RESEARCH ARTICLE

Reconstructing biochemical pathways from time course data

Jeyaraman Srividhya[1], *Edmund J. Crampin*[2, 3], *Patrick E. McSharry*[4, 5, 6] *and Santiago Schnell*[1]

[1] Indiana University School of Informatics and Biocomplexity Institute, Bloomington, IN, USA
[2] Auckland Bioengineering Institute, The University of Auckland, Auckland, New Zealand
[3] Department of Engineering Science, The University of Auckland, Auckland, New Zealand
[4] Department of Engineering Science, University of Oxford, Oxford, UK
[5] Oxford Centre for Industrial and Applied Mathematics, Mathematical Institute, Oxford, UK
[6] Oxford Centre for Integrative Systems Biology, University of Oxford, UK

Time series data on biochemical reactions reveal transient behavior, away from chemical equilibrium, and contain information on the dynamic interactions among reacting components. However, this information can be difficult to extract using conventional analysis techniques. We present a new method to infer biochemical pathway mechanisms from time course data using a global nonlinear modeling technique to identify the elementary reaction steps which constitute the pathway. The method involves the generation of a complete dictionary of polynomial basis functions based on the law of mass action. Using these basis functions, there are two approaches to model construction, namely the general to specific and the specific to general approach. We demonstrate that our new methodology reconstructs the chemical reaction steps and connectivity of the glycolytic pathway of *Lactococcus lactis* from time course experimental data.

Received: June 9, 2006
Revised: September 29, 2006
Accepted: October 10, 2006

Keywords:
Biochemical network inference / Model selection / Reverse engineering / Systems biology / Time series data

1 Introduction

In the biological sciences it is increasingly common for data to be collected in high-throughput experiments on genomic, proteomic, and metabolomic scales. These data hold great promise for enabling researchers to identify and model the components and interactions comprising regulatory biochemical networks. However, systematic and comprehensive profiling experiments produce large and complicated data sets for analysis. Therefore, these experimental advances demand parallel development of computational approaches for their analysis.

In recent years, there have been multiple attempts to map biochemical pathways from experimental data, using a variety of computational tools. Techniques which have been adopted include sequence similarity [1, 2], identification of common structural motifs [3], gene order [4], gene fusion events [5], and correlated gene expression profiles [6]. These approaches have proved to be very useful in providing a static picture of protein function in a biochemical pathway. However, biochemical systems are, by their nature, dynamic. As a consequence, the focus is changing toward the development of mathematical and computational methods to predict function based on the dynamic regulation of genes and proteins in networks [7]. The current advances in high-throughput measurement technologies, combined with high performance computing,

Correspondence: Professor Santiago Schnell, Indiana University School of Informatics, 1900 East Tenth Street, Eigenmann Hall 906, Bloomington, Indiana 47406, USA
E-mail: schnell@indiana.edu
Fax: +1-812-856-1995

Abbreviations: FBP, fructose-1,6-bisphosphate; **G6P,** glucose-6-phosphate; **IC,** information criterion; **ODE,** ordinary differential equation; **PEP,** phosphoenolpyruvate make possible the application of such methods to determine reaction pathways and kinetics from experimental data on a system-wide scale.

Information about the biochemical pathway can be obtained by studying the behavior of the system near to a steady state. Data obtained in perturbation methods, in which one or more of the species are disturbed from their steady values and the transient response of the pathway is monitored, can be used to identify the connectivity in the pathway [8, 9]. An alternative approach for probing biochemical pathways near to a kinetic steady state is to manipulate system parameters, rather than the concentrations of the reactants and reaction intermediates themselves [10, 11]. A qualitative form of impulse response analysis has also been proposed to gather information on the connectivity of a biochemical network [12]. Correlation based approaches for identifying networks have been increasingly useful in analyzing gene networks from microarray experiments [13, 14]. Genetic network information can be obtained using the reverse engineering approach [15, 16] and cellular networks can be inferred using probabilistic graph models [17]. For a comprehensive review of computational methods available for deducing the biochemical reaction mechanism, we invite the reader to consult Crampin *et al.* [18].

Experimental tools are available which provide powerful strategies for identifying the structure of metabolic and proteomic networks. Such tools include NMR [19, 20], MS, time-resolved fluorescence spectroscopy, fluorescence labeling combined with autoradiography on 2-D gels [21], protein kinase phosphorylation [22], and tissue arrays [23] for simultaneous high throughput analysis of proteins in a tissue section by means of antibody binding and MS. What is common among these techniques is that they allow the simultaneous measurement of the abundance of multiple metabolites or proteins, either at one time point [24] or as a sequence of measurements giving time series data. Currently, experiments that give the concentration of metabolites as a function of time are limited in number. However, this situation is rapidly changing, for example, with the emergence of *in vivo* ^{13}C NMR experiments [19] and microarray time series experiments [25].

Time series data can reveal transient behavior, away from chemical equilibrium, and contain information on the dynamic interactions between reacting components. But this information can be difficult to extract from time series data sets using conventional analysis techniques. There is, therefore, a compelling need for the development of computational tools to extract mechanistic information from biochemical time series data, in particular for situations in which prior information on the biochemical steps in the pathway is not available. Recently, there have been attempts to identify metabolic networks from time series using S-systems approach [26, 27].

The task for identifying biochemical pathways from time course data consists, firstly, of identifying the connectivity of the pathway – the reaction diagram relating reactants and products – and, secondly, determining and parameterizing the reaction mechanisms for each of the steps in the pathway. These two steps require a good deal of chemical knowledge about plausible interconversions for the species in the pathway. Once the reaction steps and mechanisms are known, techniques are available for the estimation of kinetic parameters [28, 29]. However, for less well characterized chemical components, or for more complicated networks, this approach is not practicable. One strategy to tackle this problem is to develop techniques which can reveal details of the molecular interactions that constitute a complex reaction mechanism or pathway by considering elementary reaction steps.

In this paper, we present a new method to infer biochemical pathway mechanisms from time course data using a global nonlinear modeling technique to identify the elementary reaction steps which constitute the pathway. A significant feature of our method is that we develop a global nonlinear modeling technique based on the law of mass action. This helps our procedure to arrive at chemically plausible reaction steps and to identify pathway connectivity. The method involves the generation of a complete dictionary of possible chemical interactions (which we will refer to as 'elementary reactions') and applies a model selection technique to deduce the reaction mechanism from the data. Model selection can be approached by two routes: the specific to general approach and the general to specific approach. The algorithm predicts the reaction mechanisms as a set of kinetic equations describing the rates of change of each chemical species in the pathway, reconstructed from the time series data. In Section 2, we discuss the methodology in detail followed, in Section 3, by some examples of its application. A preliminary study was previously published in [30].

