Published in IET Systems Biology
Received on 31st August 2007
Revised on 24th June 2008
doi: 10.1049/iet-syb:20070042

ISSN 1751-8849

# Nested uncertainties in biochemical models

*J. Schaber W. Liebermeister E. Klipp*

*Humboldt University Berlin, Institute for Biology, Theoretical Biophysics, Invalidenstr. 42, 10115 Berlin, Germany*
*E-mail: klipp@molgen.mpg.de*

**Abstract:** Dynamic modelling of biochemical reaction networks has to cope with the inherent uncertainty about biological processes, concerning not only data and parameters but also kinetics and structure. These different types of uncertainty are nested within each other: uncertain network structures contain uncertain reaction kinetics, which in turn are governed by uncertain parameters. Here, the authors review some issues arising from such uncertainties and sketch methods, solutions and future directions to deal with them.

## 1 Introduction

Mathematical modelling of dynamic biological processes is an integral part of systems biology. Mathematical models have proven to be very useful to explain physical and biological principles since over two centuries. Nowadays, models are also widely used to address more specific questions arising from biological and medical experiments. The choice of the model type crucially depends on the question we have about the system. General questions such as 'what are the possible reaction mechanisms that result in a step-like dose-response curve?' or 'what mechanism can explain hysteresis?' can be approached by mathematical models that one can solve on paper [1, 2]. However, to model specific systems in a quantitative manner, we need to confront our understanding of complex reaction networks with experimental data at hand and to use computationally more demanding approaches, like systems of differential equations. When model development is driven by experimental data, the focus is not only on qualitative types of behaviour but rather on quantitative predictions.

Despite enormous efforts in experimental research in cellular and molecular biology, there is still a substantial uncertainty in the qualitative and quantitative aspects of biochemical networks, including protein–protein interactions, transcription and translation as well as metabolic fluxes or metabolite concentrations. To model nonlinear dynamic processes in cells, one needs quantitative time-resolved data. Often, only the presence of a gene, mRNA or a protein has been demonstrated while their concentration can only be roughly estimated, and even this holds only for a limited number of compounds. Moreover, data produced by measurements may not be appropriate for the computational approaches. For example, quantities are measured *in vitro* instead of *in vivo* – which makes conclusions about the living system questionable – or in cell populations instead of single cells, which obscures cell-to-cell variability.

When constructing a biochemical model, a number of choices have to be made: we have to choose an appropriate model structure based on hypothesised interactions of biochemical components, and for the reactions, we have to choose kinetic rate laws and the corresponding enzymatic parameters. At this stage, the true wiring scheme and the true parameters are uncertain. These uncertainties need to be resolved by confronting model alternatives with experimental data. Even after considering the data, some of the uncertainty will remain, but this uncertainty can in turn be quantified: uncertainty in parameters can be described by confidence intervals or posterior probability densities, whereas uncertainty in structures can be described by probabilities or rankings of models.

However, network structure, kinetic laws and kinetic parameters cannot be determined independently. On the one hand, we need to specify model structure and kinetics in advance to determine kinetic constants; on the other, to choose between model structures, we need to judge their quality, for instance, by their ability to reproduce given data. Hence, the uncertainties on different levels are nested and have to be resolved by a combination – or rather, iteration – of model selection and parameter fitting.

In the following, we review some issues arising from uncertainties concerning kinetic constants, kinetic laws and network structure. We sketch problems, solutions and future directions for dealing with them and present workflows that handle uncertainties in a consistent and rational manner.

## 2 Parameter uncertainty

Let us first assume that the structure and kinetic laws of a biochemical network are known and just the parameter values $\theta$ need to be determined. Available parameter values may be uncertain because of experimental errors or biological variability or because they were measured *in vitro*. For other parameters, we may only have rough guesses about the order of magnitude or no information at all. A second source of information is dynamical data: if a model successfully describes, for instance, concentration time courses, the parameter values used will seem more plausible. In parameter estimation, we intend to determine the most plausible parameter values given all the available information and to assess the remaining uncertainty about them. Some methods will yield, as a result, confidence intervals or probability distributions for the parameters. This information can later be used to assess the range of model predictions via Monte Carlo sampling.

### 2.1 Maximum likelihood estimation

If parameter values are completely unknown in advance, a standard method to determine them is to fit the model to sets of experimental data $\boldsymbol{y}$. Such a parameter estimation is usually based on the principle to maximise the likelihood

$$L(\boldsymbol{y}, \boldsymbol{\theta}) = P(\boldsymbol{y} \mid \boldsymbol{\theta}_f)$$

defined as the probability to observe the data $\boldsymbol{y}$ from model $f$ with parameter vector $\boldsymbol{\theta}_f$. We shall not explicitly mention the model subscript $f$ henceforth. Under the assumption of independent standard Gaussian measurement errors, maximising the likelihood is equivalent to minimising the sum of squared residuals (SSR), that is, the squares of the distance between experimental data points and the respective results of the model simulation. With given data $\boldsymbol{y}$, the SSR is a function of the parameter vector and the estimation task boils down to finding the minima of this function.

