using systematic approaches in botany, zoology, and ecology as well as by investigating the composition and molecular biology of single cells. For a long time biologists have thoroughly investigated how parts of the cell work: they have studied the biochemistry of small and large molecules, the structure of proteins, the structure of DNA and RNA, and the principles of DNA replication as well as transcription and translation and the structure and function of membranes. In addition, theoretical concepts about the interaction of elements in different types of networks have been developed. The next step in this line of research is further effort towards a systematic investigation of cells, organs, and organisms and of (mainly) cellular processes such as cellular communication, cell division, homeostasis, and adaptation. This approach has been termed systems biology.

Now the time has come to integrate different fields of biology and natural science in order to better understand how cells work, how cellular processes are regulated, and how cells react to environmental perturbations or even anticipate those changes. The development of a more systematic view of biological processes is accompanied by and based on a revolution of experimental techniques and methodologies. New high-throughput methods allow measurement of the expression levels of all genes of a cell at the same time and with reasonable temporal resolution, although this is still very expensive. Fluorescence labeling and sophisticated microscopic techniques allow tracing individual molecules within a single cell. A fine-grained study of cell components and cell processes in time and in space is an important prerequisite for the further elucidation of cellular regulation.

Systems biology is driven partly by the curiosity of scientists, but even more so by the high potential of its applications. Biotechnological production requires tools with high predictive power to design cells with desired properties cheaply and reliably. There are many promises for health care: models of regulatory networks are necessary to understand their alterations in the case of disease and to develop methods to cure the disease. Furthermore, since there is an observable trend in health care towards individualized and predictive medicine (Weston and Hood 2004), there will be

Systems Biology in Practice. Concepts, Implementation and Application.
E. Klipp, R. Herwig, A. Kowald, C. Wierling, H. Lehrach

ISBN: 3-527-31078-9

an increasing need for the exact formulation of cellular networks and the prediction of systems behavior in the areas of drug development, drug validation, diagnostics, and therapy monitoring. For example, it has been shown that the epidermal growth factor receptor, which is targeted by a new generation of cancer drugs, belongs to a family of at least four related receptors. These receptors can be turned on by more than 30 different molecules. Thus, such a complex setup makes it necessary to derive the wiring diagram to understand how each component plays its role in responding to various stimuli and causing disease. Once a detailed model has been constructed, all effects of possible perturbations can be predicted fairly cheaply *in silico*. Furthermore, models gained by systems biology approaches can be used for prediction of the behavior of the biological system even under conditions that are not easily accessible with experiments.

Systems biology approaches offer the chance to predict the outcome of complex processes, e. g., the effect of different possible courses of cancer treatment on the tumor (how effectively the treatment eliminates the tumor as well as possible metastatic cells) and the patient (what the cancer treatment does to other rapidly growing tissues, how bad the predicted side effects of a specific treatment in a specific patient are).

These and many other problems that could have enormous effects on our survival, our health, our food supplies, and many other issues that are essential to our existence and our well being might very well be almost impossible to approach without the tools of systems biology that are currently being developed. E. g., to optimize the treatment of an individual cancer patient, we have to be able to accurately predict the outcome of the possible courses of treatment. This would be easy if we were able to understand the complex processes (drug effects, drug side effects, drug metabolism, etc.) the way that we understand some processes in physics (e. g., the famous equation $E = mc^2$ describing the dependence of mass and energy) or even some of the basic processes in biology (the genetic code). This is very unlikely for the complex, highly connected systems we are faced with in many real-world problems in biology. It is not even clear whether our current approach of studying such systems – analyzing small segments (often one or a few genes at a time) – will ever give us enough insight to be able to make useful prediction, as, at least in mathematics, many systems cannot be subdivided in that form. The only option we have might therefore very well be to generate as much information as possible on the system, using the tools of functional genomics, and to model the entire process in as much detail as necessary to allow quantitative predictions of the parameters we are interested in.

Systems biology relies on the integration of experimentation, data processing, and modeling. Ideally, this is an iterative process. Experimentally obtained knowledge about the system under study together with open questions lead to an initial model. The initial model allows predictions that can be verified or falsified in new experiments. Disagreements stimulate the next step of model development, which again results in experimentally testable predictions. This iteration continues until a good agreement is achieved between the data obtained in the experiment and the model predictions.