2 Methodology

A kinetic model for a biochemical pathway provides a description for the rate of production of each species in terms of the concentrations $\mathbf{x} = \mathbf{x}(t)$

$$\frac{\mathrm{d}x_\mathrm{i}}{\mathrm{d}t} = F_\mathrm{i}(\mathbf{x}, \mathbf{a}_\mathrm{i}) \qquad (1)$$

The net production rate of each species F_i can be expressed as a weighted sum of K basis functions, Φ_j,

$$F_\mathrm{i}(\mathbf{x}, \mathbf{a}_\mathrm{i}) = \sum_{\mathrm{j}=1}^{\mathrm{K}} a_\mathrm{ij}\Phi_\mathrm{j}(\mathbf{x}, \mathbf{b}) \qquad (2)$$

These basis functions are mathematical functions corresponding to each of the elementary reactions, as described above, and the weighted sum represents the contributions from different elementary processes. Here **b** refers to the parameters specific to elementary processes. If the basis functions are kept fixed, however, and only the weights a_ij varied (*i.e.* the parameters **b** are assumed to be known *a priori*), then the model may be fitted to the data using least squares and singular value decomposition (SVD). The model parameters $\mathbf{a}_i = \{a_{ij}\}_{j=1}^{K}$ can be determined by minimizing the sum of squared residuals

$$\chi_i^2 = \left\|\mathbf{y}_i - \Phi \cdot \mathbf{a}_i\right\|^2 \tag{3}$$

where $\mathbf{y}_i = \{dx_i(t_j)/dt\}_{j=1}^{N}$ is the derivative of the time series, and the matrix $\Phi_{jl} = \Phi_j(\mathbf{x}(t_l))$ is the model design matrix. This is described in detail in the following section.

A drawback of this approach is that SVD will find non-zero values for all weights a_{ij} (for an overdetermined system; while for an underdetermined system with N independent data points, weights up to *N-1* basis functions can be determined). In particular, for a noisy data set this approach will tend to over-fit the data. If we choose basis functions which represent elementary reactions between species, only a subset of the potential reactions should be required to model the time series data. Therefore, we wish to select only those basis functions which represent genuine interactions underlying the data. One strategy would be to perform an exhaustive search over all models using q from K basis functions, choosing the one which minimizes the model residuals, and then using a model selection criterion to determine which model size q gives the best model of the data. However, this exhaustive search quickly becomes computationally intractable as the number of species, and basis functions, increases.

Our approach to this problem is to construct models using $q \leq K$ basis functions using an iterative method [31]. This determines how to select the optimal model comprising $q+1$ basis functions, starting from the optimal model with q basis functions, and *vice versa*. Model selection can be approached by starting with one basis function and iteratively adding to the model (simple to general) or, if there are more data points than basis functions, starting from a model using the entire set of basis functions and removing basis functions (general to specific). For both methods, the optimal model size can then be selected by minimizing a cost function which penalizes the use of more basis functions unless there is sufficient payoff in reducing the model residuals. Formally, this is achieved using a cost function or a penalty term called an information criterion (IC). After considerable experimentation with different penalty terms from various ICs, such as Akaike [32] and Schwarz [33] ICs, we have adopted the following empirical IC, which was found to provide the most reliable results. The cost function to be minimized over the model size q is as follows:

$$C_{IC} = \frac{1}{N}\left(\mathbf{E}^{(q)^T} \cdot \mathbf{E}^{(q)}\right) + q \tag{4}$$

where $\mathbf{E}^{(q)} = \mathbf{y} - \Phi^{(q)} \cdot \mathbf{a}$ is the model residual vector for the optimal model constructed from q basis functions.

In Fig. 1 we outline our method schematically, showing how the different steps in the algorithm are integrated to arrive at the reaction mechanism. The sequence of steps involved in the method is (i) construction of the model design matrix, (ii) construction of the derivative matrix, (iii) model selection module, and (iv) the ordinary differential equation (ODE) reconstructor. The final output consists of the predicted reaction steps and the reconstructed ODE model for the pathway. In the following sections, each of these steps is discussed.

2.1 Construction of the model design matrix

A key aspect of our method lies in the construction of a model design matrix Φ_{ij} appropriate for biochemical pathways. The model design matrix is a matrix with columns representing unscaled velocities corresponding to all possible elementary reaction steps involving the different species in the pathway. Therefore, the first step of our method is to construct a complete dictionary of chemically feasible elementary reaction steps for a given number of species.

2.1.1 Law of mass action

Chemical reaction pathways are composed of a number of elementary steps. Let us consider the general chemical elementary reaction

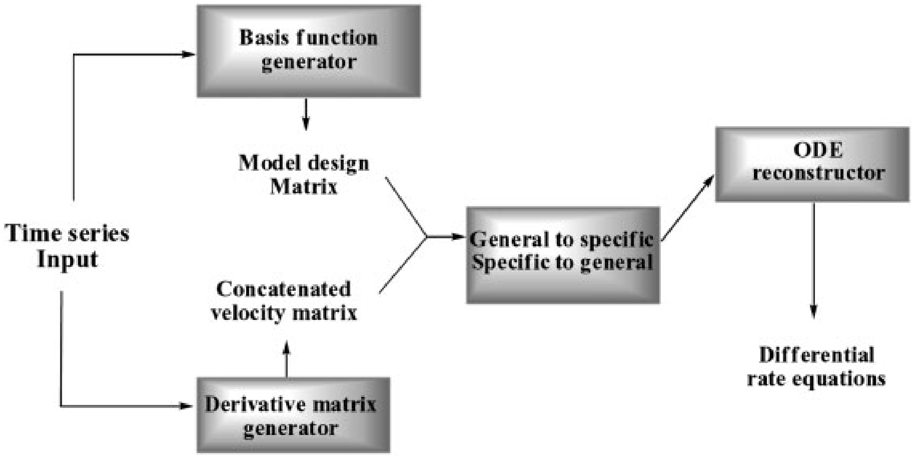

Basis function generator
Model design Matrix
Time series Input
Concatenated velocity matrix
Derivative matrix generator
General to specific Specific to general
ODE reconstructor
Differential rate equations

Figure 1. Block diagram of the model selection method. The basis function dictionary depends on the number of species in the time series and the model design matrix is constructed accordingly. Once the model is selected, the ODE reconstructor generates a set of differential rate equations as a final output of the algorithm.

$$n^{A}A + n^{B}B \xrightarrow{\lambda} n^{C}C + n^{D}D \tag{5}$$

Here, λ is the rate constant of the reaction and n^{A}, n^{B}, n^{C}, and n^{D} are the number of molecules of reactants A, B, C, and D that participate in the reaction. The velocity or rate of the above reaction is given, according to the law of mass action, by

$$\nu = \lambda\ (x_{A})^{nA}\,(x_{B})^{nB} \equiv \lambda\,\phi(x_{A}, x_{B}) \tag{6}$$

where x_A and x_B are the concentrations of species A and B, respectively. This defines the unscaled velocity ϕ which describes the functional dependence of the reaction velocity on the concentration variables and which does not depend on any further unknown parameters. The rates of change of the species are given as follows:

