

# Reconstructing biochemical pathways

## from time course data

J. Srividhya<sup>1</sup>, Santiago Schnell<sup>1\*</sup>, Edmund J. Crampin<sup>2</sup> and Patrick E.  
McSharry<sup>3,4,5</sup>

<sup>1</sup>*Indiana University School of Informatics and Biocomplexity Institute  
1900 East Tenth Street, Bloomington, IN 47406, USA*

<sup>2</sup>*Bioengineering Institute, The University of Auckland,  
Private Bag 92019, Auckland, New Zealand*

<sup>3</sup>*Department of Engineering Science, Parks Road, Oxford OX1 3PJ, UK*

<sup>4</sup>*Oxford Centre for Industrial and Applied Mathematics, Mathematical Institute, 24-29 St  
Giles', Oxford, OX1 3LB, UK*

<sup>5</sup>*Centre for the Analysis of Time Series, London School of Economics, London WC2A  
2AE, UK*

---

**\*Corresponding author:** Indiana University School of Informatics and Biocomplexity Institute,  
1900 East Tenth Street – Eigenmann Hall 906, Bloomington, Indiana 47406, USA. Tel: +1-812-  
856 1833 Fax: +1-812-856 1995.

E-mail: [schnell@indiana.edu](mailto:schnell@indiana.edu) (Santiago Schnell)

## Abstract

---

Time series data of biochemical reactions reveals transient behavior, away from chemical equilibrium, and contains information on the dynamic interactions between 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 law of mass action. Using the 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 *Lactococcus lactis* from time course experimental data.

---

## Keywords

time series data, biochemical kinetics, general to specific, specific to general, model selection, global non-linear models

## Abbreviations

G6P – Glucose-6-phosphate, FBP – Fructose-1,6-bisphosphate, DHAP – Dihydroxyacetone phosphate, Ga3P- Glyceraldehyde-3-phosphate, 3-PGA – 3-phosphoglycericacid, PEP – Phospho-enol-pyruvate

## **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 a parallel development of the computational approaches available 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, identification of common structural motifs [1], gene order [2], gene fusion events [3] and correlated gene expression profiles [4]. 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 towards the development of mathematical and computational methods to predict functions based on the role of genes and proteins in networks. The current advances in high-throughput measurement technologies, combined with high performance computing, make possible the application of such methods to determine reaction pathways and kinetics from experimental data on a system-wide scale.

Experimental tools are available which provide powerful strategies for identifying the structure of metabolic and proteomic networks. Such tools include nuclear magnetic resonance (NMR) [5, 6], mass spectrometry (MS), time resolved fluorescence spectroscopy, fluorescence labeling combined with autoradiography on 2-D gels [7], protein kinase phosphorylation [8] and tissue arrays [9] for simultaneous high throughput analysis of proteins in a tissue section by means of antibody binding and MS. What is common between these techniques is that they allow the simultaneous measurement of the abundance of multiple metabolites or proteins, either at one time point [10] or as a sequence of measurements giving time series data. Time series data 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.

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. 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 [11, 12]. 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 [13, 14]. A qualitative form of impulse response analysis has also been proposed to gather information on the connectivity of a biochemical network [15]. Correlation based approaches to identifying networks have been increasingly useful in analyzing gene networks from microarray experiments [16, 17]. For a comprehensive review of methods currently available for reconstructing the reaction mechanism from time course data we invite the reader to consult Crampin et al. [18].

Determining reaction mechanisms and parameterizing models for the kinetics of those reaction steps requires 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 estimation of kinetic parameters [19]. However, for less well characterized chemical components, or for more complicated networks, this approach is not practicable. One possibility 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. The most significant feature of our method [20] 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 (known as basis functions) and applies a model selection technique to deduce the reaction mechanism from the data. Model selection is approached by two routes: namely the top-down (specific to general) approach and the bottom-up (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 by some examples in Section 3.