The main problems in parameter optimisation are (i) the nonlinear nature of ODEs, causing numerical instabilities in the calculations, (ii) the high dimensionality of the parameter space, possibly leading to many local minima and (iii) the notorious disproportion between a small number of data points and a large number of parameters, which will result in overfitting. In summary, we face a number of risks including not to sample the parameter space appropriately, to be stuck in a local minimum, to find a global minimum that is still different from the biological reality or to proceed too slowly in approaching the minimum. If we need to rule out suboptimal local minima, we cannot use local gradient-based methods, but must use global, computationally demanding methods. An overview of the most popular parameter estimation algorithms in biology and biochemistry is given in [3]. A number of algorithms are implemented in the tools of systems biology such as Copasi [4], SBML-PET [5] or SBToolBox [6].

### 2.2 Bayesian parameter estimation

If prior information about the model parameters is available (e.g. ranges or probability distributions for kinetic constants), it should be taken into account in the parameter estimation. Bayesian parameter estimation methods [7] can determine a compromise between such prior knowledge and the information obtained from the likelihood function. The result is a posterior probability distribution over the parameter sets: a high probability shows that a certain parameter set seems plausible in the light of all the information considered.

Formally, model parameters $\boldsymbol{\theta}$ and experimental data $\boldsymbol{y}$ are described by a joint probability distribution $P(\boldsymbol{y}, \boldsymbol{\theta})$. Information about the parameter values has to be specified in the form of the prior distribution $P(\boldsymbol{\theta})$, the marginal distribution of the parameters. By combining the prior with the likelihood function $P(\boldsymbol{y}|\boldsymbol{\theta})$, one obtains the posterior probability distribution $P(\boldsymbol{\theta}|\boldsymbol{y})$ of the parameter values given in the data. According to Bayes' theorem, it can be computed by $P(\boldsymbol{\theta}|\boldsymbol{y}) = P(\boldsymbol{y}|\boldsymbol{\theta}) \cdot P(\boldsymbol{\theta})/P(\boldsymbol{y})$ where the marginal distribution $P(\boldsymbol{y})$ of the data only appears as a normalisation constant.

In simple cases (e.g. linear models and Gaussian distributions), the posterior can be computed analytically. If this is not possible, it can be characterised by sampling methods like Monte Carlo Markov chains [7] or approximated, for instance, by a Gaussian distribution [8]. A Gaussian posterior maximum represents a most plausible parameter set (the centre of the Gaussian) together with the remaining uncertainties (given by the covariance matrix). This information can be used as a starting point for further modelling. Bayesian methods in bioinformatics and computational systems biology have been reviewed recently [9].

### 2.3 Estimation of kinetic constants based on heterogeneous and uncertain data

Many kinetic parameters have been published in the literature, but they cannot be directly inserted into models: parameters measured *in vitro* or under different conditions may be unreliable or incompatible with each other [10]. To obtain a consistent, complete set of kinetic parameters, we need to guess the uncertainties of known values, make reasonable assumptions about unknown values, and find a 'balanced' parameter set that appears most plausible in the




doi: 10.1049/iet-syb:20070042

light of all the parameter values collected. Bayesian statistics provides a rational method to do that.

*2.3.1 Parameter balancing:* Statistical distributions can be used not only to describe actual parameter variability (e.g. protein expression) within cell populations, but also our ignorance or beliefs about parameter values. In Bayesian parameter balancing, measured parameter values are treated as data $\boldsymbol{y}$ and each kinetic parameter has a prior distribution that expresses our general beliefs. The empirical distribution of all known $K_M$ values, for instance, may serve as a prior for an unknown $K_M$ value [11]. Eventually, we obtain a posterior distribution that describes the remaining uncertainty owing to missing knowledge, measurement uncertainties or biological variability. The resulting posterior distribution can be used to sample parameter sets for Monte Carlo simulations, which in turn give us probabilistic statements about the dynamic behaviour of the model (analytic results within a linear approximation are described in [12]).

If we intend to describe kinetic parameters by probability distributions, a problem remains to be solved: thermodynamic laws may lead to dependencies among the parameters in a network and, hence, drawing parameters independently from a statistical distribution would almost certainly lead to thermodynamically wrong models. Thus, we need to construct a parameter distribution that satisfies the thermodynamic constraints. The key idea is to choose a different parametrisation of the model, in which any combination of parameter values is feasible: several approaches have been developed for this purpose [8, 13, 14]. As an example, we shall discuss here parameter balancing for the convenience kinetics [8].