A major topic of current systems biology is the analysis of networks: gene networks, protein interaction networks, metabolic networks, signaling networks, etc. Initially, investigation of abstract networks was fashionable. However, it has become

clear that it is necessary to study more realistic and detailed networks in order to uncover the peculiarities of biological regulation. Different theoretical attempts have been made to study the different types of networks. For example, gene regulatory networks are sometimes described by Boolean logic assigning to genes one of two states, on or off; protein relations are mainly characterized by a static view of putative interactions measured by yeast two-hybrid methods, and metabolic networks are determined by the set of catalyzing enzymes and the possible metabolic fluxes and intrinsic modes of regulation.

A unified view of a cellular network is currently emerging in the sense that each action of a cell involves different levels of cellular organization, including genes, proteins, metabolism, or signaling pathways. Therefore, the current description of the individual networks must be integrated into a larger framework.

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

Computational models serve as repositories of the current knowledge, both established and hypothetical, on how pathways might operate, providing one with quantitative codification of this knowledge and with the ability to simulate the biological processes according to this codification (Levchenko 2003). The attempt to formulate current knowledge and open problems in mathematical terms often uncovers a lack of knowledge and requirements for clarification. On the other hand, computational models can be used to test whether different hypotheses about the true process are reliable.

Many current approaches pay tribute to the fact that biological items are subject to evolution. This concerns on one hand the similarity of biological organisms from different species. This similarity allows for the use of model organisms and for the critical transfer of insights gained from one cell type to other cell types. Applications include, e.g., prediction of protein function from similarity, prediction of network properties from optimality principles, reconstruction of phylogenetic trees, or identification of regulatory DNA sequences through cross-species comparisons. On the other hand, the evolutionary process leads to genetic variations within species. Therefore, personalized medicine and research is an important new challenge for biomedical research.

1.2
Systems Biology is Modeling

Observation of the real world and, especially, of biological processes confronts us with many simple and complex processes that cannot be explained with elementary

principles and the outcome of which cannot reliably be foreseen from experience. Mathematical modeling and computer simulations can help us to understand the internal nature and dynamics of these processes and to arrive at well-founded predictions about their future development and the effect of interactions with the environment.

What is a model? The answer will differ among communities of researchers. In the broadest sense, a model is an abstract representation of objects or processes that explains features of these objects or processes. For instance, the strings composed of the letters A, C, G, and T are used as a model for DNA sequences. In some cases a cartoon of a reaction network showing dots for metabolites and arrows for reactions is a model, while in other cases a system of differential equations is employed to describe the dynamics of that network. In experimental biology, the term model is also used to denote species that are especially suitable for experiments. For example the mouse Ts65DN serves as a model for human trisomy 21 (Reeves et al. 1995).

1.2.1 Properties of Models

1.2.1.1 Model Assignment is not Unique

Biological phenomena can be described in mathematical terms. Many examples have been presented during the past few decades (from the description of glycolytic oscillations with ordinary differential equations, to populations growth with difference equations, to stochastic equations for signaling pathways, to Boolean networks for gene expression). It is important to note that a certain process can be described in more than one way.

- A biological object can be investigated with different experimental methods.
- Each biological process can be described with different (mathematical) models.
- A mathematical formalism may be applied to different biological instances.
- The choice of a mathematical model or an algorithm to describe a biological object depends on the problem, the purpose, and the intention of the investigator.
- Modeling has to reflect essential properties of the system. Different models may highlight different aspects of the same instance.

This ambiguity has the advantage that different ways of studying a problem also provide different insights into the system. An important disadvantage is that the diversity of modeling approaches makes it very difficult to merge established models (e. g., for individual metabolic pathways) into larger super-models (e. g., for the complete cellular metabolism).

1.2.1.2 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 predict the behavior of the system for all future times. The state of the system is described 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 differential 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.

1.2.1.3 Steady States

The concept of stationary states is important for the modeling of dynamical systems. *Stationary states* (other terms are *steady states* or *fixed points*) are determined by the fact that the values of all state variables remain constant in time. The asymptotic behavior of dynamic systems, 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 separation 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.