## 2 Methodology

A kinetic model for the biochemical pathway

$$\frac{dx_i}{dt} = F_i(\mathbf{x}, \mathbf{a}_i) \quad (1)$$

provides a description for the rate of production of each species in terms of the concentrations  $\mathbf{x} = \mathbf{x}(t)$ . The net production rate of each species  $F_i$  can be expressed as a weighted sum of  $K$  basis functions,  $\Phi_j$ , representing contributions from different elementary processes

$$F_i(\mathbf{x}, \mathbf{a}_i) = \sum_{j=1}^K a_{ij} \Phi_j(\mathbf{x}, \mathbf{b}) \quad (2)$$

If the basis functions are kept fixed, and only the weights  $a_{ij}$  are varied, the model may be fitted to the data using least squares and singular value decomposition. The solution for the model parameters  $\mathbf{a}_i = \{a_{ij}\}_{j=1}^K$  is achieved by minimizing the sum of squared residuals

$$\chi_i^2 = \|\mathbf{y}_i - \Phi \cdot \mathbf{a}_i\|^2 \quad (3)$$

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

For a noisy data set, this approach alone will tend to over fit the data. For basis functions which represent elementary reactions between species, only a subset of these potential reactions is expected to be required to model the time series data. However, singular value decomposition will return nonzero coefficients for most if not all of them due to noise in the data. Our approach is to construct models using  $q$  of the total set of  $K$  basis functions, for  $q=1,\dots,K$ , using an iterative method proposed by Judd and Mees [21]. They showed how the best model using  $q+1$  basis functions can be selected from the best model using  $q$  basis functions, and vice versa. Model determination is done by the two approaches, namely the specific to general method starting with 1 basis function, and the general to specific method, starting from a model using all available basis functions. For both methods, the appropriate model size can then be selected by minimizing a cost function that penalizes use of more basis functions without sufficient payoff in reducing the model residuals. Formally this is achieved using a cost function or a penalty term called as information criterion (IC). After considerable experimentation with different penalty terms from various information criteria like Akaike [22], Bayesian [23] etc, the following empirical IC was found to provide the most reliable results. The cost function to be minimized over the model size  $q$  is as follows,

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

where  $\mathbf{E}^{(q)} = \mathbf{y} - \Phi \cdot \mathbf{a}$  are the model residuals.

In Figure 1 we outline our method schematically, showing the modules are integrated to arrive at the reaction mechanism. The sequence of steps involved in the method is (a) construction of the model design matrix, (b) construction of the derivative matrix, (c) model selection module and the (d) ordinary differential equation (ODE) reconstructor. The final output comprises the possible reaction steps and the reconstructed ODE model for the pathway. In the following sections, the above steps are discussed systematically.

## 2.1 Construction of the model design matrix

The key aspect of our method lies in the construction of a model design matrix  $\Phi_{ij}$  appropriate for biochemical pathways. The model design matrix  $\Phi_{ij}$  is a velocity matrix of all possible reaction steps involving all the species. 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



Here  $\lambda$  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

$$v = \lambda (x_A)^{n^A} (x_B)^{n^B} = \lambda \phi(x_A, x_B) \quad (6)$$

where  $x_A$  and  $x_B$  are the concentrations of species  $A$  and  $B$ . The function  $\phi$ , the reaction velocity without the rate constant  $\lambda$ , will form the basis function for this elementary reaction. The rates of change of the species are given as follows

$$v(t) = \frac{-1}{n^A} \frac{dx_A}{dt} = \frac{-1}{n^B} \frac{dx_B}{dt} = \frac{1}{n^C} \frac{dx_C}{dt} = \frac{1}{n^D} \frac{dx_D}{dt} \quad (7)$$

This general framework can be used to construct a chemically feasible set of elementary reactions if we restrict the reactions to a maximum molecularity. For example, for two species the general elementary reaction (5) can produce 18 chemically realistic schemes (18 choices of the integers  $n^A, n^B, n^C$  and  $n^D$ ) up to and including bimolecular reactions, as shown in Figure 2. In the figure the indices given in square brackets label the species and zero indexes imply absence of any 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  $\Phi_{ij}^k = \sigma_i^k n_i^k \phi_i(t_j)$  where  $n_i^k$  is molecularity of species  $k$  in the  $i^{\text{th}}$  reaction and the sign of the element  $\sigma_i^k$  is positive if  $k$  is a product and negative if  $k$  is a reactant for the  $i^{\text{th}}$  reaction.  $\phi_i(t_j)$  is the unscaled velocity of the  $i^{\text{th}}$  reaction at the  $j^{\text{th}}$  time point as described in Eq. (6), and is a function of the concentrations of the reactants alone. The velocities are evaluated from each point in the time course and the model design matrix  $\Phi_{ij}^k$  is constructed for each species

$$\Phi^k = \begin{pmatrix} \sigma_1^k n_1^k \phi_1(t_1) & \sigma_2^k n_2^k \phi_2(t_1) & \dots & \sigma_M^k n_M^k \phi_M(t_1) \\ \sigma_1^k n_1^k \phi_1(t_2) & \sigma_2^k n_2^k \phi_2(t_2) & \dots & \sigma_M^k n_M^k \phi_M(t_2) \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_1^k n_1^k \phi_1(t_N) & \sigma_2^k n_2^k \phi_2(t_N) & \dots & \sigma_M^k n_M^k \phi_M(t_N) \end{pmatrix}$$

Only for the minority of 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 all the species, thereby resulting in a matrix of size  $M \times N$  where  $N$  is the number of time points in the time series and  $M$  is the number of species.

