Thesis for the Degree of Doctor of Philosophy

Modeling and identification of biological systems with emphasis on osmoregulation in yeast

Peter Gennemark

CHALMERS | GÖTEBORG UNIVERSITY

Department of Computer Science and Engineering
Chalmers University of Technology
and Göteborg University
SE-412 96 Göteborg, Sweden

Göteborg, Sweden 2005

Modeling and identification of biological systems with emphasis on osmoregulation in yeast
Peter Gennemark
ISBN 91-7291-647-8

© Peter Gennemark, 2005.

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr. 2329
ISSN 0346-718x

Technical report no. 10D
Department of Computer Science and Engineering
Research group: Bioinformatics

Department of Computer Science and Engineering
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg
Sweden
Telephone + 46 (0)31-772 1000

Göteborg, Sweden 2005

Modeling and identification of biological systems with emphasis on osmoregulation in yeast

Peter Gennemark
Department of Computer Science and Engineering
Chalmers University of Technology and Göteborg University

Abstract

This thesis deals with two topics in the area of systems biology. The first topic, model identification, concerns the problem of automatically identifying a mathematical model of a biochemical system from experimental data. We present algorithms for parameter estimation and model selection that identify both the structure and the parameters of a differential equation model from experimental data. The system is designed to handle problems of realistic size, where reactions can be non-linear in the parameters and where data can be sparse and noisy. To achieve computational efficiency, parameters are estimated for one equation at a time, giving a fast and accurate parameter estimation algorithm compared to other algorithms in the literature. The model selection is done with an efficient heuristic search algorithm, where the structure is built incrementally. The main strengths of our algorithms are that a complete model, and not only a structure, is identified, and that they are considerably faster compared to other identification algorithms.

The second topic concerns mathematical modeling of osmoregulation in *Saccharomyces cerevisiae*, budding yeast. This system involves the biophysical and biochemical responses of a cell when it is exposed to an osmotic shock. We present two different differential equation models based on experimental data of this system. The first model is a detailed model taking into account an extensive amount of molecular detail, while the second is a simple model with less detail. We demonstrate that both models agree well with experimental data on wild-type cells. Moreover, the models predict the behavior of other genetically modified strains and input signals.

Keywords: model identification, model selection, parameter estimation, ordinary differential equations, Saccharomyces cerevisiae, osmotic stress, HOG signaling pathway.

List of papers

This thesis is based on the work contained in the following papers:

1. Efficient ODE model identification for biological applications.
   Gennemark P. and Wedelin D.
   Submitted.

2. Integrative model of the response of yeast to osmotic shock.
   Klipp E., Nordlander B., Krüger R., Gennemark P. and Hohmann S.
   Nat Biotechnol. 2005, 23(8), 975-82.

3. A simple mathematical model of adaptation to high osmolarity in yeast.
   Gennemark P. and Nordlander B.

Contents

Acknowledgments

I would like to thank:

Docent Dag Wedelin (Computing science, Chalmers), who has guided me through this education in a professional and pedagogical way. Dag's curiosity and determination in problem solving have been a great source of inspiration to me.

Bodil Nordlander (Cell and Molecular Biology, Göteborg University), who has explained and discussed the biology of osmoregulation as well as the experimental issues concerning our mathematical models.

Professor Olle Nerman (Mathematical statistics, Chalmers), who has shown a great interest in the project and given me valuable advice concerning modeling in general and stochastic aspects in particular.

Professor Stefan Hohmann (Cell and Molecular Biology, Göteborg University), who has explained the biological details of osmoregulation and helped me to understand its role from a biophysical perspective.

Professor Per Sunnerhagen (Cell and Molecular Biology, Göteborg University), who has put valuable questions concerning our work on identification and who has also explained and discussed the biology of osmoregulation.

Dr. Edda Klipp (Max-Planck Institute for Molecular Genetics, Berlin), who has hosted me in Berlin and introduced me to biological modeling in general and ODEs in particular.

The people in my PhD committee: Aarne Ranta, Peter Damaschke and Per Sunnerhagen for giving feedback and input to my work and taking their time for several meetings.

Colleagues and friends at Computer Science, Mathematics and Lundberg laboratory for giving me a nice and pleasant work environment.

Tesse & Erik, for your great support and patience.

Peter Gennemark, August 2005

1 Introduction

The adaptive responses of a living cell to internal and external signals are controlled by complex networks of proteins affecting for example transcriptional responses and metabolic processes. On a basic level, the structure of such a network can be described by a graph, see e.g. Figure 1. This gives a useful overview of the network, but it is not a complete description, since concentrations and the dynamic behavior in time and space are not described. Despite this fact, this is the level of detail at which biologists traditionally model biochemical systems. In part this is due to lack of quantitative experimental data and the difficulty in manually inferring the model from such data.

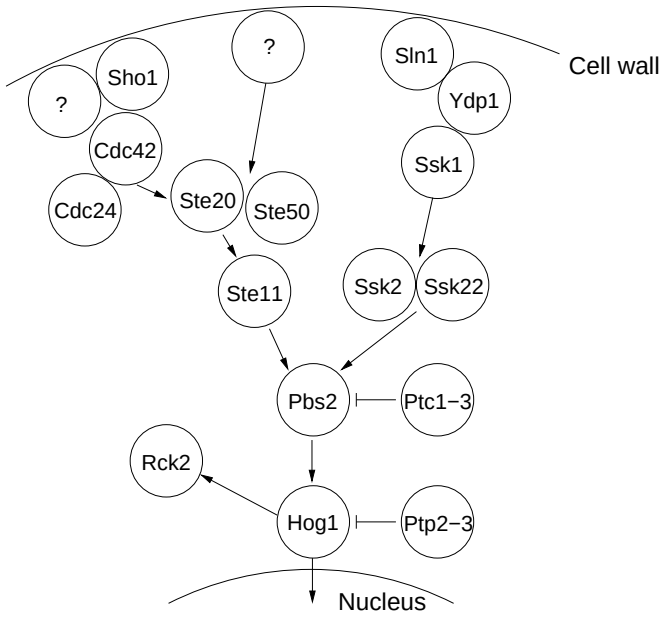

Figure 1: *Traditional pathway model of the main components in the High Osmolarity Glycerol (HOG) signaling pathway in S. cerevisiae (Hohmann 2002). The proteins (vertices) are connected by interactions (edges). Because of imprecise meaning of the interactions and lack of dynamic information, models like these only give a schematic overview of the system. Details of the HOG pathway are presented in Section 6.*

To create more powerful descriptions, dynamic mathematical models based on biochemical rate equations can be considered. One of several basic motivations for creating a more complete model is to simulate the system. For a sufficiently exact model it then becomes possible to predict the behavior of the real system as well as modified systems. In cell biology, a typical experiment involves variation of one or several input variables, such

as temperature, osmolarity and drug concentration. Besides, the cell can be genetically modified, e.g. by deletion or over-expression of a certain gene.

The use of systematic experimental technologies in order to develop and analyze mathematical models of complex biological systems constitutes the base of *systems biology*, a research field that has rapidly evolved in recent years (Kitano 2002a, 2002b). This involves a shift from studying specific cellular components like a single gene or a single protein to emphasizing systems level studies of cellular processes. The development of systems biology has been driven by the advancement of experimental methods. Several major breakthroughs like the genome sequencing and the development of high-throughput and large-scale techniques, such as micro-arrays, offer a great potential for obtaining a sufficient volume of data (Zhu et al. 2002). At the same time, methods for obtaining high-quality data, such as quantitative mass spectrometry, have become more efficient (Aebersold et al. 2003, Mckenzie et al. 2003). Another key component in systems biology is the development of systems approaches for modeling (Westerhoff et al. 2004), like Metabolic Control Analysis (Heinrich et al. 1977, Kacser et al. 1973) and Biochemical Systems Theory (Savageau 1976). In combination with the rapid increase of computational power, such approaches offer a framework for detailed mathematical modeling of complete systems. Recently, several new computational approaches, like the systems biology mark-up language SBML (Hucka et al. 2003) and the software environment for whole-cell simulation, E-cell (Tomita et al. 1999, Takahashi et al. 2003), have also been developed in this area.

This thesis deals with two separate but related topics within systems biology:

Automatic model identification. This topic concerns the problem of automatically identifying a mathematical model from data. Identification complements data simulation as illustrated in Figure 2, and closes a loop between model and data. We present efficient model identification algorithms, that reconstruct an ordinary differential equation (ODE) model from time series measurement of individual compounds (Paper 1). The performance of the algorithms has been evaluated on three previously published biological models. We show that our approach is more accurate and considerably faster compared to existing methods. Model identification involves both estimating the parameters of a model and selecting the model structure. In this introduction we consider these two issues in Sections 4 and 5, respectively.

Modeling osmoregulation in yeast. We present work on modeling of osmoregulation in the yeast *Saccharomyces cerevisiae*. This work has been done in collaboration with experimentalists at Göteborg University. Osmoregulation involves the biophysical and biochemical responses of a cell when it is exposed to an osmotic shock, see Figure 3 for an overview. We present two different ODE models based on experimental data of this sys-

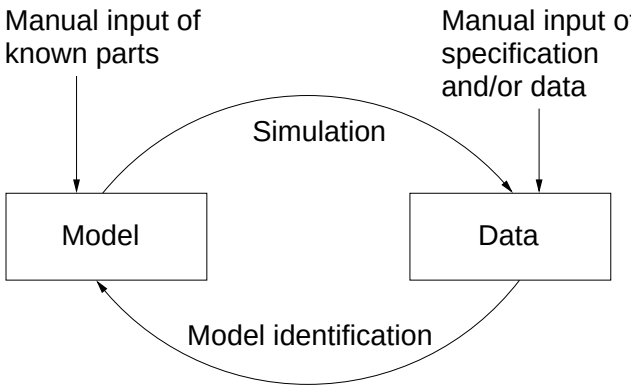

Figure 2: *The desired relationship between a model and data involves two main functionalities: data simulation and model identification.*

tem. The first model (Paper 2) is a detailed model taking into account an extensive amount of molecular detail, while the second (Paper 3) is a simple model with less detail. We demonstrate that both models agree well with experimental data on wild-type cells. Moreover, the models predict the behavior of other genetically modified strains. In this introduction, we describe osmoregulation in Section 6.

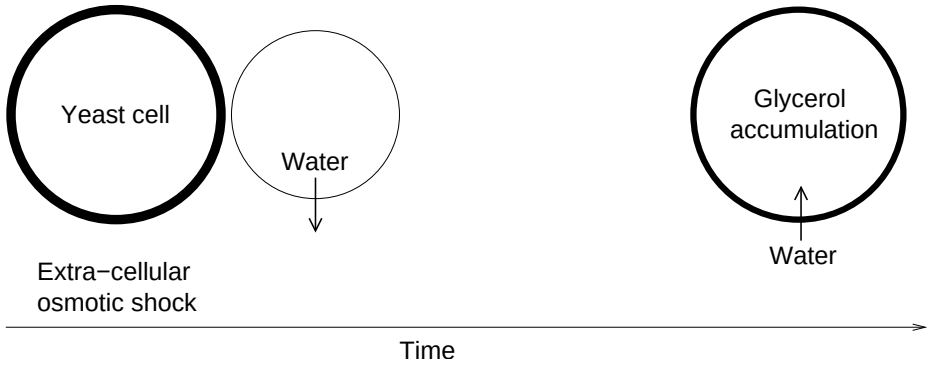

Figure 3: *A basic overview of osmoregulation in yeast (Gervais et al. 2001, Hohmann 2002). An extra-cellular osmotic shock, e.g. the addition of 0.5M NaCl to the medium, rapidly initiates a water flow out of the cell leading to loss of turgor pressure and volume decrease. The cell adapts by accumulating glycerol in order to regain water and thereby volume and turgor pressure. In the figure, turgor pressure is indicated by thickness of the cell membrane.*

Basically, the two topics of this thesis both deal with the question of how to construct a model from experimental data. This can be done either automatically using our identification algorithms or manually as mainly done for the osmoregulation system. However, also in the latter case we use auto-

matic means for estimating the parameters. Hence, parameter estimation is a recurrent theme in this thesis. Furthermore, regardless of approach used for model construction, the problem of choosing a proper level of detail, i.e. to avoid over- and underfitting when deciding on the model, is of great importance and this is also an issue in common.

This extended introduction is intended to give a background to the papers and also to provide further perspectives that are not present in the papers themselves.

The rest of the introduction is structured as follows. The next section is devoted to a brief introduction to mathematical modeling of biological systems and in Section 3 we focus on ODE models for biochemical systems. In the following two sections we consider parameter estimation and model selection, respectively. In Section 6 we describe our modeling efforts on yeast osmoregulation. Finally, in Section 7 the main contributions of this thesis are summarized.

2 Modeling biological systems

This section introduces basic concepts and approaches for the modeling of *biological systems*. In the context of this thesis we restrict ourselves to biological systems in cell biology, although most of the content also applies to other biological modeling areas, such as population dynamics. In the following sections we will also refer to *biochemical systems*, which can be considered a subclass of biological systems. In a biochemical system we only deal with molecules such as proteins and metabolites. In Paper 1 we consider biochemical systems, while the models of Paper 2 and 3 can be viewed as biological models since biochemical and biophysical modeling are combined.

An essential issue in all modeling is to define the scope of the model. This involves specifying which subsystems and which variables that should constitute the model. A natural goal is to find a system that is reasonably well isolated under the considered experimental conditions. Obviously this is a very difficult task, since all processes in the cell are more or less dependent on each other. As an example, the level of a particular enzyme can be assumed constant for a short time interval. However, for an experiment starting with some environmental stimulus, the stress may trigger changes in gene expression that alter the activity of the enzyme. This is especially important for experimental scenarios ranging in the order of hours. It is therefore natural to try to identify all variables that are adjusting to the experimental perturbation, for instance by large-scale experiments.

Another important issue in modeling is to consider what amount and quality of experimental data is available. This influences the choice of modeling approach as well as the level of detail of the model. A variety of modeling approaches with different precision are used for modeling and analysis of biological systems. In general, a more precise approach requires more precise and extensive data to be identified. Naturally, a more precise approach also offers more realistic and useful predictions. To give an overview, it is useful to distinguish three basic modeling approaches:

Boolean networks. This is the most coarse approach in which each variable is either 'on' or 'off'. For instance, when modeling a genetic network a gene is either fully expressed or not expressed at all. Using boolean functions one defines how the system deterministically goes from one state to the other. As an example, gene A is 'on' in the next state given that gene B and gene C are 'on' in the current state. In this formalism computation is obviously rapid - the updates of all variables occur synchronously and only boolean functions are evaluated. From any initial state, a boolean network reaches either a steady state or a state cycle in finite time.

Typically, boolean networks have been applied to genetic networks where the number of variables is large and where data is sparsely sampled and noisy (Huang 1999). There are several methods available for identification of boolean networks from experimental data (Liang et al. 1998, de Jong 2002). In order to decrease the complexity of identification one can set an upper bound on the number of inputs to each function. The biological interpretation of this is that each gene can only be influenced by a subset of other genes.

Obviously, boolean network models are inaccurate since variables are discrete and there is no precise notion of time. Therefore, this approach is mainly applicable to systems were a steady state is reached.

