# Prediction of Enzyme Kinetic Parameters Based on Statistical Learning

**Simon Borger** borger@molgen.mpg.de
**Wolfram Liebermeister** lieberme@molgen.mpg.de
**Edda Klipp** klipp@molgen.mpg.de
Max Planck Institute for Molecular Genetics, Ihnestraße 63–73, 14195 Berlin, Germany

**Abstract**

Values of enzyme kinetic parameters are a key requisite for the kinetic modelling of biochemical systems. For most kinetic parameters, however, not even an order of magnitude is known, so the estimation of model parameters from experimental data remains a major task in systems biology. We propose a statistical approach to infer values for kinetic parameters across species and enzymes making use of parameter values that have been measured under various conditions and that are nowadays stored in databases. We fit the data by a statistical regression model in which the substrate, the combination enzyme-substrate and the combination organism-substrate have a linear effect on the logarithmic parameter value. As a result, we obtain predictions and error ranges for unknown enzyme parameters. We apply our method to decadic logarithmic Michaelis-Menten constants from the BRENDA database and confirm the results with leave-one-out crossvalidation, in which we mask one value at a time and predict it from the remaining data. For a set of 8 metabolites we obtain a standard prediction error of 1.01 for the deviation of the predicted values from the true values, while the standard deviation of the experimental values is 1.16. The method is applicable to other types of kinetic parameters for which many experimental data are available.

**Keywords:** enzyme parameter, Michaelis-Menten constant, regression model, statistical learning

## 1 Introduction

Systems biology aims at understanding the behaviour of entire cells by combining experiments, data analysis, and computational modelling. Typical models of biochemical networks in the form of ordinary differential equations (ODEs) comprise a few up to tens of reactions, which is much less than the total number of chemical reactions in the cell. To grasp the complexity of the behaviour of living cells under different natural or experimental conditions, all major constituents of the system should be included. Such models would contain hundreds or more metabolites or proteins.

Metabolic networks have been studied for a long time and means for their analysis have been developed (e.g. Metabolic Control Theory [3, 5]). Deterministic models of metabolic networks consist of ODE systems describing the kinetics of the metabolic processes, i.e. the production and consumption of metabolites and the role of enzymes in these processes. The mathematical expressions of the kinetic laws each require parameters. In the case of a reversible enzyme catalysed reaction with one substrate and one product

$$A \longleftrightarrow B$$

and assuming Michaelis-Menten-kinetics, the net rate in the direction $A \longrightarrow B$ reads

$$v = \frac{v_{+}^{\max}(A/K_{\mathrm{A}}) - v_{-}^{\max}(B/K_{\mathrm{B}})}{1 + A/K_{\mathrm{A}} + B/K_{\mathrm{B}}}.$$

This expression contains four kinetic parameters, namely the Michaelis-Menten-constants $K_{\mathrm{A}}$ and $K_{\mathrm{B}}$ for substrate $A$ and $B$, respectively, and the maximal velocities $v_{+}^{\max}$ and $v_{-}^{\max}$. The subscripts "+"

and "−" denote the two directions of the reaction. Reactions with more substrates or products employ even more parameters, such as Michaelis-Menten constants for all reactants or dissociation constants for the modifiers. Thus, even moderately sized models with tens of components can already contain a large number of parameters.

When it comes to simulating a biological system, it is important to have the values for all parameters at hand. Often, the only data available are of poor quality because measurements have not been repeated sufficiently often or have been undertaken in conditions different from the ones considered in the model. To cope with these limitations, we have to estimate missing parameters on the basis of the best information available. A standard approach in parameter estimation is to move through the solution space and to optimise an objective function [2, 7, 8]. A prominent example are least-square fits to experimental metabolic time courses. For setting inequality constraints in parameter optimisation and for regularising ill-defined estimation problems by prior distributions, it is in a first step very helpful to know rough parameter estimates that match at least the order of magnitude.

Here we describe a way to estimate unknown enzyme parameters from the known parameter values of related enzymes. We demonstrate the approach with Michaelis-Menten constants $K_M$ and appraise the prediction errors. While Michaelis-Menten constants $K_M$ can generally vary within a broad range, we may expect that the values for homologous enzymes share some similarity. Moreover, we may also expect that a certain substrate show increased $K_\mathrm{M}$ values, at least in a certain organism.