*2.3.2 Convenience kinetics:* The convenience kinetics [8] is a generalised form of the Michaelis–Menten kinetics for arbitrary numbers of substrates and products (Fig. 1). It allows for describing enzyme saturation, inhibition and activation and can be used to model enzymes with an unknown kinetic law. For a two-substrate/one-product reaction A + B ↔ C with an activator X and an inhibitor Y (concentrations $a$, $b$, $c$, $x$, $y$), the reaction rate reads

$$v(a,b,c,x,y) = E\,\frac{\tilde{x}}{1+\tilde{x}}\,\frac{1}{1+\tilde{y}}\,\frac{k_+\,\tilde{a}\,\tilde{b} - k_-\,\tilde{c}}{(1+\tilde{a})(1+\tilde{b})+(1+\tilde{c})-1} \tag{1}$$

where $E$ denotes the enzyme concentration and $\tilde{a} = a/k_A$, $\tilde{b} = b/k_B$, $\tilde{c} = c/k_C$, $\tilde{x} = x/k_X$ and $\tilde{y} = y/k_Y$ denote scaled metabolite concentrations. Each metabolite concentration is scaled by its corresponding enzyme parameter, in this case reactant constants (corresponding to Michaelis constants) $k_A$, $k_B$ and $k_C$, the activation constant $k_X$ and inhibition constant $k_Y$. In addition, there are the two turnover rates $k_+$ and $k_-$ for the forward and backward reactions, respectively.

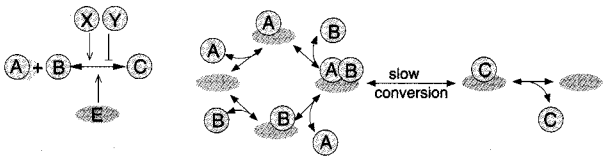


**Figure 1** *Convenience kinetics*

Left: scheme of the example reaction (see text)
Right: enzymatic mechanism
Binding of the substrates (A, B) and the product (C) to the enzyme (grey oval) is assumed to be fast, reversible, non-cooperative and in random order
Activation and inhibition effects are not shown here
For details see [8]

*2.3.3 Thermodynamically correct parameters and parameter balancing:* If chemical reactions form a reaction network, the kinetic constants of different reactions are constrained by thermodynamic laws. To yield a zero reaction rate at thermal equilibrium, the parameters have to satisfy the Haldane relation

$$k_{eq} = \frac{k_+}{k_-}\,\frac{k_C}{k_A\,k_B} \tag{2}$$

where the equilibrium constant $k_{eq} = c/(a \cdot b)$ denotes the concentration ratio in thermodynamic equilibrium. For thermodynamic reasons, the equilibrium constant is determined by the Gibbs free energies of formation of the reactants via

$$\ln k_{eq} = \frac{\left[G_A^{(0)} + G_B^{(0)} - G_C^{(0)}\right]}{RT} \tag{3}$$

where $R$ is Boltzmann's gas constant and $T$ is the temperature in Kelvin.

Equations (2) and (3) together lead to – possibly complicated – constraints between the parameters [8]. This problem – which would also arise with reversible mass-action or Michaelis–Menten kinetics – can be solved by introducing a new parametrisation. By rewriting (2) as

$$\ln k_+ - \ln k_- = \ln k_{eq} + \ln k_A + \ln k_B - \ln k_C \tag{4}$$

and inserting (3), we obtain

$$\ln k_+ - \ln k_- = \frac{\left[G_A^{(0)} + G_B^{(0)} - G_C^{(0)}\right]}{RT} + \ln k_A + \ln k_B - \ln k_C$$

If we introduce a new parameter ('velocity constant")

$$\ln k_R = \frac{1}{2}\left[\ln k_+ + \ln k_-\right]$$

we can express the turnover rates $k_+$ and $k_-$ by the remaining

parameters

$$\ln k_{\pm} = \ln k_R \pm \frac{1}{2}\left[\frac{\left[G_A^{(0)} + G_B^{(0)} - G_C^{(0)}\right]}{RT} + \ln k_A + \ln k_B - \ln k_C\right]$$

With such equations for the turnover rates, an entire network model can be parametrised in a thermodynamically correct manner: the reactant constants ($k_A$ etc.), Gibbs free energies of formation ($G_A^{(0)}$ etc.) and velocity constants ($k_R$) are regarded as model parameters, which are thermodynamically unconstrained and for which probability distributions can be specified. The turnover rates $\ln k_{\pm}$, on the other hand, are computed from the model parameters. For an entire model, we collect the independent parameters (in logarithmic form as above) in a vector $\boldsymbol{\theta}$ and all kinetic parameters (including the dependent ones, also in logarithmic form) in a vector $\boldsymbol{x}$. Both vectors are then related by the linear equation

$$\boldsymbol{x} = \boldsymbol{R}_{\theta}^{x}\theta \tag{5}$$

where the matrix $\boldsymbol{R}_{\theta}^{x}$ can be easily derived from the network structure [8].

The issue of thermodynamic correctness arises from the numerator of the kinetic term in (1) and does not regard the activation and inhibition parameters. The re-parametrisation also works for other kinetics in which the Haldane relation has a multiplicative form like (2), in particular mass-action kinetics, Michaelis–Menten kinetics and variants with different activation or inhibition mechanisms.

