Integration of Enzyme Kinetic Data from Various Sources

Simon Borger, Jannis Uhlendorf, Anselm Helbig and Wolfram Liebermeister*

Computational Systems Biology, Max Planck Institute for Molecular Genetics, Ihnestr. 63-73, D-14195 Berlin, Germany

Edited by I. Rojas and U. Wittig (guest editors); received and accepted 21 March 2007; published 27 March 2007

ABSTRACT: We describe a workflow to translate a given metabolic network into a kinetic model; the model summarises kinetic information collected from different data sources. All reactions are modelled by convenience kinetics; where detailed kinetic laws are known, they can also be incorporated. Confidence intervals and correlations of the resulting model parameters are obtained from Bayesian parameter estimation; they can be used to sample parameter sets for Monte-Carlo simulations. The integration method ensures that the resulting parameter distributions are thermodynamically feasible. Here we summarise different previous works on this topic: we give an overview over the convenience kinetics, thermodynamic criteria for parameter sets, Bayesian parameter estimation, the collection of kinetic data, and different machine learning techniques that can be used to obtain prior distributions for kinetic parameters. All methods have been assembled into a workflow that facilitates the integration of biochemical data and the modelling of metabolic networks from scratch.

KEYWORDS: Data integration, metabolic model, enzyme kinetics, convenience kinetics, kinetic parameter, Bayes statistics, posterior distribution, equilibrium constant, thermodynamics

INTRODUCTION

Kinetic modelling of biochemical networks requires a choice of the network structure, the kinetic rate laws, and the parameter values. Different reasons preclude inserting measured kinetic parameters directly into a model: on the one hand, parameters are often measured in vitro or under different conditions, so their values may be unreliable; on the other hand, thermodynamic laws lead to parameter dependencies in the model, which are easily violated if the parameter values are not correct [1].

Some biochemical quantities (in particular, protein levels) vary strongly within cell populations, which can be modelled by statistical distributions. The same probabilistic framework can also be used to handle uncertain or unknown parameters: we do not describe them by sharp values, but by a joint distribution [2]. Such distributions can be treated in a Bayes statistical model: each kinetic parameter has a prior distribution (describing, for instance, the parameter range that we expect for an unknown K_M value), and all measured parameter values are used as data. Eventually, we obtain a posterior distribution that describes a statistical ensemble of parameter sets; the parameter variances and correlations account for missing knowledge, measurement uncertainties or biological variability.

*Corresponding author. E-mail: {borger,uhlndorf,helbig,lieberme}@molgen.mpg.de.

Electronic publication can be found in *In Silico Biol.* **7**, S1, 09 <http://www.bioinfo.de/isb/2007/07/S1/09/>, 27 March 2007.

So, even if the data do not suffice for an exact parameter fit, we shall still obtain a model; the uncertainty of the parameters and correlations between them can be read directly from the posterior parameter distribution. The posterior summarises all information that has been put into the model; it can be used to sample model instances [3,4] or to provide parameter ranges or prior distributions for further modelling.

Workflow for integration of enzyme kinetic data

This paper summarises methods that we have developed for computer-assisted modelling and data integration [1,2,5–8]. Together, they form a workflow to translate a metabolic network into a dynamic model.

1. A structural model in SBML format is constructed from a list of reactions (currently, a list of KEGG reaction IDs [9]).
2. All reactions are described by the convenience rate law (see below); this choice defines a list of (thermodynamically independent) system parameters, which in turn determine all relevant kinetic parameters.
3. Data collection; values of various kinetic and thermodynamic parameters - and possibly entire kinetic laws – are retrieved from databases and collected from the literature.
4. For each system parameter, we fix a log-normal prior distribution. It represents our initial expectations based on empirical distributions of parameters, values in other species, or knowledge about the molecules.
5. Bayesian parameter estimation; to determine the model parameters, evidence from various experimental data is weighted and combined, resulting in a posterior parameter distribution.
6. Prediction of model behaviour; the posterior distribution can be used to sample model instances (Monte-Carlo sampling) and to obtain probabilistic statements about the model's dynamic behaviour (analytic results within a linear approximation are described in [2]).

