Modeling and identification of biological networks

Esa Pitkänen

Helsinki 20.2.2007

Seminar on Computational Systems Biology

UNIVERSITY OF HELSINKI
Department of Computer Science

Contents

1 Introduction

The focus of biological research has shifted from the study of individual biological components, such as a single gene or enzyme, to systems of interacting components. The term *systems biology* has been coined to describe the new field. In systems biology, the goal is to reveal interactions between biological components of interest and to construct a model of components and interactions to analyse, interpret and predict experimental outcomes [10].

Systems biology is sometimes also called "biology of networks", because a typical system studied by the discipline can be naturally seen as a network of components connected together by different types of interactions. Consequently, computational systems biology often makes use of graphs to model the phenomenon under study.

In a recent survey, Florence d'Alché-Buc and Vincent Schachter discuss modeling frameworks that are available for modeling of different biological systems [4]. This study follows roughly the same structure as theirs. First, in section 2 transcriptional regulation and metabolism are introduced as examples of biological systems of networks. Then, in section 3, modeling frameworks for systems biology are discussed.

2 Biological networks

Systems biology deals with various types of biological systems, which can be seen as networks. These systems include protein-protein interactions, signaling networks, regulation of gene expression at different levels and metabolism, for example. The last two types of networks are discussed in more detail. Figure 1 illustrates the connections between transcriptional regulation, signal transduction and metabolism at an abstract level.

2.1 Transcriptional regulation

Cells adapt to changes in the environment by adjusting the activity of processes which turn DNA into proteins. For instance, depletion of some nutrient might trigger the expression of certain genes to initiate the uptake of another nutrient. These processes include transcription of DNA into mRNA, translation of mRNA into protein and post-translational modification of proteins.

Transcriptional regulation is the process by which genes regulate the transcription