1.2.1.4 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 $N_A = 6.02 \cdot 10^{23}$ (number of molecules per mole). *Parameters* are quantities that are assigned a value, such as the K_m value of an enzyme in a reaction. This value depends on the method used and on the experimental conditions and may change. *Variables* are quantities with a changeable value for which the model establishes relations. 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 necessary 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 $V = \pi d^3/6$. π 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 enzyme concentration is frequently considered a parameter in biochemical reaction kinetics. That is no longer valid if, in a larger model, the enzyme concentration may change due to gene expression or protein degradation.

1.2.1.5 Model Behavior

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 (output). 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.

1.2.1.6 Process Classification

For modeling, processes are classified with respect to a set of criteria. *Reversibility* determines whether a process can proceed in a forward and backward direction. Irreversible means that only one direction is possible. *Periodicity* indicates that a series of states may be assumed in the time interval $\{t, t + \Delta t)$ and again in the time interval $\{t + i \cdot \Delta t, t + (i + 1) \cdot \Delta t\}$ for $i = 1, 2, \ldots$ With respect to the randomness of the predictions, deterministic modeling is distinct from stochastic modeling. A description is *deterministic* if the motion through all following states can be predicted from the knowledge of the current state. *Stochastic* description gives instead a probability distribution for the succeeding states. The nature of values that time, state, or space may assume distinguishes a *discrete* model (where values are taken from a discrete set) from a *continuous* model (where values belong to a continuum).

1.2.1.7 Purpose and Adequateness of Models

Models represent only specific aspects of the reality. The intention of modeling is to answer particular questions. Modeling is, therefore, a subjective and selective procedure. It may, for example, aim at predicting the system output. In this case it might be sufficient to obtain precise input-output relation, while the system internals can be regarded as black box. However, if the function of an object is to be elucidated, then its structure and the relations between its parts must be described realistically. One may intend to formulate a model that is generally applicable to many similar objects (e.g., Michaelis-Menten kinetics holds for many enzymes, the promoter-operator concept is applicable to many genes, and gene regulatory motifs are common) or that is specific to one special object (e.g., the 3D structure of a protein, the sequence of a gene, or a model of deteriorating mitochondria during aging). The mathematical part can be kept as simple as possible to allow for easy implementation and comprehensible results. Or it can be modeled very realistically and be much more complicated. None of the characteristics mentioned above makes a model wrong or right, but they determine whether a model is appropriate to the problem to be solved.

1.2.1.8 Advantages of Computational Modeling

Models gain their reference to reality from comparison with experiments, and their benefits are, therefore, somewhat dependent on experimental performance. Nevertheless, 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 knowledge or understanding. During the process of model formulation, unspecified components or interactions have to be determined.

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

1.2.1.9 **Model Development**

For the process of model development, we suggest the following modeling workflow:

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 results with respect to phenotypic changes such as growth and shape after system perturbations such as knockout experiments, RNAi, and variation of environmental conditions.
3. Selection of model structure: Based on the available information and on the problem to solve, the general type of the model is determined: (1) the level of description 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, schematically, 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 recommendable 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 design in biology is usually hypothesis-driven. In fact, hypotheses that state general relations can rarely be verified, but only falsified. These predictions usually concern 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 performed, and at what time intervals is not uniquely defined but depends on the researcher'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 results: 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 simplifications, wrong model structure, inadequate experimental design, or other inadequately 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.

1.2.2 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.

1.2.2.1 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.

1.2.2.2 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.

1.2.2.3 Robustness and Sensitivity are Two Sides of the Same Coin

Robustness is an essential feature of biological systems. It characterizes the insensitivity of system properties to variations in parameters, structure, and environment or to other uncertainties. Robust systems maintain their state and functions despite external and internal perturbations. An earlier notion for this observation is homeostasis. Robustness in biological systems is often achieved by a high degree of complexity 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 depending 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 variable by the change of a quantity that caused the change in the variable.