All kinetic constants in the convenience kinetics can be measured, but for a given network of interest, only some of their values may be known and the remaining ones may be uncertain. In Bayesian parameter balancing, we can use these values as clues to determine a set of 'balanced' system parameters, which is complete and thermodynamically correct. If some or all kinetic constants for a model have been measured (experimental data in a vector $\boldsymbol{y}$ corresponding to the model values $\boldsymbol{x}$), then solving the linear relation $\boldsymbol{y} \simeq \boldsymbol{R}_{\theta}^{x}\theta$ corresponding to (5) would be a simple way to compute the parameter vector $\boldsymbol{\theta}$. If enough data are available, it can be solved in the sense of least squares. A more general and statistically justified way to determine $\boldsymbol{\theta}$ is by Bayesian parameter estimation [15] as described in Section 2.2, using a prior distribution for $\boldsymbol{\theta}$.

*2.3.4 Prior distributions for kinetic parameters:* Bayesian parameter estimation relies on an appropriate choice of the prior distributions. In parameter balancing, for instance, the prior is especially important for those parameters in the vector $\boldsymbol{\theta}$ for which no data are available. To specify a prior, we need to specify both the mathematical form (e.g. log-normal distribution) and the hyperparameters (e.g. mean values and covariances) that determine its exact size and shape. Gaussian distributions for logarithmic kinetic parameters are biologically plausible and together with a linear relation like (5) lead to a simple Gaussian posterior.

To determine the prior mean and width for a certain parameter type, say $K_M$ values, one could fit a Gaussian distribution to the empirical distribution of all logarithmic $K_M$ values available in databases like Brenda [16]. More accurate priors can be derived for individual $K_M$ values based on an analysis of variance with the enzyme and the organism as factors [11]. We applied this analysis to $\log_{10}$ $K_M$ values from the enzyme database Brenda and assessed the quality of predictions with leave-one-out cross-validation. The quality of predictions depended on the studied substrate. The overall correlation coefficient between measured and predicted $\log_{10}$ $K_M$ was 0.77. The resulting predictions and error ranges for enzyme parameters can be used for defining individual priors for $k^M$ values in convenience kinetics [11].

*2.3.5 Workflow for building kinetic models:* We have assembled the methods and protocols into a modelling workflow [17] that facilitates the integration of biochemical data and the modelling of metabolic networks from scratch (Fig. 2).

In brief, it translates a given metabolic network structure into a draft version of a kinetic model with convenience kinetics; known kinetic laws [18] can also be inserted into the models. Such an automatic workflow cannot replace manual kinetic modelling, but it can ease it: effectively, the workflow provides modellers with a simple way to query several enzyme databases in the context of a specific model. The result is a posterior parameter distribution that can be used to define parameter ranges for further manual modelling, and it can be directly used as a parameter prior for model fitting. The variances of model parameters show where information is missing and can point at additional measurements that would provide the most valuable information.

# 3 Uncertainty about kinetic rate laws and network structure

Although parameter estimation is a demanding issue in itself, the main challenge in modelling is often to determine the structure of the network, in terms of choosing appropriate rate laws, that is, kinetics, as well as the wiring scheme itself. Metabolic networks are to a large extent specified by a list of enzymes, which can be obtained from genome sequencing, sequence comparison, traditional biochemical protein characterisation and function identification. Still, there is some uncertainty, since not all enzymes are identified and the role of isoenzymes is not always clear.




doi: 10.1049/iet-syb:20070042

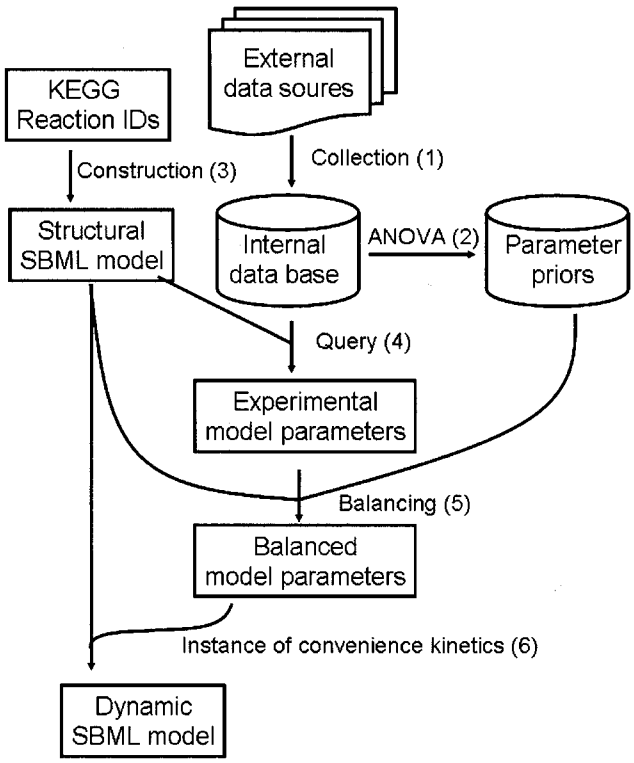


**Figure 2** *Workflow for translating a metabolic network into a dynamic model*

(1) Thermodynamic and kinetic parameters are collected from literature and databases
(2) Statistical learning methods [11] yield prior distributions for types of parameters and for individual parameter values
(3) A biochemical network of interest (in SBML format) is constructed from a list of KEGG reaction identifiers
Relevant experimental parameter values for the network are retrieved (4), balanced (5) and inserted into the model (6)
The resulting SBML model contains convenience kinetics with the balanced parameters

This is also documented by the regular updates of large-scale stoichiometric metabolic models [19, 20].

