Available online at www.sciencedirect.com
ScienceDirect
Current Opinion in Biotechnology

Model-based inference of biochemical parameters and dynamic properties of microbial signal transduction networks

Jörg Schaber[1,2] and Edda Klipp[1]

Because of the inherent uncertainty about quantitative aspects of signalling networks it is of substantial interest to use computational methods that allow inferring non-measurable quantities such as rate constants, from measurable quantities such as changes in protein abundances. We argue that true biochemical parameters like rate constants can generally not be inferred using models due to their non-identifiability. Recent advances, however, facilitate the analysis of parameter identifiability of a given model and automated discrimination of candidate models, both being important techniques to still extract quantitative biological information from experimental data.

Addresses
[1] Theoretical Biophysics, Humboldt-Universität Berlin, Invalidenstr. 42, 10115 Berlin, Germany
[2] Institute of Experimental Internal Medicine, Medical Faculty, Otto von Guericke University, Leipziger Str. 44, 39120 Magdeburg, Germany

Corresponding authors: Schaber, Jörg (schaber@med.ovgu.de) and Klipp, Edda (klipp@molgen.mpg.de)

Current Opinion in Biotechnology 2011, **22**:109–116

This review comes from a themed issue on
Analytical biotechnology
Edited by Matthias Heinemann and Uwe Sauer

Available online 20th October 2010

0958-1669/$ – see front matter

DOI 10.1016/j.copbio.2010.09.014

Introduction

The present understanding of cellular signal transduction is restricted, at the best, to the wiring schemes of signalling pathways. Little is known about the kinetic rate laws of signalling processes, let alone their biochemical parameters, which is also a consequence of their limited experimental accessibility. It is therefore of substantial interest to use computational methods that allow inferring non-measurable quantities such as rate constants, from measurable quantities such as changes in protein abundances and/or activities.

Here, we will argue that due to the substantial biological uncertainty about signalling networks biochemical signalling parameters cannot be inferred in general, because they are neither structurally nor practically identifiable. Nevertheless, being aware of these principal limitations we can still infer useful quantitative characteristics of signal transduction in microorganisms by employing appropriate mathematical and computational techniques.

We first demonstrate the general impossibility to estimate the true biochemical parameters of signalling rates, such as rate or Michaelis–Menten constants or V_{max}-values, with a simple worked example. Then we provide arguments that we can still infer qualitative as well as the quantitative aspects of signalling networks, illustrated with recent examples and approaches. Finally, we shortly discuss, where we see need for development in the field.

Nested uncertainties and the impossibility to find the truth

Parameter estimation has to face three types of challenges: firstly, ambiguity in the definition of the model structure, that is which compounds and interactions to be considered, and the appropriate choice of kinetic laws; secondly, the possibility of parameter dependencies, and thirdly, availability of sufficient and suited data for the actual parameter estimation.

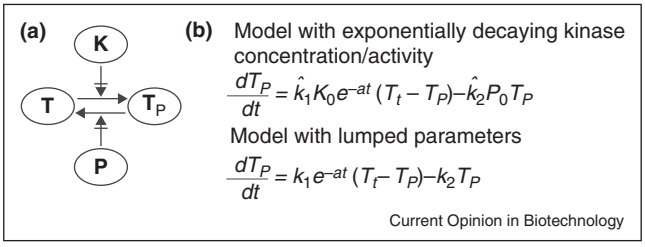
To demonstrate various aspects of model uncertainties, parameter identifiability and parameter estimation, we use a simple worked example: the presumably most universal signalling motif is a simple protein-dependent post-translational modification of a protein and its reverse reaction, which is usually also catalysed by another protein. Even though there are many kinds of modifications involved in signal transduction, we use here the classical example of a kinase-dependent phosphorylation and a phosphatase-dependent de-phosphorylation (Figure 1a).

Uncertainty in the model structure

In general, the wiring schemes of signal transduction networks are unknown. Even if we assume to know the wiring scheme, as in our example (Figure 1a), the kinetic rate laws are usually unknown. Therefore, already at the point of translating the wiring scheme into a mathematical model we are confronted with uncertainty about the model structure [1,2•]. Moreover, even if we make some justified assumptions about model structure, kinetic laws and kinetic parameters cannot be determined independently. We referred to this dilemma as nested uncertainties and reviewed and proposed strategies to address these elsewhere [3].

In case nothing is known about the kinetic rates laws, it is reasonable to assume mass action kinetics, because in

Figure 1



Simple model of kinase-dependent phosphorylation and phosphatase-dependent de-phosphorylation. **(a)** Wiring scheme according to Systems Biology Graphical Notation (SBGN), arrows between components indicate reactions, arrows on arrows indicate modifying influences. *K* kinase, *P* phosphatase, *T* target, T_P phosphorylated target. **(b)** Mathematical models: $T_t = T + T_P$, K_0, P_0 initial concentrations; $\hat{k}_1$ $\hat{k}_2$, k_1, k_2, *a* kinetic constants, *t* time.

signal transduction networks the involved reaction partners are usually in concentrations within the same order of magnitude. Here, we can describe the dynamics of the phosphorylated target T_P with one differential equation (Figure 1b), because in our simple model the total concentration of the target T_t is a conserved moiety. We further assume that the concentration or activity of the kinase K is exponentially decreasing with time, $K(t) = K_0 e^{-at}$, whereas the concentration of the phosphatase P is constant (P_0). With these assumptions we can formulate a mathematical model describing the dynamics of a signal-dependent phosphorylation and de-phosphorylation (Figure 1b, upper formula) [4].