1.3 Systems Biology is Data Integration

The information that we can gain about a biological system appears in practice as an experimental observation, and systems biology research is restricted to the granularity and the precision of the experimental techniques in use. Systems biology has evolved rapidly in the last few years, driven by the new high-throughput technologies. The most important impulse was given by the large sequencing projects such as the human genome project, which resulted in the full sequence of the human and other genomes (Lander et al. 2001; Venter et al. 2001). This knowledge builds the theoretical basis to compute gene regulatory motifs, to determine the exon-intron structure of genes, and to derive the coding sequence of potentially all genes of many organisms. From the exact sequences probes for whole-genome DNA arrays have been constructed that allow us to monitor the transcriptome level of most genes active in a given cell or tissue type. Proteomics technologies have been used to iden-

tify translation status on a large scale (2D-gels, mass spectrometry). Protein-protein interaction data involving thousands of components were measured to determine information on the proteome level (von Mering et al. 2002). Data generated by these techniques are the basis for system-wide investigations. However, to validate such data in the system-wide hierarchical context ranging from DNA to RNA to protein to interaction networks and further on to cells, organs, individuals, etc., one needs to correlate and integrate such information. Thus, an important part of systems biology is data integration.

Data integration itself cannot explain the dynamical behavior of the biological system and is not a replacement for a mathematical model. However, it is extremely useful for increasing the information content of the individual experimental observation, enhancing the quality of the data, and identifying relevant components in the model for the biological system. Both the generation and the analysis of genome, transcriptome, and proteome data are becoming increasingly widespread and need to be merged for the generation of biological models.

At the lowest level of complexity, data integration defines common schemas for data storage, data representation, and data transfer. For particular experimental techniques, this has already been established, e.g., in the field of transcriptomics with MIAME (minimum information about a microarray experiment) (Brazma et al. 2001), in proteomics with PEDRo (Proteomics Experiment Data Repository) (Taylor et al. 2003), and the HUPO (The Human Proteome Organization) consortium (Hermjakob et al. 2004). On a more complex level, schemas have been defined for biological models and pathways such as SBML (Hucka et al. 2003) and CellML (Lloyd et al. 2004). Most of these repositories use an XML-like language style.

On a second level of complexity, data integration deals with query-based information retrieval, the connection of different data types (typically stored in different databases), and the visualization and presentation of the data. Here, for example, commercial applications such as SRS (Etzold et al. 1996) are in use. SRS provides a user interface that enables access to hundreds of biological databases. The EnsMart system developed at EBI is an advanced tool for data retrieval from database networks using a powerful query system (Kasprzyk et al. 2004). Both systems allow a simple integration of additional resources and programs so that they are continuously growing.

Data integration at the next level of complexity consists of data correlation. This is a growing research field as researchers combine information from multiple diverse datasets to learn about and explain natural processes (Ideker et al. 2001; Gitton et al. 2002). For example, methods have been developed to integrate insights from transcriptome or proteome experiments with genome sequence annotations. The integration of data enables their explanation and analysis, e.g., the comparison of gene expression patterns for orthologous genes or their evaluation in light of conserved transcription factor binding sites in upstream regions of the corresponding gene sequences (Tavazoie et al. 1999). At this level of complexity, researchers typically face the fact that data from diverse experimental platforms are correlated on a much lower level than assumed. This is partially due to the fact that experimental data generation typically involves a large pipeline of experimental stages with numerous fac-

tors of influence that might affect the output. Normalization strategies are therefore indispensable for interpretation of the data. This step requires highly sophisticated analysis tools, data mining models, and algorithms. Data mining defines the process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data stored in repositories and by using pattern recognition technologies as well as statistical and mathematical techniques. Taking together, there is no doubt that data handling, storage, integration, and analysis methods and rules must be enforced in order to interpret the experimental outcomes and to transfer the experimental information into functional knowledge. Furthermore, in the case of complex disease conditions, it is clear that an integrated approach is required in order to link clinical, genetic, behavioral, and environmental data with diverse types of molecular phenotype information and to identify correlative associations. Such correlations, if found, are the key to identifying biomarkers and processes that are either causative or indicative of the disease. Importantly, the identification of biomarkers (e. g., proteins, metabolites) associated with the disease will open up the possibility to generate and test hypotheses on the biological processes and genes involved in this condition. The evaluation of disease-relevant data is a multi-step procedure involving a complex pipeline of analysis and data-handling tools such as data normalization, quality control, multivariate statistics, correlation analysis, visualization techniques, and intelligent database systems (Kanehisa and Bork 2003). Recently, several pioneering approaches have indicated the power of integrating datasets from different levels, e. g., the correlation of gene membership of expression clusters and promoter sequence motifs (Tavazoie et al. 1999); the combination of transcriptome and quantitative proteomics data in order to construct models of cellular pathways (Ideker et al. 2001); and the identification of novel metabolite-transcript correlations (Urbanczyk-Wochniak et al. 2003).