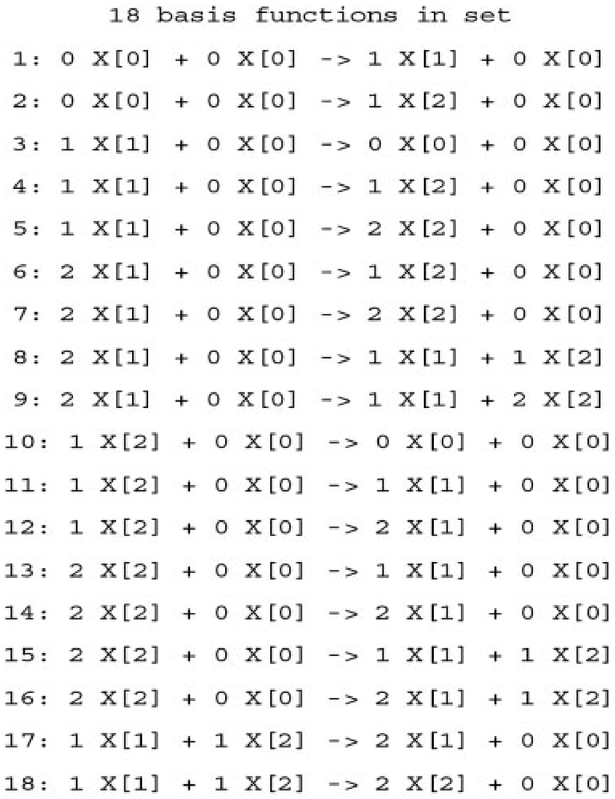
$$\nu(t) = \frac{-1}{n^{A}}\frac{dx_{A}}{dt} = \frac{-1}{n^{B}}\frac{dx_{B}}{dt} = \frac{1}{n^{C}}\frac{dx_{C}}{dt} = \frac{1}{n^{D}}\frac{dx_{D}}{dt} \tag{7}$$

This general framework can be used to construct a set of chemically feasible elementary reactions if we restrict the reactions to a maximum molecularity. For example, for two species, the general elementary reaction (5) can produce up to 18 chemically realistic schemes (18 choices of the integers n^{A}, n^{B}, n^{C}, and n^{D}) including bimolecular reactions, as shown in Fig. 2. In the figure, indices given in square brackets label the species, with zeroes indicating the absence of a species. This is called the complete dictionary of basis functions for two species. As the logic used to generate these reactions is based on mass action kinetics, it can be manipulated to generate reactions of any molecularity, for any number of species. Note that we have restricted our investigations to uni- and bimolecular elementary reactions only.

For a species k, an element of the model design matrix is then defined as $\Phi^{k}_{ij} = \sigma^{k}_{i} n^{k}_{i} \phi_{i}(t_{j})$. Here, n^{k}_{i} is the molecularity for species k in the ith reaction, and the element σ^{k}_{i} has unit magnitude with a positive sign ($\sigma^{k}_{i} = +1$) if k is a product and negative sign ($\sigma^{k}_{i} = -1$) if k is a reactant for the ith reaction. $\phi_{i}(t_{j})$ is the unscaled velocity for the ith reaction, as described in Eq. (6), evaluated at the jth time point. For example, for the set of basis functions in Fig. 2, $\phi_{4}(t_{j}) = -x_{1}(t_{j})$ and $\phi_{17}(t_{j}) = -x_{1}(t_{j}) \cdot x_{2}(t_{j})$. The velocities are evaluated from each point in the time course and the model design matrix Φ^{k}_{ij} is constructed for each species as follows:

$$\Phi^{k} = \begin{pmatrix} \sigma^{k}_{1} n^{k}_{1} \phi_{1}(t_{1}) & \sigma^{k}_{2} n^{k}_{2} \phi_{2}(t_{1}) & \ldots & \sigma^{k}_{K} n^{k}_{K} \phi_{K}(t_{1}) \\ \sigma^{k}_{1} n^{k}_{1} \phi_{1}(t_{2}) & \sigma^{k}_{2} n^{k}_{2} \phi_{2}(t_{2}) & \cdots & \sigma^{k}_{K} n^{k}_{K} \phi_{K}(t_{2}) \\ \vdots & \vdots & \vdots & \vdots \\ \sigma^{k}_{1} n^{k}_{1} \phi_{1}(t_{N}) & \sigma^{k}_{2} n^{k}_{2} \phi_{2}(t_{N}) & \ldots & \sigma^{k}_{K} n^{k}_{K} \phi_{K}(t_{N}) \end{pmatrix}$$

Only for those reactions in which species k takes part will n be nonzero. The overall matrix for the biochemical pathway is a concatenation of such matrices for each of the M species, resulting in a matrix of dimensions $NM \times K$, where N is the number of time points in the time series and K is the number of elementary reactions from which the model is to be constructed.



```
18 basis functions in set
1: 0 X[0] + 0 X[0] -> 1 X[1] + 0 X[0]
2: 0 X[0] + 0 X[0] -> 1 X[2] + 0 X[0]
3: 1 X[1] + 0 X[0] -> 0 X[0] + 0 X[0]
4: 1 X[1] + 0 X[0] -> 1 X[2] + 0 X[0]
5: 1 X[1] + 0 X[0] -> 2 X[2] + 0 X[0]
6: 2 X[1] + 0 X[0] -> 1 X[2] + 0 X[0]
7: 2 X[1] + 0 X[0] -> 2 X[2] + 0 X[0]
8: 2 X[1] + 0 X[0] -> 1 X[1] + 1 X[2]
9: 2 X[1] + 0 X[0] -> 1 X[1] + 2 X[2]
10: 1 X[2] + 0 X[0] -> 0 X[0] + 0 X[0]
11: 1 X[2] + 0 X[0] -> 1 X[1] + 0 X[0]
12: 1 X[2] + 0 X[0] -> 2 X[1] + 0 X[0]
13: 2 X[2] + 0 X[0] -> 1 X[1] + 0 X[0]
14: 2 X[2] + 0 X[0] -> 2 X[1] + 0 X[0]
15: 2 X[2] + 0 X[0] -> 1 X[1] + 1 X[2]
16: 2 X[2] + 0 X[0] -> 2 X[1] + 1 X[2]
17: 1 X[1] + 1 X[2] -> 2 X[1] + 0 X[0]
18: 1 X[1] + 1 X[2] -> 2 X[2] + 0 X[0]
```

Figure 2. Set of possible elementary reactions generated for two species for uni- and bimolecular interactions. The indices given in square brackets indicate the species and the zero index implies the absence of any species. The numbers beside the species indicate the molecularity of the species.

Selecting different sets of possible reactions (different basis sets) will, in turn, alter the model determined by the algorithm from the data. A complete dictionary is a comprehensive description of all the chemically feasible elementary reactions. If, however, one is interested only in the connectivity of a pathway, then a subset of this complete dictionary can be used. For example, the subset of the complete dictionary which consists solely of interconversions between species is particularly useful for identifying a metabolic pathway diagram. The basis set used here would be confined to interactions of the type $n_{i}X_{i} \rightarrow n_{j}X_{j}$ only. Having constructed the model design matrix, the next step is the application of the iterative model selection method to deduce the mechanism.

2.2 Construction of the derivative matrix

Time course data on the biochemical pathway are used to calculate the derivative matrix. For the time series $x(t_{j})$, the derivative vector y of the points is calculated according to

$y_j = (x(t_{j+1}) - x(t_j))/(t_{j+1} - t_j)$ for each species. The data points are interpolated if the time series contains few time points.

2.3 Model selection

The iterative scheme proposed by Judd and Mees [31] uses a sensitivity analysis to determine the basis function to add to a model that will give most improvement to the model fit to the data, and the basis function to remove from a model that will least damage the approximation. A model selection criterion such as the cost function described above can then be used to select the best model size k from a set of models constructed in this way. McSharry *et al.* [34] extended this iterative data-driven approach to extract a set of non-orthogonal empirical functions (NEFs) from multivariate data sets. NEFs have the appeal of providing a decomposition which is motivated by the problem-domain (accounting for the underlying dynamics and conservation laws) rather than the statistical convenience offered by classical decomposition techniques such as principal component analysis. Crampin *et al.* [30] demonstrated that the law of mass action can be employed to constrain the set of relevant basis functions and that the model construction can be attempted by either a specific to general or a general to specific approach.