Uncertainty in the parameters: identifiability and parameter estimation

Structural non-identifiability

Apparently, both the phosphorylation and the de-phosphorylation reactions include two constants, that is the actual rate constants, $\hat{k}_1$ and $\hat{k}_2$, and the initial concentration or activities of the kinase K_0 and the phosphatase P_0, respectively. Therefore, having independent information about neither the rate constants nor the initial concentrations, we can obtain infinitely many solutions for the constants $c_1 = \hat{k}_1 K_0$ and $c_2 = \hat{k}_2 P_0$ with respect to $\hat{k}_1$, K_0 and $\hat{k}_2$, P_0, respectively. Thus, the parameters $\hat{k}_1$, K_0 $\hat{k}_2$, and P_0 are non-identifiable, more specifically, they are structurally non-identifiable, because this is independent of data. This is a very common dilemma, which is usually resolved by lumping non-identifiable parameters together thereby obtaining structurally identifiable parameters (Figure 1b, lower formula).

Note that to obtain a structurally identifiable mathematical model for this simple signalling motif, we already abstracted from the true biochemical processes at various levels [2•]:

(a) *the wiring scheme*; we abstracted from the detailed reaction scheme like kinase–substrate binding and dissociation and the third reaction partner ATP, which provides the phosphate and was assumed not to be rate limiting and, therefore, omitted.
(b) *the kinetic rate law*; we abstracted from detailed kinetic rate laws by assuming mass action kinetics.
(c) *the parameters*; we abstracted from true biochemical parameters by lumping parameters together. In fact, parameters of biochemical models are models themselves in the sense that they represent lumped biochemical processes, which are assumed to be constant in the considered system and the given time scale.

By definition, all models abstract from underlying biochemical processes and therefore rate constants or other biochemical parameters of the model do usually not represent *in vivo* counterparts, but rather lumped processes.

A number of reviews and original work nicely describe how a model should be set up, unfold pitfalls and emphasise various aspects of the modelling process including identifiability and parameter estimation [1,2•,5–10]. Notably, there is a software tool that allows automatic checking of global parameter identifiability of (linear and) nonlinear dynamic models that does not require understanding of the underlying mathematical principles and can be used by researchers in applied fields with a minimum of mathematical background [11,12].

Parameter estimation and practical non-identifiability

The structurally identifiable model can be used to extract quantitative information from biological data. To this end, we have to estimate the parameters such that the model simulations give an optimal approximation to the data. In the remaining part of this section we shortly review and illustrate how one can analyse whether estimated parameters are actually good enough to extract quantitative information from biological data.

Parameters can be estimated, for example, by minimizing the sum of squared residuals (SSR) (Box 1) between data and model simulation [13]. There is a wealth software packages that provide methods and algorithms to minimize the SSR and estimate parameters. An overview of parameter estimation algorithms used in systems biology is given in [13,14]. Table 1 shows a list of freely available and commonly used software packages with implemented algorithms for parameter and confidence interval estimation. For a review of modelling tools also see [15•].

Minimizing the SSR for parameter identification is so popular, because under certain regularity conditions the

Box 1 Concepts of parameter estimation and identifiability

Sum of squared residuals (*SSR*): The SSR measures the distance between the simulated and the measured data as a function of the parameter vector p

$$SSR(p) = \sum_{i=1}^{n} \left(\frac{y_i - f(t_i, p)}{\sigma_i} \right)^2,$$

where $p = (k_1, k_2)$ for the model in Figure 1, y_i are the $i = 1, \ldots, n$ data points with standard deviation σ_i, here for the model components T_P, $f(t_i, p)$ is the numerically calculated solution of a model (e.g. Figure 1) at time points t_i with the parameter vector p.

Maximum Likelihood estimator (*MLE*): Under the assumptions that model f is the true model that generated the data and normally distributed errors, $\varepsilon \propto N(0, \sigma^2)$, the parameter vector $\hat{p}$ that minimizes $SSR(p)$, i.e.

$$\hat{p} = \arg\min_{p}(SSR(p)),$$

corresponds to the maximum likelihood estimate (MLE) of p. This means that $\hat{p}$ maximizes the probability to observe the data y given parameter vector p under a certain model f.