Ordinary differential equations (ODEs). ODEs deal with continuous variables that typically assume real-valued concentrations. In general, a system is described as

$$X_i'(t) = f_i(\mathbf{X}, \mathbf{I}), \quad i = 1...n \tag{1}$$

where $\mathbf{X} = [X_1 \dots X_n]$ is the state vector of concentrations and $\mathbf{I} = [I_1 \dots I_m]$ is a vector of input variables, and f_i are typically non-linear functions. These functions usually include several parameters (rate constants) that can either be experimentally determined or estimated from various data. We generally note that it is difficult to measure kinetic rate constants experimentally and that parameter estimation is a complex optimization problem for ODE models of realistic size.

The main feature of a system of ODEs is that it can be simulated in order to obtain deterministic time series for the variables. Standard numerical methods exist for this purpose. The input to such a simulation is the ODEs, values for the parameters and initial values for all variables.

ODEs have been widely used to model biological systems. For instance, the metabolism and the cell cycle regulation have been extensively modeled using ODEs (Chen et al. 2004, Rizzi et al. 1997). We note, however, that there are few identification methods available for ODE models. In Section 3 we give a more comprehensive introduction to ODEs.

Stochastic models. This is a very detailed modeling approach, in which each variable represents the number of molecules. The state changes discretely, but how and when is determined stochastically. There are standard methods to perform stochastic simulation, although they are typically very computer intensive (Gillespie 1976, Gibson et al. 2000, Meng et al. 2004). One such simulation gives one potential behavior of the system. By repeating the simulation many times, we obtain an approximation to the probability distribution of the system over time. Hence, we can tell the probability of having exactly n_i molecules of variable X_i at time t.

Stochastic modeling is typically applied when the number of molecules is low and the assumption of continuously varying concentrations becomes too inexact. In signaling pathways, for example, stochastic fluctuations may be large enough to affect the system. To measure the average value of several cells leads to a more deterministic shape of the experimental time series, but a systematic error may be present. This is especially true, if there are non-linear reactions in the system. As an example, the distribution of individual cells with different swimming behaviors could be predicted by introducing stochasticity into a model of signaling proteins in *E. coli* (Morton-Firth et al. 1998, Levin et al. 1998, Abouhamad et al. 1998). Naturally, another way of dealing with the problem of inhomogeneous cell populations is to consider single-cell experiments (Peng et al. 2004).

Compared to ODE models, stochastic models are better approximations of the biochemical reality, but also require considerably more computational effort to be simulated. Furthermore, ODE models give a deterministic answer that may involve a systematic error, while stochastic models give a probabilistic answer. The choice between ODEs and the stochastic approach is therefore partly a trade-off between computational efficiency and accuracy in the simulations. We note, however, that for many systems the accuracy obtained by ODEs is a good approximation, since the effects of stochasticity do not influence the behavior of the system at the observed level of detail. In addition, it is often the case that the stochasticity itself is not essential to the biological functionality of the system.

We finally note that there are also several intermediate approaches between the basic ones presented above (de Jong 2002, Bower et al. 2001).

3 ODE models of biochemical systems

The most widespread formalism to model dynamical systems in science and engineering is ODEs and in this thesis we only consider this approach. Therefore, we will introduce the use of ODEs in biochemical modeling in more detail.

Consider the following biochemical reaction for the transition of compound S to P with rate constant k

$$S \xrightarrow{k} P$$

The rate of the reaction is obtained by the mass action law as $k[S]$, where $[S]$ denotes the concentration of S. For simplicity, we will from now on skip the brackets for denoting concentration. The ODEs for the variables can then be obtained as

$$P'(t) = -S'(t) = kS(t). \tag{2}$$

Similarly, if the reaction is assumed catalysed by enzyme E, the bilinear reaction mechanism is the simplest possible

$$P'(t) = -S'(t) = kS(t)E(t). \tag{3}$$

However, a more detailed analysis is often required in order to model an enzymatic reaction. In particular, Michaelis-Menten accounts for the kinetic properties of many enzymes (Stryer 1995). In this approach, a substrate S is turned into a product P by an enzyme E according to the following reaction

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} ES \xrightarrow{k_2} E + P$$

where ES is a transition state complex, k_1 and k_{-1} are the forward and backward reaction constants of the first step, respectively, and k_2 is the reaction constant of the second step of the reaction. By assuming that $S \gg E$, which is usually valid for metabolic systems, and by assuming catalytic steady state, that is $ES'(t) = 0$, we obtain (Stryer 1995)

$$P'(t) = -S'(t) = \frac{V_{max} S(t)}{S(t) + K_M} \tag{4}$$

where V_{max} and K_M are constants. We note that a linear approximation of the same form as (3) is obtained if $S \ll K_M$.

To add one more level of complexity, we introduce the mechanism of non-competitive inhibition, which will be used as an example in Sections 4 and 5. This is also an enzymatic reaction, but here the enzyme has two binding sites: one active site for the substrate and one regulatory site for the non-competitive inhibitor (Stryer 1995), see Figure 4. The enzyme can bind substrate at the active site and catalyze the production of product as long as the non-competitive inhibitor is not bound to the regulatory site. However, once the non-competitive inhibitor binds at the regulatory site, the shape of the active site changes so that it can no longer catalyze the reaction. The enzyme will remain inhibited until the non-competitive inhibitor leaves the regulatory site. Using similar assumptions as for the Michaelis-Menten reaction, the following ODE can be derived

$$P'(t) = -S'(t) = \frac{V_{max} S(t)}{(S(t) + K_D)\left(1 + \frac{I(t)}{K_I}\right)} \tag{5}$$

where I is the inhibitor concentration and V_{max}, K_D and K_I are constants.

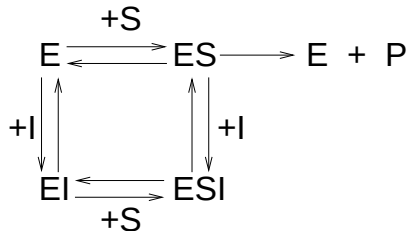

Figure 4: *Reaction mechanism for an enzymatic reaction with non-competitive inhibition (Stryer 1995). S denotes the substrate, P denotes the product, E denotes the enzyme and I denotes the inhibitor.*

Finally, we want to point out that there are several other biochemical reactions that can be modeled in a similar way as described here. One example is reactions having several substrates and/or products.

3.1 Model examples

By combining a set of compounds with reactions (like the reactions presented in the previous section), an ODE model of a biochemical system can be constructed in the form of (1).

We exemplify by two test models, which will also be used to illustrate certain concepts in Sections 4 and 5. The first model contains one compound that

exists in two states, A_1 and A_2, and where the transition from A_1 to A_2 is catalysed by the input signal I, while the reverse transition occurs spontaneously, see the left part of Figure 5. Assuming simple linear kinetics, the system of ODEs is obtained as

$$A_2'(t) = -A_1'(t) = k_1 A_1(t) I(t) - k_2 A_2(t) \tag{6}$$

where each term on the right hand side corresponds to one reaction. From now on we will denote the form of the ODEs as the structure of the model. The model structure in combination with values for the parameters, k_1 and k_2 in this case, define the complete model.

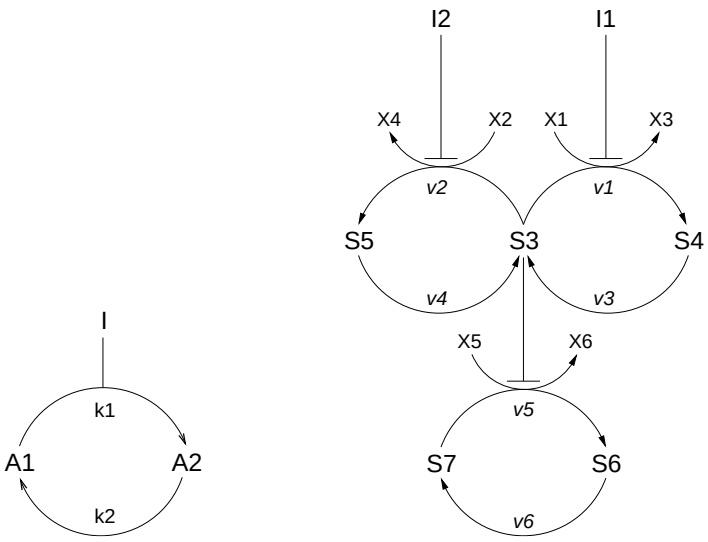

Figure 5: *Left: Simple model of two variables and one input variable.* k_1 *and* k_2 *are rate constants. Right: The metabolic test system in Paper 1.* I_1 *and* I_2 *are input variables,* $S_3 - S_7$ *are measured variables,* $X_1 - X_6$ *are variables corresponding to metabolites assumed buffered at constant levels and reactions* $v_1 - v_6$ *are catalysed by different enzymes which also are present at constant levels. All reactions follow Michaelis-Menten kinetics and* v_1*,* v_2 *and* v_5 *are non-competitively inhibited.*

The second model that we consider is the metabolic test system in Paper 1, which is originally taken from Arkin et al. (1995). This system has two input variables, I_1 and I_2, and five variables $S_3 - S_7$. The kinetic equations all follow Michaelis-Menten kinetics and inhibition is non-competitive. The right part of Figure 5 depicts the model.

The system of ODEs is given in Paper 1 and here we simply illustrate by

giving the ODE for variable S_4

$$S_4'(t) = v_1 - v_3 = \frac{S_3(t)V_{max1}}{(S_3(t) + K_{D1})\left(1 + \frac{I_1(t)}{K_{I1}}\right)} - \frac{S_4(t)V_{max3}}{S_4(t) + K_{D3}} \tag{7}$$

where V_{max1}, K_{D1}, K_{I1}, V_{max3} and K_{D3} are rate constants.

Given a model, one of the fundamental things to do is to simulate it in order to study the dynamic behavior of the variables.

3.2 Simulation

Systems of differential equations are often difficult to solve analytically, but can be simulated by numerical methods. The simplest method is *Euler's method*. The formula for this method is

$$\mathbf{X}(t + \Delta t) = \mathbf{X}(t) + \Delta t \mathbf{X}'(t) \tag{8}$$

repeated for the desired number of iterations (time). Here, Δt is a constant, typically much smaller than the simulation interval, and again, the vector $\mathbf{X}$ corresponds to the concentration of all compound states. We note that the formula is asymmetrical since it advances the solution through an interval Δt, but uses derivative information only at the beginning of that interval. For more accurate integration we can consider the Runge-Kutta method (see e.g. Press et al. 1993) and for even better accuracy and efficiency standard methods exist (Lambert 1991, Shampine et al. 1997).

As an example, we consider the model given in (6) and set the parameters as $k_1 = 0.05$ and $k_2 = 0.02$. Furthermore, we let the total concentration of A_1 and A_2 be 1 and consider the following input function

$$I(t) = \begin{cases} 1, & t \geq 20 \\ 0.01, & \text{otherwise} \end{cases} \tag{9}$$

Before simulating, it is often useful to calculate the initial steady states of the variables by setting all derivatives to zero and solve for the state variables. The steady state values for A_1 and A_2 are obtained from (6) as 200/205 and 5/205, respectively. Then, using a standard integration method (ode15s in Matlab), we obtain simulated time series data as shown in Figure 6.

For an example of simulated time series data for the metabolic test system we refer to Figure 3 of Paper 1.

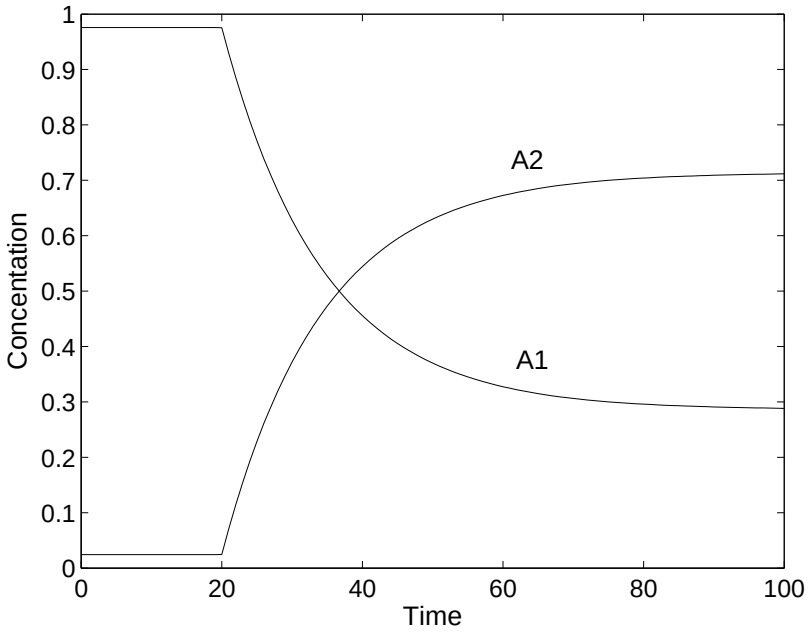

Figure 6: *Time series data of the model given in the left part of Figure 5. The ODEs are given in (6) and the step input function is given in (9).*

3.3 S-systems

Besides the metabolic test system in Paper 1, we also consider an ODE model of a genetic network. This model is taken from the literature (Kikuchi et al. 2003) and it is specified as a so-called S-system model (Savageau 1976, Voit 2000). S-systems are represented by a fixed formal structure and the generic form of equation i reads

$$X_i'(t) = \alpha_i \prod_{j=1}^{n+m} X_j^{g_{ij}}(t) - \beta_i \prod_{j=1}^{n+m} X_j^{h_{ij}}(t) \tag{10}$$

where $\mathbf{X}$ is a vector (length $n+m$) of both dependent and independent variables, α and β are vectors (length n) of non-negative rate constants and g and h are matrices ($n \times n+m$) of kinetic orders, that can be negative as well as positive. Concerning $\mathbf{X}$, the first n positions contain the dependent variables, while the remaining m positions contain the independent variables, which were denoted by $\mathbf{I}$ in (1). For an example, we refer to Figure 2 and Table 2 in Paper 1.

Basically, the S-systems formalism is derived from (1) by splitting f_i into two functions as (Voit 2000)

$$X_i'(t) = f_i^+(\mathbf{X}, \mathbf{I}) - f_i^-(\mathbf{X}, \mathbf{I}), \quad i = 1...n \tag{11}$$

where f_i^+ reflects all processes of production of variable i and f_i^- reflects all processes of degradation of variable i. We note that these functions typically are very complicated and unknown.

The functions f_i^+ and f_i^- are assumed differentiable and positive-valued and are specified as power-law functions using non-linear approximations. This is achieved by first representing the functions and variables in logarithmic coordinates. Then, the functions are approximated by Taylor series, where only the constant and linear terms are retained. The linearized functions are finally translated back into Cartesian coordinates. The result of this process is the generic formula as given in (10).

Because of the first order Taylor's approximation it is difficult to judge the validity of an S-system model. In principle, the validity can be improved by considering additional terms in the Taylor's approximation. However, that would increase the number of parameters and give a less compact form of the equations, why analysis and identification would become much more difficult (Guebel 2004).

4 Parameter estimation in ODEs

In this section we consider the problem of assigning values to the parameters in a known model structure. For instance, in the model of (6) we want to assign values to k_1 and k_2.