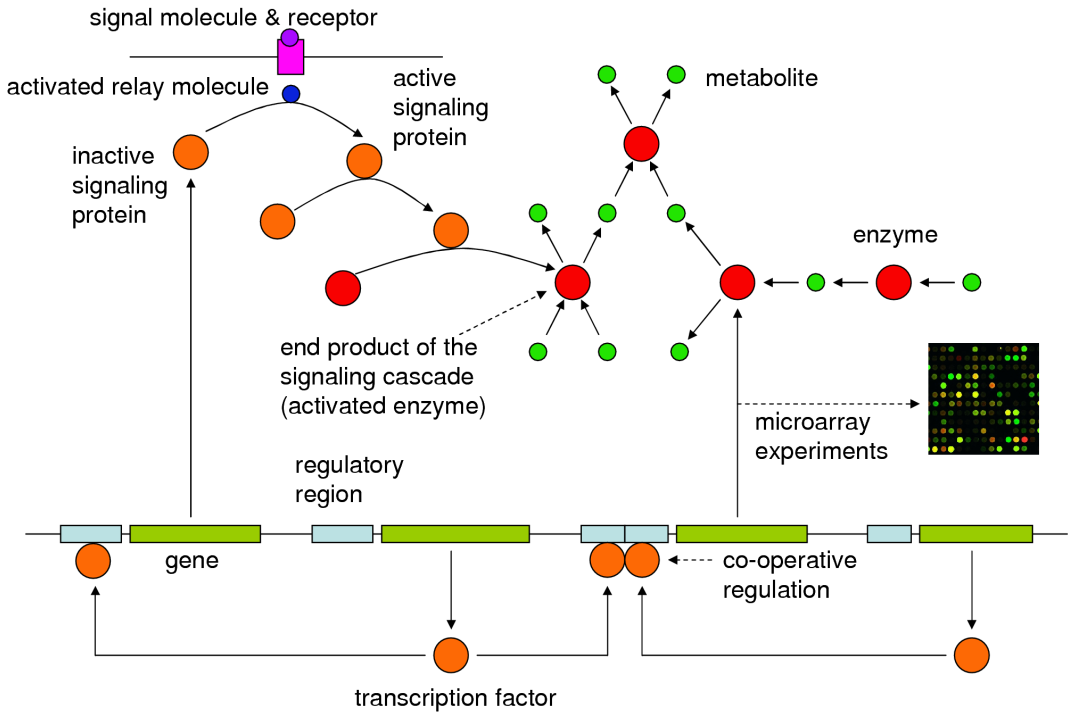

Figure 1: Transcriptional regulation, signal transduction and metabolism illustrated in a simple example system. The lower part of the figure depicts four genes (green boxes) with associated regulatory regions (blue boxes). Protein (orange circles) from the first gene on the sequence (from left to right) participates in signal transduction, while the product of the last gene is an enzyme. Gene products from the middle genes are transcription factors to the first and third gene. The top right part has four enzymes (red circles) with metabolites (small green circles). Finally, the top left part shows a signalling cascade of three proteins. The signal molecule (small blue circle) originating from the signal receptor in the cell wall activates the first protein in the cascade. This protein then activates the second protein, which in turn activates the third. The third protein acts as an enzyme participating in the example metabolic network.

of other genes. Genes have regulatory effect on other genes via their gene products, or proteins. In *direct regulation*, the gene product of gene A binds to a *regulatory region* of gene B. The binding can have either activating or inhibiting effect on the transcription of gene. The protein acts as a *repressor* if it binds to a DNA region close to the gene blocking the function of the RNA polymerase, which is the cellular machine responsible for transcription. This results in inhibition of transcription. It is also possible for the protein to alter the specificity of the RNA polymerase to either activate or inhibit transcription.

The lower part of figure 1 depicts four genes (green boxes) with associated regulatory regions (blue boxes). Protein from the first gene on the sequence (from left to right) participates in signal transduction, while the product of the last gene is an enzyme. Gene products from the middle genes are transcription factors to the first and third gene.

Regulation can also be *indirect*: gene A can regulate gene B which regulates gene C. Moreover, two or more genes may have a more complex regulatory effect on some gene than just a simple linear additive relationship. The activation of the gene may require the presence of proteins from two genes at the same time, for instance. Such arrangement can be seen to implement an AND logic gate.

Transcriptional regulation can be measured using many different technologies. Some of the methods can be considered to be *high-throughput* in the sense of being able to produce data from a large number of genes at a time. The main high-throughput measurement technology is the DNA chip, or DNA microarray, which measures the concentrations of mRNAs corresponding to the set of genes under study. A single microarray can cover all genes of an organism providing a snapshot of genome-wide gene expression at a specific time point. Furthermore, microarray experiments in sequence can be used provide a time series of expression measurements.

Microarray measurements can thus be used to give a "snapshot" of the state of the transcriptional regulation system. Nevertheless, it should be noted that it is usually not feasible, economically or otherwise, to perform more than tens of microarray measurements in a study. This has important implications by restricting the modeling frameworks feasible for the modeling task at hand.

Another technology which can be used to provide information on transcriptional regulation is chromatin immunoprecipitation (ChIP). A ChIP-chip assay identifies the binding of a protein to a DNA sequence. This allows detection of regulatory relationships by identifying proteins binding to the regulatory regions of a gene.

2.2 Metabolism

Metabolism is the set of cellular processes which transform nutrients into energy and precursor molecules, which are in turn converted into more complex biomolecules such as amino acids and lipids. Metabolism can be seen as an assembly line: small biomolecules, or metabolites, are processed along a series of enzymes, which catalyse the biochemical reactions transforming the metabolites. The cell is able to regulate metabolism at a detailed level, because enzymes are very specific. An enzyme usually catalyses only a single reaction or a small number of related reactions. By adjusting the concentration of an enzyme, or altering its enzymatic activity, the cell can activate or disable metabolic capabilities.

In Figure 1, the top right part shows four enzymes (large red circles) with metabolites (small green circles).