Asymptotic confidence intervals for MLEs: the $100(1-\alpha)\%$ asymptotic confidence intervals are defined by

$$\hat{p}_i \pm \sqrt{s^2 C_{i,i}}\, t^{\alpha/2}_{[n-m]}$$

where $C_{i,t}$ is the ith diagonal element of the covariance matrix C of the model $f(t,p)$, $s^2 = (SSR(\hat{p}))/(n-m)$ is an estimate of the error variance σ^2, n is the number of data points, m is the number of parameters and $t^{\alpha/2}_{[n-m]}$ is the $(1-\alpha/2)$ quantile of a t-distribution with $n-m$ degrees of freedom. Asymptotic confidence intervals can be readily computed and are standard in many software packages (Table 1). However, asymptotic confidence intervals only converge to the true confidence intervals for $n \to \infty$. They are obviously symmetric around $\hat{p}$ and have to be taken with care for small n and may even give meaningless results (Figure 2b and d). On the basis of the eigenvectors and eigenvalues of the covariance matrix C asymptotic confidence regions can also be easily computed (Figure 2b and d).

Exact confidence intervals for MLEs: The set

$$P_{CR} = \left\{ p : SSR(p) \leq SSR(\hat{p}) \left(1 + \frac{m}{n-m} F^{\alpha}_{m,n-m} \right) \right\}$$

is also an approximate ellipsoidal $100(1-\alpha)\%$ confidence region for $\hat{p}$. $F^{\alpha}_{m,n-m}$ is the α upper a critical value for the $F_{m,n-m}$ -distribution. P_{CR} also converges to the true confidence region only for $n \to \infty$. However, P_{CR} is considered to be more realistic than asymptotic confidence regions, because they are based on the likelihood contours and do not give meaningless results as they are not necessarily symmetric around $\hat{p}$. Finding the set P_{CR} is an inverse problem and for a parameter space of more than two dimensions it is difficult to compute and therefore rarely employed.

Monte-Carlo confidence intervals: Are derived from empirical parameter distributions obtained by re-sampling and re-fitting the data. With $\hat{p}$ being a MLE, the assumed model $f(p)$ and an estimation of the error variance s^2, it is possible to generate an arbitrary number of new artificial datasets by

$$y^k = f(t, \hat{p}) + \varepsilon, \quad (k = 1, 2, \ldots, K),$$

where $\varepsilon \propto N(0, s^2)$. These new datasets are all equivalent to the original dataset with respect to the error distribution and the number of observed data points. For the new datasets, we can estimate the set of corresponding MLEs $\hat{p}^k$ from which we can deduce a empirical probability density of the parameter $\hat{p}$ (Figure 2b,d). This method has a solid mathematical foundation and it has been shown to be useful in practice. However, it may be computationally time consuming, depending on the optimisation algorithm.

Structural non-identifiability: Structural non-identifiability is a property of the parameters of a model, which is independent of data. It relates to a redundant parameterization of the model and is characterized by a subset of parameters that can be varied without changing the solution. The confidence interval of a structurally non-identifiable parameter is infinite [17•].

Practical non-identifiability: We follow the definition of Raue *et al.* [17•] and define a parameter as practically non-identifiable when its likelihood-based confidence region is infinite with respect to a certain confidence level (Figure 2b). This characterizes a parameter vector that is close to arbitrary in a certain region. Practical non-identifiability can usually, but not necessarily, be resolved by using more data for the parameter estimation (Figure 2d).

resulting parameter vector is a maximum likelihood estimator (MLE) (Box 1), which comes with a large body of useful theory [13,16], in particular the concept of parameter confidence regions. Parameter confidence regions provide a measure of reliability of the obtained parameter estimate. Moreover, confidence regions can be used to decide whether a parameter is practically identifiable [17•]. Practical identifiability is a property of a parameter that depends on available data, opposed to structural identifiability introduced above (see Box 1). Intuitively, a parameter can be practically identified, if its estimated confidence interval is small, that is if it is reliably estimated. The parameters of a structurally identifiable model might still be practically unidentifiable because of limited amount of data. To illustrate this, we generated artificial data using the lumped parameter model from Figure 1b with rate constants $k_1 = 2$ and $k_2 = 1$ and added a random measurements error to each data point that was generated from a normal distribution with zero mean and a standard deviation of 0.05 (Figure 2a, points are the generated data and the assumed standard deviation is indicated by the error bars). In addition, the data were scaled to unity. Data scaling is very common as many biochemical measurements provide only arbitrary units. Therefore, not only the rate constants have to be estimated from the data but also a scaling factor in case we have no information about how much of our target T is phosphorylated for a given stimulus. In Figure 2a we show the generated data (points), the 'true' model from which they are generated (dashed line) and a fitted curve (solid line), obtained from a parameter estimation with the Differential Evolution algorithm [19].

The 'true' time course and parameters are obscured by the measurements error and cannot be recovered by parameter estimation in general. In fact, the true parameters can give a worse fit than the estimated parameters (see SSR in Figure 2a and c). In Figure 2b and d we plot confidence regions of the estimated parameter vector (k_1, k_2). We display three different types of commonly used confidence regions, that is the asymptotic confidence region (dashed