2.3.1 Specific to general approach

This approach expands the model size, starting from a single basis function and then adding basis functions iteratively until the stopping criterion, minimization of the cost function (Eq. (4)) is reached. Selection of the basis function to be used to increase the model size is determined by considering $\mu = -\Phi^T E^K$, the projection of the vector of the residuals onto the model design matrix. The largest positive element in μ is selected as the first basis function and subsequently basis functions are added, subject to the minimization of the cost function. Additionally, the algorithm uses a non-negative constraint for obtaining positive coefficients in the least square method.

2.3.2 General to specific approach

Alternatively, all of the basis functions from the complete dictionary are used to form the initial model, which is then simplified by discarding terms iteratively, until the same cost function is minimized. The algorithm requires an initial selection of coefficients to start with. The least squares solution of $\mathbf{y}$ and Φ with non-negative constraint was selected as initial coefficients. This was then followed by application of the same IC (Eq. (4)) alternately to eject and then to add a basis function until the same basis function is chosen and is removed from the subset, reducing the model size by one.

2.4 ODE reconstruction

Having identified the mechanism using either of the above approaches, the differential rate equations can be reconstructed from the basis functions and the coefficients inferred. The reconstruction is simply based on chemical kinetics. The ODE reconstructor gives the differential equations as the final output of the algorithm.

3 Results

3.1 Method validation

We tested the performance of our methodology using an approach based on calculating the sensitivity of the model inference. The sensitivity here accounts for two aspects, namely, (i) the correctness of the reaction structure and (ii) the correctness of the parameters, in comparison with the true "generative" system. The former can give topological information about the mechanism and the latter can yield a measure of the goodness-of-fit of the parameters. We define the topological sensitivity index S_I [16] of the inferred model as

$$S_I = \frac{T_C}{T_C + T_F + T_U} \tag{8}$$

where T_C is the total number of correctly identified reactions, T_F is total number of falsely identified reactions and T_U is the total number of unidentified reactions. Here the term T_U refers to a case where the number of identified reactions is less than that of the total number of reactions used to generate the time series. We note that the differential rate equations, and not the reaction mechanisms, are used to calculate this sensitivity.

To test the ability of our approach to correctly infer biochemical networks, we calculated the sensitivity index for a wide range of chemical reactions using simulated time series data sets. In order to facilitate comparison of performance across a range of reactions, we quantified the complexity of the chemical reactions underlying these data sets based on an approach from graph theory [35]. A brief account of evaluating the complexity of the chemical reactions is given as follows: for any given chemical reaction or a set of chemical reactions, it is possible to construct a network diagram as a directed bipartite graph linking reactants *via* reaction fluxes. A typical bipartite graph for a simple reaction network $X_1 \longrightarrow X_2 \longrightarrow X_3$ is shown in Fig. 3. From the bipartite graph for the chemical reaction the complexity index I_C can then be calculated using the formula

$$I_C = m\, z \sum_{i=1}^{m} T_i \tag{9}$$

where m is the number of elementary chemical reaction steps in the mechanism, z is the number of species and T_i is the total number of branches emanating from and

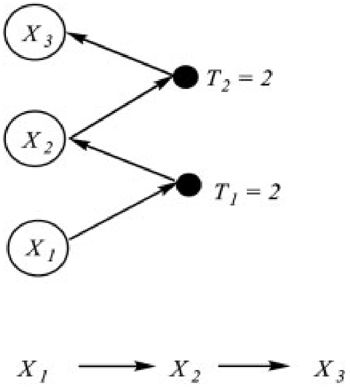


Figure 3. Bipartite graph for a chemical transformation. The reactants are indicated in circles and the reactions are represented as filled circles. Here, the number of elementary chemical reaction steps, $m = 2$, and the number of species, $z = 3$. The number of branches originating from and ending in each reaction point $T_1 = T_2 = 2$. The overall complexity index is $I_C = 24$.

Table 1. Complexity index I_C for some unimolecular and bimolecular reactions calculated on the basis of bipartite graphs. m is the number of reaction steps, z is the number of species, T_i is the number of branches emanating from the reaction points in the corresponding bipartite graph

	m	z	T_i	$I_C = m\,z\sum_{i=1}^{m} T_i$
Unimolecular reactions				
$X_1 \rightarrow X_2$	1	2	2	4
$X_1 \rightarrow X_2 \rightarrow X_3$	2	3	2, 2	24
$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow$	3	3	2, 2, 1	45
$\rightarrow X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow$	4	3	1, 2, 2, 1	72
$\rightarrow X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow$ $\uparrow$ $\rightarrow X_4$	5	4	1, 2, 2, 2, 1	160
$X_1 + X_2 \rightarrow X_3 \rightarrow X_2 + X_4$	2	4	3, 3	48
Bimolecular reactions				
$X_1 + X_2 \rightarrow$	1	2	2	4
$2X_1 \rightarrow X_1 + X_2$	1	2	4	8
$X_1 + X_2 \rightarrow 2X_1$	1	2	4	8
$X_1 + X_2 \rightarrow X_3$	1	3	3	9
$X_1 + X_2 \rightarrow X_3 + X_4$	1	4	4	16

ending in each reaction node (dark circles, Fig. 3). Complexity indices for some simple reactions are given in Table 1.

We simulated time series data for a range of pathways of unimolecular and bimolecular reactions of varying complexity in order to measure the network inference performance as a function of network complexity. Figures 4a and b show the performance of the specific to general and general to specific approaches in terms of topological sensitivity indices for increasing reaction complexity, for unimolecular and bimolecular reactions, respectively. Figure 4b shows that for the specific to general approach, the sensitivity decreases with complexity for pathways involving bimolecular reactions. However, that is not so with the general to specific approach, for which the topological sensitivity of the method remains fairly high for reactions of increasing complexity. Thus the general to specific approach outperforms the specific to general approach.

The topological sensitivity index S_I gives a measure of the inferred reaction topology only. Total sensitivity T_S is a measure of both the topological accuracy and parametric accuracy and can be defined as follows:

$$T_S = S_I - e_P \tag{10}$$

$$e_P = \frac{1}{\beta}\sum_{i=1}^{M}\left(\sum_{j}^{M}\left(a_{ij} - a_{ij}^{*}\right)^2\right) \tag{11}$$

where e_P is the error associated with the parametric estimations. Here β is the total number of nonzero terms in $F(x_i, \mathbf{a}_i)$, a_{ij} are the coefficients identified by the model and a_{ij}^{*} are the true coefficients used to generate the time series. A plot of the total sensitivity for pathway inference with the general to specific approach, for unimolecular reactions of varying complexity index, is given in Fig. 4c. We see that the total sensitivity decreases to 0.6 when the complexity index reaches 60 and then remains fairly constant for the general to specific approach. While we suppose that this is one way to quantify the model errors, model accuracy depends on many other factors, for example the time interval between data points, parameters of the system itself, and so on, whose evaluation complicates the error variables. Nevertheless, the sensitivity measures described by Eqs. (8) and (10) are adequate to support the validity of our method.