If such similarities hold true, we can roughly predict an unknown $K_\mathrm{M}$ value from known $K_\mathrm{M}$ values in other species and from $K_\mathrm{M}$ values of other enzymes (acting the same substrate). Technically, we quantify our expectation in form of a statistical regression model. In the model, a Michaelis-Menten constant is influenced by three factors: the substrate, the enzyme, and the organism in which the enzyme is found.

## 2 Method

### 2.1 Data Retrieval

To predict $K_\mathrm{M}$ values, we retrieve a set of measured $K_\mathrm{M}$ values from the BRENDA database [11, 14]. For a chosen metabolite, we search the data for all $K_\mathrm{M}$ values related to this metabolite and store them together with their associated enzymes (denoted by EC numbers) and the organisms in which the values were measured. The result is a set of data triples ($K_M$ value, EC number, organism).

Next, we arrange these data (still for a single metabolite) in a matrix $X$ with rows corresponding to the EC numbers and columns corresponding to the organisms. A matrix element $x_{ij}$ contains the logarithmic $K_\mathrm{M}$ value for the respective combination (EC number, organism). We sometimes find several $K_\mathrm{M}$ values for the same pair (EC number, organism): in this case, we compute the mean of the logarithmic values and take this as the matrix element. Because the values are logarithmic, this accounts to the geometric mean of the real $K_M$ values. However, many of the elements will remain empty either because the respective $K_\mathrm{M}$ value has not been measured yet or because the enzyme simply does not exist in the organism. Our aim is to fill the missing values for the biologically meaningful combinations (EC number, organism) with a prediction based on the known elements.

### 2.2 Linear Regression Model

To this end, we fit the logarithmic $K_\mathrm{M}$ values contained in the data matrix $(x_{ij})$ by a linear statistical model. The EC numbers and the organism names are regarded as qualitative data (factors) defining classes with associated effects $\alpha_i$ and $\beta_j$. If we denote the logarithmic $K_\mathrm{M}$ value of enzyme $i$ in organism $j$ by $x_{ij}$, the model reads

$$x_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \tag{1}$$

where $\mu$ is the general mean, $\alpha_i$ denotes the effect of enzyme $i$, $\beta_j$ is the effect of the organism $j$, and the $\epsilon_{ij}$ are independent identically distributed Gaussian random numbers: $\epsilon ij = \mathcal{N}(0, \sigma^2)$. We shall take Equation 1 as a purely statistical model and study whether it is supported by the data. Given experimental data for some of the $x_{ij}$, we compute the effects $\mu$, $\alpha_i$, and $\beta_j$ by a least squares fit, minimising the sum of quadratic residuals. For the actual calculations, we use the function `lm()` from the R language for statistical computing [15].

It is known [9] that only certain linear combinations of $\mu$, $\alpha_i$, and $\beta_j$ can be estimated. Another possibility in order to identify meaningful values $\mu$, $\alpha_i$, and $\beta_j$ is to set restrictions such as $\langle \alpha_i \rangle = \langle \beta_j \rangle = 0$ on the effects. This last method we employ.

Predictions $x^*_{ij}$ for unknown $K_\mathrm{M}$ values are, in the best case, obtained from the fitted model by computing

$$x^*_{ij} = \mu + \alpha_i + \beta_j. \tag{2}$$

This requires that $\alpha_i$ and $\beta_j$ be well defined.

There are four cases. In the most undetermined case, we want to predict a $K_\mathrm{M}$ value where neither $K_\mathrm{M}$ values are known for the same enzyme in other organisms nor for other enzymes in the same organism. This corresponds to the case where in the data matrix $(x_{ij})$ the row $i$ and the column $j$ are empty. The linear model does not yield any result neither for $\alpha_i$ nor for $\beta_j$, so the prediction will only be the overall mean

$$x^*_{ij} = \mu. \tag{3}$$

In a second case, there are known $K_\mathrm{M}$ values for other enzymes in the same organism, but no known $K_\mathrm{M}$ values the same enzyme in other organisms. In the data matrix $(x_{ij})$ row $i$ is empty but the column $j$ is not.There is only a result of the linear model for $\beta_j$ and the prediction is

