AUTOMATICALLY GENERATED MODEL OF A METABOLIC NETWORK

SIMON BORGER
borger@molgen.mpg.de

WOLFRAM LIEBERMEISTER
lieberme@molgen.mpg.de

JANNIS UHLENDORF
uhlndorf@molgen.mpg.de

EDDA KLIPP
klipp@molgen.mpg.de

Max Planck Institute for Molecular Genetics, Berlin, Germany

We demonstrate an approach to automatically generating kinetic models of metabolic networks. In a first step, the metabolic network is characterised by its stoichiometric structure. Then to each reaction a kinetic equation is associated describing the metabolic flux. For the kinetics we use a formula that is universally applicable to reactions with arbitrary numbers of substrates and products. Last, the kinetics of the reactions are assigned parameters. The resulting model in SBML format can be fed into standard simulation tools. The approach is applied to the sulphur-glutathione-pathway in Saccharomyces cerevisiae.

Keywords: Metabolic networks; systems biology; parameter estimation; data integration; sulphur-glutathione pathway; Bayesian data analysis.

1. Introduction

Studying the regulation of metabolic reaction networks is an important task in systems biology and functional genomics. A complete understanding of metabolic regulation requires quantitative information about kinetic laws and the concentrations of metabolites and enzymes. This quantitative knowledge in combination with the known network of metabolic pathways allows the construction of mathematical models that describe the dynamic changes in metabolite concentrations over time. The models are high-dimensional systems of ordinary, non-linear differential equations. The main problems of the approach are the setup of the equations that describe the metabolic pathways in form of kinetic rate equations and the identification of the system parameters. To solve these problems, a variety of pathway modeling tools such as Copasi [6], CellDesigner [15], and others have been developed which simplify model construction and analysis. Most of these tools are able to store and exchange models in the Systems Biology Markup Language (SBML, [21]) and to fit parameters for a given set of experimental data. A long-term goal is the construction of genome-scale metabolic models. For various model organisms stoichiometric genome-scale models have been constructed. They have been shown to be useful for the investigation of steady state fluxes in wild type cells and in mutants. Large-scale dynamic models would be very useful to predict the effect of transient perturbations, for instance by gene regulation, or to apply powerful analysis tools such as metabolic control analysis (MCA), but is still hampered by a lack of systematically retrieved data. Facing the need of a more systematic construction and parameterisation of metabolic models, we present an approach to (i) automatically construct an SBML model from a list of reactions, (ii) automatically associate kinetic expressions to all reactions and (iii) automatically assign the parameter values based on available information and on statistically based estimates for missing information.

2. Methods

We have set up a workflow to automatically generate models of metabolic networks from a given set of reactions; it consists of the following steps:

(i) set up a structural model
(ii) assign a kinetic law to each reaction
(iii) collect kinetic and thermodynamic data
(iv) determine a feasible set of parameters
(v) construct the kinetic model in SBML format

The parameter estimation is based on Bayes statistics; the result is a posterior distribution of parameter sets that reflects the most probable values and their remaining uncertainties. This result can serve as a prior in further modelling steps. We shall now describe the single steps of the workflow; a more detailed description of it is given in [4].

Structural model

We base our models on knowledge contained in the Kyoto Encyclopedia of Genes and Genomes (KEGG) [16]. KEGG is a set of databases that constitute a computer representation of biological knowledge at different levels, i.e. pathways, reactions, enzymes, compounds and genes. These levels are interconnected.
Our metabolic networks are built from chemical reactions stored in KEGG. We start with a set of reactions; in practice we map the reactions to their identifiers in KEGG. With these reaction identifiers we retrieve information on the relevant compounds (in the form of KEGG compound indentifiers) and enzymes (EC numbers). Each reaction in the model has a reaction identifier, an enzyme identifier and a metabolite identifier associated with it. In the resulting SBML file, each element is described by a MIRIAM-compliant annotation [8, 18], which points to the respective KEGG identifer.