3 Modeling biological networks

Graphs are natural models for systems described above. At the simplest level, a graph can be used to model static relationships between biological components. The graph can then be either undirected or directed, depending on whether the relation is symmetric or not. For instance, graphs modeling protein-protein interactions are undirected, because the actual physical mechanism, protein binding to another protein, is symmetric. Directionality of graph edges may be used to encode different properties, such as causal relations.

This section divides the models discussed with respect to two attributes. First, whether the model includes discrete or continuous variables, and second, whether the model deals with the notion of time. Models taking time into account are called dynamic, and static otherwise. The modeling frameworks discussed in this study are shown in Figure 2 in terms of this two-way division.

Typical use for both undirected and directed graphs is to answer questions about static network structure. Consider the following graph model for transcriptional regulation. A *gene regulatory network* is a directed graph, where vertices represent genes and edges regulatory interactions. Since regulatory interactions can be either activating or inhibiting, the graph is enriched by labeling the edges to distinguish between the two cases. Such model could then be used to find nodes with a high out-degree corresponding to genes regulating many other genes. Figure 3 shows an

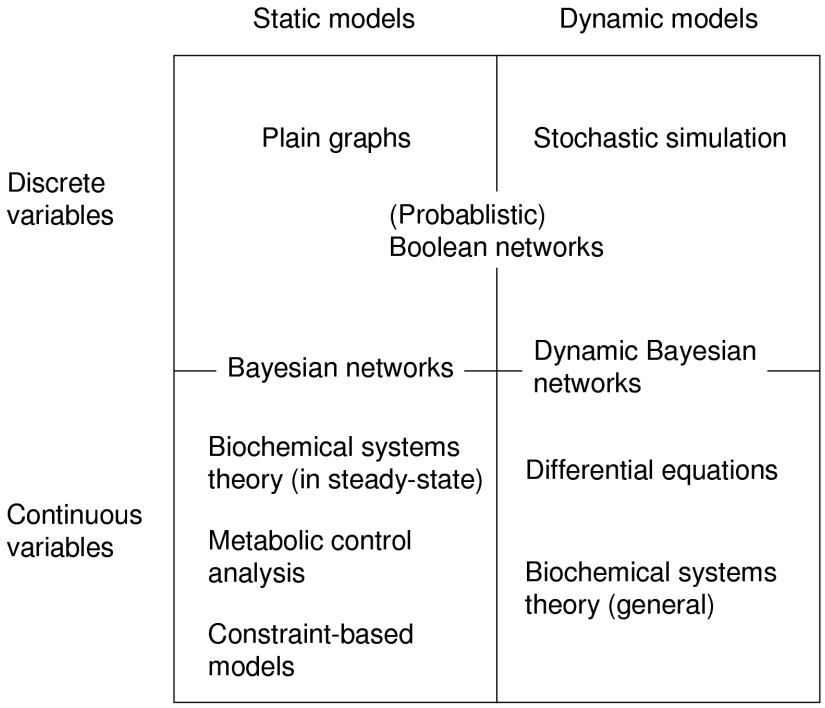

Figure 2: Modeling frameworks discussed in this study divided into static and dynamic models, and models including discrete and continuous variables.

example of a gene regulatory network.

In general, the vertices or edges of the model can be enriched to encode relevant properties of the system. To construct a model of metabolism, one could use a bipartite directed graph, where both reactions and metabolites are nodes of the graph. Edges would then correspond to the consumption and production relations between reactions and metabolites. Again, it would be easy to find high degree metabolite nodes. These hub nodes correspond to reactants that participate in many reactions. We could then argue that these metabolites might have a central role in metabolism [8]. Figure 4 shows an example of a metabolic network.

3.1 Dynamic models

Often biological questions concern the dynamics of the system under study. In other words, we would like to find out the system behaviour over some time period. For instance, we could ask how the concentrations of regulatory proteins in transcriptional regulation or metabolites in metabolism fluctuate in given conditions.