We are currently implementing and combining the parts of this workflow. The final automated version will provide modellers with a simple way to query various databases in the context of a specific model; the posterior distribution can be used to define parameter ranges for further manual modelling, and it can be directly used as a parameter prior for model fitting. Finally, the variances of model parameters show where information is missing and can point at additional measurements that would be most valuable.

BIOCHEMICAL MODELS WITH CONVENIENCE KINETICS

Convenience kinetics

The convenience kinetics [1] is a generalised form of Michaelis-Menten kinetics, covers all possible stoichiometries, and describes enzyme regulation by activators and inhibitors. For a reaction with stoichiometry

$$\alpha_1 A_1 + \alpha_2 A_2 + \ldots \leftrightarrow \beta_1 B_1 + \beta_2 B_2 + \ldots,$$

it reads

$$v(a,b) = E_{\text{tot}} f_{\text{reg}} \frac{k_{+}^{\text{cat}} \prod_i \tilde{\alpha}_i^{\alpha_i} - k_{-}^{\text{cat}} \prod_j \tilde{b}_j^{\beta_j}}{\prod_i (1 + \tilde{\alpha}_i + \ldots + \tilde{\alpha}_i^{\alpha_i}) + \prod_j (1 + \tilde{b}_j + \ldots + \tilde{b}_j^{\beta_j}) - 1} \tag{1}$$

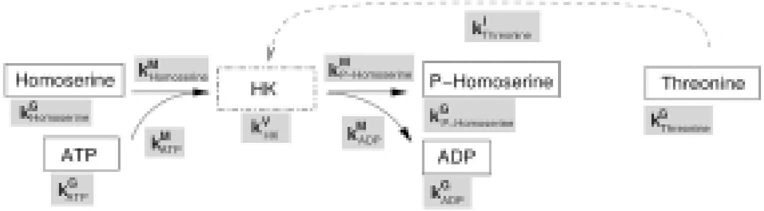

Fig. 1. Independent system parameters for convenience kinetics. If a reaction network (here: the homoserine kinase reaction) is displayed as a bipartite graph of metabolites and reactions, each of the nodes and each of the arrows is characterised by one of the system parameters. In addition, each node can carry an enzyme concentration E_l or a metabolite concentration c_i.

with enzyme concentration E_{tot} and turnover rates k_+^{cat} and k_-^{cat} (measured in s^{-1}). Variables with a tilde denote the normalised reactant concentrations $\tilde{\alpha}_i = a_i/k_{a_i}^M$ and $\tilde{b}_j = b_j/k_{b_j}^M$; the reactant constants $k_{a_i}^M$ and $k_{b_j}^M$ (in mM) correspond to the well-known K_M values in Michaelis-Menten kinetics. The regulatory prefactor f_{reg} is a product of terms $\frac{d}{k^A+d}$ or 1 $+d/k^A$ for activators and $\frac{k^I}{k^I+d}$ for inhibitors. Activation constants k^A and inhibition constants k^I are measured in mM, and d is the concentration of the modifier.

In analogy to Michaelis-Menten kinetics, k^M values denote substrate concentrations at which the reaction rate is half-maximal (or 1/(1 $+ \alpha_i$)-maximal for $\alpha_i \neq 1$) if the reaction products are absent; k^I and k^A values denote concentrations at which the inhibitor or activator has its half-maximal effect. In this respect, many parameters in convenience kinetics are comparable to the kinetic constants measured in enzyme assays. This is important for parameter estimation: for example, we shall claim below that measured K_M values can be used as data for the estimation of k^M.

Thermodynamic correctness

To facilitate parameter estimation and optimisation, we introduce system parameters that can be varied independently, without violating any thermodynamic constraints (see Fig. 1). For each reaction, we define the velocity constant $k^V = (k_+^{cat}\ k_-^{cat})^{1/2}$ (geometric mean of the turnover rates in both directions). Given the equilibrium and velocity constants, the turnover rates can be written as