Kinetic laws

In a next step each reaction is assigned a kinetic expression. We use the convenience kinetics, a rate law that assumes a random-order enzyme mechanism and is applicable to reactions with any number of substrates and products [9]:

$$v = E \cdot \left(\prod_a \frac{\bar{c}_a}{\bar{c}_a + k_a^A} \right) \cdot \left(\prod_i \frac{k_i^I}{\bar{c}_i + k_i^I} \right) \cdot \frac{k_+^{cat} \prod_s \bar{c}_s^{\,n_s} - k_-^{cat} \prod_p \bar{c}_p^{\,n_p}}{\prod_s \left(\sum_{m=0}^{n_s} \bar{c}_s^{\,m} \right) + \prod_p \left(\sum_{m=0}^{n_p} \bar{c}_p^{\,m} \right) - 1} \; . \quad (1)$$

The index variables a, i, s and p run over the sets of activators, inhibitors, substrates and products of the reaction, respectively. The concentration of the enzyme catalysing the reaction is denoted by E. The variables k_+^{cat} and k_-^{cat} stand for the turnover rates of the enzyme in the forward (+) and the backward (−) direction, $\bar{c}_s = c_s/k_s^M$ and $\bar{c}_p = c_p/k_p^M$ are the ratios of the substrate and product concentrations with their k^M values, K_i^I denotes the inhibition constant of inhibitor i and K_a^A the activation constant of activator a. Finally, n_s and n_p are the stoichiometric coefficients of substrate s and product p, respectively.

This formula is directly applicable once the stoichiometric structure of the model - i.e., the substrates and products of all reactions - and the regulatory structure - i.e. activators and inhibitors of an enzyme - are known.

The parameters that enter this kinetic expression are a k^M value for each reactant, a k^A value for each activator of the enzyme, a k^I value for each inhibitor and two turnover rates $k_\pm^{cat}$ for the enzyme. The total number of kinetic parameters entering the formula is $N_s + N_p + N_i + N_a + 2$ as indicated in Table 1. In order to ensure thermodynamic consistency of the parameter set, we actually regard the turnover rates $k_\pm^{cat}$ as dependent quantities and express them by two different kinds of parameters, one k^V value for each reaction and one k^G value for each metabolite [9].

Table 1. Types of kinetic parameters and their numbers entering the kinetic formula Eq.(1).

Parameter type	number required
k^M	$N_s + N_p$
k^I	N_i
k^A	N_a
k^{cat}	2

Note: N_s: number of substrates, N_p: number of products, N_i: number of inhibitors, N_a: number of activators of a reaction.

Data collection

We use two types of data available. First, we search literature and databases for thermodynamic, kinetic, metabolomic and proteomic data. The thermodynamic data include Gibbs free energies of formation [2, 5, 11], and equilibrium constants [19]. The kinetic data comprise k^M values [14, 17], k^I values [14, 17], $ic50$ values [17] and turnover rates k^{cat} [14, 17]. Metabolomic data sources are metabolite concentrations [1]. Protein concentrations come from [23].