To answer these questions, the model has to deal with the change of molecular

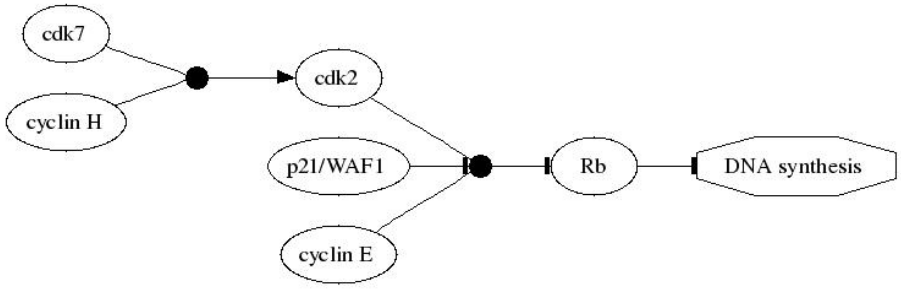

Figure 3: Example gene regulatory network (a simplified part of cell cycle regulation) [12]. Arrows represent activation, lines with bar represent inhibition.

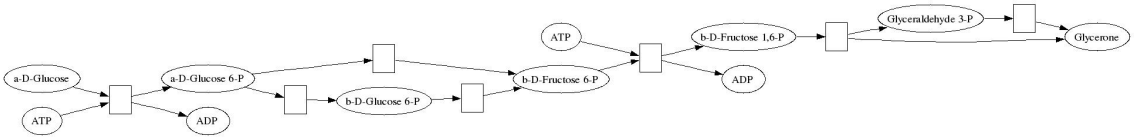

Figure 4: Example metabolic network (a part of glycolysis in *S. cerevisiae*). Ellipses represent metabolites, boxes represent reactions. Nodes for ATP and ADP have been duplicated to increase readibility.

concentrations over time, either in continuous or discrete manner. This allows simulating the model and experimenting with different model parameters. A dynamic model can be used to either explain observed system behaviour or predict behaviour in conditions from which no experimental data available.

The most common approach to model a dynamical system of biomolecules is to use differential equations. In a differential equation model, variables correspond to the concentrations of biological molecules. In general, the model consists of rate equations

$$\frac{dx_i}{dt} = f_i(x), 1 \leq i \leq n,$$

where $\mathbf{x} = (x_1, \ldots, x_n)$ is the vector of concentrations and f_i is an arbitrary function.

The crucial step in building the model is choosing appropriate functions f_i, balancing the level of detail and complexity. A complex model might be necessary to correctly describe the interactions in the system. However, the number of parameters needed to specify the model grows with the level of detail. It is often difficult to obtain data to estimate parameters for a large model with the current measurement technology.

Finally, analytical solutions for differential equation models are not known in the general case and we have to resort to simulation to solve the model.

In a gene regulation network, we could have x_i represent the concentrations of mRNAs and proteins. The available knowledge on reaction mechanisms between molecules is then encoded into functions f_i. The level of detail can be adjusted for example by simplifying the reaction mechanisms (e.g., whether to include mRNA and protein degradation), or taking into account the time delays associated with processes.

In a metabolic network, we usually consider the concentrations of metabolites and enzymes as variables x_i. The rate of a reaction depends both on the concentration of reactant metabolites and the enzyme catalysing the reaction. The rate functions are derived from enzyme kinetics [1]. For instance, consider the following reaction equation,

$$E + S \underset{k_1'}{\overset{k_1}{\rightleftharpoons}} ES \overset{k_2}{\rightarrow} E + P,$$

specifying a reaction where enzyme E and metabolite S first bind together and form the complex ES. Then, the metabolite S either is transformed into product metabolite P or the complex ES reverts back to the original, unbound state. Constants k_1, k_1' and k_2 indicate maximum reaction rates in arrow directions. The rate function corresponding to this mechanism, called Michaelis-Menten equation, can be shown to be

$$f = \frac{V_{max}[S]}{K_M + [S]},$$

where $[S]$ is the concentration of S, V_{max} is the maximum reaction rate and $K_M = \frac{k_1' + k_2}{k_1}$ is the Michaelis-Menten constant. Constants V_{max} and K_M need to be measured experimentally for the enzyme in question. While for this particular reaction mechanism the number of parameters is two, a function for a more complex mechanism might need 10-20 parameters. Hence, the construction of a reasonable differential equation model is sometimes prohibitive, particularly when dealing with very large models and/or incomplete data.