The highest level of data integration is the mapping of the integrated experimental data from multiple sources into networks in order to model interactions of the biological objects of the system. These networks represent qualitative models for the biological system. For example, Ideker et al. (2001) studied the galactose utilization pathway in yeast. The authors employed several strains of yeast, each with a different galactose gene knocked out, and a wild type and monitored changes in the levels of yeast genes using DNA arrays with the system in the presence and absence of galactose. Together with known data such as protein-protein interactions and protein-DNA interactions, they were able to construct an entire physical interaction network of that pathway. Davidson and colleagues (2002) studied endomesoderm specification in sea urchin and constructed a large gene regulatory network model comprising 60 genes. Most of the network architecture is based on perturbation experiments and expression data. Several conclusions can be drawn from these and other studies (Lee et al. 2002; Shen-Orr et al. 2002). There appears to be a variety of small modules similar to those found in engineering (feed-forward loops, single-input motifs). Such motifs can be found through different organisms (sea urchin, yeast, *E. coli*). Thus, current research tries to classify motifs into a kind of lexicon for higher-order functioning. By topological analysis, genes can be identified in these networks that may change fundamental properties of the system (hubs, articulation points, etc.)

and give rise to suggestions for further perturbation experiments. Thus, these qualitative models provide fundamental new strategies for systems biology research.

It should be pointed out that the current state of data integration is well elaborated at the lower levels of complexity, in particular with the database networks, whereas the higher stages need far more development. This is due to the fact that system-wide approaches are rare at the current state. These would require a guided and planned set of interacting experimental techniques on a defined experimental model, which is hard to realize. Instead, many data available for computational research are generated under varying experimental conditions with different experimental platforms and without any serious attempt at standardization. For example, it is a well-known fact that DNA array data from different platforms correlate at a very low level (Kuo et al. 2002; Tan et al. 2003), and the same phenomenon is observable with protein-protein interactions (Ito et al. 2000; Uetz et al. 2000). The lack of standardization remains the most important limiting factor of data integration and has to be tackled by future system-wide approaches.

1.4 Systems Biology is a Living Science

Systems biology comprises experimentation and computational modeling. To this end, it integrates approaches from diverse areas of science such as biology, chemistry, physics, mathematics, applied science, engineering, cybernetics, and computer science. By demanding new strategies, it also stimulates their further development and contributes to new solutions.

The integrative and interdisciplinary nature of systems biology necessitates the exchange of information among scientists from different fields. This means, for example, that mathematical formulas have to be made understandable for biologists and that people acquainted with the exact world of computers have to understand the diversity of biological objects and the uncertainty in the outcome of experiments. In the long term, these problems may be solved by education. In the short term, they require presentation of results from different perspectives and at different levels of accuracy.

Information exchange necessitates a *common language* about biological aspects. One seminal example is the gene ontology (GO, see Chapter 13, Section 13.1), which provides a controlled vocabulary that can be applied to all organisms, even as knowledge about gene and protein roles in cells is accumulating and changing. Another example is the Systems Biology Markup Language (SBML, see Chapter 14, Section 14.2.2) as an exchange language for models of biochemical reaction networks.

In addition to statements in mathematical terms or detailed verbal explanations, information and knowledge exchange demand *visualization* of concepts, perceptions, and insights, since it enhances understanding. Important fields for visualization are (1) the spatial organization of cell components and of cellular processes, (2) the representation of complex dynamics, and (3) interactions and regulatory patterns in

networks. A traditional, well-known example is the Boehringer chart (Michal 1999), which can be found in the majority of biological labs.