Signalling pathways and regulatory networks are usually less well defined. First, it is often not clear if all relevant compounds have been identified. Even then, the precise order of interaction and modification events within the pathway is often not known. For example, although MAP kinase cascades have been studied intensively, it is often not clear if the multiple phosphorylation observed at various levels occurs upon a single or successive binding events of the upstream kinase. Moreover, even if the substrates, modifiers and products of a reaction are known, the exact kinetic law usually remains elusive.

In computational modelling of biochemical networks, we can use different strategies to cope with the uncertain network structure. First, one may include all possible reactions related to the set of proteins in the network. An example is the pheromone pathway model of [21]. Since the precise binding order of the MAP kinases Ste11p, Ste7p and Fus3p to the scaffold protein Ste5p is not known and the order of phosphorylation events is not resolved, all potential combinations of binding and phosphorylation sequences have been considered in the model. Obviously, this approach can lead to an enormous number of models that are all compatible with the existing knowledge. Another possible choice is to include exactly those compounds into the model for which quantitative data are available, and to describe their connections with 'black box' functions instead of the traditional biochemical reaction kinetics. For yeast MAP kinase cascades, this approach would basically reduce the number of model compounds to two: the added stimulus and the activated MAP kinase. All other players are currently hardly accessible. Finally, one may choose a compromise and describe the network on an intermediate level of granularity based on the available information and on intuition [22, 23]. A model of this type represents the compounds or processes that are considered important or relevant; in a way, it summarises the literature and expert knowledge (and convention) and employs it to answer specific questions, for instance, to predict the behaviour of a compound that is experimentally not accessible.

## 3.1 Automatic construction of candidate models

There are several practical issues related to the modelling of uncertainty of network structures. Possible combinations of uncertain structures and kinetics directly translate into alternative mathematical models. Generating and handling such alternative models poses a considerable challenge to the modeller for several reasons. First, each model has to be implemented, simulated and analysed separately. Often, model alternatives vary only slightly in structure and/or kinetics. This may seduce the modeller to copy-paste the original model and then introduce the modifications by hand. This is an error-prone process. Secondly, changes that affect the whole family of models have to be updated in each model separately, which is also an error-prone and tedious task, especially when the number of models is high. The combinatorial complexity often renders it impossible to implement and handle each model individually. Several formalisms and tools were developed to address these problems. We can roughly distinguish between two groups: some of them help to automatically generate models and others handle the combinatorial complexity.

As examples for the first group, the tools Cellerator [24] and MMT2 [25] can be mentioned. Cellerator is a Mathematica® package designed to facilitate biological modelling via automated equation generation. Reactions are specified by an arrow-based reaction notation from which a single model is created. MMT2 is a software tool that produces all possible network variants of metabolic networks by switching reactions off and on. Subsequently,

unrealistic models are sorted out by analysis of extreme pathways and elementary flux modes.

As examples for the second group, BioNetGen [26] and Moleculizer [27] can be mentioned. BioNetGen is a language that focuses on protein–protein interactions. The user can specify rules from which one single model is created automatically. BioNetGen is designed to handle the combinatorial complexity of protein complex formation with modifications, for example, a receptor with $n$ phosphorylation sites can occupy $2^n$ different states; if it dimerises $2^2n$, different states are allowed and so on. Moleculizer is a stochastic simulator for intracellular biochemical systems, with special treatment for protein complexes. It resembles BioNetGen but avoids using the network of all possible complexes and reactions, based on their probability of occurrence. Only one single model is generated and simulated.

These state-of-the-art approaches have a major shortcoming: even though most tools aim at handling combinatorial complexity, they produce only one model at the end, which includes all or a reduced number of possible molecular interactions generated from certain rules. Currently, there is no tool that automatically implements and manages different models, which differ in the number of components, reactions and kinetics. However, this is what modellers in systems biology are confronted with. MMT2 aims into that direction but it falls short in the ability to actually control the kind of generated models, because it does not allow for alternatively removing components or employing alternative kinetics.

In our daily work and discussions with the community, we see that it is not so much the combinatorial complexity that creates problems for the modellers but rather its management and handling of a specific set of candidate models. Often, modellers have a very clear idea what different versions of a model they want to implement, simulate and fit to data. It is just tedious and error-prone to it all by hand.