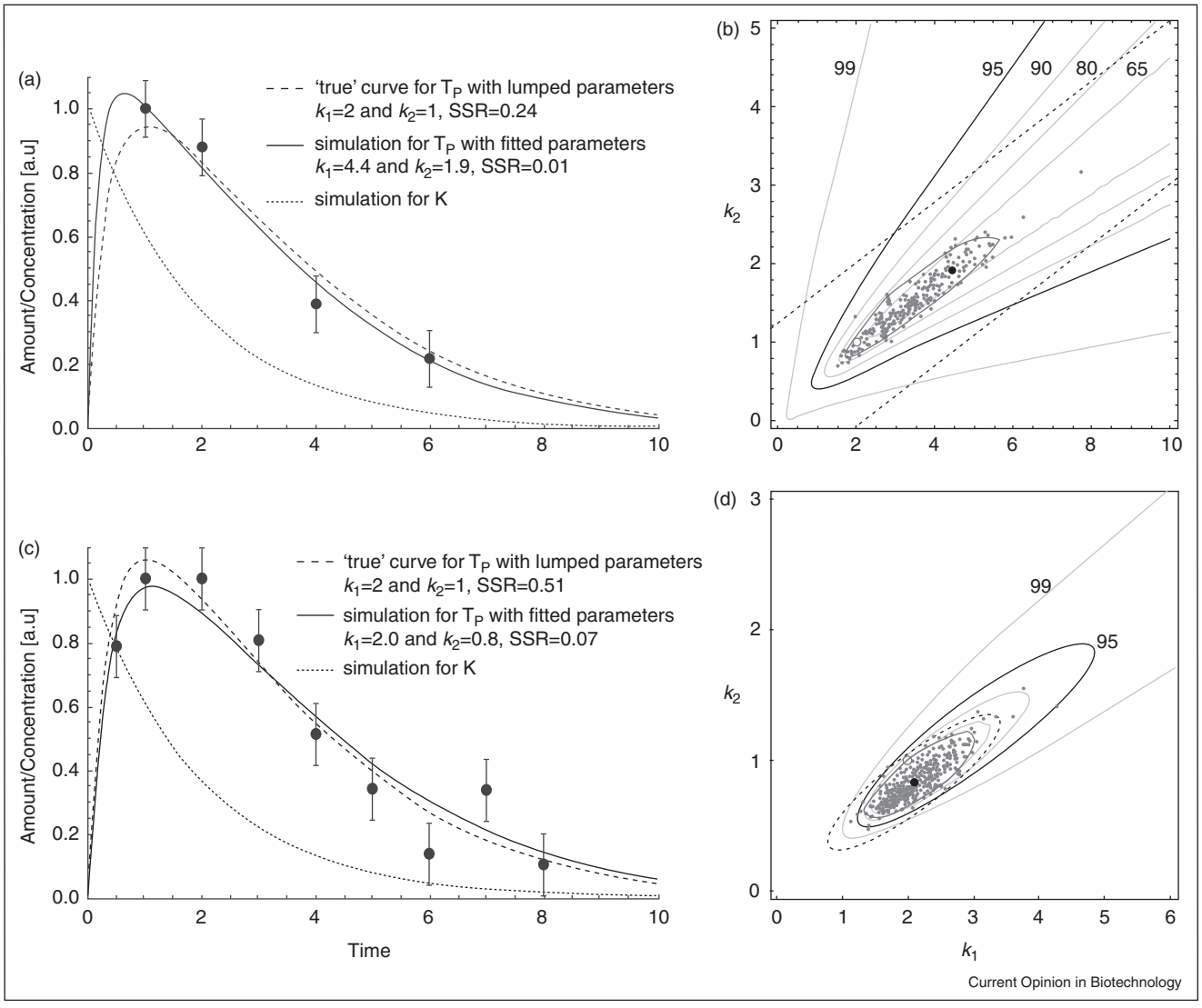
Table 1

Freely available software for simulation and parameter estimation of dynamic models

Software	URL	Reference	Remark
Copasi	www.copasi.org	[46]	Stand-alone
SBToolBox	www.sbtoolbox.org	[47]	Matlab® toolbox
SBML-PET	sysbio.molgen.mpg.de/SBML-PET	[48]	Stand-alone
PottersWheel	www.potterswheel.de	[18]	Matlab® toolbox

line), exact confidence regions (thick solid lines) and Monte-Carlo confidence regions (dots and thin solid line). Please refer to Box 1 for more information. Asymptotic and Monte-Carlo confidence intervals are finite by definition (Box 1); however, in Figure 2b the asymptotic confidence region expands into the negative range of the parameters, which makes no biological sense. The 95% exact confidence region based on the likelihood contours of SSR is infinite, indicating that SSR does not exceed a confidence threshold into a certain direction of parameter space. In practice, this has a similar effect as a structurally non-identifiable parameter in the sense that varying the

Figure 2



Data, fitted curves and parameter confidence regions for the lumped parameter model from Figure 1. **(a)** Dots: artificial data generated from the dashed time course with an added random error taken from a normal distribution with zero mean and a standard deviation of 0.05. Solid line: fitted time course. Dotted line: time course of the input stimulus *K*. SSR: sum of squared residuals, that is the log-likelihood function. **(b)** Solid lines: exact parameter confidence regions (Box 1). The black solid line indicates the 95% exact confidence region. The dotted line indicates the 95% asymptotic confidence region (Box 1). The thin solid line is the 95% polytop quantile calculated from the grey dots. Grey dots: estimated MLEs for k_1 and k_2 using Monte-Carlo re-sampling of the data in (a). **(c)** Same as (a) only with more data. **(d)** Same as (b) but for the parameter estimates and the data from (c).

parameter vector within the infinite 95% exact confidence region has very little effect on the SSR and therefore the simulation result [17•]. Usually, but not necessarily for a particular case, practical non-identifiability can be resolved using more data. In Figure 2c and d we display generated data and fitted curve and corresponding confidence regions using more data points. Apparently, the three different confidence regions are now much more similar than using fewer data and all confidence regions are bounded, characterising an identifiable parameter. Theory states that with increasing n all three confidence regions converge. Note that all of the above only holds provided our model is the true model, which can never be verified, only falsified [1,20].