Selecting different sets of possible reactions will, in turn, alter the model determined by the algorithm from the data. A complete dictionary is a comprehensive description of all chemically feasible elementary reactions. If, however, one is interested only in the connectivities of a pathway, then a subset of this complete dictionary can be used. For example, the subset of the complete dictionary which consists of single species interconversions alone is particularly useful for identifying a 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 a global non-linear method to deduce the mechanism.

## 2.2 Construction of the derivative matrix

Time course data on the biochemical pathway is used to calculate the derivative matrix. For the time series  $x(t_j)$  derivative vector  $\mathbf{y}$  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***

An exhaustive search across all models using  $k$  from  $K$  basis functions would be computationally expensive. An alternative approach was proposed by Judd and Mees [21], who use an iterative scheme to increment model size, identifying the ‘best’ model using that number of basis functions. A sensitivity analysis determines the basis function to add that will most improve the model fit to the data, and the basis function to remove that will least damage the approximation. These can be used iteratively to select the best model of size  $k$ . McSharry et al. [24] extended this iterative data-driven approach to extract a set of non-orthogonal empirical functions (NEFs) from multivariate datasets. 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. [20] 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 model size is achieved by determining  $\mu = -\Phi^T E^K$ , the projection of the vector of the residuals on to the model design matrix. The largest positive element in  $\mu$  is selected as the first basis function and subsequently basis functions are added, subject to the minimization of the cost function. 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 square solution of  $\mathbf{y}$  and  $\Phi$  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 reconstructor

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

## 3 Method verification

The efficiency of the above method has been tested using an approach based on calculating the *sensitivity* of the model inference. The sensitivity here accounts for two aspects namely (a) the correctness of the reaction structure and (b) the correctness of the parameters. The former can give topological information about the mechanism and the later can yield a measure of the parameters. We therefore define a topological sensitivity index  $S_I$  [25] of the inferred model as

$$S_I = \frac{T_C}{T_C + T_F + T_U} \quad (8)$$

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.

We calculated the sensitivity index for a wide range of chemical reactions using simulated time series data. In order to facilitate comparison of performance across a

range of reactions we quantified the complexity of the chemical reactions used. Here we categorize the elementary chemical reactions based on an approach from graph theory [26]. 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 *bipartite graph*. From the bipartite graph for the chemical reaction the complexity index is evaluated. A typical bipartite graph for a simple reaction network  $X_1 \longrightarrow X_2 \longrightarrow X_3$  is shown in Figure 3.

The complexity index  $I_C$  can then be calculated using the formula

$$I_C = m z \sum_{i=1}^m T_i \quad (9)$$

Here  $m$  is the number of elementary chemical reaction steps in the mechanism,  $z$  is the number of species in the mechanism and  $T_i$  is the total number of branches emanating from and ending in each reaction point (dark circles). This index was calculated for a wide range of chemical reaction mechanisms for which the corresponding sensitivities for the model selection methods described above were evaluated. Table (1) lists some of the reaction sets along with their complexity indices.

Figures 4a and 4b show the performance of the algorithms in terms of topological sensitivity indices for increasing reaction complexity, for unimolecular and bimolecular reactions respectively. The different colors indicate the sensitivities for the different approaches. In Figure 4b, for bimolecular reactions, we see that the sensitivity decreases with complexity for the specific to general approach. However, that is not so with the general to specific approach. The sensitivity is fairly high for complex reactions. It should be remembered that the above discussed sensitivity is a measure of the inferred reaction topological only. *Total* sensitivity  $T_S$  is a measure of both the topological accuracy and parametric accuracy and can be defined as follows

$$T_S = S_I - e_p \quad (10)$$

$$e_p = \frac{1}{\beta} \sum_{i=1}^M \left( \sum_j (a_{ij} - a_{ij}^*)^2 \right) \quad (11)$$

where  $e_p$  is the error associated with the parametric estimations. Here  $\beta$  is the total number of non-zero terms in  $F(x_i, \mathbf{a}_i)$  and  $a_{ij}$  are the coefficients identified by the model

and  $a_{ij}^*$  are the true coefficients used to generate the time series. Since the applicability of the specific to general approach is limited, we show the evaluated total sensitivities for the general to specific approach only. A plot of the total sensitivity with the complexity index for unimolecular reactions with the general to specific approach is given in Figure 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 of the ways 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 described by Eqs. (8) and (10) are adequate to support the validity of the method.