It is technically difficult to measure kinetic rate constants in experiments. The origin of such existing data is often *in vitro*[1] experiments and it cannot be generally assumed that the corresponding values *in vivo*[2] are the same. Besides, different laboratory conditions makes it difficult to compare data from the literature. Because of these difficulties, only the order of magnitude of parameters are usually available. We note that it is particularly difficult to obtain this kind of data for signaling pathways, mainly because of the low number of molecules and the fast kinetics.

If the parameters of a model cannot be directly measured or found in the literature, their values can be indirectly found by fitting the model as well as possible to existing data, e.g. time series measurements of concentrations.

In general, the parameter estimation problem can be formulated as a minimization of an error function over the parameters. This function is typically a measure of goodness-of-fit to data. In Paper 1, we use the following error function for a single time series X_j,

$$\frac{1}{2}\sum_i \left(\frac{X_j(t_i) - \hat{X}_j(t_i)}{\sigma_j(t_i)} \right)^2 \tag{12}$$

where i indexes the measurement points, where X_j denotes values obtained from the model, where $\hat{X}_j$ denotes experimental values, and where σ_j is the standard deviation modeling the inaccuracy in the experimental values. The total error of the model is calculated by summing the errors for all variables in all experiments. Assuming independent and normally distributed measurement errors, (12) corresponds to the negative log likelihood, L, of observing the data given the model.

The input to a parameter estimation method is typically a set of time series data. We distinguish between complete data, where data for all variables is available, and incomplete data, where data for some variables is missing. In Paper 1, we consider artificial complete time series data from one or several experiments. As an example, one experiment for the metabolic test system is specified by the input functions I_1 and I_2 and includes time series for

[1] Literally "in glass." Refers to tests or reactions taking place outside a living organism, on a microscope slide, in a test tube, etc.

[2] Literally "in life." Refers to tests or reactions taking place in a living organism.

$S_3 - S_7$. In Papers 2 and 3, on the other hand, we use real data that is incomplete.

A potential problem in parameter estimation is that it may be impossible to unambiguously determine all parameters from the considered data set. One source for problem of ambiguity is incomplete data. Using an algorithm for algebraic observability (Sedoglavic 2002) we can test whether the parameters of an ODE model in theory can be identified for different sets of in- and output parameters/variables. For a model that cannot be identified, infinitely many values of the parameters can fit the observed data. Hence, an extended set of input and/or output variables or parameters is required to obtain observability. On the other hand, if the observability test suggests that the parameters are observable, it is important to note that this holds for ideal data, but may not hold for a realistic data set.

Another source for problem of ambiguity is noisy data. For instance, the parameters V_{max} and K_M in a Michaelis-Menten reaction (4) are difficult to estimate from a noisy data set in which the substrate concentration, S, is much lower than K_M. One way to solve this is to include additional experiments where higher substrate concentrations are considered. This can be achieved either by a different input signal or by employing genetic modifications.

Finally, it is the case that minimization of (12) is a hard optimization problem for models of realistic size and complexity, especially when the ODEs are non-linear in the parameters. In particular, the error function typically has several local minima. However, the complexity of the search can be reduced by considering parameter bounds and/or constraint functions. For instance, in Papers 1 and 3 we constrain the parameters by lower and upper bounds.

In the next section we will discuss different ways of estimating parameters.

4.1 The basic method

A general method to minimize (12) is:

1. Try a parameter set.
2. Evaluate the error function.
3. Update the parameters according to some rule and then repeat from step 2 until termination according to some criterion, e.g. that the error is sufficiently stable.

This method follows the standard way of minimizing a function, although we note that the derivatives of the function with respect to the parameters are also required by some methods. Typically, a local minimum of the error function is found, since the parameters are iteratively modified in small steps. Hence, only if the initial parameters can be sufficiently well guessed we can expect to find a global minimum. An example of a local method is the steepest descent method.

For ODE models the evaluation of the error function is usually slow, since it requires the entire model to be simulated for each experiment. Since this has to be repeated many times, the overall method is computationally intensive for realistic problems. Here we also note that the derivatives of the error function with respect to the parameters can not be derived analytically.

In addition to local methods, there are global methods, which are designed to avoid local minima of the error function. We note, however, that no optimization method can guarantee finding a global minimum and that global methods typically require more computational time than local methods. Some examples of global methods are simulated annealing and evolutionary algorithms (see also Pintér (1996) and Press et al. (1993)).

Concerning the particular application to biochemical modeling, Moles et al. (2003) evaluate seven different global methods on a biochemical model including 36 parameters and simulated data from that model. The difficulty of this particular problem is that the search space is large and that the ODEs are highly non-linear in the variables as well as in the parameters. Of the seven methods used, one was deterministic and the remaining six were stochastic methods. Only two of the methods obtained parameters close to the true values. Both these methods are based on evolutionary computation. Basically, in evolutionary computation, a population of parameter vectors (individuals) are maintained. For each individual the error is calculated and a new population of the same size is created by recombining the best individuals of the current population. This procedure is then repeated according to the basic algorithm. In the study by Moles et al. the best method, Evolution Strategy using Stochastic Ranking (Runarsson et al. 2000), obtained the true parameters within 16% relative error using about 39 hours computational time (Pentium III, 866MHz).

In Papers 1-3 we use several different approaches to estimate the parameters. The choice of method is largely dependent on the complexity of the models and the requirements on computational efficiency. We first consider the two osmoregulation models in Papers 2 and 3:

- The simple model in Paper 3 contains ten parameters. We use various experimental data to constrain the search space by lower and upper

parameter bounds. This gives a parameter estimation problem of relatively low complexity and we can use a global minimization technique. Since we only do this once, the computational efficiency of the method is not a major issue.

- The situation is much worse for the detailed model of Paper 2 because of the high dimension (70 parameters) of the search space. To partly overcome this difficulty we study subparts of the model in isolation. As an example, the steady state characteristics of one sub-model may indicate what parameter values that result in a realistic signal amplification. Besides, for many of the parameters plausible values can be found in the literature. The manually selected parameters are then fine-tuned with respect to time series experimental data. Specifically, the parameters were randomly perturbed using a normal distribution with mean at the manually selected values. Several such perturbed parameter sets were evaluated and the set resulting in lowest error was chosen. Due to the complexity of the model, the standard deviation of the perturbations must be selected relatively small. For that reason, this method falls in between local and global optimization. As in Paper 3, we only estimate the parameters once.

In Paper 1 we have a different and more challenging situation since the model structure is unknown. We then have to estimate the parameters of many different model structures to find the best one. It is therefore difficult to use the general parameter estimation method and at the same time obtain a realistic computational time. To overcome this problem, we applied a decomposition approach of considering one equation at a time. A method that completely follows this approach is the so-called derivative approach. Since this approach has been an important starting point for our work on Paper 1, we describe it in detail in the following section.

4.2 The derivative approach

Under certain conditions one can speed up the parameter estimation dramatically by considering one equation at a time and not performing any simulations at all. This simplified approach, the *derivative approach* (see e.g. Englezos et al. (2001) and Voit (2000)), is based on the least-squares method (see e.g. Johnson et al. 1992). The method has one advantage - its computational speed, but several disadvantages:

- It is only working for complete data sets, that is, every single variable must be measured.

- The method requires estimates of not only variables but also derivatives of the variables at arbitrary time-points. We note that this problem can be reduced by considering different types of data preprocessing like spline methods (de Boor 1978, Voit et al. 2004).
- The function it minimizes is usually not the function that we want to minimize, e.g. (12). Instead, the residual of the least-squares is minimized as will be further explained below.

To illustrate the derivative approach we consider the linear model (6) with the input signal (9). The parameters to estimate are k_1 and k_2. Given a complete data set, that is time series data for A_1 and A_2, we can apply the derivate approach.

In principle, by estimating $A_2'(t)$ and all concentrations on the right hand side from experimental data, (6) gives us a number of linear equations. Each time point in the experiment where A_2 is measured gives one such equation. We let $\widehat{A_i}(t)$ denote experimental data of variable A_i at time t. The full system can then be written

$$\underbrace{\begin{pmatrix} \widehat{A_1}(t_1)I(t_1) & -\widehat{A_2}(t_1) \\ \vdots & \vdots \\ \widehat{A_1}(t_m)I(t_m) & -\widehat{A_2}(t_m) \end{pmatrix}}_{M} \underbrace{\begin{pmatrix} k_1 \\ k_2 \end{pmatrix}}_{\mathbf{k}} = \underbrace{\begin{pmatrix} \widehat{A_2'}(t_1) \\ \vdots \\ \widehat{A_2'}(t_m) \end{pmatrix}}_{\mathbf{b}} \tag{13}$$

where t_1 and t_m refer to the first and last experimental time point respectively. The system of equations is over-determined and can be solved by the least-squares method, which minimizes the Euclidean norm between $M\mathbf{k}$ and $\mathbf{b}$, that is

$$\min_{\mathbf{k}} \sum_i \left(\widehat{A_2'}(t_i) - (k_1 \widehat{A_1}(t_i)I(t_i) - k_2 \widehat{A_2}(t_i)) \right)^2 \tag{14}$$

If the column vectors of M are linearly independent, the solution to the least-squares problem is obtained from the linear system

$$M^T M\mathbf{k} = M^T \mathbf{b}. \tag{15}$$

For models including several ODEs, we repeat the least squares method for each individual variable in order to estimate all parameters of the model.

Figure 7 illustrates how the parameter estimates of (6) are dependent on the number of data-points in the time series. We note that although we have

noise-free data, the least-squares method fails to estimate the parameters correctly when we have few data-points. In this test, the derivatives were estimated by the central difference as

$$\widehat{A_2'}(t_i) = \frac{\widehat{A_2}(t_{i+1}) - \widehat{A_2}(t_{i-1})}{t_{i+1} - t_{i-1}}. \tag{16}$$

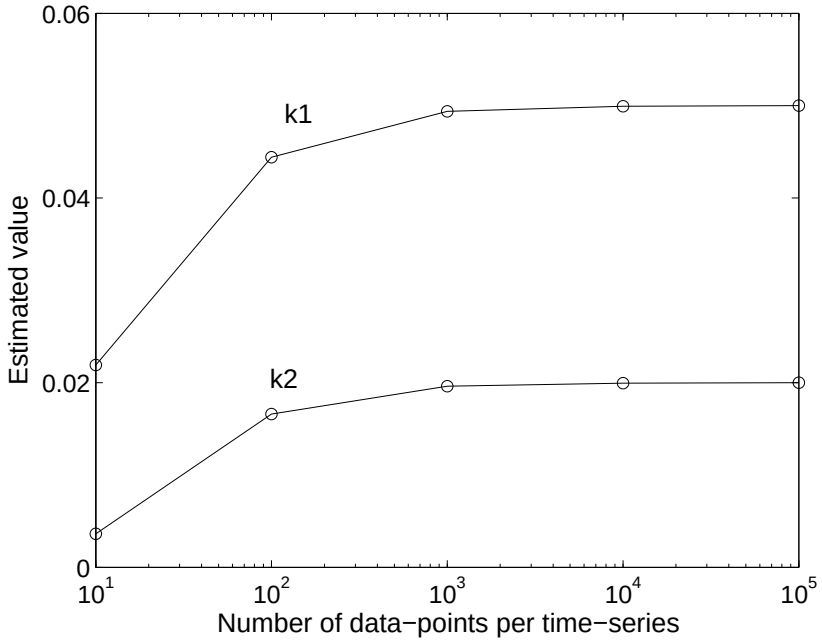

Figure 7: *Parameters estimated using the derivative approach on the model of (6) with different amounts of simulated data that is uniformly distributed. The true values are* $k_1 = 0.05$ *and* $k_2 = 0.02$.

Although the precision can be increased by using more accurate interpolation methods, such as smoothing spline interpolation (de Boor 1978), the general behavior of this plot will remain.

Besides the problem of estimating the derivatives, we have the problem that the minimization function (14) is not the same as our original minimization function (12). Only for perfect data both (14) and (12) evaluate to zero for the correct parameters. However, for noisy data the two functions typically evaluate to different values and hence are minimized for different parameters.

We finally note that the derivative approach can be generalized by considering non-linear least-squares (Marquardt 1963, Press et al. 1993). This is needed if there are reactions that are non-linear in the parameters, like the

Michaelis-Menten kinetics. Non-linear least-squares algorithms require an initial guess of the parameters and it is therefore common to re-start the procedure with different initial guesses.

4.3 Our method for parameter estimation

The particular method we apply for parameter estimation in Paper 1 tries to combine the computational efficiency of the derivative method with the high accuracy of the basic method. Our method is based on two main ideas:

- Each ODE is considered separately as in the derivative method. This increases the computational efficiency compared to the basic method.
- Simulation is employed as in the basic method. However, we only simulate the single variable under consideration and not the complete model. This increases accuracy compared to the derivative method.

When simulating a single ODE, all variables on the right-hand side of the equation except the one that is simulated must be determined in some way. A natural first approach is to employ interpolated data. However, in an iterative search for the parameters (as the basic method) it can happen that simulated data from the best model gives a better performance than interpolated data. Ideally, we can then estimate the parameters with high accuracy. This idea is used in Paper 1 and it is the main reason why the parameter estimates are so good given the relatively short computational time.

Using our algorithm for the parameter estimation problem of the linear model (6) with the input signal (9) considered in the previous section (see Figure 7) we can obtain the correct parameters with only few (< 10) data-points per time-series. For more advanced examples we refer to Paper 1.

For biological systems, it is common that experimental time series data is not available for all variables in the model, while our approach requires a complete data set. Missing data is a fundamental algorithmic difficulty and we are typically referred to the basic method for parameter estimation. However, using methods conceptually based on the Expectation-Maximization (EM) algorithm (Dempster et al. 1977), which is a standard statistical algorithm for treating incomplete data problems, we are able to estimate the parameters for certain incomplete data sets and still keeping the strategy of considering one variable at a time. To exemplify this, we consider the model presented in Figure 8 and a data set including three time series experiments with 8 data-points per variable and experiment. By removing all data from e.g. variables B_1 and B_2 we can still estimate the 16 parameters using our strategy.

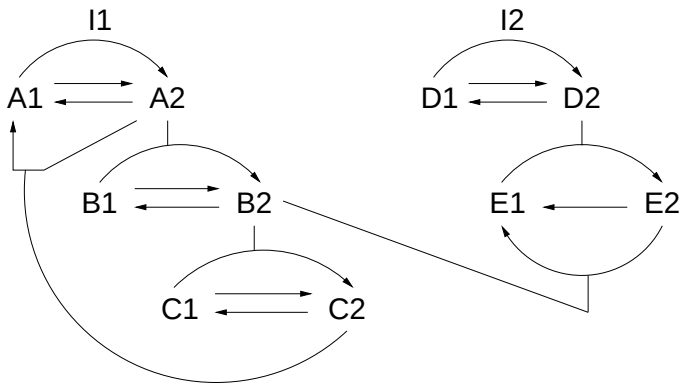

Figure 8: *Model of a signaling network with two input functions and ten variables. All reactions follow linear (2) or bilinear kinetics (3). The total number of parameters is 16.*

4.4 Parameter sensitivity

To evaluate the reliability of the parameters obtained by a parameter estimation method it can be useful to perform a sensitivity analysis. The sensitivity of the error function to a given parameter can be calculated as the partial derivative of the error function with respect to that parameter. A sensitivity analysis can reveal parameters that are undetermined from the considered data set. For instance, some parameters in the model of Paper 3 could not be estimated with a high degree of confidence. In principle, the sensitivities can also be used in an estimation method in order to more efficiently search the feasible region.