$$k_\pm^{cat} = k^V (k^{eq})^{\pm 1/2} \tag{2}$$

Next, the equilibrium constants k^{eq} have to be expressed by independent parameters; here we can choose between (i) Gibbs free energies of formation or (ii) a set of independent equilibrium constants. For each substance i, we define the dimensionless energy constant $k_i^G = e_i^{G_i^{(0)}/(RT)}$ with Boltzmann's gas constant $R \approx$ 8.314 J/(mol K) and absolute temperature T. The equilibrium constants then satisfy ln $k^{eq} = -N^T$ ln k^G. Instead of the energy constants, we can also choose a subvector k^{ind} of independent equilibrium constants as independent parameters; they have to be thermodynamically independent and determine all other equilibrium constants in the model via a linear equation ln $k^{eq} = R_{ind}^{eq}$ ln k^{ind}. A choice of independent equilibrium constants and the corresponding matrix R_{ind}^{eq} can be computed from the stoichiometric matrix N (see [1]).

By taking the logarithm in both sides of Eq. (2), we obtain a linear equation. We can express various kinetic parameters by the system parameters: let θ denote the vector of logarithmic system parameters and x a vector containing various derived parameters in logarithmic form. It can be computed from θ by the linear relationship

$$x(\theta) = R^x_\theta \theta \tag{3}$$

The sensitivity matrix R^x_θ is sparse and can be constructed from the network structure (see [1]).

Enzyme mechanism

The convenience kinetics represents a simple molecular enzyme mechanism: (i) the substrates bind to the enzyme in arbitrary order and are converted into the products, which then dissociate from the enzyme in arbitrary order; (ii) binding of substrates and products is reversible and much faster than the conversion step; (iii) the binding energies of individual reactants do not depend on other reactants already bound to the enzyme (see [1]).

In the context of the enzyme mechanism, all system parameters can be expressed in terms of Gibbs free energies: the k^M, k^A, and k^I values represent binding energies, and the energy constants k^G are defined by the Gibbs free energy of formation. The velocity constants k^V represent an energy barrier in the catalysed reaction.

Parameter estimation for convenience kinetics

Many parameters in convenience kinetics – independent and dependent ones – can be measured in experiments. The data may be heterogenous, incomplete, and uncertain; we shall use them to determine "balanced" system parameters for a given metabolic network. In practice, we mine the literature for thermodynamic and kinetic data (see below) and merge their logarithms in a large vector x*. The vector can contain multiple values for a parameter, it can contain thermodynamically dependent parameters, and of course, many parameters from the model will be missing.

Our goal is to determine a vector θ of logarithmic system parameters such that x* $\approx R^x_\theta\ \theta$. We could solve this by least squares, but we prefer Bayes estimation because we can then include prior information in order to ensure that θ is well-determined.

Bayesian parameter estimation

In Bayesian parameter estimation, model parameters are described by a posterior probability distribution; it scores the potential parameter sets, showing how well each of them agrees with the data and with the prior assumptions made. Our prior distribution of θ is a multivariate Gaussian distribution $\mathcal{N}(\bar{\theta}_{(0)}, C_{(0)})$ with mean vector $\bar{\theta}_{(0)}$ and a diagonal covariance matrix $C_{(0)}$. For the likelihood function $p(y$*$|\theta)$, we assume that the experimental values in x* equal the values predicted by the model plus uncorrelated additive Gaussian noise, hence x* $= \mathcal{N}(x(\theta)$,$C_x)$. We assume a diagonal covariance matrix $C_x =$ diag$(\sigma_x)^2$, where the vector σ_x contains a noise level for each single measurement.

For the parameter estimation, we need experimental values and uncertainties. We collected thermodynamic, kinetic, and metabolic data from different sources; the thermodynamic data include standard Gibbs free energies of formation $G^{(0)}$ [10] and equilibrium constants k^{eq} [11]. Among the kinetic data are Michaelis Menten constants K_M, turnover rates K^{cat} and inhibition constants K^I [12]. The metabolic data contain metabolite concentrations [13] and protein concentrations [14]. In addition, we employ data from a comprehensive text-mining screen of PubMed abstracts (see http://sysbio.molgen.mpg.de/KMedDB).