We encounter another problem with differential equation models if we try to model concentrations of molecules present in the cell in very low quantities with continuous variables. For instance, in a signaling network there might be only a couple of some molecular species (i.e., a type of molecule) present at any given time. A possible alternative approach is to use stochastic modeling, where the fate of individual molecules is decided stochastically [9].

The general differential equation modeling approach can be simplified by restricting the choice of rate functions appropriately. One option is to use piecewise-linear differential functions. This approach has been utilised in modeling both gene regulation [5] and metabolism [6].

Another framework which restricts the class of reaction rate functions is *Biochemical Systems Theory* [11], where the functions f_i are expressed in a power-law form,

$$\frac{dx_i}{dt} = \alpha_i \prod_{j=1}^{n} x_j^{g_{ij}},$$

where α_i is the rate constant for metabolite i and g_{ij} is a kinetic constant for the metabolite-reaction pair (i, j). The power-law representation approximates the kinetic system while requiring smaller number of parameters.

3.2 Static models

If the model is overly complex with respect to the available data, we can try to simplify it by giving up the dynamics. This section discusses options available if we are content with static models that cannot reveal or predict dynamic behaviour of the system, but can still perhaps reveal useful information about the system in a fixed state.

A popular approach in metabolic modeling is to study the system in or in the vicinity of a *steady-state*, where the concentrations of molecules do not change over time,

$$\frac{dx_i}{dt} = 0.$$

In many cases, this assumption turns the model into a form which is easier to solve and analyse than the original form. For instance, it can be shown that in Biochemical Systems Theory, by assuming steady-state, analytical solutions can be derived to the rate functions. The steady-state approach is less useful in transcriptional regulation and signaling models, where the change of concentrations in response to stimuli is of essence.

In metabolic modeling, other steady-state simplifications of the kinetic framework include Metabolic Control Analysis and constraint-based modeling.

Metabolic Control Analysis (MCA) approximates the differential equation system in the neighborhood of steady-state [14]. In MCA, one is interested in the sensitivity of different parameters and variables to perturbations. For instance, the question

"how much would the metabolite concentrations x_i be changed, if the activity of enzyme A was increased by 5%" could be tackled with the approach.

Constraint-based modeling is a linear framework, where metabolic processes are modelled in steady-state [13]. In contrast to MCA, where the effect of perturbations on metabolite concentrations and reaction rates is investigated, constraint-based modeling focuses on steady-state reaction rates, or *fluxes*. The metabolic model is usually represented by a *stoichiometric matrix* S with a row for each metabolite and a column for each reaction in the system. Stoichiometric coefficient S_{ij} then is the number of metabolites i produced in reaction j in a time unit, with $S_{ij} < 0$, if metabolites i are consumed by the reaction.

In terms of graphs, it is intuitive to enrich the edges of a metabolic network with coefficients S_{ij}: each edge from reaction j to metabolite i is given label S_{ij}, and edge from metabolite i to reaction j label $-S_{ij}$ to keep the labels positive.

One of the basic questions in constraint-based modeling is to characterise solutions to the equation

$$S\mathbf{v} = 0,$$

where $\mathbf{v}$ is the vector of fluxes. The equation states the steady-state condition in matrix form. In other words, the net production of each metabolite is zero. The equation system can be constrained further, if additional data is available.