However, we would also like to point out that biological systems tend to be robust with respect to parameter variations (Eldar et al. 2002). Therefore, it can be difficult to estimate parameters with high accuracy from only wild-type experiments. Instead, various system modifications, like deletions in order to break up feedback loops, can be useful in order to find the parameters.

5 Model selection

Model selection is the problem of how to select the structure, i.e. the form of the ODEs. We will assume that we can estimate the parameters in any model structure using one of the methods discussed previously. As in parameter estimation, we minimize a function, e.g. (12), but we now minimize it over both the structures and the parameters. We will refer to the problem of finding both the structure and the parameters of a model as *model identification*.

We would like to point out that it is generally much more challenging to identify the structure of a model than to estimate the parameters in a known model structure. There are several reasons why this is a difficult problem. One reason is the difficulty to define the problem in such a way that a model with reasonable complexity is selected. We discuss this topic in Section 5.1. Another reason is the problem of model ambiguity as will be discussed in Section 5.2. A third reason concerns the problem of performing the selection, mainly due to the combinatorial increase in possible model structures when increasing the number of variables. We discuss manual and automatic model selection in Sections 5.3 and 5.4, respectively.

5.1 Model complexity

The purpose of a model is usually to explain available data sufficiently well, and to predict the behavior of the real system. When manually building a model one usually starts from a simple model and then incrementally adds details to the model, intuitively matching the complexity of the model with its purpose and available data.

In general, a too simple model lacks validity and fails to capture the trends in data. We refer to this as underfitting to data. On the other hand, a too complex model, e.g. including several parameters, tends to have a good fit to data, since it has many degrees of freedom and can be fitted to noise as well as to regularities in data. We refer to this as overfitting to data and note that these models typically give weak predictions.

There are different ways of dealing with model complexity. Cross-validation and bootstrapping are both methods for estimating the error based on resampling (Zucchini 2000). In k-fold cross-validation, the data set is divided into k subsets of equal size. The model error is then calculated k times, each time leaving out one of the subsets in the parameter estimation, but using only the omitted subset to calculate the error function. In bootstrapping, instead of repeatedly analyzing subsets of the data, we repeatedly analyze subsamples of the data. Here, each subsample is a random sample with

replacement from the complete data set.

A different approach to avoid unnecessarily complex models is to penalize complexity in the error function. A common way is to add a penalty term that is typically a function of the number of parameters and/or the number of data-points (Zucchini 2000). It is an open research question how to choose this function in a best way for a particular application (Crampin et al. 2004). Common examples include Akaike Information Criteria (AIC, Akaike 1973), Minimum Description Length (MDL, Rissanen 1978) and Bayesian Information Criteria (BIC, Schwarz 1978). In Paper 1, we use the following error function for a single time series

$$-L + \lambda K \tag{17}$$

where L is the log likelihood according to (12), λK is the penalty term including a problem-specific parameter λ and the number of model parameters K. In model selection, the effect of this penalty can be observed by assigning a very low or high value to λ, typically resulting in over- or underfitting, respectively.

5.2 Model structure ambiguity

In model selection it is important to be aware of the problem of ambiguity in the model structure. We illustrate this point by presenting examples when two different biological models create the same or similar experimental data.

The first example considers the biochemical models presented in Figure 9. In model I, two compounds (A_2 and B_2) both activate compound C, while in model II only B_2 activates C. As indicated in the figure, the parameter (k) of the catalysed reaction from C_1 to C_2 in model II is the sum of the corresponding parameters (k_3 and k_4) in model I. All other parameters are the same in the two models. If we consider a wild-type experiment, the two models will produce exactly the same experimental data for the variables. That is, the data does not unambiguously derive from one model and it is impossible to distinguish the true model.

However, by including an additional experiment where either A or B is deleted, the set of all data will unambiguously derive from either model I or model II. This is a powerful experimental technique that for instance was used to reveal the basic structure of the HOG signaling pathway in yeast (Maeda et al. 1995).

Another example of models that output similar data for an experiment are models that differ in reaction mechanisms. For instance, we reconsider the

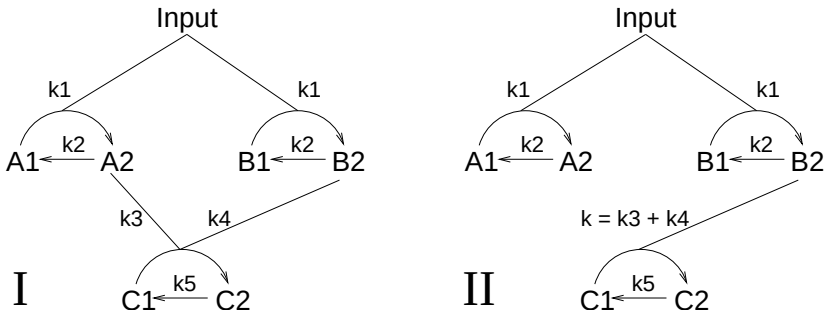

Figure 9: *Experimental data of models I and II are identical. Compounds A, B and C all exist in two different states and all reactions are assumed linear or bilinear and the reaction constants are indicated as k's.*

model given in (6) together with a modified version of that model where I is squared. The modified model reads

$$A_2'(t) = -A_1'(t) = k_1 A_1(t) I^2(t) - k_2 A_2(t) \tag{18}$$

where $k_1 = 0.05$ and $k_2 = 0.02$ are the same for both models.

These two models output very similar data for certain input signals. For instance, the input function (9) is depicted as Input I in Figure 10. Data is similar but there are actually two kind of differences: the initial steady states and the form of the rising curves of the two models differ slightly. However, for moderate levels of measurement noise, it becomes very difficult to uniquely distinguish them from each other.

On the other hand, by applying a different input signal we can obtain data with much better discriminating power. As an example, an input function that steps from 0.01 to 0.2 is illustrated as Input II in Figure 10. In this case, the separation of the curves is evident, and at the end of the simulation A_2 of the original model (6) has more than four times higher concentration than A_2 of the modified model (18).

To conclude, we have shown two examples where models output equal or similar data and cannot be distinguished from each other using the given data. One possible solution is to provide a more extensive data set, for instance by using a different choice of input function and/or a modified system. However, we note that it is non-trivial to determine how much and what kind of data is needed to give uniqueness.

In the first example, the ambiguity could also be resolved mathematically by using an error function penalizing model complexity such as (17). In the

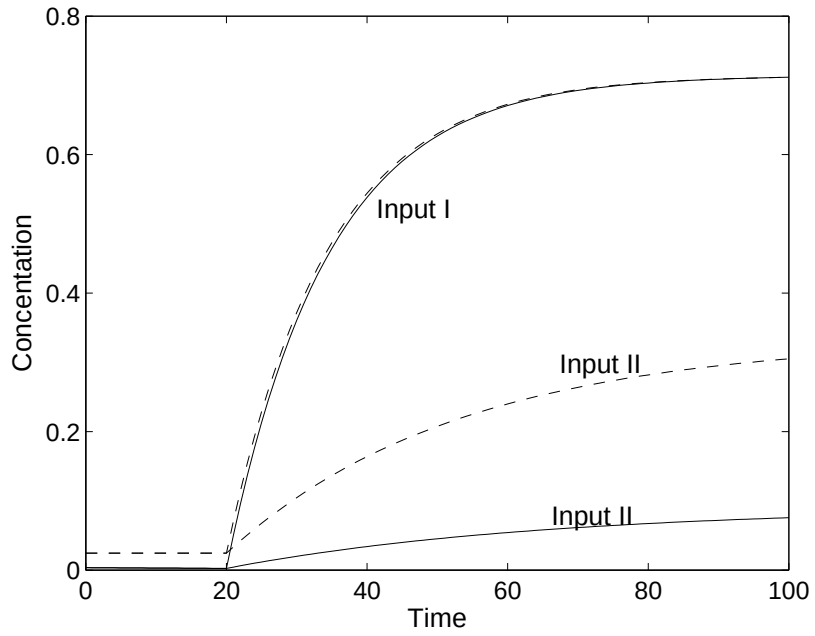

Figure 10: *Simulation of variable A_2 in the models given in (6) (dashed lines) and (18) (solid lines) using two different input signals: Input I steps from 0.01 to 1 at $t = 20$ and Input II steps from 0.01 to 0.2 at $t = 20$.*

second example, however, this is hardly sufficient since both models have the same complexity according to (17).

5.3 Manual model selection

A common approach in modeling is to manually select the model structure and parameter constraints and then estimating the parameters automatically. In principle, when manually creating an ODE model of a biochemical system, one can consider any form of equations in the model structure. However, usually the modeler tries to employ standard reaction types (like the Michaelis-Menten kinetics) that can be derived from plausible reaction mechanisms of the considered interactions. We note that S-systems are an exception to this.

It is difficult to present a general methodology for how to construct a model of a particular system. Instead, we exemplify by giving a brief description of the methodology used when modeling the osmoregulation system. In very simple terms, this system involves a signaling pathway working as an

information carrier in the cell. The sensor of the pathway is activated by reduced turgor pressure and the output of the pathway initiates glycerol production that works as a feedback loop and causes turgor pressure to regain. For an overview of our models we refer to Section 6 and for a more elaborate description we refer to Papers 2 and 3.

Initially, our objective was to model the signaling pathway in isolation. The basic structure of this signaling pathway was described in the literature. Besides, we found models in the literature of similar systems in evolutionary closely related species. Given this information we could assign plausible kinetic equations to the reactions. However, since the signaling pathway interacts with other systems it was difficult to model the pathway in isolation and we therefore had to extend our modeling scope.

One point of interaction involves an environmental stimulus that serves as input signals to the signaling pathway. In the beginning of the project, the exact nature of this environmental stimulus was not known. Among others, Gustin (1998) speculated in turgor pressure and this was later experimentally verified by Reiser et al. (2003). From thermodynamics it is known that turgor pressure is related to osmotic pressure and volume. Hence, in order to model the input signal in a realistic way, a biophysical model including at least these variables should be considered.

Another point of interaction exists between the output of the signaling pathway and metabolic pathways of glycerol production. In principle, it is easy to include metabolic pathways in a model, since a lot of modeling efforts have been done in that field. A challenge, however, is to select a proper modeling scope and level of detail. In Paper 2 we use an existing model from the literature.

To conclude, in order to model the signaling pathway it was necessary to extend the modeling scope by including two additional modules: one representing the biophysical changes of the cell and one representing the glycerol production. Given the biophysical description we could explicitly link glycerol production to the turgor pressure and consequently to the input signal of the pathway. Once this was established we could simulate various experiments in the computer.

Since we propose two different models of the same system it is interesting to compare these with each other. The models share the main characteristics and give the same qualitative predictions. Instead, the difference lies in the level of detail at which the processes are modeled. In the simple model hardly no molecular details are included, while the detailed model takes into account a considerable amount of available structural information of the pathways. We refer to Paper 3 for a discussion of qualitative aspects of the models with respect to their different complexities.

In principle, it would be interesting to perform quantitative comparative analyses of different models, using a complexity measure like AIC, BIC or MDL. Such a measure would reveal to what price of increased complexity it is reasonable to increase the goodness-of-fit to data. Unfortunately, there is no single accepted measure for this kind of problems. Besides, it is very difficult to compare models when different data sets have been employed in the construction of the models. For instance, the detailed model of Paper 2 is based on the currently identified structure of the system, and hence, implicitly uses the data from which the structure is determined. Such data is typically not obtained from time-series experiments of protein concentrations but rather from protein-protein interaction experiments and experiments measuring cell growth in various mutant strains. Although this kind of data can be directly employed in model identification, it may be difficult and tedious to extract the data from the literature. We would also like to point out that the structural information obtained from these experiments only tells whether variables interact, not the mechanism of interaction.

5.4 Automatic model selection

We will now describe some general principles for automatic model selection algorithms. Typically, the inputs to a model selection algorithm are:

- Time course data for the variables. In Paper 1, we consider the same input of data as previously described for the parameter estimation. However, we note that different types of data, like steady-state data or protein interaction data, can be employed as well.
- An initial structure including all variables and potentially known interactions. As a base case, the initial structure can be assumed empty, and hence, all interactions should be identified. As an example, for the metabolic test system we would have $S_3' = S_4' = S_5' = S_6' = S_7' = 0$.

To define the model selection problem we must also specify an error function, like (17), and a *search domain* or *model space*, which defines the space of possible models. The search domain can be obtained by defining reaction building blocks that may be used in the model. For instance, the metabolic test system in Figure 5 contains two different types of reactions: the Michaelis-Menten reaction (4) and the Michaelis-Menten reaction with non-competitive inhibition (5). Therefore, to identify correctly the metabolic test system the search domain must at least contain these two reaction types. In Paper 1, we consider not only these two types, but also a spontaneous state transition with linear kinetics (2) and an enzymatic reaction with bilinear kinetics (3). The resulting search domain is given in Table

1. We note that the identification problem becomes more difficult for a large search domain. In a real situation, the true reaction types are unknown and a plausible guess of the search domain must be made.

Possible reactions for S_3, S_4 and S_5					
S(S_3)	S(S_4)	S(S_5)	B(S_3,I_1)	B(S_3,I_2)	B(S_3,S_4)
B(S_3,S_5)	B(S_3,S_6)	B(S_3,S_7)	B(S_4,I_1)	B(S_4,I_2)	B(S_4,S_3)
B(S_4,S_5)	B(S_4,S_6)	B(S_4,S_7)	B(S_5,I_1)	B(S_5,I_2)	B(S_5, S_3)
B(S_5, S_4)	B(S_5, S_6)	B(S_5,S_7)	M(S_3)	M(S_4)	M(S_5)
I(S_3,I_1)	I(S_3,I_2)	I(S_3,S_4)	I(S_3,S_5)	I(S_3,S_6)	I(S_3,S_7)
I(S_4,I_1)	I(S_4,I_2)	I(S_4,S_3)	I(S_4,S_5)	I(S_4,S_6)	I(S_4,I_7)
I(S_5, I_1)	I(S_5,I_2)	I(S_5, S_3)	I(S_5, S_4)	I(S_5, S_6)	I(S_5,S_7)

Possible reactions for S_6 and S_7					
S(S_6)	S(S_7)	B(S_6,I_1)	B(S_6,I_2)	B(S_6,S_3)	B(S_6,S_4)
B(S_6,S_5)	B(S_6,S_7)	B(S_7,I_1)	B(S_7,I_2)	B(S_7,S_3)	B(S_7,S_4)
B(S_7,S_5)	B(S_7,S_6)	M(S_6)	M(S_7)	I(S_6,I_1)	I(S_6,I_2)
I(S_6,S_3)	I(S_6,S_4)	I(S_6,S_5)	I(S_6,S_7)	I(S_7,I_1)	I(S_7,I_2)
I(S_7,S_3)	I(S_7,S_4)	I(S_7,S_5)	I(S_7,S_6)		

Notation	
$S(A)$	Linear transition (2) with substrate A
$B(A, E)$	Bilinear reaction (3) with substrate A and enzyme E
$M(A)$	Michaelis-Menten reaction (4) with substrate A
$I(A, B)$	Michaelis-Menten reaction (5) non-competitively inhibited by B having the substrate A

Table 1: *The search domain for the metabolic test system. We assume that mass conservation constraints are known, that is the sum of S_3, S_4 and S_5 as well as the sum of S_6 and S_7 are constant. In this notation, the true ODE of variable 4 is $S_4'(t) = I(S_3, I_1) - M(S_4)$.*

For problems of realistic size an exhaustive search over all possible model structures is not feasible due to the combinatorial explosion of possible model structures. For instance, given an upper limit of four reactions per variable (something that we do not assume in Paper 1) there are about $2.7 * 10^6$ possible structures only for variable S_3 in the metabolic test system. For this reason, it is very difficult to find algorithms that solve the model selection problem in realistic time. To make the best of the situation, one typically employs heuristic algorithms that at least are able to propose a model that is close to the real system.