In most software packages it is standard to report asymptotic confidence regions, because they are easily and efficiently calculated. However, as shown above, they may grossly overestimate the true confidence region and may even give meaningless results, especially for low number of data points. Monte-Carlo and exact confidence regions are time consuming and difficult to compute, respectively, and therefore usually not reported. Recently, a computationally feasible method has been proposed that computes exact confidence intervals based on likelihood contours [17•] and comes with an implementation for Potters-Wheel fitting toolbox [18].

Above we have illustrated that in order to avoid arbitrary parameter combinations (structurally non-identifiable parameters), we have to lump processes and parameters together and abstract from detailed biochemical processes. We have also demonstrated that even with a structurally identifiable model, parameter combinations can still be close to arbitrary (practically non-identifiable parameters), if we want to model quantitative data.

Only in the rare case of a completely defined system like an *in vitro* experiment with only two components and a known reaction mechanism, we can image that true biochemical rate constants can be inferred from modelling and parameter estimation. However, the validity of such parameters for the *in vivo* situation is questionable.

Nested uncertainties and what we can still learn about the truth

What and how do we learn from models?

Being aware of the general impossibility to infer non-measureable biochemical reaction rates from measurable quantities, dynamic models with lumped parameters can still be very useful to learn something about the system behaviour for several reasons:

(1) *Modelling requires that verbal hypotheses are made specific and conceptually rigorous*. Formulating the mathematical model requires, first, to consider the system boundaries, that is to decide which components are actually necessary to address the scientific question. Second, the interconnections, that is the wiring scheme, have to be carefully considered. Often already during this process, gaps in our knowledge and weaknesses of our verbal concepts are highlighted, when it turns out that the model cannot show even the expected qualitative behaviour. The important steps in model development are reviewed in [7,9].

(2) *Modelling provides qualitative as well as quantitative predictions*. Once we have a parameterised model, quantitative predictions can be made about systems behaviour that has not been measured yet. Indeed, in case a model prediction can experimentally be verified, we gain confidence in our model in the sense that it captures the most important processes to explain a certain phenomena, which augments our biological understanding [21–25]. Another important aspect is qualitative predictions about possible processes that explain observed behaviour. By testing several candidate models and by model discrimination analysis, predictions about yet unknown processes, for example, feedbacks, can be made [21,26•,27,28•,29] or the important processes to explain a certain phenomena can be identified (J Schaber *et al.*, Automated ensemble modeling with modelMaGe: analyzing feedback mechanisms in the Sho1 branch of the HOG pathway. Unpublished data) [24]. Even non-parameterized models can be used to elucidate principal properties of biochemical reaction networks [30–32]. Recently, a tool has been developed that facilitates automatic generation and discrimination of candidate models [33] and further development of such techniques will certainly enhance biological knowledge generation by mathematical modelling.

(3) *Modelling facilitates the analysis of complex interactions before experimental tests*. Apparently, it is much easier to change a parameter in a mathematical model than to tune a biochemical property *in vivo*. Therefore, models are useful to predict experimental designs, which yield most information. In fact, it is the paradigm of systems biology that models are a low-cost rapid test of candidate interventions and optimal experimental design [5]. One particularly useful type of experiment that can be planned using models is perturbation experiments.

Model structure and parameter inference from perturbation experiments

When we want to learn something about the dynamic behaviour of a signal transduction network, clearly, we have to stimulate it in such a way that it performs its biological function. Moreover, the more different stimuli or perturbations we apply to a pathway without inter-

rupting it, the more we learn about its function for two reasons. First, we can infer or learn more precisely pathway structure and dynamic regulation principles from perturbation experiments, because they help to discriminate between alternative network topologies and kinetics (J Schaber *et al.*, Automated ensemble modeling with modelMaGe: analyzing feedback mechanisms in the Sho1 branch of the HOG pathway. Unpublished data) [2•,26•,34]. Second, the more data we have, the more probable it is that we can identify our parameters, as explained above.

A set of perturbations has shown to be useful in signalling networks, which are:

- Stimulus variations such as
  - Different strengths of stimuli [35],
  - Different temporal profiles of stimulus application (e.g. square or sinusoidal pulses of different frequency and/or amplitude),
  - Stimulus combinations.

- Protein abundance manipulations [36] such as
  - Knock-out,
  - Knock-down,
  - Over-expression using GAL or TET promotors.

- Protein activity manipulations [37] by
  - Binding property modification,
  - Inhibition by chemical substances.