Because no parameters besides stoichiometric coefficients is required, it is possible to build genome-scale constraint-based models. The major drawback is that the model is restricted to steady-state, which is not always a realistic assumption.

3.3 Discrete models

Modeling frameworks discussed above, with the exception of directed graphs, were examples of frameworks including continuous variables. Discrete frameworks have been proposed for modeling biological networks as well, especially transcriptional regulation. Examples of numerous discrete frameworks include boolean networks, generalized logical networks, petri-nets, process algebrae and rule-based formalisms. This section discusses two continuous frameworks, Boolean networks and Bayesian networks.

Boolean networks have been widely used in modeling gene regulation. The cell exhibits switch-like behaviour during regulation, and this behaviour is more or less

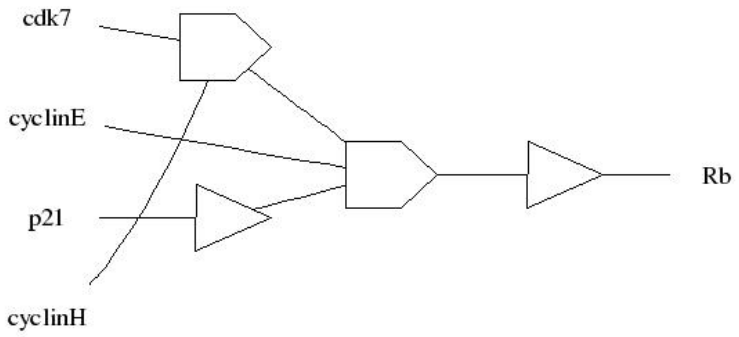

Figure 5: A logic diagram corresponding to the gene regulation network in Figure 3. Triangle is a NOT gate and pentagons are AND gates.

naturally modelled using boolean networks [12]. A boolean network $G(V, F)$ is defined by a set of nodes $V = \{x_1, \ldots, x_n\}$ and a list of Boolean functions $F = (f_1, \ldots, f_n)$. A Boolean function $f_i(x_1, \ldots, x_n)$ is assigned to node x_i.

In gene regulation, each node x_i corresponds to the state of gene i. If $x_i = 1$, gene i is active, or expressed, and if $x_i = 0$ the gene is not active. Boolean functions represent the regulatory interactions between genes.

Dynamic behaviour can be simulated with Boolean networks by considering the transition between two consequtive time steps. In the first step, the system is in a state defined by nodes $x_1, \ldots, x_n$. Next, the Boolean function f_i for each node is evaluated with $x_1, \ldots, x_n$. The resulting values are then assigned as $x_1, \ldots, x_n$ of the second time step. Because Boolean functions are deterministic, also the dynamic behaviour in Boolean networks is deterministic: the initial state of the model completely determines the end state. Figure 5 depicts a logic diagram corresponding to the gene regulation network in Figure 3.

Although Boolean models are simple to understand and they can be inferred from data effectively in some problem settings [12], the model framework has serious limitations. First, a Boolean variable having only two possible states is not a realistic way to model many biological properties. Second, a deterministic Boolean model does not cope well with noisy or missing data. Even though there is the "correct" Boolean function for some variable, noise in the data might make it impossible to infer it without probabilistic considerations. Third, specifying large Boolean models requires a lot of data which may not be available. This is particularly true with gene regulation networks.

Many probabilistic discrete modeling frameworks have been developed, which can be used to alleviate problems encountered with Boolean networks and real-world modeling situations.

Probalistic Boolean networks extend Boolean networks by accomodating more than one possible function for each node [12]. In addition, each node is given a probability distribution over the possible functions. The dynamics of the probabilistic Boolean network are same as for Boolean networks, except when determining the next value for a variable, the function is chosen randomly from the possible functions according to the associated probability distribution.

As in Boolean networks, the complexity of individual Boolean functions can be limited. For instance, the maximum number of variables affecting the outcome could be set to some small number. In addition, in probabalistic Boolean network, one has to choose the number of possible functions for each variable. This choice increases both flexibility and the danger of overfitting the model to data. Ideally, these choices should reflect the amount of data available.