In order to reduce the complexity of model identification we can constrain the problem in different ways. For instance, we can include verified interactions in the initial structure of the model and we can restrict the search domain in different ways. Besides, the search space for the parameters can be restricted as discussed in the previous section. We also note that identification becomes easier the more the system has been experimentally disturbed by various input signals and system modifications.

A general heuristic way of searching the best model is to divide the search into two steps: (1) a structure search and (2) a parameter estimation method for a given structure. In this way, we obtain the following approach:

1. Try a structure from the search domain.
2. Estimate the parameters in this structure and evaluate the error function.
3. Update the structure according to some rule and then repeat from step 2 until termination according to some criterion.

The model selection algorithm that we propose in Paper 1 is based on this approach.

Finally, we want to remind the importance of distinguishing between what information is possible to extract from a given data set and how well the algorithm performs on that data set. In particular, given sufficient data to unambiguously define the correct model and an ideal identification algorithm, one can find the correct model. However, for this kind of problems, a heuristic approach may fail since it does not perform an exhaustive search. Hence, the heuristic nature of an algorithm may give an attractive computational time but also limits the performance on data sets that are small but nevertheless unambiguously define the correct model.

5.5 Our model selection algorithm

In Paper 1 we suggest a model selection algorithm, in which we employ the general heuristic method presented in 5.4 and consider one variable at a time, as also done in our parameter estimation algorithm. We build the structure incrementally and always maintain a current model with structure and parameters. As a base case, the initial model is trivial with all variables independent of each other. Our model selection algorithm can be described as follows.

For each variable we do the following:

1. Calculate the error of the initial model.
2. For each possible test reaction from the search domain:
   (a) Temporarily add the reaction to the model.
   (b) Estimate the parameters.
   (c) Calculate the error.
3. If a better model was found in step 2, use this model as the new best model.
4. Remove reactions if this results in a lower error.

This process of considering all variables in turn is repeated until no better model is obtained. Hence, for each iteration over all variables, a reaction may be added to each equation and any of the existing reactions might be removed. A reaction is removed if it improves the fit to data (measured by the first term of (17)) less than it increases the complexity of the model (measured by the penalty term of (17)). We note that this heuristic algorithm can not guarantee a global minimum of the error function, and hence, as in the parameter estimation we may obtain a local minimum, where the structure and/or parameters are incorrect.

In an attempt to illustrate the progress of the model selection algorithm we consider the metabolic test system for the noise-free data set of 12 experiments employed in Paper 1. We consider the search domain given in Table 1 and we also use the same notation as in that table. The true model structure that we search for is

$$S_3'(t) = M(S_4) + M(S_5) - I(S_3, I_1) - I(S_3, I_2) \tag{19}$$

$$S_4'(t) = -M(S_4) + I(S_3, I_1) \tag{20}$$

$$S_5'(t) = -M(S_5) + I(S_3, I_2) \tag{21}$$

$$S_6'(t) = -S_7'(t) = -M(S_6) + I(S_7, S_3) \tag{22}$$

However, this model is from now considered unknown to the algorithm, and the initial structure is empty, that is $S_3' = S_4' = S_5' = S_6' = S_7' = 0$. The only information we use is the set of time series experiments (data from simulation of the true model) with various input functions I_1 and I_2.

After one iteration over all variables the following model is obtained:

$$S_3'(t) = -B(S_3, S_7) \tag{23}$$

$$S_4'(t) = I(S_3, I_1) \tag{24}$$

$$S_5'(t) = -B(S_5, I_2) \tag{25}$$

$$S_6'(t) = I(S_7, S_3) \tag{26}$$

$$S_7'(t) = -I(S_7, S_3) \tag{27}$$

We note that the ODE of S_3 includes a bilinear reaction that does not belong to the true structure. The same holds for the ODE of S_5 where $B(S_5, I_2)$ is a false positive reaction, while true positive reactions are added to all other ODEs.

We repeat the procedure for all variables and obtain:

$$S_3'(t) = -B(S_3, S_7) + B(S_4, I_1) \tag{28}$$

$$S_4'(t) = I(S_3, I_1) - M(S_4) \tag{29}$$

$$S_5'(t) = -B(S_5, I_2) + I(S_3, I_2) \tag{30}$$

$$S_6'(t) = I(S_7, S_3) - M(S_6) \tag{31}$$

$$S_7'(t) = -I(S_7, S_3) + M(S_6) \tag{32}$$

Hence, after the second iteration, the ODE of S_3 contains two false positive reactions, the ODE of S_5 contains one false positive and one true positive reaction and the structure of the other ODEs are correctly identified.

Iteration 3 gives:

$$S_3'(t) = B(S_4, I_1) - I(S_3, I_2) \tag{33}$$

$$S_4'(t) = I(S_3, I_1) - M(S_4) \tag{34}$$

$$S_5'(t) = I(S_3, I_2) - M(S_5) \tag{35}$$

$$S_6'(t) = I(S_7, S_3) - M(S_6) \tag{36}$$

$$S_7'(t) = -I(S_7, S_3) + M(S_6) \tag{37}$$

Here we make two observations: First, the addition of the true positive reaction $I(S_3, I_2)$ to the ODE of S_3 results in a model in which the previously added reaction $B(S_3, S_7)$ was unnecessary and could be removed. This is due to the non-greedy strategy of the search: a reaction that has been added might fall off in later stages. Similarly, the addition of the true positive reaction $M(S_5)$ to the ODE of S_5 pushed out the false reaction $B(S_5, I_2)$. Second, no reactions were added to the ODEs of S_4, S_6 and S_7. In other words, the cost in increased complexity of an additional reaction was higher than the (potential) gain in goodness-of-fit due to more parameters.

In the following iterations only the structure of S_3' is modified. The true model is obtained after a total of 7 iterations.

This example illustrates how the heuristic search incrementally builds up the true structure of the metabolic test system. We note that the search route is not only dependent on the error function and parameter estimation routine but also on the specific data set employed.

5.6 Model identification algorithms in experimental planning

In this section we consider the potential use of model identification algorithms in experimental planning. Specifically, we outline a computer-based planning methodology where a model identification algorithm plays an important role.

In the area of molecular biology experimental plans are traditionally made manually by professionals with great biological insight and experience. Basically, the next experiment is determined by the current knowledge of the system, the current hypotheses about the system and the currently available experimental techniques. Based on the outcome of the experiment, the knowledge of the system as well as the hypotheses are modified and new experiments are thus iteratively proposed and executed. Our computer-based experimental planning method mimics this iterative exploration of a biological system.

We assume the functionality illustrated in Figure 2, in which we can both simulate data from a model and identify a model from data. Before we discuss experimental planning we note that this functionality also offers several more elementary operations:

- Simulation of one or several models, e.g. for manual evaluation of their quality.
- Evaluation of the error function to determine to what extent a new experiment provides new information.

- Using the identification method we can ask what type of experiments and what amount and accuracy of data are needed in order to identify a certain model.

Our experimental planning method more or less includes these elementary operations and we will now describe the method in more detail.

We exemplify the method on an artificial cell signaling pathway presented in Figure 11. This model corresponds to the true system that we aim at finding.

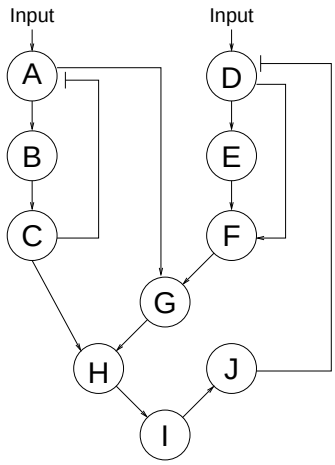

Figure 11: *Model of a signaling pathway used to illustrate the experimental planning method. The model includes ten proteins, each of them existing in two different states (inactive and active), and several interactions including positive and negative feedback loops.*

We consider the following scenario:

- We have a minimal base model, M^0, containing all compounds of interest connected by previously verified interactions. This model corresponds to our current knowledge of the system. In our example, M^0 does not contain any reactions at all, see the left part of Figure 12.

- We have performed a set of experiments, E^0. Typically, an experiment is a certain genomic background in combination with a certain input function. For instance, we have performed one experiment on a wild-type cell using a step input signal.

- We have the potential of executing several experiments, denoted E, see Table 2.

- We have a hypothetical model, M, that differs from M^0. In our example, the hypothetical model is given in the right part of Figure 12. We want to test this hypothesis experimentally.

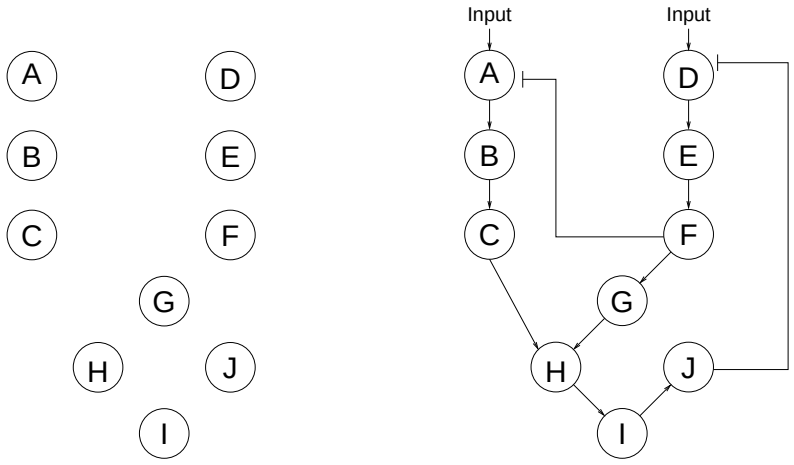

Figure 12: *Left: base model M^0. Right: Hypothetical model M.*

Experiment	**Genomic background**	**Input signal**
1	wild-type	pulse
2	ΔA	step
3	ΔB	pulse
4	ΔD	step
5	$\Delta C \Delta J$	pulse
$\vdots$	$\vdots$	$\vdots$

Table 2: *An example of E, the set of possible experiments to perform.*

We specify the experimental planning problem as follows: extend the data set E^0 by the smallest possible set of experiments from E such that we can reject or verify our hypothesis M. To reject M it is enough to execute an experiment for which data is sufficiently distinct, measured by some metric, from simulated data of M. Verifying that M is the unique model from a given data set is a much more difficult task.

A powerful approach to deal with this problem is to employ the model identification algorithm. In this way, we can search a set of experiments that uniquely identifies the hypothetical model. The method works as follows.

For every possible experiment $e \in E$, we do the following:

1. Simulate e from M.

2. Temporarily merge simulated data from 1 with the real data set E^0.

3. Run the model identification algorithm with the initial model and the extended data set.

4. Evaluate how close the output model from step 3 is to M. A very simple measure of similarity is the number of similar interactions minus the number of non-similar interactions.

The experiment from E corresponding to the best result in step 4 is the suggested experiment to perform. Ideally, this experiment is executed in the laboratory and we can either reject M or we have uniquely identified M. We note that any simulated experiment in step 1 obviously implies M, but that only some experiments may uniquely identify M.

The procedure above can be generalized in the following ways:

- It cannot generally be assumed that one single experiment from E is enough to uniquely identify M. Therefore, the test set E can be extended by including not only single experiments but also combinations of for example two experiments from E. One test case could then include both experiment 1 and experiment 2 from Table 2. However, if all possible combinations are to be tested, this also implies an exponential increase in the number of test cases and thereby computational time.

- The procedure can be repeated for a number of hypothetical models. In this case, the experiments suggested is the union of the experiments suggested for each hypothetical model.

- We can assign costs to the experiments (e.g. corresponding to time or labor) and also include that in the evaluation function of step 4.

To conclude, we suggest a way of generating more efficient experimental plans by including a model identification algorithm in an automatic decision process. In principle, such a planning method would take advantage of the integrated data simulation and model identification functionality presented in Figure 2. We also note that other functionality, such as testing for algebraic observability, would be a valuable complement to simulation and identification. The planning method would probably prove helpful not only in research, but also as a pedagogical tool in education for biologists, as well as for mathematicians and computer scientists. Besides, it could help people from these disciplines to learn more about the other subjects and also facilitate communication between these groups when exchanging ideas.

6 Modeling osmoregulation in yeast

In this section we present our modeling work on osmoregulation in the yeast *Saccharomyces cerevisiae*, which is one of the most well-studied eukaryotic organisms (Sherman 2002).

To understand osmoregulation it is useful to consider a simplified cell, containing a water solution of large molecules (e.g. proteins and sugars) and small inorganic ions. We further assume that the cell membrane is semi-permeable, such that the large molecules are unable to pass the membrane, while water and the small ions can freely pass. In principle, the ions would then have equal concentration inside and outside the cell at equilibrium. However, the large molecules in the cell are often highly charged and attract many small inorganic ions. Therefore, the concentration of ions is greater inside than outside the cell at equilibrium (the Donnan effect, see e.g. Alberts et al. 1994).

Based on this simple cell model we can give a conceptual explanation of two fundamental variables in osmoregulation: osmotic pressure and turgor pressure. On a basic level, osmotic pressure is proportional to the *concentration* of molecules other than water in a solution. Hence, a large protein contributes as much as a small ion to the osmotic pressure. Since the concentration of ions is greater inside than outside the cell at equilibrium, the cell has a higher intra-cellular than extra-cellular osmotic pressure. This causes an outward pressure on the plasma membrane. Due to this difference water will flow into the cell. In isolation, this would cause the cell to swell and potentially lead to cell rupture. This is a fundamental problem that any cell must master. Basic solutions are to actively pump out ions, to actively extrude water or to prevent the cell to swell by a cell wall.

The yeast cell uses the latter solution and has a cell wall with less elasticity than the plasma membrane. Basically, the cell wall resists the expansion of the cell, and creates an inward pressure on the cell contents. This pressure is called the turgor pressure, defined as the difference in the hydrostatic pressure between the inside and the outside of the cell. At equilibrium, the osmotic pressure difference is balanced by the turgor pressure and the cell volume is constant with no net flow of water.

An *osmotic shock* is a sudden increase in the extra-cellular osmotic pressure, for instance due to the addition of salt to the cell medium. The immediate effect on yeast to an osmotic shock involves water outflow and decreasing volume. In this way, a new equilibrium is reached, in which the higher extra-cellular osmotic pressure is balanced by an increased intra-cellular osmotic pressure (due to the reduced volume), and reduction of turgor pressure (due to reduced size of the cell wall). We will refer to these processes as the

biophysical system of the cell.