Another hotly debated issue is model documentation [28, 29]. It is not only the successful models that are of interest to the research community, but also those that failed. Usually, in the course of a modelling project, many unsuccessful model versions are tested but only the successful one is finally published. The unsuccessful versions, even though of interest, are never documented, because this is also a laborious task but not rewarded. Finally, having several model alternatives at hand, it is again painful to simulate each one individually and to compare the results to select and refine the best ones.

To handle uncertainty in kinetics and model structure, we developed a tool that automatically generates candidate models based on a root or master model and specified modifications and directives (36) (http://modelMaGe.org). The generated models are automatically documented such that it is always comprehensible how they were derived from the master model, thereby keeping track of model versions and alternatives. This facilitates that common parameters, modifications or directives are changed in only one place and automatically updated in each model, which removes the errors introduced by modifying each model individually. Finally, all generated models are simulated, fitted to data and compared automatically. At the end, the user is provided with a ranking of the model fits and statistical measures that enables him to discriminate between model alternatives.

## 3.2 Model selection and nested uncertainty

To cope with the uncertainty in structure or kinetic laws, one may either choose the most plausible or most interesting network structure or one may fit all model alternatives, including alternative structures and kinetics, to the available experimental data. There exist a number of criteria to discriminate or select among alternative fitted candidate models. For nested models, that is where a model is a restricted version of the other one, statistical measures like an $F$-test or likelihood-ratio test can be used. Nested and non-nested models can be ranked according to the Akaike information criterion (AIC) [30, 31], the minimum description length (MDL) [32] or others criteria that are based on information theory. The model with the lowest AIC or MDL is selected (for a review on model selection see [33]).

Alternative model structures can be generated from a master model, which includes all possible components and reactions, by leaving out sets of specified components and/or reactions. In this way, an acyclic graph of models is created, which are partly nested within each other. From these alternative model structures, alternative kinetic variants are derived, which in turn can be nested within each other. These model alternatives are eventually fitted to data. Fig. 3 shows a graph of models that have been generated from a master model $M_1$. Leaving out component A or C or both generates different models. Each structural alternative has two kinetic alternatives, indicated by $k_1$ and $k_2$, where $k_2$ is assumed to be nested in $k_1$. When each model is fitted to data, a simple heuristic of traversing the graph and comparing models by, for example, their likelihood ratio for nested models and by the AIC for non-nested models eventually yields one final model.

If the precise structure of the model network cannot be defined on the basis of established knowledge but needs to be determined through the model selection processes, we must be aware that parameters for different networks have different meanings. For example, if a double phosphorylation in a MAP kinase cascade is described with either one or two individual reactions, then the kinetic constants of both scenarios have different meanings and may have different units too.




doi: 10.1049/iet-syb:20070042

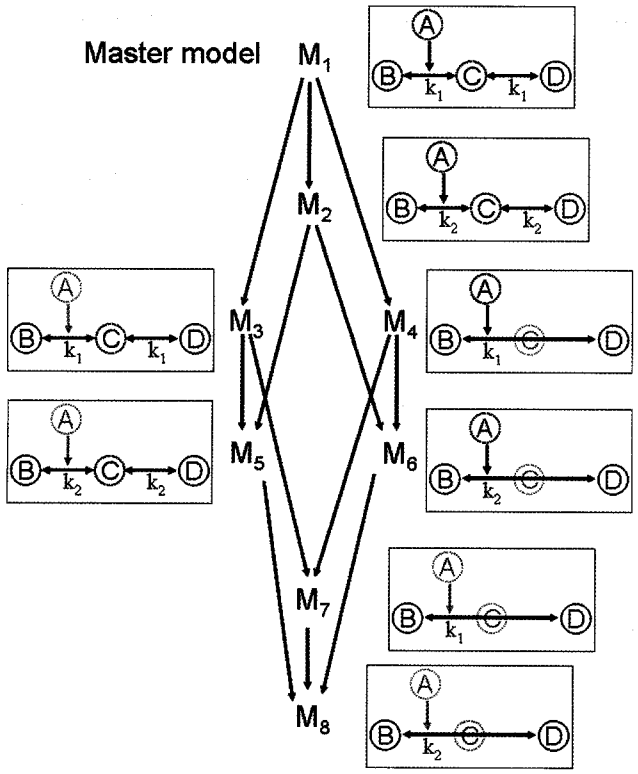


**Figure 3** *An acyclic graph of models that are generated from a master model $M_1$*

Models are generated by leaving out component A or C or both Each structural alternative has two kinetic alternatives, indicated by $k_1$ and $k_2$

Of course, none of these measures is made to evaluate how good a model in fact describes the essence of a biological process or a biological experiment. The value of a quantitative model is indicated by how good it can predict experimental results that have not been used to construct the model. Those tests include simulation and measurement of the effect of mutants, be it gene deletion, loss or gain of function mutants or gene over-expression. This statement again relates to the problem of model structure and parameter value estimation: if knowledge about a mutant has been used to define the network structure, then correct prediction of the mutant behaviour is a necessary condition to accept the model, not a new prediction. Another useful test is the prediction of the effect of different external stimuli [34].