$$x^*_{ij} = \mu + \beta_j. \tag{4}$$

In a third case, $K_\mathrm{M}$ values for the same enzyme have been measured in other organisms, but no $K_\mathrm{M}$ value for different enzymes is known for the same organism. In the data matrix $(x_{ij})$ the row $i$ is not, but the column $j$ is empty. The linear model yields a result for the effect $\alpha_i$ only and the prediction is

$$x^*_{ij} = \mu + \alpha_i. \tag{5}$$

Finally, there are known $K_\mathrm{M}$ values for different enzymes in the same organism and for the same enzyme in other organisms. The prediction will be according to Equation 2

$$x^*_{ij} = \mu + \alpha_i + \beta_j.$$

Figure 1 illustrates this method for some logarithmic $K_\mathrm{M}$ values of enzymes measured in 9 different organisms for some substrate. In this example all empty fields can be calculated according to Equation 2. Thus for every matrix entry, there is information from its column and its row contributing to its prediction.

### 2.3 Cross-Validation

The linear model yields predictions of the missing $K_\mathrm{M}$ values. In order to check the quality of these predictions, we compared them to known $K_\mathrm{M}$ values. This test is known as *leave-one-out cross-validation*. That is, the fit to the linear model 1 is done with the respective logarithmic $K_\mathrm{M}$ value $x$ omitted from the data set. We compute a prediction $x^*$ of the specific $x$ value by the linear fit of the reduced data set and compare to its original value. To assess the quality of the prediction, we plot $x$ and $x^*$ against each other and compute the root mean square error $\sigma_{pred} = \sqrt{\langle (x - x^*)^2 \rangle}$.

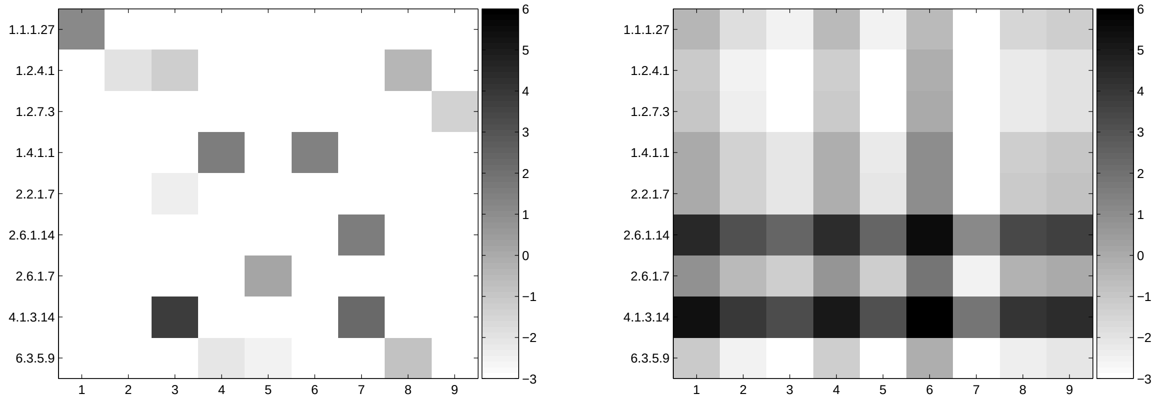

Figure 1: A regression model for hypothetical $K_{\mathrm{M}}$ values. Here we compare values $x$ and predicted values $x^*$. Enzymes (denoted by EC numbers) are listed in the rows, while the numbers of the columns columns refer to different organisms. The left matrix corresponds to the true values, the right one to the predicted ones. The greyscales indicate numerical values as can be seen in the grayscale bars to the right of the figures.The original data are sparse(see Table 2). That is why so many fields in the left matrix are empty (white). The right matrix shows values $x^*$ predicted from the regression model. As a result of the fitting procedure, rows and columns that correspond to strong values in the true data become also strong, and vice versa for weak values.

## 3 Results

The number of $K_M$ values in the BRENDA database varies greatly for the different metabolites. For NAD+ for instance, we found 1216 hits in the BRENDA database. These were reduced to 581 cases after the classification according to enzymes and organisms, i.e. several values for the same enzyme and the same organism were averaged over arithmetically . Acetyl-CoA yielded 487 search results in BRENDA and we were left with 269 data points.