Generally, the cell strives to keep volume, turgor pressure and relative water content constant and independent of environmental changes. It therefore has a control system responding to these changes by accumulating glycerol and thereby increasing the intra-cellular osmotic pressure in order to regain its previous size (Gervais et al. 2001, Hohmann 2002, de Nadal et al. 2002). This process is called *osmoregulation*.

The control system consists of two main components, as illustrated in Figure 13. First, the aquaglyceroporin Fps1 closes upon hyper-osmotic shock preventing the outflow of glycerol (Tamas et al. 1999, Tamas et al. 2000). Second, the glycerol production is increased in the following way: The osmotic shock activates the High Osmolarity Glycerol (HOG) pathway, see Figure 1. This pathway belongs to the class of Mitogen Activated Protein Kinase (MAPK) pathways that are found in all eukaryotic organisms and are important for transmitting and processing signals from the cell membrane into the cell. Typically, a MAPK pathway consists of a sensing system, a cascade of three tiers of protein kinases and output systems such as transcriptional regulators. Upon activation the MAPK, i.e. the last kinase in the pathway, enters the nucleus and induces transcription. For the HOG pathway, there are at least two independent sensors and one of them, Sln1, has been shown to respond to changes in turgor pressure (Reiser et al. 2003). The other sensor of the HOG pathway, the so-called Sho1-branch, is not identified. Active Hog1 accumulates in the nucleus where it interacts with transcription factors and actively participates in transcriptional activation of target genes. One effect of HOG pathway activity is a metabolic shift towards production of glycerol to balance osmotic changes.

To analyze the different aspects of osmoregulation, genetics and molecular biology are used in numerous ways. Cells are exposed to high osmolarity medium and the response to the hyper-osmotic stress is analyzed. The phosphorylation (activation) state of Hog1 is measured to elucidate the kinetics and the duration of the response. mRNA expression patterns of a few genes dependent on activated Hog1 (such as *GPD1* and *STL1*) are also studied. In order to understand the physiological response to the stress, intra-cellular and total amount of glycerol are measured.

The osmoregulation system in yeast is an interesting target for mathematical modeling for several reasons:

- The system is relatively well-characterized. Several key components are identified, e.g. in the HOG signaling pathway, although we note that other parts are described in less detail, e.g. the transcriptional response.

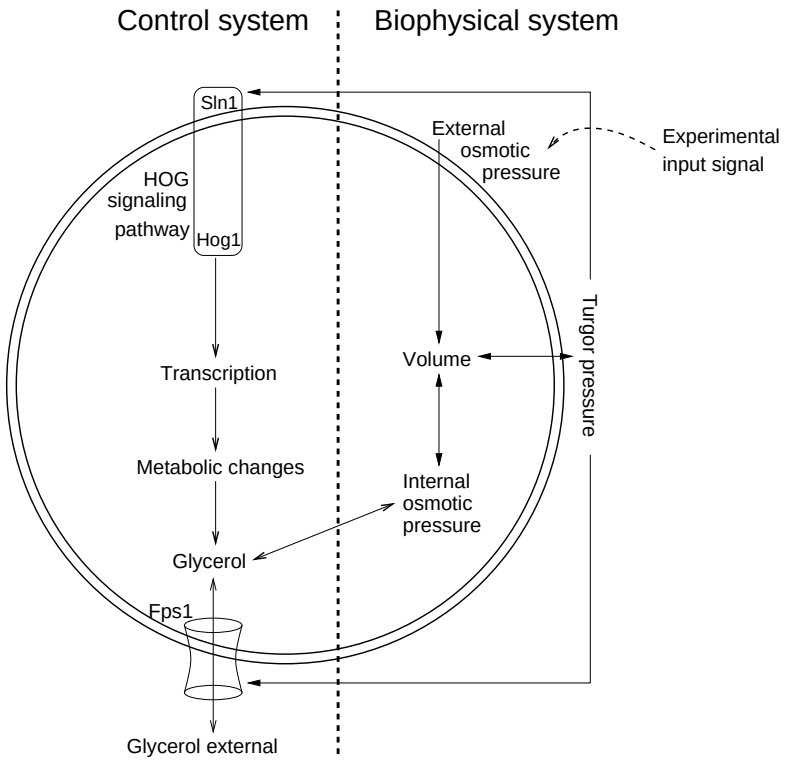

Figure 13: *Key components in osmoregulation in S. cerevisiae. The biophysical system accounts for changes in volume and osmotic pressure. This system can be experimentally disturbed by adding an osmolyte (e.g. NaCl) to the medium. The control system involves the activation of the HOG signaling pathway and the closure of Fps1. Subsequent glycerol accumulation constitutes a feedback loop to balance osmotic changes.*

- The complexity of the system is non-trivial and challenging for modeling studies. In particular, the down-regulation of the HOG pathway has been an open question.
- Basic strategies of cellular adaptation are conserved from bacteria to humans (Somero et al. 1997). Therefore, the system involves several components that are of general biological and medical interest (osmosensors, signaling pathways et.c.).

To mathematically model yeast osmoregulation, it is natural to divide the model into two components. The first component considers the *biophysical system*, involving the changes in osmotic pressure and volume. This component is described in the two following sections. The second component involves the *control system*, such as the HOG signaling pathway and glycerol accumulation, and is covered in Sections 6.3 and 6.4.

6.1 Physics behind osmoregulation

In order to describe the biophysical system mathematically we need proper definitions of the different pressures involved. Therefore, this section introduces the physics behind osmoregulation in a formal way and thus serves as a background to Papers 2 and 3.

The chemical potential of water can be seen as a measure of the effective water concentration in a given area. The value of the water potential is influenced by two factors: (1) the osmotic potential and (2) the pressure potential. The first is affected by the concentration of dissolved molecules of solutes. As the concentration of solute molecules increases, the water potential decreases. The latter takes into account the hydrostatic pressure. If a solution is put under pressure, the water potential increases.

Formally, the chemical potential of a compound describes how the Gibbs energy[3] changes in a system when the compound is added to it (Atkins 1994). The chemical potential for water can be derived as (Levin 1979)

$$\mu_w = \mu_w^*(T) + \overline{v}_w \, p + RT \, ln \, a_w \tag{38}$$

where $\mu_w^*(T)$ is the chemical potential of pure water at temperature T, $\overline{v}_w$ is the apparent molar volume of water [dm^3 mol^{-1}], p is the hydrostatic pressure [Pa], R the universal molar gas constant [J K^{-1} mol^{-1}], T is the temperature [K], and a_w is the water activity in the solution. The latter is defined as (Atkins 1994)

$$a_w = \frac{p_w}{p_w^*} \tag{39}$$

where p_w^* is the vapor pressure of pure water and p_w is vapor pressure of water when it is a component of a solution.

If two regions of water with different potential are separated from each other by a membrane permeable to water but not to the solute (a semi-permeable membrane), there will be a water flow to the region of lower potential (Atkins 1994). This process is called *osmosis* and the water flow, J_w [mol dm^{-2} s^{-1}], is given as (Levin 1979)

$$J_w = \frac{L_p}{\overline{v}_w^2} \left(\mu_w^i - \mu_w^o \right) \tag{40}$$

[3]Gibbs energy is defined as $G = H - TS$ where H is the enthalpy, S is the entropy and T is temperature (Atkins 1994).

where L_p is the hydraulic water permeability coefficient [dm^2 s kg^{-1}], μ_w^i and μ_w^o are, respectively, the chemical potentials of water on the inside and outside of the membrane [kg dm^2 s^{-2} mol^{-1}].

The osmotic pressure, Π, of a solution is the force per unit of surface exerted by the flow of water moving by osmosis from a region containing distilled water to a region containing the solution, the two regions being separated by a semi-permeable membrane (Eckert et al. 1997). For very dilute solutions in which ideal behavior can be assumed, van't Hoff equation relates Π for a solute in a solution to solute concentration and water activity as (Levin 1979, Atkins 1994)

$$\Pi = RT\ \Phi\ B\ n = -\frac{RT}{\overline{v}_w} \ln\ a_w \tag{41}$$

where Φ is the osmotic coefficient, B is the concentration of the solute, and n is number of particles that dissociated from the solute molecule. Taking more than one solute into consideration gives (Eckert et al. 1997)

$$\Pi = RT \sum_j \Phi_j\ B_j\ n_j \tag{42}$$

where j indexes the solutes and Φ_j is the osmotic coefficient of solute j. By dividing the above equation by RT we obtain the osmotic pressure in the unit Osm instead of Pa. For example, a solution containing 0.1 M glucose, 0.3 M KCl, and 0.4 M $MgCl_2$ has an approximate osmolarity of $0.1 + 0.3 * 2 + 0.4 * 3 = 1.9$ Osm assuming osmotic coefficients of 1.

Given the above expressions for chemical potential, flow of water and osmotic pressure, we can derive an expression for the flow of water over a cell membrane in terms of osmotic pressure and turgor pressure (Levin 1979). First of all, (38), (40) and (41) can be combined and simplified to

$$\overline{v}_w J_w = L_p \left(\Pi_t + \Pi_e - \Pi_i\right) \tag{43}$$

where Π_e and Π_i are, respectively, the external and internal osmotic pressure and Π_t is the difference in hydrostatic pressure over the membrane $(p_i - p_e)$, also called turgor pressure.

Turgor pressure can be seen as the outward hydrostatic pressure exerted against the inside surface of a cell wall as water tries to flow into the cell by osmosis. If the cell membrane is not stabilized by the presence of a cell wall, the cell will expand and eventually burst. For a walled cell (like *S. cerevisiae*)

at equilibrium (eq), the turgor pressure is balanced by the osmotic pressure difference between the internal and external medium (Smith et al. 2000)

$$\Pi_t^{eq} = -\left(\Pi_e^{eq} - \Pi_i^{eq}\right). \tag{44}$$

If the concentration of the external medium is increased, its osmotic pressure increases, and water flows out of the cell. A new equilibrium is established and the cell turgor pressure is reduced. At a certain point the external concentration will be large enough to abolish the cell turgor pressure ($\Pi_t = 0$), and hence (Smith et al. 2000)

$$\Pi_e^{\Pi_t=0} = \Pi_i^{\Pi_t=0}. \tag{45}$$

If Π_e is increased further, the turgor pressure is assumed to remain negligible. In an ideal and dilute system[4] the cell will behave as an ideal osmometer and the van't Hoff relationship holds, so that at constant temperature (Smith et al. 2000)

$$\Pi_i \left(V - b\right) = \Pi_i^{\Pi_t=0} \left(V^{\Pi_t=0} - V_b\right) \tag{46}$$

where V is the volume of the cell and V_b is the so-called intra-cellular non-osmotic volume, which is the sum of the volumes of hydrophobic cellular components (such as lipid bilayers) that are osmotically unresponsive.

To obtain an explicit expression for the transient behavior of turgor pressure under a varying volume, we assume that changes in p_i are related to the fractional changes in cell volume (dV/V) by a volumetric elastic modulus ϵ as (Levin 1979)

$$\epsilon = V \frac{dp_i}{dV}. \tag{47}$$

By integrating the above equation and approximating $\ln(V(t)/V^0)$ by the linear expression $(V(t)/V^0 - 1)$, we obtain (Levin 1979)

$$\Pi_t(t) = \epsilon \left(\frac{V(t)}{V^0} - 1\right) + \Pi_t^0. \tag{48}$$

[4] $\Pi_e < 13$ MPa (Martinez de Marañon 1997).

6.2 The biophysical model

We obtained the biophysical model for osmoregulation in Paper 2 and 3 from (42), (43), (46) and (48) in combination with the following assumptions:

- The cell volume only changes due to in- and outflow of water. This is a reasonable assumption for the rapid changes upon osmotic shock and a first approximation for longer time intervals. Besides, for simplicity, the cell surface area is assumed constant.
- Other variables than volume and pressure are assumed constant. Examples of such possible variables are cell surface area, cell wall thickness, membrane composition and vacuole volume, which all are affected by osmotic stress, see e.g. Hohmann (2002). However, the importance of these responses is difficult to judge and these processes are typically non-trivial to include in a model, mainly due to lack of data. Therefore, we disregard them in our current models.
- We consider glycerol as the sole osmolyte and, hence, ions and other small molecules that have been reported to change upon osmotic shock (see e.g. Sunder et al. 1996) are not considered. This simplification is to a certain extent motivated by experimental results from Reed et al. (1987), who found that glycerol counter-balances in the order of 80% of applied stress of NaCl in *S. cerevisiae*.

In particular, we multiply (43) by the cell surface area and obtain a relation for the cell volume as

$$V'(t) \propto \Pi_i(t) - \Pi_e(t) - \Pi_t(t). \tag{49}$$

The intra-cellular osmotic pressure is calculated from (42) and (46) according to

$$\Pi_i(t) = \frac{n + Gly(t)}{V(t) - V_b} \tag{50}$$

where Gly [mol] is the main osmolyte glycerol and n [mol] is the number of other osmotically active compounds in the cell.

The natural input variable of the osmoregulation system is Π_e. A typical experiment involves adding 0.5M NaCl to the medium, thereby increasing Π_e by 0.93 Osm ($\Phi_{NaCl} = 0.93$, $n_{NaCl} = 2$).

The turgor pressure is obtained from (48) as

$$\Pi_t(t) = \begin{cases} \Pi_t^0 \frac{V(t) - V^{\Pi_t=0}}{V^0 - V^{\Pi_t=0}}, & V(t) > V^{\Pi_t=0} \\ 0, & \text{otherwise} \end{cases} . \tag{51}$$

by restricting Π_t to positive values. Here $V^{\Pi_t=0}$ is a constant for the volume when $\Pi_t = 0$.

We finally note that the order of magnitude of the parameters in the biophysical model can be found directly or indirectly in the literature.

6.3 A first control model

In order to get an intuitive understanding of the osmoregulation system, we now describe a first simple model of how the cell controls the biophysical system when exposed to an osmotic shock. This first control model is simpler than the models presented in Paper 2 and 3.

As illustrated in Figure 13, the trans-membrane sensor proteins Sln1 and Fps1 are dependent on the biophysical variable Π_t. In this first model we only consider Sln1 and its effect on intra-cellular glycerol production. We note that accumulation of glycerol works as a feedback response to osmotic shock, since the biophysical variable Π_i is dependent on intra-cellular glycerol, as given by (50). An overview of the model is given in Figure 14.