The posterior distribution is multivariate Gaussian $\mathcal{N}(\bar{\theta}_{(1)}, C_{(1)})$ with mean and covariance matrix (see [15])

$$\begin{aligned} \theta_{(1)} &= \left(C_{(0)}^{-1} + (R_{\theta}^{x})^{T} C_{x}^{-1} R_{\theta}^{x}\right)^{-1} \left((R_{\theta}^{x})^{T} C_{x}^{-1} x * + C_{(0)}^{-1} \theta_{(0)}\right) \\ C_{(1)} &= \left(C_{(0)}^{-1} + (R_{\theta}^{x})^{T} C_{x}^{-1} R_{\theta}^{x}\right)^{-1} \end{aligned} \tag{4}$$

The parameter vector $\theta_{(1)}$ maximises the posterior; inserting it into Eq. (3) leads to consistent, balanced values of all kinetic parameters.

Mixed models including detailed kinetic laws

Detailed kinetic laws for many enzymes have been collected in Sabio-RK (http://sabio.villa-bosch.de/SABIORK/; we shall call them "known kinetics" here). A given network can be filled with known kinetics; reactions with unknown kinetic laws can be completed with convenience kinetics. To make sure that the resulting model is thermodynamically correct, the equilibrium constants of all reactions have to be matched: first we have to check whether the eqilibrium constants of all known kinetics are compatible: they have to satisfy log $k_{known}^{eq} = N_{known}^{T}\, \mu$ for a vector μ of chemical potentials. If the equilibrium constants are feasible, they are used as data (with zero variance) in the balancing procedure for the convenience kinetics parameters.

PARAMETER RANGES OBTAINED FROM STATISTICAL LEARNING

Choosing plausible parameter priors

Many parameters needed for a model will be missing or uncertain; therefore, we need to use realistic and accurate priors. The prior is especially important for parameters that have not been measured. To describe K_M values, for instance, we could simply choose the empirical distribution of all K_M values (or a log-normal distribution with the same mean and width) as a prior for each single K_M value. But we can also employ a more accurate, individual prior for each single K_M value; such a prior could represent additional knowledge that we can obtain without actually measuring the parameter. One possibility would be ab-initio calculations, but we can also try to obtain better estimates from machine learning. We describe here two approaches: (i) correlations between values for same metabolite, enzyme, organism; (ii) prediction from molecule structure (see [5]).

Parameter similarities across species

We explored a statistical approach that infers K_M values across species and enzymes [7]. Each K_M value is characterised by a triple (enzyme, organism, metabolite). For a fixed choice of the metabolite, we describe all corresponding logarithmic K_M values by a regression model in which the enzyme and the organism appear as qualitative factors with linear effects. We applied our method to Michaelis-Menten constants from the enzyme database Brenda [12] and assessed the quality of predictions with leave-one-out crossvalidation. The resulting predictions and error ranges for enzyme parameters can be used for defining individual priors for k^M values in convenience kinetics.

Quantitative structure-property relationships

We used linear regression to predict physiological concentrations of metabolites from their molecule structure [5]. The model relates logarithmic concentrations to feature vectors describing the chemical groups contained in a molecule. In order to focus on chemical groups that clearly affect the concentration, we used a regularisation term (lasso regression). In our study, the physical concentrations were increased by the occurrence of amino and hydroxyl groups, while aldehydes, ketones, and phosphates show decreased concentrations. Thus given the structure of a small molecule, the model can predict a confidence interval for physiological concentrations; again, this information can be used to define an individual prior for each metabolite concentration.

DISCUSSION

We presented a workflow to collect, predict, and integrate kinetic data for a given biochemical network. In particular, we

1. use convenience kinetics, at least if the actual kinetic law is not known;
2. eliminate thermodynamic dependencies by introducing independent parameters;
3. use measured kinetic parameters as data for estimation and integrate various pieces of evidence from different kinds of kinetic data;
4. employ prior distributions reflecting the distribution of known kinetic parameters or guesses based on machine learning.

Our main aim is to face the large number, but also the poor quality of kinetic data that are currently accessible: Bayesian estimation allows us to handle data with different levels of accuracy and to combine them with relatively loose prior assumptions. The log-normal parameters distributions of convenience kinetics are biologically plausible, fit well with the mathematical relationships between parameters, and lead to simple Gaussian distributions in the maximum-posterior calculations.