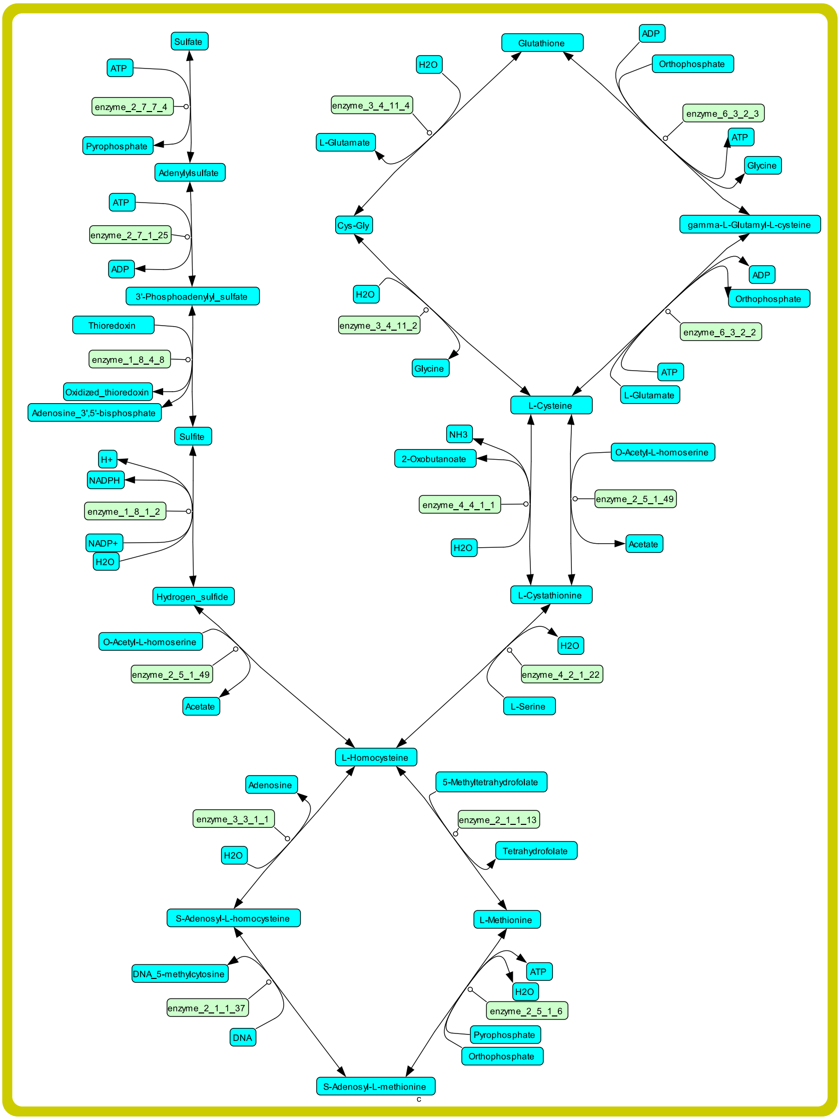

Fig. 1. An automatically generated metabolic network set up starting with the KEGG reaction identifiers R00529, R00509, R02021, R00858, R01287, R00192, R00380, R00177, R00946, R01290, R01001, R03217, R00894, R00497, R00494, R00899.

Furthermore, we use predicted k^M values based on a statistical linear model [3]. The idea behind is that there are different statistical factors that explain the logarithm of a k^M value, $\ln(k^M)$. The first of the three factors is the substrate contribution μ determined by the substrate's chemical properties. Secondly, there is the substrate-enzyme contribution α reflecting evolutionary conservation across organisms. Finally, there is the substrate-organism contribution β stemming from the adjustment of k^M values to typical concentrations of the respective metabolite in a certain organism. The sum will be the value of the logarithm of the k^M value: $\ln(k^M) = \mu + \alpha + \beta$.

We map all the collected experimental data to one of the entities `reaction`, `enzyme` and `compound` from KEGG, or to a combination of them, according to the type of data. The entities are represented by their respective KEGG identifiers. With the KEGG identifiers the data are written into a database. By searching the database for the KEGG identifiers present in the model, data can be retrieved for the kinetic expression Eq.(1) of each reaction.

The data for the model are searched in the following way: first we look for data that are identified by a reaction identifier like equilibrium constants k^{eq} or Michaelis Menten constants k^M. If for a certain reaction no such data are found by the reaction identifier, we search by the enzyme identifier that is associated with the reaction. An equilibrium constant is already completely determined by a reaction identifier, for a k^M value the metabolite identifier also has to be extracted. In the next step, data that require either only a metabolite identfier, like concentrations and Gibbs free energies of formation, are searched by the the respective identifier.