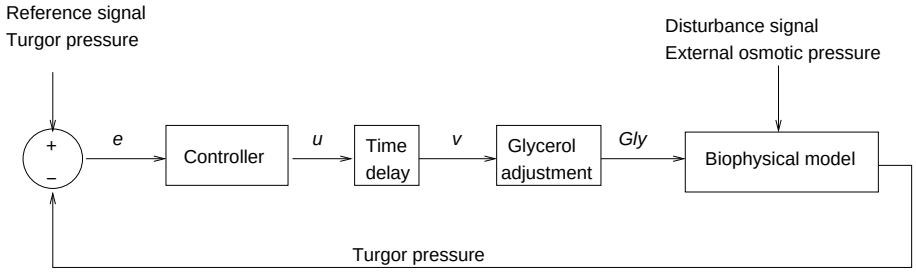

Figure 14: *A first model controlling the biophysical model. The glycerol level is adjusted by a proportional and time-delayed controller.*

To model the turgor sensor Sln1, the HOG pathway, transcription and translation in a very simple way, we consider a single time-delayed control function corresponding to all these steps. We let the difference between Π_t and a reference level Π_t^0 be the input (e) to this controller as

$$e(t) = \Pi_t^0 - \Pi_t(t). \tag{52}$$

We consider the simplest possible controller (u) that adjusts e by a constant K as

$$u(t) = K\ e(t). \tag{53}$$

To make the model more realistic, we also include a time-delay (t_d) corresponding to the time it takes to initiate glycerol accumulation, e.g. transcription and translation of enzymes. The time-delayed control signal (v) is obtained as

$$v(t) = u(t - t_d) \tag{54}$$

Finally, we let the rate of change of glycerol, Gly, be dependent on the control signal as

$$Gly'(t) = v. \tag{55}$$

We use this model to simulate an experiment where the input signal is an osmotic shock of 0.5M NaCl, see Figure 15. We note the input signal of increased Π_e at $t = 0$, followed by the rapid changes towards a new equilibrium in the biophysical variables. First, the imbalance in (49) causes a drop in volume, which leads to a decrease in turgor pressure (51) and an increase in intra-cellular osmotic pressure (50). Turgor pressure is abolished and the system reaches a new equilibrium where $\Pi_i = \Pi_e$ only a few seconds after the applied stress. The control model initiates glycerol production immediately after the time-delay has expired (10 minutes after the stress in this simulation), which in turn results in increasing intra-cellular osmotic pressure. As a consequence, water flows back into the cell and both volume and turgor pressure are slowly increasing to their original values. In particular, about 33 minutes after stress volume is recovered above $V^{\Pi_{t=0}}$ and turgor pressure starts to increase, while the increase in volume slows down. At this point we also see a slight increase in the rate of glycerol accumulation. This is because glycerol is plotted as concentration and therefore is dependent on volume.

The model presented in Paper 3 involves further refinements of this first control model:

- Both intra- and extra-cellular glycerol are considered in the model. Diffusion of glycerol molecules over the cell membrane is assumed to follow Fick's law (Gervais et al. 2001). Hence, the glycerol diffusion rate is proportional to the difference between intra-cellular and extra-cellular glycerol concentration.

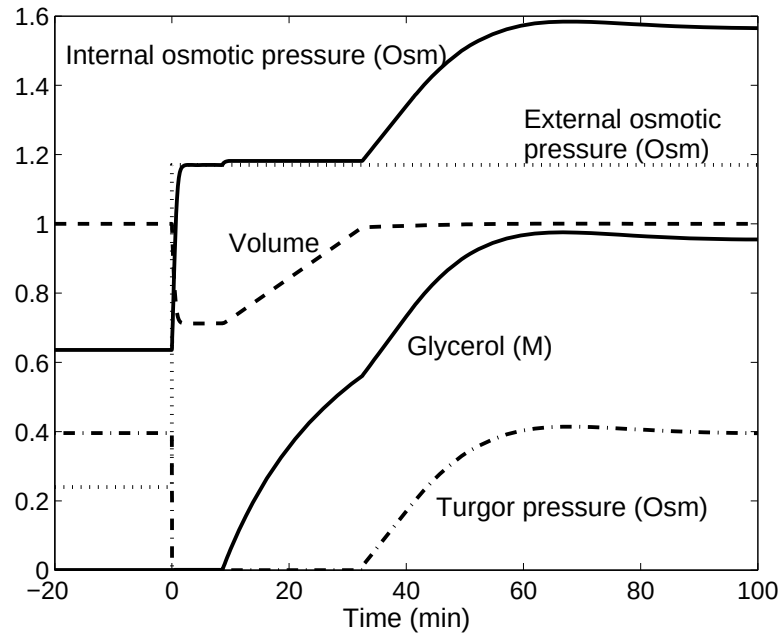

Figure 15: *Simulated data from the biophysical model in combination with (55). The input signal is an osmotic shock of 0.5M NaCl and reasonable model parameters are taken from Paper 3.*

- Changes in turgor pressure is independently sensed by Sln1 and Fps1. The second affects the diffusion constant for glycerol over the cell membrane.

Using the simplest possible sensor mechanisms (linear dependence on the difference between Π_t and Π_t^0) we obtain reasonable time series for both intra- and extra-cellular glycerol. Hence, the model in Paper 3 captures the fundamental processes that are discovered in yeast osmoregulation hitherto.

6.4 A more detailed control model

We continue to refine the model we have developed in order to approach the more detailed model considered in Paper 2. In particular, the molecular details of the control system are further analyzed, starting with the HOG signaling pathway and continuing with transcription/translation, metabolism and glycerol production.

The key components and interactions of the HOG signaling pathway have been identified during the last decade, see Figure 1 for an overview. To

model the HOG pathway we make the following assumptions:

- Reactions are modeled with linear (2) and bilinear kinetics (3). We note that detailed analyses of the mechanisms of isolated signaling reactions have been presented (Ferrell et al. 1997), while linear and bilinear kinetics usually have been employed in models of entire signaling pathways (Schoeberl et al. 2002, Swameye 2003). The main reason for this is that data typically is sparse and incomplete.

- The Sln1-branch of the pathway in isolation gives a similar response as the complete pathway including the Sho1-branch (O'Rourke 2004). We can therefore exclude the Sho1-branch from the model, something that is very useful since the sensor protein of that branch is not identified.

- The cell contains two compartments, the nucleus and the cytosol. Double phosphorylated Hog1 may enter the nucleus and is considered to be a transcription factor in the nucleus compartment. Furthermore, dephosphorylated Hog1 can leave the nucleus.

- The transmembrane protein Sln1 senses turgor pressure by adjusting its rate of auto-phosphorylation as

$$\text{rate of Sln1 auto-phosphorylation} \propto \left(\frac{\Pi_t(t)}{\Pi_t^0}\right)^{\beta} \tag{56}$$

  where β is a constant.

  We note that the auto-phosphorylation of Sln1 is needed to keep the HOG pathway inactive under normal conditions and that the exact sensor mechanism of Sln1 is unknown.

- All phosphorylated compounds are dephosphorylated by protein phosphatases. The rate of dephosphorylation is dependent on Hog1-induced protein synthesis of phosphatases. This realizes a negative feedback loop on the activation of the HOG signaling pathway. However, we note that there is always a basal level of phosphatases (Ghaemmaghami et al. 2003).

- This has to do with so called scaffold proteins, which are able to bind several (different) other proteins. They might facilitate signal transduction by forming multi-molecular complexes that can be rapidly activated by an incoming signal. In the HOG pathway, Pbs2 is believed to act as a scaffold protein (Posas et al. 1997). One detailed way of modeling scaffold complexes is discussed in Levchenko et al. (2000). However, due to lack of data we have not been able to include this aspect in our model.

- The number of signaling molecules are assumed to be sufficient for allowing deterministic simulation. A recent study suggests that the number of signaling molecules ranges from about 300 (Ssk2) to about 7000 (Hog1) for the proteins in the HOG pathway (Ghaemmaghami et al. 2003).

The parameters of the HOG signaling pathway were obtained taking into account experimental data on the response time of the pathway and the amplification of the signal. Here we note that the structure of a signaling MAPK cascade allows for signal amplification (Heinrich et al. 2002) as well as switch-like response of the kinases in the end of the cascade (Huang et al. 1996, Ferrell 1998). In our model, we also note that the sensor contributes to the switch-like behavior when $\beta > 1$ in (56).

The HOG signaling pathway triggers transcription and translation of several genes as indicated in Figure 13. The biochemical details of this activation are not understood to the same degree as the HOG signaling pathway, although there are ongoing research in this area (de Nadal et al. 2004). These processes are therefore simplified and we consider only two types of mRNA species and two types of proteins. The first type corresponds to metabolic enzymes, such as *GLK1*, *GPD1* and *GPD2*, and the second corresponds to phosphatases, such as *PTP2* and *PTP3*. Transcription is assumed linearly dependent on active Hog1 in the nucleus and translation is assumed linearly dependent on mRNA in the cytoplasm.

To model carbohydrate metabolism and glycerol production we considered previously published models (Hynne et al. 2001, Teusink et al. 2000, Rizzi et al. 1997) and adjusted the kinetics to allow for stable steady state concentrations and flows as determined by Rizzi et al. (1997) and Theobald et al. (1997). The dependence of carbohydrate metabolism and glycerol production on the HOG signaling pathway was included by letting the rates of several reactions be linearly dependent on the Hog1-induced protein. Besides, in order to include the dependence of glycerol transport on Fps1, we assume the following sensor mechanism

$$\text{Fickean diffusion coefficient} \propto \left(\frac{\Pi_t(t)}{\Pi_t^0}\right)^{\gamma} \tag{57}$$

where the exponent γ is a constant.

Finally, in order to obtain a complete model we also let the concentrations of all species in the cytosol be dependent on the cell volume (which is a dependent variable in the biophysical model). The complete model as given in Paper 2 includes 35 ODEs and 70 parameters.

6.5 Discussion

The detailed model of Paper 2 and the simpler model of Paper 3 share some main characteristics. Both models include two parallel ways of control in the cell, since these seem to be necessary to explain experimental data. The first control way is the ability of the cell to increase the intra-cellular concentration of glycerol, and the second control way is the ability to control the glycerol diffusion rate over the membrane. If any of these two control ways is absent, the cell fails to counter-balance an osmotic shock in an efficient way. A slight difference between the models is that the detailed model takes Hog1-induced up-regulation of phosphatases into account and thereby closes a negative feedback loop on the HOG signaling pathway. However, considering realistic induction of the phosphatases, this feedback plays no important role in pathway down-regulation.

To realize the models mathematically it is essential to combine a biophysical description with a description of the cellular control mechanisms. We generally note that our mathematical models are important not only for simulations but also for communicating the system in a compact and precise way.

In combination with new experimental results, our models have improved the biological understanding of osmoregulation in yeast and we exemplify this in two different ways. The first example concerns glycerol accumulation and Fps1. It is generally assumed that stimulated expression of *GPD1* and *GPP2* and the resulting increased glycerol production capacity accounts for the increase in intra-cellular glycerol level upon osmotic shock. However, our results indicate that this effect is only important for the long-term accumulation of glycerol. We suggest that a rapid closure of Fps1 leads to an initial glycerol accumulation that, in turn, accounts for HOG pathway down-regulation. This also implies that down-regulation of the HOG pathway occurs before intra-cellular glycerol peaks and hence before cells have fully adapted to the osmotic stress. Consequently, a strain expressing an Fps1 that cannot close should result in a strongly prolonged HOG pathway activation. This has also been experimentally verified.

The second example concerns feedback control of the HOG pathway in osmotic adaptation. It has been suggested that enhanced expression of genes encoding phosphatases accounts for feedback control (Hohmann 2002). However, our data suggests that an increase in the level of phosphatases is not necessary to down-regulate the pathway. Instead, the input signal to the HOG pathway is decreasing as turgor pressure is recovered. The phosphorylated kinases of the pathway are then dephosphorylated by phosphatases at a basal level. This view is supported by experimental results indicating that the pathway can be fully reactivated by a second osmotic shock.

The simple and detailed models of osmoregulation have been constructed in parallel. Notably, it can be useful to consider a simple model when developing a more detailed model, since the main characteristics of the system can more easily be observed and since the simple model can be parameterized with higher confidence than the detailed model. For instance, data from the simple model has suggested how to adjust the detailed model to give realistic output on intra-cellular glycerol.

In order to further develop our models of osmoregulation several experiments could be performed. Naturally, quantitative time series data are of particular interest. Below we give some examples of potential experiments for future studies:

- To investigate the roles of the two input branches of the HOG signaling pathway one can consider mutants with only one branch active and an input signal of various salt concentrations. In this way we obtain the dose-response characteristics for the different branches. This has already been done for one branch, but can be repeated for the other. Also for these experiments it can be useful to follow glycerol in time series.

- Concerning the mutant with an open Fps1 one could think of experiments with different combinations of salt and glycerol/sorbitol stress, e.g. 25% salt and 75% glycerol. This kind of experiments can be important in order to reveal the exact relationship between the two control functions.

- The osmotic pressures and turgor pressure of the biophysical model are difficult to measure experimentally. However, the volume can be measured by different techniques. Ideally, one could follow one individual cell in time series using state-of-the-art micro-fluid systems. This would significantly increase the measurement precision compared to data on a cell population.

A general observation of the experiments that have been performed hitherto is that the collection of possible system modifications using genetically modified strains is very rich and advanced. Such modifications give valuable insights into the system and can actually be necessary in order to completely understand certain systems, e.g. systems with mixed fast and slow kinetics and systems including feedback loops. However, one should not forget that variations of the input signal can be employed in combination with these modifications in order to identify the system. The standard step function of 0.5M NaCl could be complemented by other functions, e.g. a steady increase in NaCl from 0 to 1M.

7 Main contributions

In this section the main contributions of the three papers are listed. Besides, my contribution to each paper is listed.

Paper 1

Efficient ODE model identification for biological applications.
Gennemark P. and Wedelin D.

A parameter estimation algorithm. An algorithm that estimates the parameters of an ODE model from time series data has been devised. It considers one equation at a time and combines least-squares estimation with simulation of a single ODE to obtain both computational efficiency and accuracy. Our results suggest that the method is more accurate and considerably faster compared to other published methods.

A model selection algorithm. An algorithm that identifies both structure and parameters of an ODE model from time series data has been devised. It is designed to handle problems of realistic size, where reactions can be non-linear in the parameters and where data can be sparse and noisy. The model selection is done in an efficient heuristic way, where the structure is built incrementally. The method is evaluated on two previously published models using artificial data. In comparison to other methods that were used for these test systems, the main strength of the algorithm is that a complete model, and not only a structure, is identified, and that it is more accurate and considerably faster compared to other identification algorithms.

My contribution: literature studies and all implementation. Development of the basic ideas for both parameter estimation and model selection in cooperation with DW.

Paper 2

Integrative model of the response of yeast to osmotic shock.
Klipp E., Nordlander B., Krüger R., Gennemark P. and Hohmann S.

A mathematical model of yeast osmoregulation. The ODE model includes receptor stimulation, a MAP kinase cascade, activation of gene expression and adaptation of cellular metabolism as well as a biophysical description of volume regulation and osmotic pressure. Simulations agree well with experimental results obtained under different stress conditions or

with certain mutants. The model is predictive since it suggests previously unrecognized features of the system with respect to osmolyte accumulation and feedback control, which we confirm experimentally.

Improved understanding of Glycerol accumulation and Fps1. It is generally assumed that stimulated expression of *GPD1* and *GPP2* and the resulting increased glycerol production capacity accounts for the increase in intra-cellular glycerol level upon osmotic shock. However, our results indicate that this effect is only important for the long-term accumulation of glycerol. We suggest that a rapid closure of Fps1 leads to an initial glycerol accumulation that, in turn, accounts for HOG pathway down-regulation. This also implies that down-regulation of the HOG pathway occurs before intra-cellular glycerol peaks and hence before cells have fully adapted to the osmotic stress. Consequently, a strain expressing an Fps1 that can not close should result in a strongly prolonged HOG pathway activation.