3.2 Examples

Two examples that highlight the features of our approach are presented below. The first example is a typical enzyme kinetics reaction which we use to illustrate the application of the algorithm to predict a reaction mechanism from a single time series only. The second example uses a data set measured from the glycolytic pathway of *Lactococcus lactis*. In this example we show how our method can predict the topology of this metabolic pathway.

3.2.1 Example 1

The following reaction mechanism is a typical example containing a bimolecular reaction step [36]:

$$E + S \xrightarrow{k_1} C \tag{12}$$

$$C \xrightarrow{k_2} P + E \tag{13}$$

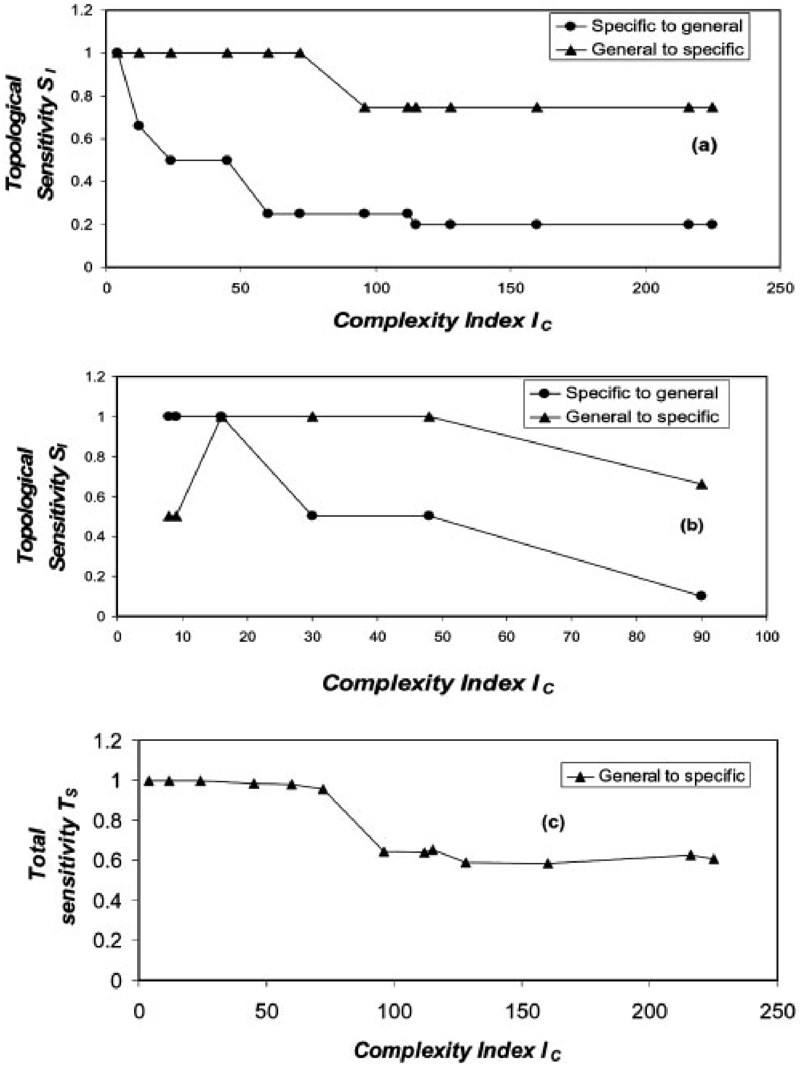


Figure 4. Variation of topological sensitivity with complexity index I_C for (a) unimolecular reactions and (b) bimolecular reactions. In both (a) and (b), the sensitivity decreases as the complexity index increases, and in both cases the sensitivity of the general to specific approach is greater than that of the specific to general approach. (c) Variation of total sensitivity with complexity index I_C for general to specific approach for unimolecular reactions.

Writing $S \equiv X_1, E \equiv X_2, C \equiv X_3$ and $P \equiv X_4$, the corresponding rate equations for all the four species are as follows:

$$\frac{dX_1}{dt} = -k_1X_1X_2 \tag{14a}$$

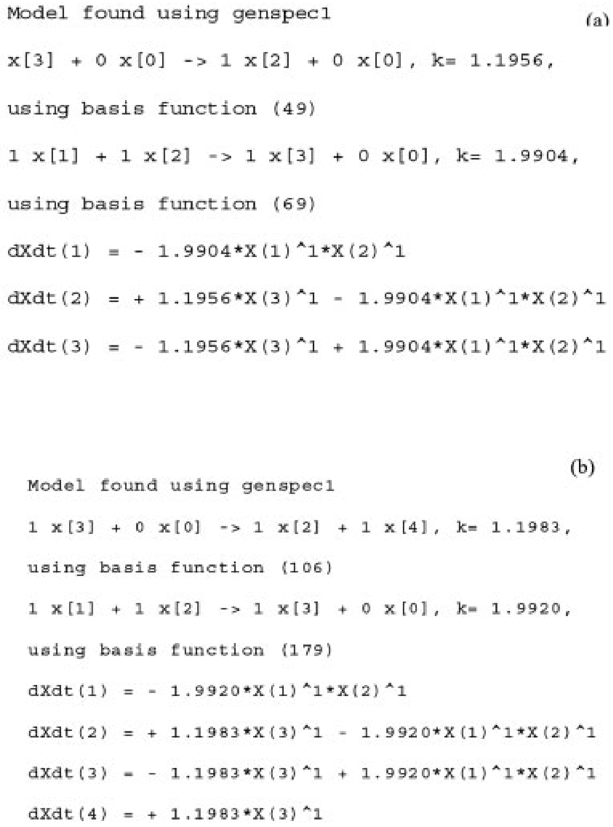
$$\frac{dX_2}{dt} = -k_1X_1X_2+k_2X_3 \tag{14b}$$

$$\frac{dX_3}{dt} = k_1X_1X_2 - k_2X_3 \tag{14c}$$

$$\frac{dX_4}{dt} = k_2X_3 \tag{14d}$$

The complexity index of the above reaction scheme (Eqs. (12) and (13)) is 48. A time series was generated for the above set of reactions with initial conditions {1, 0.1, 0, 0} for *E*, *S*, *C* and *P*, respectively, with $k_1 = 2$ and $k_2 = 1.2$. Initially time series for the three species X_1, X_2, and X_3 were used. The algorithm generated a dictionary of 86 basis functions for these three species, including uni- and bimolecular terms. The basis functions selected by the general to specific algorithm are listed in Fig. 5a. The algorithm predicts two basis functions which correspond to three differential rate equations. Comparison of the predicted differential rate equations, Fig. 5a, with Eqs. (14a–c) shows that the predicted model coincides with the generative model, giving a topological sensitivity index of 1. The inferred rate coefficients also closely match the generative model, with total sensitivity index of 0.988.