Table 2. Prior distributions.

quantity	no. of data	5% quantile	median	95% quantile
thermodynamic				
equilibrium constant	2088	0.000001	0.119	162.0
Gibbs energy of formation (kJ/mol)	9804	-1522.6	-331.0	324.3
kinetic				
k^M (mM)	90240	0.00098	0.14	20
k^I (mM)	21092	0.000003	0.016	14.0
$ic50$ (mM)	8324	0.000002	0.002	0.67
k^{cat} (mM)	22587	0.008	6.0	1100
metabolomic				
metabolite concentration (mM)	225	0.0018	0.122	4.9
proteomic				
protein abundance (molecules/cell)	10141	279	2939	33502

Note: The parameter types used for data collection for the model. Shown are the number of available data and properties of their distributions.

Distribution of feasible parameter sets

The collected experimental data cannot be directly written into the kinetics of the reactions of the metabolic network. The reasons are: (i) experimental data are noisy because of measurement errors. Sources of noise are biological variability, measurement errors, measurement in vitro or in other species; (ii) we may find different contradicting values for the same parameter; (iii) due to thermodynamical constraints [9] (iv) many parameters will not be available and therefore remain undetermined.

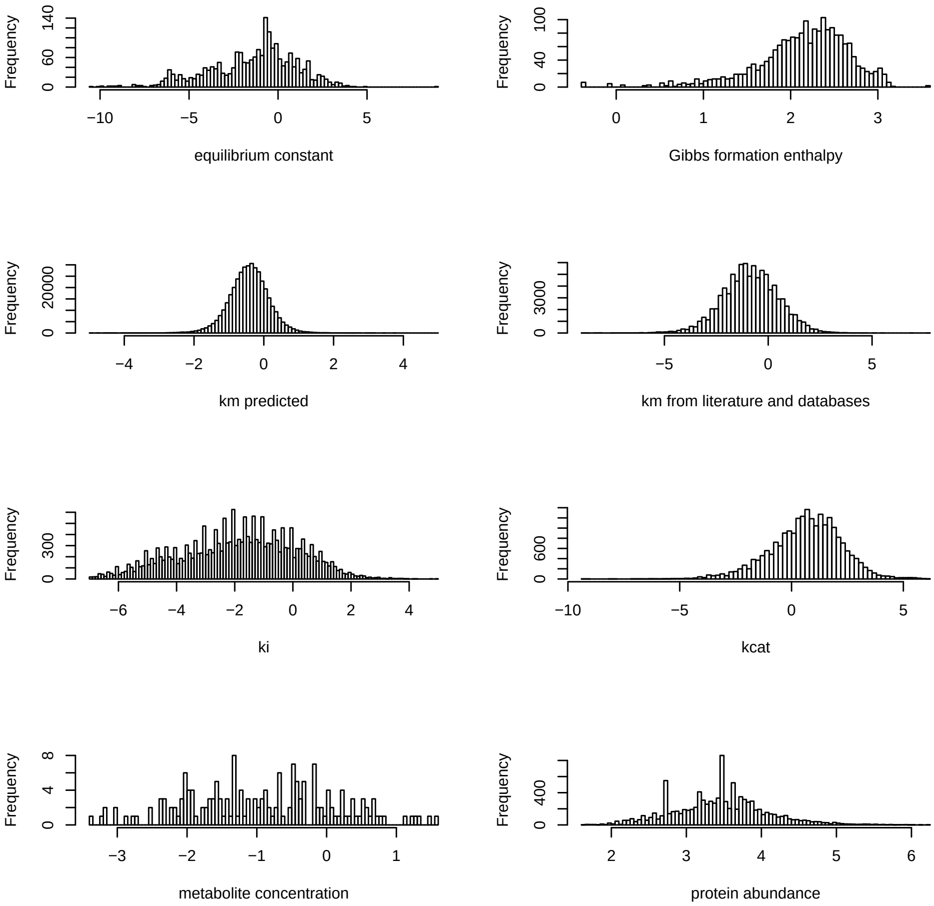

Fig. 2. Prior distributions of logarithms of different data types used in the Bayesian parameter estimation approach.