Improved understanding of feedback control of the HOG pathway. It has been suggested that enhanced expression of genes encoding phosphatases accounts for feedback control (Hohmann 2002). However, our data suggests that an increase in the level of phosphatases is not necessary to down-regulate the pathway. Instead, the input signal to the HOG pathway decreases as turgor pressure is recovered. The phosphorylated kinases of the pathway are then dephosphorylated by phosphatases at basal level. This view is supported by experimental results indicating that the pathway can be fully reactivated by a second osmotic shock.

My contribution:

1. Original idea and first models of combining a biophysical description with a control model of osmoregulation. This idea was presented on a talk and poster together with BN at the Functional Genomics conference in Göteborg 2001. This biophysical model has been further developed in collaboration with EK.

2. Work on the basic model of the HOG signaling pathway (including the two-compartment model cytosol/nucleus) together with RK and EK.

3. My results from the simple model in Paper 3 have suggested how to adjust the detailed model to give realistic output on intra-cellular glycerol.

4. Suggestion of an experiment with different magnitude of osmotic shock in order to study the pathway sensor mechanism.

Paper 3

A simple mathematical model of adaptation to high osmolarity in yeast.
Gennemark P. and Nordlander B.

A mathematical model of yeast osmoregulation. This model complements the detailed model of Paper 2. Compared to the detailed model, the main strength of this model is its lower complexity, contributing to a better understanding of osmoregulation by focusing on relationships which are obscured in the more detailed model. The ten parameters of this simple model were constrained by data from various literature sources as well as our own data and estimated from absolute time series data on glycerol. The low complexity makes it possible to parameterize the model from absolute data. The qualitative behavior of the model has been successfully tested on data from other genetically modified strains as well as data for different input signals.

Improved understanding of osmoregulation. The model strengthen the hypothesis that at least two ways of control are required in order to efficiently counter-balance an osmotic shock in the cell. The first control way is the ability of the cell to adjust the intra-cellular concentration of glycerol, and the second control way is the ability to control the glycerol diffusion rate over the membrane.

My contribution: All work, based on experimental data supplied by BN.

References

Abouhamad W.N., Bray D., Schuster M., Boesch K.C., Silversmith R.E. and Bourret R.B. 1998. Computer-aided resolution of an experimental paradox in bacterial chemotaxis. *J Bacteriol.* 180(15), 3757-64.

Aebersold R. and Mann M. 2003. Mass spectrometry-based proteomics. *Nature* 422, 198-207.

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory*, Petrov B.N. and Csaki F. (eds.), Akademiai Kiado, Budapest, 267-281.

Alberts B., Bray D., Lewis J., Raff M., Roberts K. and Watson J.D. 1994. *Molecular Biology of the Cell.* Garland Publ. Inc., New York, 3rd ed.

Arkin, A.P. and Ross, J. 1995. Statistical Construction of Chemical Reaction Mechanisms from Measured Time-Series. *J. Phys. Chem.* 99, 970-979.

Atkins P.W. 1994. *Physical chemistry.* 5th. Oxford University Press. Walton Street, Oxford OX2 6DP.

Bower J.M. and Bolouri H. 2001. *Computational Modeling of Genetic and Biochemical Networks.* MIT Press.

Chen K.C., Calzone L., Csikasz-Nagy A., Cross F.R., Novak B. and Tyson J.J. 2004. Integrative analysis of cell cycle control in budding yeast. *Mol Biol Cell.* 15(8), 3841-62.

Crampin E.J., Schnell S. and McSharry P.E. 2004. Mathematical and computational techniques to deduce complex biochemical reaction mechanisms. *Prog. Biophy. Mol. Biol.* 86, 77-112.

de Boor C. 1978. *A practical guide to splines.* Springer-Verlag, New York, 235-43.

de Jong H. 2002. Modeling and simulation of genetic regulatory Systems: A literature review. *Journal of Computational Biology*, 9(1), 69-105.

de Nadal E, Alepuz P.M. and Posas F. 2002. Dealing with osmostress through MAP kinase activation. *EMBO Rep.* 3(8), 735-40.

Dempster A.P., Laird N.M. and Rubin D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc.* B 39, 1-38.

Eckert R., Randall D.J., Burggren W. and French K. 1997. *Animal Physiology*, 4th edition. W. H. Freeman Company, New York.

Eldar A., Dorfman R., Weiss D., Ashe H., Shilo B.Z. and Barkai N. 2002. Robustness of the BMP morphogen gradient in Drosophila embryonic patterning. *Nature.* 419(6904), 304-8.

Englezoz P. and Kalogerakis N. 2001. *Applied parameter estimation for chemical engineers.* Marcel Dekker, Inc., New York, NY.

Ferrell JE Jr. 1998. How regulated protein translocation can produce switch-like responses. *Trends Biochem Sci.* 23(12), 461-5.

Ferrell J.E. Jr and Bhatt R.R. 1997. Mechanistic studies of the dual phosphorylation of mitogen-activated protein kinase. *J Biol Chem.* 272(30), 19008-16.

Gervais P. and Beney L. 2001. Osmotic mass transfer in the yeast Saccharomyces cerevisiae. *Cell Mol Biol (Noisy-le-grand).* 47(5), 831-9.

Ghaemmaghami S., Huh W.K., Bower K., Howson R.W., Belle A., Dephoure N., O'Shea E.K. and Weissman J.S. 2003. Global analysis of protein expression in yeast. *Nature.* 425(6959), 737-41.

Gibson M.A. and Bruck J. 2000. Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels. *J. Phys. Chem. A.* 104, 1876-1889.

Gillespie D.T. 1976. A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions. *J. Comp. Phys.* 22, 403-434.

Guebel D.V. 2004. Canonical sensitivities: A useful tool to deal with large perturbations in metabolic network modeling. *In Silico Biology* 4, 0015.

Gustin M.C., Albertyn J., Alexander M. and Davenport K. 1998. MAP kinase pathways in the yeast Saccharomyces cerevisiae. *Microbiol Mol Biol Rev.* 62(4), 1264-300.

Heinrich, R., Neel, B.G. and Rapoport, T.A. 2002. Mathematical models of protein kinase signal transduction. *Mol. Cell.* 9, 957-970.

Heinrich, R., Rapoport, S. M. and Rapoport, T. A. 1977. Metabolic regulation and mathematical models. *Progr. Biophys. Mol. Biol.* 32, 1-82.

Hohmann S. 2002. Osmotic stress signaling and osmoadaptation in yeasts. *Microbiol Mol Biol Rev.* 66(2), 300-72.

Huang C.Y. and Ferrell JE Jr. 1996. Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proc Natl Acad Sci U S A.* 93(19), 10078-83.

Huang S. 1999. Gene expression profiling, genetic networks, and cellular

states: an integrating concept for tumorigenesis and drug discovery. *J Mol Med.* 77(6), 469-80.

Hucka M., Finney A., Sauro H.M. et al. 2003. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 19 (4), 513-523.

Hynne F., Dano S. and Sorensen P.G. 2001. Full-scale model of glycolysis in Saccharomyces cerevisiae. *Biophys Chem.* 94(1-2), 121-63.

Johnson M.L. and Faunt L.M. 1992. Parameter estimation by least-squares methods. *Methods In Enzymology.* 210, 1-37.

Kacser, H. and Burns, J.A. 1973. *In Rate Control of Biological Processes (Davies, D.D., ed.)* 65-104, Cambridge University Press, London.

Kikuchi S., Tominaga D., Arita M., Takahashi K. and Tomita M. 2003. Dynamic modeling of genetic networks using genetic algorithm and S-system. *Bioinformatics.* 19(5), 643-50.

Kitano H. 2002a. Computational Systems Biology, *Nature.* 420, 206-210.

Kitano H. 2002b. Systems Biology: A Brief Overview. *Science*, 295, 1662-1664.

Lambert, J. D. 1991. *Numerical methods for ordinary differential systems the initial value problem.* John Wiley.

Levchenko A., Bruck J. and Sternberg P.W. 2000. Scaffold proteins may biphasically affect the levels of mitogen-activated protein kinase signaling and reduce its threshold properties. *Proc Natl Acad Sci U S A.* 97(11), 5818-23.

Levin, R. L. 1979. The water permeability of yeast cells at sub-zero temperatures. *J. Mem. Biol.* 46, 91-124.

Levin M.D., Morton-Firth C.J., Abouhamad W.N., Bourret R.B. and Bray D. 1998. Origins of individual swimming behavior in bacteria. *Biophys J.* 74(1), 175-81.

Liang S., Fuhrman S., and Somogyi R. 1998. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. *Pacific Symposium on Biocomputing*, 3, 18-29.

Maeda T., Takekawa M. and Saito H. 1995. Activation of yeast PBS2 MAPKK by MAPKKKs or by binding of an SH3-containing osmosensor. *Science.* 269(5223), 554-9.

Martinez de Marañon, I., Gervais P. and Molin P. 1997. Determination of cells' water membrane permeability: unexpected high osmotic permeability of Saccharomyces cerevisiae. *Biotechnol. Bioeng.* 56, 63-70.

Marquardt D.W. 1963. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11, 431-441.

Mckenzie J.A. and Strauss P.R. 2003. A quantitative method for measuring protein phosphorylation. *Anal. Biochem.* 313, 9-16.

Meng T.C., Somani S. and Dhar P. 2004. Modeling and simulation of biological systems with stochasticity. *In Silico Biology* 4, 0024.

Moles C.G., Mendes P. and Banga J.R. 2003. Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Res.* 13(11), 2467-74.

Morton-Firth C.J. and Bray D. 1998. Predicting temporal fluctuations in an intracellular signalling pathway. *J Theor Biol.* 192(1), 117-28.

O'Rourke, S.M. and Herskowitz, I. 2004. Unique and redundant roles for HOG MAPK pathway components as revealed by whole-genome expression analysis. *Mol. Biol. Cell* 15, 532-542.

Peng X.Y.and Li P.C.H. 2004. A Three-Dimensional Flow Control Concept for Single-Cell Experiments on a Microchip. 2. Fluorescein Diacetate Metabolism and Calcium Mobilization in a Single Yeast Cell As Stimulated by Glucose and pH Changes. *Anal. Chem.* 76(18) 5282-92.

Pintér. 1996. Continuous global optimization software: a brief review. *Optima*, 52, 1-8.

Posas F. and Saito H. 1997. Osmotic activation of the HOG MAPK pathway via Ste11p MAPKKK: scaffold role of Pbs2p MAPKK. *Science.* 276(5319), 1702-5.

Press W.H., Teukolsky S.A., Vetterling W.T. and Flannery B.P. 1993. *Numerical Recipes in C : The Art of Scientific Computing*, Cambridge University Press.

Reed R.H., Chudek J.A., Foster R. and Gadd G.M. 1987. Osmotic significance of glycerol accumulation in exponentially growing yeasts, *Appl Environ Microbiol.* 53(9), 2119-23.

Reiser V., Raitt D.C. and Saito H. Yeast osmosensor Sln1 and plant cytokinin receptor Cre1 respond to changes in turgor pressure. 2003. *J Cell Biol.* 161(6), 1035-40.

Rissanen J. 1978. Modeling by shortest data description. *Automatica.* 14, 465-471.

Rizzi M., Baltes M., Theobald U. and Reuss M. 1997. In Vivo Analysis of Metabolic Dynamics in Saccharomyces cerevisiae: II. Mathematical Model.

Biotechnology and Bioengineering 55, 592-608.

Runarsson T.P. and Yao X. 2000. Stochastic Ranking for Constrained Evolutionary Optimization. *IEEE Transactions on Evolutionary Computation.* 4(3), 284-294.

Savageau, M. A. 1976. *Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology.* Addison-Wesley, Reading, Mass.

Schoeberl B., Eichler-Jonsson C., Gilles E.D. and Muller G. 2002. Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nat. Biotech*, 20(4), 370-5.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6, 461- 464.

Sedoglavic, A. 2002. A probabilistic algorithm to test local algebraic observability in polynomial time. *J. Symb. Comp.* 33(5), 735-755.

Shampine L. F. and Reichelt M. W. 1997. The MATLAB ODE Suite. *SIAM Journal on Scientific Computing.* 18, 1-22.

Sherman F. 2002. Getting started with yeast. *Methods Enzymol.* 350, 3-41.

Smith A.E., Zhang Z. and Thomas C.R. 2000. Wall materials properties of yeast cells: Part 1. Cell measurements and compression experiments. *Chemical Engineering Sciences.* 55, 2031-41.

Somero, G.N. and Yancey, P.H. 1997. *Handbook of Physiology.* (eds. Hoffmann and Jamieson) Oxford University Press, Oxford, New York. 441-484.

Stryer L. 1995. *Biochemistry.* 4th ed. WH Freeman and Company. New York.

Sunder S., Singh A.J., Gill S. and Singh B. 1996. Regulation of intracellular level of Na+, K+ and glycerol in Saccharomyces cerevisiae under osmotic stress. Mol Cell Biochem. 158(2), 121-4.

Swameye I., Muller T.G., Timmer J., Sandra O. and Klingmuller U. 2003. Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling. *Proc Natl Acad Sci U S A.* 100(3), 1028-33.

Takahashi, K., Ishikawa, N., Sadamoto, Y., et al. 2003. E-CELL2: Multiplatform E-CELL Simulation System. *Bioinformatics.* 19(13), 1727-1729.

Tamas, M.J., Luyten K, Sutherland F.C., Hernandez A., Albertyn J., Valadi H., Li H., Prior B.A., Kilian S.G., Ramos J., Gustafsson L., Thevelein J.M. and Hohmann S. 1999. Fps1p controls the accumulation and release of the compatible solute glycerol in yeast osmoregulation. *Mol Microbiol.* 31, 1087-1104.

Tamas M.J., Rep M., Thevelein J.M. and Hohmann S. 2000. Stimulation of the yeast high osmolarity glycerol (HOG) pathway: evidence for a signal generated by a change in turgor rather than by water stress. *FEBS Lett.* 472, 159-165.

Teusink, B., Passarge J., Reijenga C.A. et al. 2000. Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. *Eur J Biochem.* 267, 5313-5329.

Theobald U., Mailinger W., Baltes M., Rizzi M. and Reuss M. 1997. In vivo analysis of metabolic dynamics in Saccharomyces cerevisiae: I. Experimental observations. *Biotechnology and Bioengineering*, 55(2), 305-316.

Tomita, M., Hashimoto, K., Takahashi, K., et al. 1999. E-CELL: software environment for whole-cell simulation. *Bioinformatics.* 15(1), 72-84.

Voit, E.O. 2000. *Computational analysis of biochemical systems. A practical guide for biochemists and molecular biologists.* Cambridge University Press, Cambridge.

Voit E.O. and Almeida J. 2004. Decoupling dynamical systems for pathway identification from metabolic profiles. *Bioinformatics.* 20(11), 1670-81.

Westerhoff H.V. and Palsson B.O. 2004. The evolution of molecular biology into systems biology. *Nature Biotechnology.* 22, 1249-52.

Zhu H. and Snyder M. 2002. "Omic" approaches for unraveling signaling networks. *Curr Opin Cell Biol.* 14(2), 173-9.

Zucchini, W. 2000. An Introduction to Model Selection. *J. Math. Psych.* 44, 41-61.