An interesting concept is, for example, the creation of kinases with a mutated ATP binding site that can specifically be inhibited by cell-permeable small molecules, so-called ATP-analog-sensitive mutants [28•,38].

When performing additional experiments in order to obtain more data for model structure identification and parameter estimation, one must make sure that the experimental conditions for these experiments are still compatible, that is with respect to temperature, pH-value, analysed strain, growth media and growth state, and other relevant conditions [39••].

Where else can we get parameters apart from estimation?

Models can only give meaningful predictions for actual experiments when they are parameterised in a sensible way. Parameterisation is usually the most time consuming and laborious task in model development, mainly due to the difficulties in parameter estimation described above. For the sake of identifiability and reduction in complexity of the parameter estimation task, it is advisable to get hold on information about as many parameters as possible before parameter estimation from data. There are in principle two possibilities, that is original literature and databases. Finding parameters in publications has been eased to some extent by comprehensive text mining approaches. Database curators employ both text mining (which is usually restricted to abstracts) and careful reading of full-text publications. In Box 2 we provide a crisp overview of available databases.

Box 2 Tools and databases for kinetic parameters of biochemical reaction networks

SABIO-RK: System for the Analysis of Biochemical Pathways — Reaction Kinetics http://sabio.h-its.org [44].

SABIO-RK is a web-based application employing the SABIO relational database which contains information about biochemical reactions, their kinetic equations with their parameters, and the experimental conditions under which these parameters were measured. It is able to export SBML format files of selected reactions sets together with kinetic information.

BRENDA: The Comprehensive Enzyme Information System http://www.brenda-enzymes.org [45].

The database covers information on classification and nomenclature, reaction and specificity, functional parameters, occurrence, enzyme structure and stability, mutants and enzyme engineering, preparation and isolation, the application of enzymes, and ligand-related data. The data in BRENDA are manually curated from more than 79 000 primary literature references.

Employing published parameter values may simplify the parameter estimation task; however, like in the case of new experiments, it must be carefully considered whether the respective experimental conditions are compatible.

Conclusions

We demonstrated that it is in general impossible to infer non-measurable biochemical quantities such as rate constants, from measurable quantities such as changes in protein abundances, due to the nested uncertainties are various levels. At first, non-measurable biochemical parameters are concealed by the uncertainty about structure and kinetic rate laws that force the modeller to simplify the considered processes and lump processes that can be assumed constant into parameters. Moreover, there might be several possibilities how to simplify and lump processes and parameters. Then, even if the modeller has only one model, the lumped parameters may still not be identifiable due to limited data. In fact, we are not aware of any study, where a model was explicitly used to quantitatively infer signalling rates. There are, of course, many studies where signalling rates were estimated (see references above), but only to explain other biological behaviour or to infer regulation principles, like feedbacks. Signalling rates in models are not actual biochemical signalling rates, but simplified representations of complicated biochemical processes which are lumped together for the sake of simplicity and identifiability.

For model-based inference of non-measurable features from biochemical data, we see two major bottlenecks and room for improvement for the modelling side. First,

parameter estimation algorithms still tend to be time-consuming. Of course, this will be partly mitigated by steadily increasing computer capacities and the use of computer clusters. Still, for biochemical reaction networks usually global optimization algorithms are employed [14] for which no theory is available about convergence properties. Fast and efficient optimization algorithms would clearly also enhance progress in the field. Second, multiple model testing and discrimination is recognised to be an issue in systems biology [40]. There are first attempts to provide modelling frameworks that facilitate multiple model testing; however, model discrimination should be more widely applied and accepted as a useful method to infer pathway structures [1].

We would like to emphasise a common pitfall in model development. Often, there is much more qualitative information put into a model of a signalling network, for example, the wiring scheme, than is actually supported by quantitative data. This leads to overparameterised models with non-identifiable parameters. Overparameterised models tend to show spurious effects and artefacts and may lead to wrong conclusions, especially concerning quantitative aspects. Here, a rigorous model discrimination analysis in combination with dedicated experiments is useful (J Schaber *et al.*, Automated ensemble modeling with modelMaGe: analyzing feedback mechanisms in the Sho1 branch of the HOG pathway. Unpublished data). It is advisable to try to simplify the original model and dissect in a systematic manner, what kind of quantitative information about the signalling network is actually supported by the data.

Reporting of models and parameter estimation results should be standardized or at least be more rigorously required by journals. Published model simulations and parameter estimation results must be completely comprehensible by interested readers. Otherwise the credibility of modelling approaches is seriously impaired. Fortunately, there are emerging standards and approaches into this direction [41–43].

Finally, we have the feeling that in the rather new interdisciplinary research field of systems biology both advantages and limitations of modelling approaches are not yet fully appreciated by the community. On the short term standards can mitigate this problem and on the long term students should be taught in an interdisciplinary manner.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Anderson J, Papachristodoulou A: **On validation and invalidation of biological models**. *BMC Bioinformatics* 2009, **10**:132.

2. • Marucci L, Santini S, di Bernardo M, di Bernardo D: **Derivation, identification and validation of a computational model of a novel synthetic regulatory network in yeast**. *J Math Biol* 2010.
A nice worked example of the iterative cycle of model development. Good for people beginners.