We thus regard the information gained for the parameters as uncertain data and take a Bayesian approach to find a complete set of parameters that is thermodynamically consistent [10]. From a statistics over all collected data of certain parameter types, we derive prior distributions for these parameter types as indicated in Table 2. The distributions of the logarithmic parameters are also shown in Fig. 2. For instance, a log-normal distribution fitted to k^M values in the database Brenda [14] is used as a prior for each k^M value in a model. The parameter values we retrieve for our specific model are used as data that have to be explained by the parameter set of the model; this determines a likelihood function. The prior and the likelihood function combined yield a posterior distribution of the model parameters given the found data points for the model [10].

By random sampling from the posterior distribution we can find distributions of the behaviour of the model.

Kinetic model

Once the entities of the metabolic network, i.e. reactions, metabolites and enzymes, are assigned their parameters, the result is written to an output file in SBML format. The annotations in the SMBL file accord with the MIRIAM standard [8, 18].

Table 3. Statistics of data retrieved for model.

quantity	no. of data	thereof for Sacc. Cerev.	no. of data retrieved for model
thermodynamic			
equilibrium constant	2088	2088	for 2 of 16 reactions
Gibbs energy of formation	9696	9696	for 22 of 36 metabolites
kinetic			
k^M	90240	2475	for 34 of 36 substrate-reaction-pairs
k^I	21092	158	2
$k^{cat}_{\pm}$	22587	144	3 of 74 metabolite-enzyme-pairs
metabolomic			
metabolite concentration	225	30	for 7 of 36 metabolites
proteomic			
protein abundance	10141	10141	for 13 of 15 enzymes

Note: The number of data of different parameter types in the database, how many of them apply to Saccharomyces cerevisiae and how many have been extracted for the model (not only kinetic parameters).

3. Results

As a test case, we applied the described automatic model generation to the sulphur assimilation and the glutathione biosynthesis pathways in the yeast Saccharomyces cerevisiae. These pathways play an important role in the buffering of arsenic in order to avoid toxic effects: the cell increases the uptake of sulphur, leading to a raised glutathione level. Glutathione, having a high reduction potential, forms a complex with arsenic and the complex then is disposed in the vacuoles. The expression of the enzymes involved in these pathways is enhanced upon exposure to arsenic [12]. From a manually sketched metabolic network of the sulphur assimilation and the glutathione biosynthesis pathways, an enhanced version of the model in [12], we looked up the KEGG reaction identifiers. With these identifiers, information about the reactants and enzymes is fetched from the KEGG database. The result is the metabolic network shown in Fig. 1.

In the data retrieval step we could find 131 entries; some of them refer to the same parameter of the model and are averaged. After averaging and balancing we are left with 49 parameters.

The prior distributions of the logarithms of the parameters are shown in Fig. 2. In some cases they can be well described by a normal distributions (i.e. the parameter itself is log-normal distributed). Especially where the number of collected data (see. Table 2) are large the normal distribution is a good description of the respective actual distribution. This is especially the case for the k^M, k^{cat}, k^i values and the Gibbs formation enthalpies.