Thus, model selection or system identification should be in fact an iterative procedure of repeated cycles of producing candidate models, parameter fitting and predictions [35].

# 4 Discussion

Building of dynamical cell models is hampered by difficulties on various levels: despite strong efforts in high-throughput and high-quality data generation, we still face a lack of good quantitative time-resolved data and derived kinetic parameters. In principle, parameters can also be estimated from model fitting, but parameter estimation methods are often not powerful enough to cope with large nonlinear differential equation systems and sparse data. Finally, we have to manage the nested uncertainties in model structure, rate laws and parameter values. Although the systems biology community is developing methods and tools to handle these issues, they are still largely unresolved and need further efforts.

Here, we have discussed practical and theoretical issues related to the construction of biochemical models. To integrate diverse data into models with given network structure, one may explicitly consider parameter distributions instead of fixed parameter values. After collecting original data about kinetic constants, Bayesian balancing allows to combine them with relatively loose prior assumptions. The result is a kinetic model with the convenience rate law and balanced, consistent parameters. The parameter posterior can also be used as a prior for further rounds of Bayesian estimation in which dynamic quantities – such as measured metabolite concentrations and fluxes – are integrated.

Statistical and information theory provides some theoretical approaches to discriminate between rival model formulations, concerning both structure and kinetics. Generally, we gain more confidence in our understanding of biological processes by iteratively constructing model variants, estimating parameters and predicting independent experiments. However, we still see a lack of practical tools that facilitate management and handling of alternative model formulations.

Dynamic modelling of cellular networks often has to cope with data of poor quality, such as missing or contradictory parameters, data obtained from different experiments or *in vitro* data. Even the best modelling workflow and model management tool cannot construct a reliable model from poor data.

We expect that larger amounts of specific, high-quality, quantitative, time-resolved, single-cell data will be available in the near future. This will help us to understand the organisation and regulatory principles of metabolic or signalling pathways and will enable us to derive useful predictions from the models.

# 5 Acknowledgments

This work was supported by the Max Planck Society and through funds of the European commission to EK (Network of Excellence ENFIN, Project number LSHG-CT-2005-518254; the Yeast Systems Biology Network, Project number LSHG-CT-2005-018942; IP BaSysBio, Project number LSHG-CT-2006-037469, STP EU-project CellComput; Project number 043310). We thank

all members of the Theoretical Biophysics group at Humboldt University, Berlin.

## 6 References

[1] FERRELL J.E. JR.: 'Tripping the switch fantastic: how a protein kinase cascade can convert graded inputs into switch-like outputs', *Trends Biochem. Sci.*, 1996, **21**, (12), pp. 460–466

[2] ELOWITZ M.B., LEIBLER S.: 'A synthetic oscillatory network of transcriptional regulators', *Nature*, 2000, **403**, (6767), pp. 335–338

[3] MOLES C.G., MENDES P., BANGA J.R.: 'Parameter estimation in biochemical pathways: a comparison of global optimization methods', *Genome Res.*, 2003, **13**, (11), pp. 2467–2474

[4] HOOPS S., SAHLE S., GAUGES R., *ET AL.*: 'COPASI: a COmplex PAthway Simulator', *Bioinformatics*, 2006

[5] ZI Z., KLIPP E.: 'SBML-PET: a systems biology Markup language-based parameter estimation tool', *Bioinformatics*, 2006, **22**, (21), pp. 2704–2705

[6] SCHMIDT H., JIRSTRAND M.: 'Systems biology toolbox for MATLAB: a computational platform for research in systems biology', *Bioinformatics*, 2006, **22**, (4), pp. 514–515

[7] GELMAN A., CARLIN J.B., STERN H.S., RUBIN D.B.: 'Bayesian Dara analysis' in CHATFIELD C., ZIDEK J.V. (EDS.): 'Texts in statistical science', (Chapman and Hall, Boca Raton, 1995), p. 526

[8] LIEBERMEISTER W., KLIPP E.: 'Bringing metabolic networks to life: convenience rate law and thermodynamic constraints', *Theor. Biol. Med. Model*, 2006, **3**, p. 41

[9] WILKINSON D.J.: 'Bayesian methods in bioinformatics and computational systems biology', *Brief. Bioinform.*, 2007, **8**, (2), pp. 109–116

[10] CHASSAGNOLE C., RAIS B., QUENTIN E., FELL D.A., MAZAT J.P.: 'An integrated study of threonine-pathway enzyme kinetics in Escherichia coli', *Biochem. J.*, 2001, **356**, (Pt 2), pp. 415–423

[11] BORGER S., LIEBERMEISTER W., KLIPP E.: 'Prediction of enzyme kinetic parameters based on statistical learning', *Genome Inform.*, 2006, **17**, (1), pp. 80–87