Modeling of biological processes drives the development of concepts. The necessity for specific and mathematically exact formulation has stimulated the development of common model exchange languages (Chapter 14, Section 14.2), metabolic control theory (Chapter 5, Section 5.3), and clustering algorithms (Chapter 9, Section 9.3).

Standardization of experimental conditions and model approaches seems to restrict freedom of research and is hard to achieve. But standardization is essential for comparability of results, for the integration of the efforts of several labs, and for fast exchange of information between theoretical and experimental groups. Promising examples include MIAME and SBML.

The new paradigm of integrated and concerted efforts also demands open access to information. This is given on one hand by the exchange of data via Internet databases and by the exchange of modeling facilities as in SBW (Systems Biology Workbench). On the other hand, published results must be quickly available for the community.

Systems biology might also be the key to publication in biology in the future. Instead of, or in addition to, extensive descriptions of a biological system as text, we might "publish" our view of the biological object we are describing in the form of a working "computer object", which can be "published" over the Internet. This can then be tested by other scientists, in combination with other "computer objects", to see whether the object correctly predicts all aspects of the system, which can be observed experimentally. In many cases, complete agreement of predictions and all experimentally observable parameters of a system might be as close to the "truth" about a complex process in biology as we will be able to get.

References

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F.C., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J. and Vingron, M. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data (2001) Nat. Genet. *29*, 365–71.

Davidson, E.H., Rast, J.P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C.H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., Otim, O., Brown, C.T., Livi, C.B., Lee, P.Y., Revilla, R., Rust, A.G., Pan, Z., Schilstra, M.J., Clarke, P.J., Arnone, M.I., Rowen, L., Cameron, R.A., McClay, D.R., Hood, L. and Bolouri, H. A genomic regulatory network for development (2002) Science *295*, 1669–78.

Etzold, T., Ulyanov, A. and Argos, P. SRS: information retrieval system for molecular biology data banks (1996) Methods Enzymol. *266*, 114–28.

Gitton, Y., Dahmane, N., Baik, S., Ruiz i Altaba, A., Neidhardt, L., Scholze, M., Herrmann, B.G., Kahlem, P., Benkahla, A., Schrinner, S., Yildirimman, R., Herwig, R., Lehrach, H. and Yaspo, M.L. A gene expression map of human chromosome 21 orthologues in the mouse (2002) Nature *420*, 586–90.

Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., Roechert, B., Poux, S., Jung, E., Mersch, H., Kersey, P., Lappe, M., Li, Y., Zeng, R., Rana, D., Nikolski, M., Husi, H.,

Brun, C., Shanker, K., Grant, S.G., Sander, C., Bork, P., Zhu, W., Pandey, A., Brazma, A., Jacq, B., Vidal, M., Sherman, D., Legrain, P., Cesareni, G., Xenarios, I., Eisenberg, D., Steipe, B., Hogue, C. and Apweiler, R. The HUPO PSIs molecular interaction format–a community standard for the representation of protein interaction data (2004) Nat. Biotechnol. *22*, 177–83.

Hofmeyr, J.H. and Westerhoff, H.V. Building the cellular puzzle: control in multi-level reaction networks (2001) J. Theor. Biol. *208*, 261–85.

Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., Arkin, A.P., Bornstein, B.J., Bray, D., Cornish-Bowden, A., Cuellar, A.A., Dronov, S., Gilles, E.D., Ginkel, M., Gor, V., Goryanin, II, Hedley, W.J., Hodgman, T.C., Hofmeyr, J.H., Hunter, P.J., Juty, N.S., Kasberger, J.L., Kremling, A., Kummer, U., Le Novere, N., Loew, L.M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E.D., Nakayama, Y., Nelson, M.R., Nielsen, P.F., Sakurada, T., Schaff, J.C., Shapiro, B.E., Shimizu, T.S., Spence, H.D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J. and Wang, J. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models (2003) Bioinformatics *19*, 524–31.

Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Bumgarner, R., Goodlett, D.R., Aebersold, R. and Hood, L. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network (2001) Science *292*, 929–34.

Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S. and Sakaki, Y. Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins (2000) Proc. Natl. Acad. Sci. USA *97*, 1143–7.

Kanehisa, M. and Bork, P. Bioinformatics in the post-sequence era (2003) Nat Genet *33 Suppl*, 305–10.

Kasprzyk, A., Keefe, D., Smedley, D., London, D., Spooner, W., Melsopp, C., Hammond, M., Rocca-Serra, P., Cox, T. and Birney, E. EnsMart: a generic system for fast and flexible access to biological data (2004) Genome Res. *14*, 160–9.

Kitano, H. Computational systems biology (2002) Nature *420*, 206–10.

Kuo, W.P., Jenssen, T.K., Butte, A.J., Ohno-Machado, L. and Kohane, I.S. Analysis of matched mRNA measurements from two different microarray technologies (2002) Bioinformatics *18*, 405–12.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al. Initial sequencing and analysis of the human genome (2001) Nature *409*, 860–921.

Lauffenburger, D.A. Cell signaling pathways as control modules: complexity for simplicity? (2000) Proc. Natl. Acad. Sci. USA *97*, 5031–3.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.B., Volkert, T.L., Fraenkel, E., Gifford, D.K. and Young, R.A. Transcriptional regulatory networks in *Saccharomyces cerevisiae* (2002) Science *298*, 799–804.

Levchenko, A. Dynamical and integrative cell signaling: challenges for the new biology (2003) Biotechnol. Bioeng. *84*, 773–82.

Lloyd, C.M., Halstead, M.D. and Nielsen, P.F. CellML: its future, present and past (2004) Prog. Biophys. Mol. Biol. *85*, 433–50.

Michal, G. Biochemical pathways (1999) Spektrum Akademischer Verlag, Heidelberg.

Reeves, R.H., Irving, N.G., Moran, T.H., Wohn, A., Kitt, C., Sisodia, S.S., Schmidt, C., Bronson, R.T. and Davisson, M.T. A mouse model for Down syndrome exhibits learning and behaviour deficits (1995) Nat. Genet. *11*, 177–84.

Shen-Orr, S.S., Milo, R., Mangan, S. and Alon, U. Network motifs in the transcriptional regulation network of *Escherichia coli* (2002) Nat. Genet. *31*, 64–8.

Tan, P.K., Downey, T.J., Spitznagel, E.L., Jr., Xu, P., Fu, D., Dimitrov, D.S., Lempicki, R.A., Raaka, B.M. and Cam, M.C. Evaluation of gene expression measurements from commercial microarray platforms (2003) Nucleic Acids Res. *31*, 5676–84.

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. and Church, G.M. Systematic determination of genetic network architecture (1999) Nat. Genet. *22*, 281–5.

Taylor, C.F., Paton, N.W., Garwood, K.L., Kirby, P.D., Stead, D.A., Yin, Z., Deutsch, E.W., Selway, L., Walker, J., Riba-Garcia, I., Mohammed, S., Deery, M.J., Howard, J.A., Dunkley, T., Aebersold, R., Kell, D.B., Lilley, K.S., Roepstorff, P., Yates, J.R., 3rd, Brass, A., Brown, A.J., Cash, P., Gaskell, S.J., Hubbard, S.J. and Oliver, S.G. A systematic approach to modeling, capturing, and disseminating proteomics experimental data (2003) Nat. Biotechnol. *21*, 247–54.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. and Rothberg, J.M. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae* (2000) Nature *403*, 623–7.

Urbanczyk-Wochniak, E., Luedemann, A., Kopka, J., Selbig, J., Roessner-Tunali, U., Willmitzer, L. and Fernie, A.R. Parallel analysis of transcript and metabolic profiles: a new approach in systems biology (2003) EMBO Rep. *4*, 989–93.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., et al. The sequence of the human genome (2001) Science *291*, 1304–51.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. and Bork, P. Comparative assessment of large-scale data sets of protein-protein interactions (2002) Nature *417*, 399–403.

Weston, A.D. and Hood, L. Systems biology, proteomics, and the future of health care: Toward predictive, preventative, and personalized medicine (2004) J. Prot. Res. *3*, 179–196.