3. Schaber J, Liebermeister W, Klipp E: **Nested uncertainty in biochemical models**. *IET Syst Biol* 2009, **3**:1-9.

4. Klipp E, Herwig R, Kowald A, Wierling C, Lehrach H: *System Biology in Practice: Concepts, Implementation and Application* Weinheim: Wiley–VCH; 2005.

5. Kreutz C, Timmer J: **Systems biology: experimental design**. *FEBS J* 2009, **276**:923-942.

6. Chou IC, Voit EO: **Recent developments in parameter estimation and structure identification of biochemical and genomic systems**. *Math Biosci* 2009, **219**:57-83.

7. Klipp E, Liebermeister W: **Mathematical modeling of intracellular signaling pathways**. *BMC Neurosci* 2006, **7(Suppl 1)**:S10.

8. Klipp E, Schaber J: Modelling of signal transduction in yeast — sensitivity and model analysis. In *Conference Proceedings of the International Symposium on System Biology: Understanding and Exploiting Systems Biology in Biomedicine and Bioprocesses*. Edited by: Fundación CajaMurcia; 2006:15–30. [Cánovas, M. Iborra, J.L., Manjón, A. (Series Editors)].

9. Klipp E, Schaber J: **Modeling the dynamics of stress activated protein kinases (SAPK) in cellular stress response**. *Top Curr Genet* 2008:205-224.

10. Hasenauer J, Waldherr S, Wagner K, Allgower F: **Parameter identification, experimental design and model falsification for biological network models using semidefinite programming**. *IET Syst Biol* 2010, **4**:119-130.

11. Bellu G, Saccomani MP, Audoly S, D'Angio L: **DAISY: a new software tool to test global identifiability of biological and physiological systems**. *Comput Methods Prog Biomed* 2007, **88**:52-61.

12. Saccomani MP, Audoly S, Bellu G, D'Angio L: **Examples of testing global identifiability of biological and biomedical models with the DAISY software**. *Comput Biol Med* 2010, **40**:402-407.

13. Ashyraliyev M, Fomekong-Nanfack Y, Kaandorp JA, Blom JG: **Systems biology: parameter estimation for biochemical models**. *Febs J* 2009, **276**:886-902.

14. Moles CG, Mendes P, Banga JR: **Parameter estimation in biochemical pathways: a comparison of global optimization methods**. *Genome Res* 2003, **13**:2467-2474.

15. • Klipp E, Liebermeister W, Helbig A, Kowald A, Schaber J: **Systems biology standards — the community speaks**. *Nat Biotechnol* 2007, **25**:390-391.
Proposes an algorithm to analyse parameter identifiability in a multi-dimensional parameter space.

16. Seber GAF, Wild CJ: *Nonlinear Regression*. Wiley–Interscience; 2003.

17. • Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmuller U, Timmer J: **Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood**. *Bioinformatics* 2009, **25**:1923-1929.
Proposes an algorithm to analyse parameter identifiability in a multi-dimensional parameter space.

18. Maiwald T, Timmer J: **Dynamical modeling and multi-experiment fitting with PottersWheel**. *Bioinformatics* 2008, **24**:2037-2043.

19. Storn R, Price K: **Differential evolution — a simple and efficient heuristic for global optimization over continuous spaces**. *J Global Optim* 1997, **11**:341-359.

20. Roberts MA, August E, Hamadeh A, Maini PK, McSharry PE, Armitage JP, Papachristodoulou A: **A model invalidation-based approach for elucidating biological signalling pathways, applied to the chemotaxis pathway in *R. sphaeroides***. *BMC Syst Biol* 2009, **3**:105.

21. Klipp E, Nordlander B, Kruger R, Gennemark P, Hohmann S: **Integrative model of the response of yeast to osmotic shock**. *Nat Biotechnol* 2005, **23**:975-982.

22. Adiamah DA, Handl J, Schwartz JM: **Streamlining the construction of large-scale dynamic models using generic kinetic equations**. *Bioinformatics* 2010, **26**:1324-1331.

23. Rust MJ, Markson JS, Lane WS, Fisher DS, O'Shea EK: **Ordered phosphorylation governs oscillation of a three-protein circadian clock**. *Science* 2007, **318**:809-812.

24. Brettschneider C, Rose RJ, Hertel S, Axmann IM, Heck AJ, Kollmann M: **A sequestration feedback determines dynamics and temperature entrainment of the KaiABC circadian clock**. *Mol Syst Biol* 2010, **6**:389.

25. Zi Z, Liebermeister W, Klipp E: **A quantitative study of the Hog1 MAPK response to fluctuating osmotic stress in *Saccharomyces cerevisiae***. *PLoS One* 2010, **5**:e9522.

26. • Kollmann M, Lovdok L, Bartholome K, Timmer J, Sourjik V: **Design principles of a bacterial signalling network**. *Nature* 2005, **438**:504-507.

Nice example of a successful combination of modelling, parameter estimation and dedicated experiments.

27. Malleshaiah MK, Shahrezaei V, Swain PS, Michnick SW: **The scaffold protein Ste5 directly controls a switch-like mating decision in yeast**. *Nature* 2010, **465**:101-105.