To test the quality of our data prediction, we performed a cross-validation. To this end, we dropped, one after the other, each experimental value from our dataset and predicted it from the remaining data. The results are shown in scatter plots in Figure 2. In three of the four cases the root mean square error of the deviation of the predicted from the original value $\sigma_{pred} = \sqrt{\langle (x^* - x)^2 \rangle}$ is smaller than root mean square value of the original data $\sigma_x = \sqrt{\langle (x - \overline{x})^2 \rangle}$.

In Table 1 we list for different metabolites the root mean square errors. The one in brackets is $\sigma_x$, the other the root mean square error of the deviation of the predicted values from the original values $\sigma_{pred}$. We list those values for the different prediction cases according to the Equations 2, 3, 4 and 5 and for all the values together. In the bottom line we show the values for the data of all metabolites put together.

For the entire data set we get a prediction error of $\sigma_{pred} = 1.01$ that is smaller than the uncertainty $\sigma_x$. In three of the four different cases corresponding to the Equations 3, 4 5 and 2 the prediction error is smaller than the uncertainty of the experimental values. Only for the case 2, in which the prediction is based on non-related enzymes, the prediction error is bigger, $\sigma_{pred} = 1.51$ compared to $\sigma_x = 1.32$.

## 4 Discussion

**Motivation.** Modelling of biochemical networks with ordinary differential equations is a classical approach to understand their dynamics and to study the effect of experimental intervention or envi-