Bayesian networks has been a popular tool in gene regulation network analysis [7]. A Bayesian network is a directed acyclic graph which encodes the joint probability distribution over a set of random variables $x_1, \ldots, x_n$. The joint probability distribution can be expressed as

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid pa[x_i]),$$

where $pa[x_i]$ is the set of parent variables of x_i in the graph. This decomposition gives a space-efficient method of representing the joint probability distribution compared to the general case.

To study system dynamics, one can extend Bayesian networks by "unfolding" the graph structure [7]. For each time point under consideration, a nodes of the Bayesian network are duplicated. The probability distribution for a variable in time step $t + 1$ is then given in terms of its parents in time step t. The resulting graph is also a Bayesian network, and it can be used to make inferences involving dynamic behaviour of the variables.

Bayesian networks can be learned from data. Unfortunately, learning the network structure is computationally hard [2]. Therefore, simulation methods are usually used to infer the structure. The methods are also suspectible to the choice of prior probability distributions given to variables with no parents in the graph. Intuitively,

as with other frameworks, finding the correct model from the data gets more difficult as the amount and quality of data decrease.

4 Conclusion

Graphical modeling frameworks are useful tools for systems biology. Many approaches have been developed in the recent years to deal specifically with problems rising from the domain of systems biology. In particular, one of the most prominent questions is how to deal with the lack and uncertainty of data. Choosing an appropriate modeling framework is an important part of the answer.

Related to the two systems discussed here, transcriptional regulation and metabolism, an interesting research direction is building a model combining the two systems [3]. A central issue would be how to integrate in the same model aspects from logic-oriented discrete gene regulation networks and perhaps constraints-based continuous metabolic networks, and from different time scales.

References

1 J. Berg, J. L. Tymoczko, and L. Stryer. *Biochemistry.* W. H. Freeman, New York, 5th edition, 2002.

2 D. M. Chickering, D. Geiger, and D. Heckerman. Learning Bayesian Networks is NP-Hard. Technical Report MSR-TR-94-17, November 1994.

3 M. W. Covert and B. Ø. Palsson. Transcriptional regulation in constraints-based metabolic models of escherichia coli. *J Biol Chem*, 277(31):28058–28064, 2002.

4 F. d'Alché Buc and V. Schachter. Modeling and identification of biological networks. In *Proc. Intl. Symposium on Applied Stochastic Models and Data Analysis (ASMDA 2005)*, pages 167–179, 2005.

5 H. de Jong. Modeling and simulation of genetic regulatory systems: a literature review. *J Comput Biol*, 9(1):67–103, 2002.

6 J. S. Edwards, R. U. Ibarra, and B. Ø. Palsson. In silico predictions of escherichia coli metabolic capabilities are consistent with experimental data. *Nature Biotechnology*, 19:125–130, 2001.

7 D. Husmeier. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic bayesian networks. *Bioinformatics*, 19(17):2271–2282, 2003.

8 H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature*, (407):651–654, 2000.

9 E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach. *Systems Biology in Practice: Concepts, Implementation and Application*. Wiley-VCH, 2005.

10 B. Ø. Palsson. *Systems Biology: Properties of Reconstructed Networks*. Cambridge University Press, 2006.

11 M. A. Savageau. Biochemical systems theory: operational differences among variant representations and their significance. *J Theor Biol*, 151(1):509–530, 1991.

12 I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang. Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18(2):261–274, 2002.

13 G. N. Stephanopoulos, A. A. Aristidou, and J. Nielsen. *Metabolic Engineering, Principles and Methodologies*. Academic Press, 1998.

14 H. V. Westerhoff, J. H. Hofmeyr, and B. N. Kholodenko. Getting to the inside of cells using metabolic control analysis. *Biophys Chem*, 50(3):273–283, 1994.