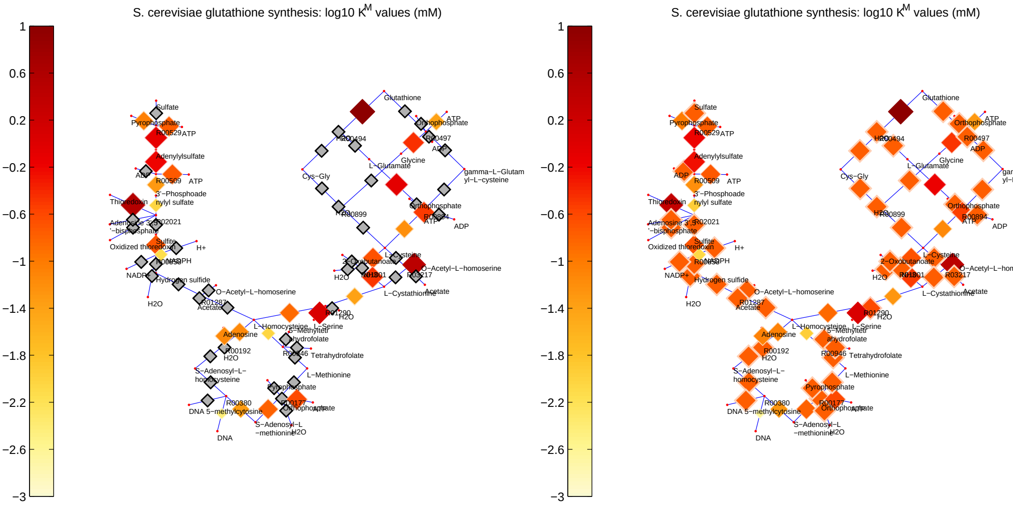

Fig. 3. Michaelis-Menten constants in the sulphur-glutathione model. Left: k^M values retrieved from the database Brenda [14]. Some of the values are missing (grey diamonds with black border). Right: balanced, complete set of k^M values for the model.

After averaging we are left with 49 experimental values useful for assigning values to the kinetic parameters of the model (s. Table 3). The kinetic parameters of the model are determined by 127 independent kinetic and thermodynamic values. Those values that cannot be extracted from the database are mainly determined by the mean values of their prior distributions (seeTable 2) and then undergo the thermodynamical adjustment and Bayesian procedure. In Fig. 3 we show as an example the k^M values of the model. To the left we display the number of extracted k^M values from the database (missing data are indicated by black borders of the diamonds) and their numerical values. To the right we show the model parameters after the adjustment to thermodynamical constraints and the replacement of missing values in the course of the Bayesian procedure. High numerical values tend to stay high, missing ones are replaced by "average" numerical values.

When simulated with initial concentration values of the metabolites in the range of 0.1 to 10 mM, and holding the concentrations of the cofactors constant, the model yields concentrations in the range of 1μM to 1mM. The fluxes obtain values in the range of 1nM/s to 1μM/s.

4. Discussion

We have presented a workflow to automatically generate metabolic models. We begin with a set of reactions, set up the structure of the model, search data that determine the kinetic parameters of the model as well as concentrations of enzymes and metabolites. These data cannot simply be written into the kinetic expression of Eq.(1), but have to be averaged and adjusted to thermodynamic constraints [9]. The data are taken as hints that determine distributions of the model parameters: In a Bayesian approach we draw values from a posterior distribution of the model parameters given the values extracted from the database [10]. In further analyses we can thus assess the distribution of models and their dynamic behaviours.

We have applied this approach to a medium-scale model, the sulphur assimilation and glutathione biosynthesis pathways. We fed an enhanced version of the model in [12] into the workflow. The outcome is a parametrised model in the SBML format with annotations according to the MIRIAM standard. The model can be fed into simulation tools for further analysis of its dynamical properties, for instance for comparison with metabolite time courses.

The parameter balancing, i.e. the adjusting to thermodynamical constraints and combining prior and likelihood, results in a joint posterior distribution of all parameters; in particular, it ensures that we obtain a mean value for each parameter, even if no data about this parameter are available. The posterior distribution of an "unknown" parameter will reflect two sources of knowledge: (i) its prior (e.g., the posterior of an unknown k^I value is just its prior, that is, the distribution of all known k^I values); (ii) its relationships with other, possibly known, parameters (e.g., as $v^{max} = Ek^{cat}$, experimental data of E and v^{max} will affect the posterior of k^{cat}).