Figure 5b gives the model output when time series data for all four species X_1, X_2, X_3, and X_4 were given to the algorithm. The model generated 248 basis functions as a complete dictionary for uni-and bimolecular interactions between four species. From the output we see that the model is



```
Model found using genspec1

x[3] + 0 x[0] -> 1 x[2] + 0 x[0], k= 1.1956,

using basis function (49)

1 x[1] + 1 x[2] -> 1 x[3] + 0 x[0], k= 1.9904,

using basis function (69)

dXdt(1) = - 1.9904*X(1)^1*X(2)^1

dXdt(2) = + 1.1956*X(3)^1 - 1.9904*X(1)^1*X(2)^1

dXdt(3) = - 1.1956*X(3)^1 + 1.9904*X(1)^1*X(2)^1
```

```
Model found using genspec1

1 x[3] + 0 x[0] -> 1 x[2] + 1 x[4], k= 1.1983,

using basis function (106)

1 x[1] + 1 x[2] -> 1 x[3] + 0 x[0], k= 1.9920,

using basis function (179)

dXdt(1) = - 1.9920*X(1)^1*X(2)^1

dXdt(2) = + 1.1983*X(3)^1 - 1.9920*X(1)^1*X(2)^1

dXdt(3) = - 1.1983*X(3)^1 + 1.9920*X(1)^1*X(2)^1

dXdt(4) = + 1.1983*X(3)^1
```

Figure 5. Model output for time series from Eqs. (14a) to (14d) with $k_1 = 2.0$ and $k_2 = 1.2$ using data for (a) X_1, X_2, and X_3, and (b) X_1, X_2, X_3, and X_4. In both cases, the predicted model and coefficients are close to the real ones.

able to predict the standard enzyme kinetics scheme with rate parameters close to the generative values. The algorithm was tested with time series generated with different time steps varying from 0.01 to 0.5 and with different initial conditions. The topological and total sensitivities did not change significantly for these variations.

3.2.2 Example 2

The glycolytic pathway in *L. lactis* has been explored experimentally in great detail [19, 37]. Time series data from these experiments have been used recently by Voit *et al.*, [38] for testing a reconstruction pathway methodology based on the power law approximation. We have used the data obtained from ^{13}C NMR experiments for our analysis [19]. These data comprise 54 time points, measured for each component over a period of 108.9 min, at time intervals of 2.2 min.

The glycolytic pathway involves the conversion of glucose to pyruvate, and comprises eight reaction steps, illustrated in Fig. 6a. In the first step, glucose is converted into glucose-6-phosphate (G6P) (X_2). Phosphoenolpyruvate (PEP) (X_5) also contributes to this step. G6P is converted into fructose-1,6-bisphosphate (FBP) (X_3), then sequentially to glyceraldehyde-3-phosphate (Ga3P) (X), 3-phosphoglyceric acid (3-PGA) (X_4) and PEP (X_5). Glucose and G6P, along with PEP, are involved in the conversion of PEP to pyruvate (X_6). This step is activated by a positive feedback from FBP, which also exerts a positive feedback on the conversion of pyruvate to lactate (X_7) [38]. In Fig. 6a, reactions are indicated with solid arrows, and the feedback interactions are indicated with dotted arrows along with the signs (+) or (–) to indicate positive or a negative feedback. Time series data were unavailable for two of the intermediate components, namely the Ga3P (X) and dihydroxyacetonephosphate (DHAP) (X'). Since we are interested in determining the pathway structure, we use a subset of the complete dictionary of basis functions which are confined to $n_iX_i \rightarrow n_jX_j$ reactions only.

Figure 6b gives the topology of the pathway predicted by our method from the data. The components inside the dotted circle were not available as inputs. Our method predicted a reaction step linking the components FBP (X_3) and 3-PGA (X_4). Figure 6b shows that the algorithm has correctly predicted most of the reaction steps. The network described in Fig. 6b is written down in differential rate equation form in Fig. 7. Firstly, we see from the differential rate equations that the basic linear skeleton of the pathway starting from $X_1 \rightarrow .... \rightarrow X_7$ is clearly predicted. Also several regulatory interactions, which have been drawn into the predicted topology, are also identified (Fig. 6b).

Our method has predicted that G6P (X_2) is produced independently by glucose and PEP (X_5), *i.e.* $2X_5 \rightarrow X_2$ and $X_1 \rightarrow X_2$. However, in reality, glucose (X_1) and PEP (X_5) are involved in the production of G6P (X_2). Similarly in the predicted model, pyruvate (X_6) is produced by reaction steps involving glucose and G6P ($X_{1,}$ X_2) and separately from FBP (X_3). In reality, pyruvate (X_6) is produced by the interaction of glucose (X_1) and G6P (X_2) and also independently from PEP (X_5). The conversion of PEP into pyruvate ($X_5 \rightarrow X_6$) is activated by FBP (X_3). Even though our basis functions do not represent feedback reactions, the predicted model indicated the involvement of FBP (X_3) in the synthesis of pyruvate (X_6). This is also the case for the involvement of FBP in the synthesis of lactate (X_7). The method predicts an additional reaction $X_1 \rightarrow X_3$ which is not originally seen in the glycolytic pathway [29].

4 Discussion

We have presented a new approach for identifying biochemical reaction mechanisms from time series data based on global nonlinear modeling. We have demonstrated that our method can give information on pathway connectivity and chemical reaction steps using simulated data and data measured on the glycolytic pathway of *L. lactis*.

Biological interactions are confined to follow the laws of chemistry. We used this information to construct the basis functions, elementary reactions from which a model will be reconstructed, based on the principle of mass action (Sec-

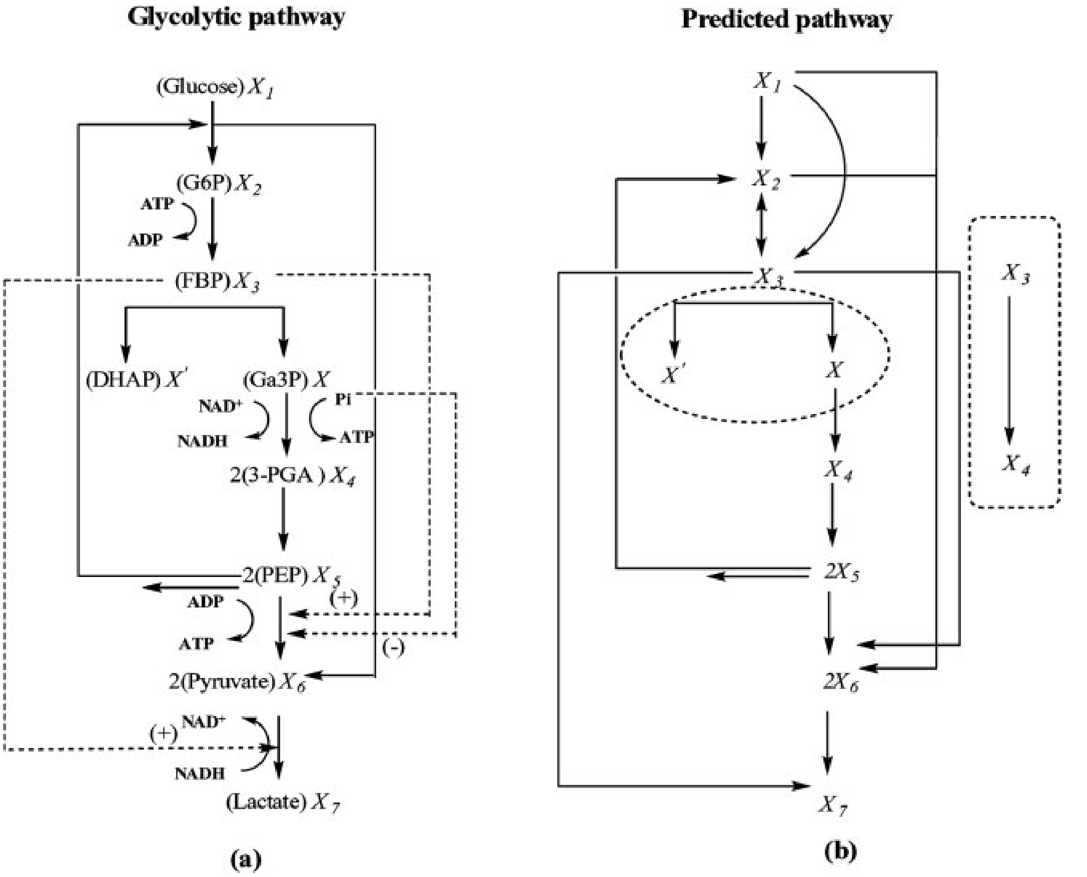
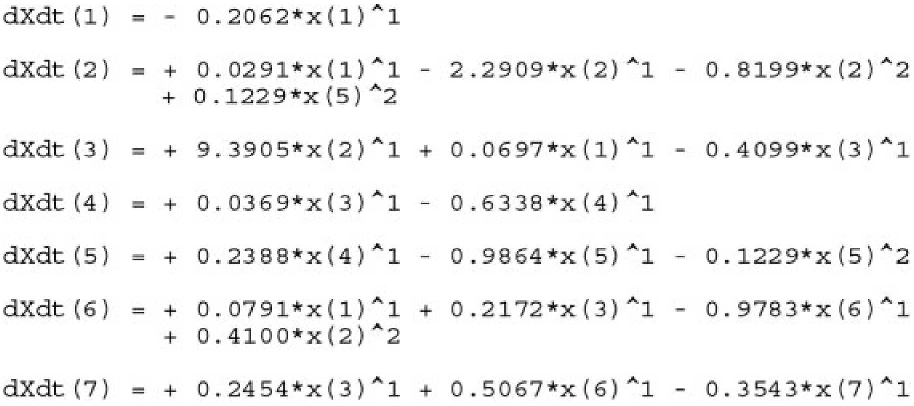