28. • Macia J, Regot S, Peeters T, Conde N, Sole R, Posas F: **Dynamic signaling in the Hog1 MAPK pathway relies on high basal signal transduction**. *Sci Signal* 2009, **2**:ra13.

Nice example of a successful combination of modelling and dedicated experiments.

29. Stricker J, Cookson S, Bennett MR, Mather WH, Tsimring LS, Hasty J: **A fast, robust and tunable synthetic gene oscillator**. *Nature* 2008, **456**:516-519.

30. Schaber J, Kofahl B, Kowald A, Klipp E: **A modelling approach to quantify dynamic crosstalk between the pheromone and the starvation pathway in baker's yeast**. *FEBS J* 2006, **273**:3520-3533.

31. Heinrich R, Neel BG, Rapoport TA: **Mathematical models of protein kinase signal transduction**. *Mol Cell* 2002, **9**:957-970.

32. Kapuy O, Barik D, Sananes MR, Tyson JJ, Novak B: **Bistability by multiple phosphorylation of regulatory proteins**. *Prog Biophys Mol Biol* 2009, **100**:47-56.

33. Flöttmann M, Schaber J, Hoops S, Klipp E, Mendes P: **ModelMage: a tool for automatic model generation, selection and management**. *Genome Inform* 2008, **20**:52-63.

34. Hao N, Behar M, Parnell SC, Torres MP, Borchers CH, Elston TC, Dohlman HG: **A systems-biology analysis of feedback inhibition in the Sho1 osmotic-stress-response pathway**. *Curr Biol* 2007, **17**:659-667.

35. Clausznitzer D, Oleksiuk O, Lovdok L, Sourjik V, Endres RG: **Chemotactic response and adaptation dynamics in *Escherichia coli***. *PLoS Comput Biol* 2010, **6**:e1000784.

36. Cantone I, Marucci L, Iorio F, Ricci MA, Belcastro V, Bansal M, Santini S, di Bernardo M, di Bernardo D, Cosma MP: **A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches**. *Cell* 2009, **137**:172-181.

37. Westfall PJ, Patterson JC, Chen RE, Thorner J: **Stress resistance and signal fidelity independent of nuclear MAPK function**. *Proc Natl Acad Sci U S A* 2008, **105**:12212-12217.

38. Bishop AC, Ubersax JA, Petsch DT, Matheos DP, Gray NS, Blethrow J, Shimizu E, Tsien JZ, Schultz PG, Rose MD *et al.*: **A chemical switch for inhibitor-sensitive alleles of any protein kinase**. *Nature* 2000, **407**:395-401.

39. •• van Eunen K, Bouwman J, Daran-Lapujade P, Postmus J, Canelas AB, Mensonides FI, Orij R, Tuzun I, van den Brink J, Smits GJ, *et al.*: **Measuring enzyme activities under standardized in vivo-like conditions for systems biology**. *FEBS J* 2010 277:749–760

The authors developed a single assay medium for determining enzyme-kinetic parameters in baker's yeast under conditions as close as possible to the in vivo situation instead of situations optimal for a single enzyme. Seminal paper since it takes the challenge to experimentally determine parameters crucial for quantitative systems biology approaches, in a field where everything seemed to been analysed already in the last century.

40. Muzzey D, Gomez-Uribe CA, Mettetal JT, van Oudenaarden A: **A systems-level analysis of perfect adaptation in yeast osmoregulation**. *Cell* 2009, **138**:160-171.

41. Le Novere N, Finney A, Hucka M, Bhalla US, Campagne F, Collado-Vides J, Crampin EJ, Halstead M, Klipp E, Mendes P *et al.*: **Minimum information requested in the annotation of biochemical models (MIRIAM)**. *Nat Biotechnol* 2005, **23**:1509-1515.

42. Le Novere N, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A, Demir E, Wegner K, Aladjem MI, Wimalaratne SM *et al.*: **The systems biology graphical notation**. *Nat Biotechnol* 2009, **27**:735-741.

43. Finney A, Hucka M: **Systems biology markup language: level 2 and beyond**. *Biochem Soc Trans* 2003, **31**:1472-1473.

44. Rojas I, Golebiewski M, Kania R, Krebs O, Mir S, Weidemann A, Wittig U: **Storing and annotating of kinetic data**. *In Silico Biol* 2007, **7**:S37-44.

45. Chang A, Scheer M, Grote A, Schomburg I, Schomburg D: **BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009**. *Nucleic Acids Res* 2009, **37**:D588-592.

46. Hoops S, Sahle S, Gauges R, Lee C, Pahle J, Simus N, Singhal M, Xu L, Mendes P, Kummer U: **COPASI — a COmplex PAthway SImulator**. *Bioinformatics* 2006, **22**:3067-3074.

47. Schmidt H, Jirstrand M: **Systems Biology Toolbox for MATLAB: a computational platform for research in systems biology**. *Bioinformatics* 2006, **22**:514-515.

48. Zi Z, Klipp E: **SBML-PET: a Systems Biology Markup Language-based parameter estimation tool**. *Bioinformatics* 2006, **22**:2704-2705.