The posterior parameter distribution can be used as a prior for subsequent modelling in which new data, e.g., metabolic timecourses, are incorporated (see [10]). Hence, our approach takes the form of an iterative learning tool.

The fact that we could not find data for all the parameters of the model is partly due to the simple lack of available data. Another reason is incomplete mapping of reaction names and metabolite names to the appropriate KEGG entities. Furthermore, also in KEGG there are cases that two entities appear to the knowledgeable user as the same, but not to computer tools. Hence, a standard for names of biological entities is desirable. It would greatly improve approaches like the presented automated modelling approach and help in paving the way towards genome scale models.

References

[1] Albe, K.R., Butler, M.H., and Wrigth, B.E., Cellular concentrations of enzymes and their substrates, *J. Theor. Biol.* 143(2):163–195, 1990.

[2] Alberty, R.A., Equilibrium compositions of solutions of biochemical species and heats of biochemical reactions, *Proc. Natl. Acad. Sci. USA*, 88(8):3268–3271, 1991.

[3] Borger, S., Liebermeister, W., and Klipp, E., Prediction of Enzyme Kinetic Parameters Based on Statistical Learning, *Genome Inform*, 17(1): 80–87, 2006.

[4] Borger, S., Uhlendorf, J., Helbig, A., and Liebermeister, W., Integration of enzyme kinetic data from various sources, *In Silico Biol.*, 7(S1):09, 2007.

[5] Hartmann, K. and Schomburg, D., GibbsPredictor: Predicting Gibbs energies from molecular structures, *Bioinformatics*, submitted.

[6] Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P., and Kummer, U., COPASI - a COmplex PAthway SImulator, *Bioinformatics*, 22(24):3067–3074, 2006.

[7] Klipp, E., Herwig, R., Kowald, A., Wierling, C., and Lehrach, H., *Systems Biology in Practice. Concepts, Implementation and Application*, Wiley-VCH Verlag GmbH and Co. KGaA, Weinheim, 2005.

[8] Le Novere, N., Finney, A., Hucka, M., Bhalla, US., Campagne, F., Collado-Vides, J., Crampin, E.J., Halstead., M., Klipp, E., Mendes, P., Nielsen, P., Sauro, H., Shapiro, B., Snoep, J.L., Spence, H.D., and Wanner, B.L., Minimum information requested in the annotation of biochemical models (MIRIAM), *Nat. Biotechnol.*, 23(12):1509–1515, 2005.

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

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

[11] Mavrovouniotis, M.L., Estimation of standard Gibbs energy changes of biotransformations, *J. Biol. Chem.*, 266:14440–14445, 1991.

[12] Thorsen M., Lagniel, G., Kristiansson, E., Junot, C., Nerman, O., Labarre, J., and Tamas, M.J., Quantitative transcriptome, proteome and sulfur metabolite profiling of the Saccharomyces cerevisiae response to arsenite, *Physiol Genomics*, 30(1):35–43, 2007.

[13] Schomburg, I., Chang, A., and Schomburg, D., BRENDA, enzyme data and metabolic information, *Nucleic Acids Res.* 30(1):47–49, 2002.

[14] `http://www.brenda.uni-koeln.de/`

[15] `http://www.celldesigner.org/`

[16] `http://www.genome.ad.jp/kegg/`

[17] `http://sysbio.molgen.mpg.de/KMedDB`

[18] `http://www.ebi.ac.uk/compneur-srv/miriam/`

[19] `http://xpdb.nist.gov/enzyme_thermodynamics/`

[20] `http://www.r-project.org/`

[21] `http://sbml.org/`

[22] `http://sysbio.molgen.mpg.de/KMedDB/`

[23] `http://yeastgfp.ucsf.edu/`