Figure 6. Glycolytic pathway topology: (a) A simplified topology of the glycolytic pathway of the *L. lactis* as given by Hoefnagel *et al.* [37]. Solid lines indicate reaction steps and dotted lines indicate regulatory influences: activation (positive sign) and inhibition (negative sign); (b) Predicted topology by our method using the time series data available for seven of the species (labeled X_1–X_7). Time series data for those components inside the dotted circle were not available; the dotted rectangle shows the predicted reaction step by our method.

tion 2.1). The set of basis functions, which provides a description of all feasible chemical interactions between the set of species, is a key aspect of the method. This approach rules out the identification of chemically impossible combinations. This chemical reaction dictionary can be made as comprehensive as required, for example using unimolecular and bimolecular interactions as we have done here. While bimolecular (or higher) terms are nonlinear in the reactant concentrations, the model selection method is linear in the coefficients to be determined. This dramatically simplifies the determination of model coefficients from the data to a linear optimization problem. The iterative approach to model selection allows the algorithm to determine the appropriate model size – the number of basis functions necessary to model the data. This iterative scheme suggests both specific to general and general to specific approaches to the analysis of the data. We have found that the general to specific method out-performs the specific to general approach, which appears to suffer from a divergent approach in minimizing the cost function. This may result in the identification of local minima, instead of attaining a global minimum [39].

We have introduced the use of a sensitivity-based analysis for model verification. This gives insight into the applicability of our method over a variety of chemical reactions based on their complexities. Topological sensitivity measures the correctness of the connectivity of the inferred pathway, and total sensitivity measures the accuracy of the inferred model as a quantitative model of the pathway kinetics. For unimolecular reactions, the topological sensitivity (Fig. 4a) and the total sensitivity (Fig. 4c) are almost the same for the general to specific approach; however, a small decrease is seen in the corresponding total sensitivities. This corresponds to the errors in the inferred parameters, and not in the topological sensitivity. The algorithm is more efficient in identifying the mechanism than in identification of the parameters. Further refinements are underway to improve the efficiency of the method.

The data used in Example 1 were simulated using a mathematical model of the reaction. The mechanism was inferred using a single data set consisting of time series of four species. This demonstrates that our method is efficient in deducing the reaction mechanism of the enzyme kinetics



```
dXdt(1)  =  -  0.2062*x(1)^1

dXdt(2)  =  +  0.0291*x(1)^1  -  2.2909*x(2)^1  -  0.8199*x(2)^2
            +  0.1229*x(5)^2

dXdt(3)  =  +  9.3905*x(2)^1  +  0.0697*x(1)^1  -  0.4099*x(3)^1

dXdt(4)  =  +  0.0369*x(3)^1  -  0.6338*x(4)^1

dXdt(5)  =  +  0.2388*x(4)^1  -  0.9864*x(5)^1  -  0.1229*x(5)^2

dXdt(6)  =  +  0.0791*x(1)^1  +  0.2172*x(3)^1  -  0.9783*x(6)^1
            +  0.4100*x(2)^2

dXdt(7)  =  +  0.2454*x(3)^1  +  0.5067*x(6)^1  -  0.3543*x(7)^1
```

Figure 7. Differential rate equations predicted by our method corresponding to the predicted network described in Fig. 6b. Note that the algorithm has predicted the direct conversion of Glucose (X_1) to FBP (X_3) which does not appear in the original pathway.

with a single set of time series data. Traditional methods used by enzymologists to distinguish reaction mechanisms, typically need steady state kinetics data, or data from a variety of perturbations in order to decipher the mechanism [40]. The time series data in Example 2 are experimentally measured data, which prove to be an ideal candidate to test the efficiency of the method. The set of time series data of seven metabolites in the glycolytic pathway was used as inputs to the algorithm. A subset of the complete dictionary of reaction steps was used to obtain a picture of the reaction pathway topology. Even with the use of this restricted subset, the method predicted a good amount of information about the pathway. This would be valuable information when analyzing time series data on biochemical pathways with little prior knowledge.

Our method is significantly different from the already available S-systems method [27] for inferring reaction mechanisms from time series data. S-systems do not follow mass action and the reactants can have fractional powers [41]. This empirical approach has been argued to provide a good alternative for modeling reactions in nonideal environments [18, 42]. Using this approach, parameters that are equivalent to rate constants as well as fractional exponents of the reactants need to be optimized, and these exponents naturally appear nonlinearly, making parameter estimation potentially a much more challenging problem. Kikuchi *et al.* [26] also use the S-systems approach but with a modified genetic algorithm to optimize several parameters at once. We have presented arguments as to which rate laws are applicable in intracellular metabolic reactions elsewhere [43, 44]; here we use the law of mass action to describe elementary reaction steps to give sufficient information about the system. Our methodology involves the estimation of the rate constants alone, since it follows from the law of mass action that the exponents are fixed according to the elementary chemical steps, and are not unknown parameters to be determined. Instead, we perform a model selection to identify which elementary reaction steps to include in our model of the data. In this method, fewer parameters are therefore required to be estimated, and furthermore there is no need for an initial guess of the values of the rate parameters as they are estimated using linear regression.