The result is (i) a kinetic model with convenience rate law, (ii) a collection of original data retrieved, and (iii) a complete set of balanced, thermodynamically consistent parameters with confidence intervals and correlations; other kinetic laws can be retrieved from Sabio-RK (http://sabio.villa-bosch.de/SABIORK/) and be embedded into the model. An additional second estimation step (see [8]) can integrate dynamic quantities such as measured metabolite concentrations and fluxes.

Kinetic modelling 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 (whether manual or automatic) cannot construct a reliable model from poor data. Nevertheless, we can find out parameter ranges that agree with the data available, and Bayes estimation is a natural framework for doing this. We expect that integrating various kinds of data will narrow down the posterior, that is, improve the model accuracy; even data points with weak evidence can contribute to the model - just with a lower weight, reflecting for instance a larger measurement error.

But how can we use measured kinetic parameters if the kinetic law is not known? We assume here that parameters like maximum turnover rates or K_M values have a similar meaning in different kinetic laws. Even if the enzyme mechanism does not follow exactly the convenience rate law, we claim that K_M values reported in databases can give us some clue about the k^M values in a convenience kinetics model; this claim still has to be tested by modelling.

We expect that larger amounts of kinetic information will be available in the near future, from which our workflow would probably profit; in particular, we would appreciate (i) agreed and precise annotations for biological entities, (ii) knowledge about allosteric regulation that is easily accessible in databases, and (iii) prediction of parameter values (e.g., binding Gibbs free energies) from *ab-initio* calculations.

ACKNOWLEDGEMENTS

This work has been funded by the European commission (Integrated project BaSysBio and Network of Excellence ENFIN) and by the Berlin Center for Genome Based Bioinformatics.

REFERENCES

[1] Liebermeister, W. and Klipp, E. (2006). Bringing metabolic networks to life: convenience rate law and thermodynamic constraints. Theor. Biol. Med. Mod. **3**, 41.

[2] Liebermeister, W. and Klipp, E. (2005). Biochemical networks with uncertain parameters. IEE Proc Systems Biology **152**, 97-107.

[3] Small, J. R. and Fell, D. (1990). Metabolic control analysis. Sensitivity of control coefficients to elasticities. Eur. J. Biochem. **191**, 413-420.

[4] Klipp, E., Liebermeister, W. and Wierling, C. (2004). Inferring dynamic properties of biochemical reaction networks from structural knowledge. Genome Informatics **15**, 125-137.

[5] Liebermeister, W. (2005). Predicting physiological concentrations of metabolites from their molecular structure. J. Comp. Biol. **12**, 1307-1315.

[6] Schulz, M., Uhlendorf, J., Klipp, E. and Liebermeister, W. (2006). SBMLmerge, a system for combining biochemical network models. Genome Informatics Series **17**, 62-71.

[7] Borger, S., Liebermeister, W. and Klipp, E. (2006). Prediction of enzyme kinetic parameters based on statistical learning. Genome Informatics Series **17**, 80-87.

[8] Liebermeister, W. and Klipp, E. (2006). Bringing metabolic networks to life: integration of kinetic, metabolic, and proteomic data. Theor. Biol. Med. Mod. **3**, 42.

[9] Kanehisa, M., Goto, S., Kawashima, S. and Nakaya, A. (2002). The KEGG databases at genomenet. Nucleic Acids Res. **30**, 42-46.

[10] Mavrovouniotis, M. (1990). Group contributions for estimating standard gibbs energies of formation of biochemical compounds in aqueous solution. Biotechnol. Bioeng. **36**, 1070-1082.

[11] Goldberg, R. N. (1999). Thermodynamics of enzyme-catalyzed reactions: Part 6 - 1999 update. J. Phys. Chem. Ref. Data **28**, 931.

[12] Schomburg, I., Chang,A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G. and Schomburg, D. (2004). BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res. **32**, D431-433.

[13] Albe, K. R., Butler, M. H. and Wright, B. E. (1990). Cellular concentrations of enzymes and their substrates. J. Theor. Biol. **143**, 163-195.

[14] Huh, W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S. and O'Shea, E. K. (2003). Global analysis of protein localization in budding yeast. Nature **425**, 686-691.

[15] Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1997). Bayesian Data Analysis. Chapman & Hall, New York.