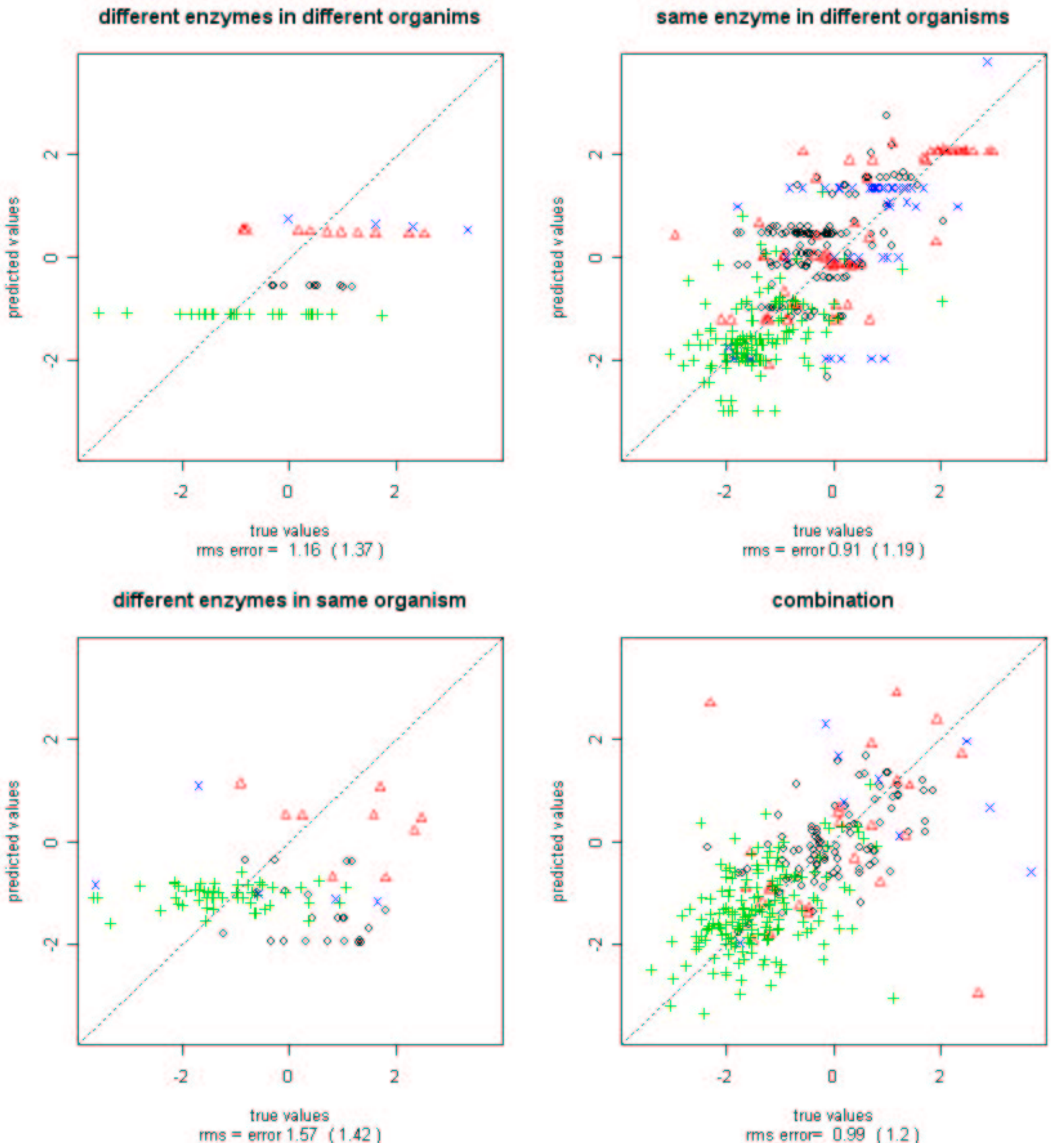

Figure 2: Scatter plots of the predicted against the true logarithmic $K_M$ values for four different metabolites and for the different prediction cases 1, 2, 3 and 4 according to the Equations 3, 4, 5 and 2. The four metabolites are D-Glucose ($\triangle$), Ethanol ($\times$), NADP+ ($+$) and Pyruvate ($\circ$). Below each plot two root mean square errors are indicated. The one in brackets is the root mean square error of the original data $\sigma_x = \sqrt{\langle (x - \overline{x})^2 \rangle}$, the other the root mean square error of the deviation from the predicted values of the true values $\sigma_{pred} = \sqrt{\langle (x^* - x)^2 \rangle}$. In the case of perfect prediction the points would all lie on the dotted lines.

Table 1: Quality of predicted $K_{\mathrm{M}}$ values. We calculated the root mean square error $\sigma_{pred} = \sqrt{\langle(x^* - x)^2\rangle}$ where $x^*$ denotes the predicted value of $x$ for the different cases denoted by 1, 2, 3 and 4 that correspond to the different prediction cases according to the Equations 3, 4, 5 and 2. The last column contains the overall root mean square error for each metabolite regardless of the 4 cases. The bottom line integrates the data from all the metabolites.

| metabolite | $\sigma_{pred,1}$ | $\sigma_{pred,2}$ | $\sigma_{pred,3}$ | $\sigma_{pred,4}$ | $\sigma_{pred,1+2+3+4}$ |
|---|---|---|---|---|---|
| D-Glucose | 1.17 (1.14) | 1.39 (1.09) | 0.92 (1.39) | 1.65 (1.33) | 1.22 (1.35) |
| Ethanol | 1.3 (1.22) | 2.53 (2.15) | 1.16 (1.01) | 1.9 (1.61) | 1.56 (1.33) |
| NADP+ | 1.22 (1.21) | 1.03 (1.02) | 0.8 (0.74) | 0.88 (0.83) | 0.91 (0.88) |
| Pyruvate | 0.51 (0.5) | 1.14 (0.85) | 0.79 (0.75) | 0.75 (0.88) | 1.06 (0.83) |
| D-Fructose-6-phophate | 1.44 (0.84) | 0.55 (0.56) | 0.7 (0.68) | 0.5 (0.62) | 0.88 (0.74) |
| D-Glucose-6-phosphate | 0.03 (0.24) | 0.76 (1.04) | 0.57 (0.79) | 0.65 (0.9) | 0.79 (0.86) |
| Acetyl-CoA | 0.87 (0.86) | 0.94 (0.85) | 0.71 (0.69) | 0.88 (0.8) | 0.83 (0.77) |
| Succinyl-CoA | 0.55 (0.52) | 1.42 (0.45) | 1.23 (0.6) | 1.18 (0.71) | 1.41 (0.67) |
| all | 1.25 (1.31) | 1.51 (1.32) | 0.87 (1.1) | 0.94 (1.11) | 1.01 (1.16) |

ronmental changes [4, 6]. The system dynamics are determined by (i) the network structure and (ii) by the kinetic parameters in the kinetic rate laws. In this paper, we focus on the parameters and acknowledge that it is important to know the values of such parameters. On the other hand, it is well-known from parameter sensitivity studies [1] and from bifurcation analysis that some parameter values may change over smaller or wider ranges without qualitative changes of the system behaviour. Other parameters may have strong effect and thus need to be determined accurately. In any case, even if we cannot learn the exact parameter values, it is worth to have the best possible guess for such values.