The computational time taken in the inference process for these examples is minimal. However, there are computational limitations to this approach. Assuming that sufficient data points can be collected, this approach can be used for data sets with large numbers of chemical species. However, the number of basis functions generated increases significantly with the number of species, and therefore, so does the computation time. Another limitation of our method is the choice of basis functions. It is possible to include trimolecular interactions and other types of basis functions in the model design matrix. Further refinement of the method applicable to complex mechanisms (such as Michaelis–Menten kinetics and Hill functions) requires the use of nonlinear optimization techniques for parameter estimation. We are currently investigating this direction.

We have developed a new approach to infer reaction mechanisms and pathway connectivity from biochemical time series data. We tested our method with several types of chemical interaction and pathway data, and used a complexity index to determine the sensitivity of our approach for different pathways. We showed that the topological sensitivity for inferred pathways is high for complex mechanisms. We demonstrated this by testing our method with a real experimental data on the glycolytic pathway.

S. S. would like to acknowledge support from National Science Foundation (USA) Grant IIS-0513701. E. J. C. acknowledges support from the University of Auckland Research Committee grant 3606318, the New Zealand Institute for Mathematics and its Applications (NZIMA) and the Maurice Wilkins Centre for Molecular Biodiscovery, University of Auckland.

P. E. M. acknowledges the support of the Royal Academy of Engineering, the Engineering and Physical Sciences Research Council and the European Union's Sixth Framework Programme. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the above funding agencies.

5 References

[1] Needleman, S. B., Wunsch, C. D., *J. Mol. Biol.* 1970, *48*, 443–453.

[2] Smith, T. F., Waterman, M. S., *J. Mol. Biol.* 1981, *147*, 195–197.

[3] Hutchinson, E. G., Thornton, J. M., *Protein Sci.* 1996, *5*, 212–220.

[4] Kosak, S. T., Groudine, M., *Science* 2004, *306*, 644–647.

[5] Enright, A. J., Iliopoulos, I., Kyrpides, N. C., Ouzounis, C. A., *Nature* 1999, *402*, 86–90.

[6] Ng, S. K., McLachlan, G. J., Wang, K., Ben-Tovim Jones, L., Ng, S. W., *Bioinformatics* 2006, *22*, 1745–1752.

[7] Crampin, E. J., Schnell, S., *Prog. Biophys. Mol. Biol.* 2004, *86*, 1–4.

[8] Fresht, A. R., *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*, W. H. Freeman and Co., New York 1999.

[9] Sontag, E. D., Kiyatkin, A., Kholodenko, B. N., *Bioinformatics* 2004, *20*, 1877–1886.

[10] Eiswirth, M., Freund, A., Ross, J., *Adv. Chem. Phys.* 1991, *80*, 127–199.

[11] Chevalier, T., Schreiber, I., Ross, J., *J. Phys. Chem.* 1993, *97*, 6776–6787.

[12] Vance, W., Arkin, A., Ross, J., *Proc. Natl. Acad. Sci. USA* 2001, *99*, 5816–5821.

[13] Arkin, A., Ross, J., *J. Phys. Chem.* 1995, *99*, 970–979.

[14] Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D., *Proc. Natl. Acad. Sci. USA* 1998, *95*, 14863–14868.

[15] Gardner, T. S., di Bernardo, D., Lorenz, D., Collins, J. J., *Science* 2003, *301*, 102–105.

[16] Wildenhain, J., Crampin, E. J., *IEE Proc. Sys. Biol.* 2006, *153*, 247–256.

[17] Friedman, N., *Science* 2004, *303*, 799–805.

[18] Crampin, E. J., Schnell, S., McSharry, P. E., *Prog. Biophys. Mol. Biol.* 2004, *286*, 77–112.

[19] Neves, A. R., Ventura, R., Mansour, N., Shearman, C. *et al.*, *J. Biol. Chem.* 2002, *277*, 28088–28098.

[20] Szyperski, T., *Quart. Rev. Biophys.* 1998, *31*, 41–106.

[21] Gerner, C., Vejda, S., Gelbmann, D., Bayer, E. *et al.*, *Mol. Cell Proteomics* 2002, *1*, 528–537.

[22] McKenzie, J. A., Strauss, P. R., *Anal. Biochem.* 2003, *313*, 9–16.

[23] Alizadeh, A. A., Ross, D. T., Perou, C. M., van de Rijn, M., *J. Pathol.* 2001, *195*, 41–52.

[24] Goodenowe, D., in: Goodacre, R., Harrigan, G. G. (Ed.), *Metabolomic Analysis with Fourier Transform Ion Cyclotron Resonance Mass Spectrometry*, Kluwer Academic Publishing, Dordrecht, The Netherlands 2003, pp. 125–139.

[25] Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L. *et al.*, *Mol. Cell* 1998, *2*, 65–73.

[26] Kikuchi, S., Tominaga, D., Arita, M., Takahashi, K., Tomita, M., *Bioinformatics* 2003, *19*, 643–650.

[27] Marino, S., Voit, E. O., *J. Bioinform. Comp. Biol.* 2006, *4*, 665–691.

[28] Moles, C. G., Mendes, P., Banga, J. R., *Genome Res.* 2003, *13*, 2467–2474.

[29] Ronen, M., Rosenberg, R., Shraiman, B., Alon, U., *Proc. Natl. Acad. Sci. USA* 2002, *99*, 10555–10560.

[30] Crampin, E. J., McSharry, P. E., Schnell, S., *Lec. Notes Artif. Intell.* 2004, *3214*, 329–336.

[31] Judd, K., Mees, A., *Physica D* 1995, *82*, 426–444.

[32] Akaike, H., *IEEE Trans. Automat. Contr.* 1974, *19*, 716–723.

[33] Schwarz, G., *Annl. Stat.* 1978, *6*, 461–464.

[34] McSharry, P. E., Ellepola, J. H., von Hardenberg, J., Smith, L. A. *et al.*, *Intl. J. Heat Mass Transfer* 2002, *45*, 237–253.

[35] Temkin, O. N., Zeigarnik, A. V., Bonchev, D. G., *J. Chem. Inf. Comput. Sci.* 1995, *35*, 729–737.

[36] Schnell, S., Maini, P. K., *Comments Theor. Biol.* 2003, *8*, 169–187.

[37] Hoefnagel, M. H. N., Hugenholtz, J., Snoep, J. L., *Mol. Biol. Rep.* 2002, *29*, 157–161.

[38] Voit, E. O., Almeida, J., Marino, S., Lall, R. *et al.*, *IEE Proc. Sys. Biol.* 2006, *153*, 286–298.

[39] Hendry, D. F., Krolzig, H. M., in: Stigum, B. P. (Ed.), *Econometrics and Philosophy of Economics*, Princeton University press, Princeton 2003, pp. 379–422.

[40] Schnell, S., Chappell, M. J., Evans, N. D., Roussel, M. R., *C. R. Biol.* 2006, *329*, 51–61.

[41] Savageau, M. A., *J. Theor. Biol.* 1995, *176*, 115–124.

[42] Grima, R., Schnell, S., *Chem. Phys. Chem.* 2006, *7*, 1422–1424.

[43] Grima, R., Schnell, S., *Biophys. Chem.* 2006, *124*, 1–10.

[44] Schnell, S., Turner, T. E., *Prog. Biophys. Mol. Biol.* 2004, *85*, 235–260.