[12] LIEBERMEISTER W., KLIPP E.: 'Biochemical networks with uncertain parameters', *IEE Syst. Biol.*, 2005, **152**, (3), pp. 97–107

[13] ROHWER J.M., HANEKOM A.J., HOFMEYR J.-H.S.: 'A universal rate equation for systems biology'. ESCEC 2006, Ruedesheim/Rhein, Germany

[14] EDERER M., GILLES E.D.: 'Thermodynamically feasible kinetic models of reaction networks', *Biophys. J.*, 2007, **92**, (6), pp. 1846–1857

[15] LIEBERMEISTER W., KLIPP E.: 'Bringing metabolic networks to life: integration of kinetic, metabolic, and proteomic data', *Theor. Biol. Med. Model*, 2006, **3**, p. 42

[16] SCHOMBURG I., CHANG A., EBLING C., *ET AL.*: 'BRENDA, the enzyme database: updates and major new developments', *Nucleic Acids Res.*, 2004, **32**, (Database issue), pp. D431–D433

[17] BORGER S., UHLENDORF J., HELBIG A., LIEBERMEISTER W.: 'Integration of enzyme kinetic data from various sources', *In Silico Biol.*, 2007, **7**, (S1), p. 9

[18] WITTIG U., GOLOBIEWSKI M., KANIA R.: 'SABIO-RK: integration and curation of reaction kinetics data'. 3rd Int. Workshop on Data Integration in the Life Sciences, 2006, Hinxton, UK

[19] FEIST A.M., HENRY C.S., REED J.L., *ET AL.*: 'A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information', *Mol. Syst. Biol.*, 2007, **3**, article ID: 121

[20] HERRGÅRD M.J., SWAINSTON N., DOBSON P., *ET AL.*: 'A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology', *Nat. Biotechnol.*, 2008, **26**, pp. 1155–1160

[21] SHAO D., ZHENG W., QIU W., OUYANG Q., TANG C.: 'Dynamic studies of scaffold-dependent mating pathway in yeast', *Biophys. J.*, 2006, **91**, (11), pp. 3986–4001

[22] KOFAHL B., KLIPP E.: 'Modelling the dynamics of the yeast pheromone pathway', *Yeast*, 2004, **21**, (10), pp. 831–850

[23] 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**, (15), pp. 3520–3533

[24] SHAPIRO B.E., LEVCHENKO A., MEYEROWITZ E.M., WOLD B.J., MJOLSNESS E.D.: 'Cellerator: extending a computer algebra system to include biochemical arrows for signal transduction simulations', *Bioinformatics*, 2003, **19**, (5), pp. 677–678

[25] HAUNSCHILD M.D., FREISLEBEN B., TAKORS R., WIECHERT W.: 'Investigating the dynamic behavior of biochemical networks using model families', *Bioinformatics*, 2005, **21**, (8), pp. 1617–1625

[26] BLINOV M.L., FAEDER J.R., GOLDSTEIN B., HLAVACEK W.S.: 'BioNetGen: software for rule-based modeling of signal




doi: 10.1049/iet-syb:20070042

transduction based on the interactions of molecular domains', *Bioinformatics*, 2004, **20**, (17), pp. 3289–3291

[27] LOK L., BRENT R.: 'Automatic generation of cellular reaction networks with Moleculizer 1.0', *Nat. Biotechnol.*, 2005, **23**, (1), pp. 131–136

[28] KLIPP E., LIEBERMEISTER W., HELBIG A., KOWALD A., SCHABER J.: 'Systems biology standards – the community speaks', *Nat. Biotechnol.*, 2007, **25**, (4), pp. 390–391

[29] LE NOVERE N., FINNEY A., HUCKA M.: 'Minimum information requested in the annotation of biochemical models (MIRIAM)', *Nat. Biotechnol.*, 2005, **23**, (12), pp. 1509–1515

[30] AKAIKE H.: 'A new look at the statistical model identification', *IEEE Trans Autom. Control*, 1974, **AC-19**, pp. 716–723

[31] BURNHAM K.P., ANDERSON D.R.: 'Model selection and multimodel inference: a practical information-theoretic approach' (Springer, 2002), p. 496

[32] DE RIDDER F., *ET AL.*: 'Modified AIC and MDL model selection criteria for short data records', *IEEE Trans. Instrum. Meas.*, 2005, **54**, (1), pp. 144–150

[33] JOHNSON J.B., OMLAND K.S.: 'Model selection in ecology and evolution', *Trends Ecol. Evol.*, 2004, **19**, pp. 101–108

[34] KLIPP E., NORDLANDER B., KRÜGER R., GENNEMARK P., HOHMANN S.: 'Integrative model of the response of yeast to osmotic shock', *Nat. Biotechnol.*, 2005, **23**, (8), pp. 975–982

[35] KLIPP E., SCHABER J.: 'Modelling of signal transduction in yeast – sensitivity and model analysis', 'Understanding and exploiting systems biology in biomedicine and bioprocesses' (Fundación CajaMurcia, Murcia, Spain, 2006), pp. 15–30