**Summary.** We presented an approach to deduce the values of kinetic parameters from experimentally determined values that are stored in databases. We used kinetic parameters from the database BRENDA and compared the parameter values for specific reactions that are determined by (i) a metabolite involved in that reaction, (ii) the enzyme catalysing the reaction, and (iii) the organism for which the value was measured. We assume correspondence for all parameter values that belong to the same EC number and metabolite, but originate from different organisms and are measured in different experiments for possibly different enzymes. By applying a linear regression model to the logarithmic data, we implicitly assume that these factors have multiplicative effects on the (non-logarithmic) $K_{\mathrm{M}}$ values.

The $K_{\mathrm{M}}$ values for each metabolite are first arranged in a matrix. The rows are determined by the enzymes, and the columns are defined by the species for which at least one parameter value has been measured. An example is shown in Figure 1. The resulting matrices are very sparse (see Table 2), since not in all species the parameters for each enzyme have been measured. Moreover, not all positions in the matrix must be realised in the biological reality: it is possible that for a certain metabolite there are experimental values for enzyme A and B in species X and experimental values for enzymes B and C in species Y. Thus both species X and Y as well as all enzymes A, B, and C give rise to rows and columns in the matrix, although species X does not necessarily also have enzyme C.

Our statistical approach uses a linear regression model that is also the basis for analysis of variance (ANOVA) [9]. As explained above, we assign values to empty entries according to the values in non-empty entries in the matrix. As a consequence, a high experimental value in one entry causes high values in the unknown fields in the same row and column. Therefore, measurement errors or errors in data handling (writing errors, mismatch of units) may drastically influence the prediction (see Figure 1).

Table 2: The size of the data matrices and a measure of their sparseness for 8 different metabolites. $I$ is the number of enzymes found that catalyse the metabolite. $J$ is the number of organisms found in which the metabolite is catalysed by an enzyme. $N_{data}$ is the number of data entries for a given metabolite.

| metabolite | $I \times J$ | $N_{data}/(I \cdot J)$ |
|---|---|---|
| D-Glucose | $39 \times 87$ | 114/3393 |
| Ethanol | $17 \times 56$ | 66/952 |
| NADP+ | $140 \times 222$ | 382/31080 |
| Pyruvate | $68 \times 181$ | 260/12308 |
| D-Fructose-6-phosphate | $17 \times 74$ | 88/1258 |
| D-Glucose-6-phosphate | $14 \times 28$ | 39/392 |
| Acetyl-CoA | $78 \times 154$ | 269/12012 |
| Succinyl-CoA | $16 \times 19$ | 25/304 |

We tested the quality of our parameter prediction by leave-one-out cross-validation. The results are summed up in Figure 2 and Table 1. For the entire data set the prediction error $\sigma_{pred} = 1.01$ is smaller than the uncertainty in the experimental values $\sigma_x = 1.16$.

After collecting the error ranges from many metabolites, we can use them to appraise the expected error when predicting new, unknown $K_{\mathrm{M}}$ values. In doing so, we need to assume that there is no systematic bias between the experimentally known and unknown parameters, in other words, that the missing mechanism is at random.

**Biological interpretation of the statistical model.** The Michaelis-Menten constants in cells are an outcome of evolution: the $K_{\mathrm{M}}$ values may be conserved between related species, and most probably they have also been shaped by functional requirements. Both factors would lead to similarities between related $K_{\mathrm{M}}$ values in different species. Our cross-validation results show that part of the statistical variation in $K_{\mathrm{M}}$ values can indeed be predicted from combinations of metabolite, EC number, and organism. But how can the influence of these factors on the $K_{\mathrm{M}}$ values be interpreted in biological terms?

According to our regression model, a $K_{\mathrm{M}}$ value depends (i) on the substrate (the total mean $\mu$), (ii) on the combination substrate-enzyme (effect $\alpha_i$), and (iii) on the combination substrate-organism (effect $\beta_j$). The effect $\alpha_i$ allows some enzyme-substrate pairs to have higher or lower $K_{\mathrm{M}}$ values, irrespective of the organism. This can reflect conservation in evolution, or to functional requirements within conserved pathways. The effect $\beta_j$ captures increased $K_{\mathrm{M}}$ values of a metabolite in a certain organism, irrespective of the enzyme. To justify this, we may hypothesize that $K_{\mathrm{M}}$ values are adjusted to the order of magnitude of the substrate concentration. If this is the case and if a metabolite exhibits a particularly high concentration in a certain organism, then all corresponding $K_{\mathrm{M}}$ values should also tend to be increased.

**Application to other kinetic parameters.** To illustrate the approach, we restricted ourselves here to $K_{\mathrm{M}}$ values. The presented approach can as well be applied to other types of parameters that are stored in databases, such as maximal velocities or catalytic constants, inhibition constants, Hill coefficients, or association and dissociation constants. The hand-curated database BRENDA hosts currently data of different type for approximately 83,000 different enzymes. In the future, we plan to extend our analysis to more data from various sources. In particular, we shall also use the about 200,000 parameter values that have been obtained by an automatic search in all abstracts contained in PubMed, using a text-mining approach. These data are stored in the database KMedDB [13].

In systems biology, it would be desirable to obtain all parameters for a specific model from experiments that are performed to support this model and to determine its parameters. Ideally, these experiments should be undertaken in compatible conditions, equal cell lines or strains, and so on. In reality, such data are typically not available. The presented method is an approach to integrate data from different experiments and to extract information about parameter values and ranges, which can then be readily used for dynamical modelling. By running the analysis on large data sets, we may also study conservation of kinetic parameters between related species, which could then be used to improve the prediction scheme.

## Acknowledgments

Simon Borger thanks the Graduate Program Berlin "Dynamics and Evolution of Cellular and Macromolecular Processes" (the predecessor of the International Research Training Group "Genomics and Systems Biology of Molecular Networks").Wolfram Liebermeister is supported by the European Comission, grant 503296, Edda Klipp by the German Federal Ministry for Education and Research (BMBF, grant 031U109C).

## References

[1] Bluthgen, N. and Herzel, H., How robust are switches in intracellular signaling cascades?, *J. Theor. Biol.*, 225(3):293–300, 2003.

[2] Gadkar, K.G., Gunawan, R., and Doyle, III F.J., Iterative approach to model identification of biological networks, *BMC Bioinformatics*, 6:155, 2005.

[3] Heinrich, R. and Rapoport, T.A., A linear steady-state treatment of enzymatic chains. General properties, control and effector strength, *Eur. J. Biochem.*, 42(1):89–95, 1974.

[4] Heinrich, R. and Schuster, S., *The Regulation of Cellular Systems*, Chapman and Hall, 1996.

[5] Kacser, H. and Burns, J.A., The control of flux, *Symp. Soc. Exp. Biol.*, 27:65–104, 1973.

[6] 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, 2005.

[7] Mendes, P. and Kell, D., Non-linear optimization of biochemical pathways: Applications to metabolic engineering and parameter estimation, *Bioinformatics*, 14(10):869–883, 1998.

[8] Moles, C.G., Mendes, P., and Banga, J.R., Parameter estimation in biochemical pathways: A comparison of global optimization methods, *Genome Res.*, 13(11):2467–2474, 2003.

[9] Rencher, A.C., *Linear Models in Statistics*, John Wiley & Sons, Inc., 2000.

[10] Schmeier, S., Kowald, A., Hakenberg, J., Leser, U., and Klipp, E., KMedDB, a tool for searching kinetic data in PubMed abstracts, submitted 2006.

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

[12] Tyson, J.J., Chen, K.C., and Novak, B., Sniffers, buzzers, toggles and blinkers: Dynamics of regulatory and signaling pathways in the cell, *Curr. Opin. Cell Biol.*, 15(2):221–231, 2003.

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

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

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