#### 4 Examples

Two examples highlight the features of our approach are presented below. The first example is a typical enzyme kinetics reaction to illustrate the use of the algorithm in predicting the reaction mechanism using 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.

##### *Example 1*

The following reaction mechanism is a typical example [27] containing a bimolecular reaction step



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

$$\frac{dX_1}{dt} = -k_1 X_1 X_2 \quad \frac{dX_2}{dt} = -k_1 X_1 X_2 + k_2 X_3 \quad (14a, 14b)$$

$$\frac{dX_3}{dt} = k_1 X_1 X_2 - k_2 X_3 \quad \frac{dX_4}{dt} = k_2 X_3 \quad (14c, 14d)$$

The complexity index of the above reaction scheme (Eq. (12)-(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 bi-molecular terms. The final basis functions selected by the bottom-up algorithm are given in Figure 5a. The algorithm predicts two basis functions which correspond to three differential rate equations. We see that the predicted models coincide very well with the generative model by comparing the differential rate equations in Figure 5a and Eqs. (14a)-(14c). The rate coefficients predicted also coincide very well with the model, resulting in a topological sensitivity index of 1 (Figure 4b for bottom up approach) and total sensitivity index of 0.988 (Figure 4c).

Figure 5b gives the model output when time series data for  $X_1, X_2, X_3$  and  $X_4$  were given to the algorithm. The model generated 248 basis functions as a complete dictionary for uni- and bi-molecular interactions between four species. From the output we see that the model is able to predict the standard enzyme kinetics scheme with rate parameters close to the real ones. 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.

### Example 2

Time series data for the glycolytic pathway of *Lactococcus lactis* has been explored in detail in experiments [5, 28]. These data have been used recently by Voit and co-workers [29] for testing a reconstruction pathway methodology based on the power law approximation. The pathway essentially involves the conversion of glucose to pyruvate. The pathway consists of eight reaction steps, illustrated in Figure 6a.

In the first step, glucose is converted into G6P ( $X_2$ ). PEP ( $X_5$ ) also contributes for the production of G6P ( $X_2$ ) along with glucose. G6P ( $X_2$ ) is then converted into FBP ( $X_3$ ) and then sequentially converted into Ga3P ( $X$ ), 3-PGA ( $X_4$ ) and PEP ( $X_5$ ). Glucose ( $X_1$ ) and G6P ( $X_2$ ) along with PEP ( $X_5$ ) is involved in the conversion of PEP

to pyruvate ( $X_6$ ). This step is activated by a positive feedback from FBP ( $X_3$ ). Further FBP ( $X_3$ ) also exerts a positive feedback on the conversion of pyruvate ( $X_6$ ) to lactate ( $X_7$ ) [29]. In Figure 6a, all the reactions are indicated with solid arrows, and the feedbacks are indicated with dotted arrows along with the signs (+) or (-) to indicate positive or a negative feedback. We can see that the complexity of this pathway is far higher than that represented in Figure 4.

We designate the components from  $X_1$  to  $X_7$ , as above. Time series data were unavailable for two of the intermediate components, namely the Ga3P ( $X$ ) and DHAP ( $X'$ ). Since we are interested in determining the pathway structure, here we use a subset of the complete dictionary of basis functions which are confined to  $n_i X_i \rightarrow n_j X_j$  reactions only.

Figure 6b gives the topology of the predicted pathway. 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 ( $X_4$ ). From Figure 6b we see that our method has correctly predicted most of the reaction steps. The network described in Figure 6b is written down in differential rate equation form in Figure 7. Firstly we see from the differential rate equations that the basic linear skeleton of the pathway starting from  $X_1 \rightarrow \dots \rightarrow X_7$  is clearly predicted. Also several side reactions, which have been drawn into the predicted topology, are also identified (Figure 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 our predicted model, we see the involvement of glucose, G6P ( $X_1, X_2$ ) and FBP ( $X_3$ ) in the production of pyruvate ( $X_6$ ). In reality pyruvate ( $X_6$ ) is produced by the interaction of glucose ( $X_1$ ) and G6P ( $X_2$ ) and also independently from PEP ( $X_5$ ). Also conversion of PEP into pyruvate ( $X_5 \rightarrow X_6$ ) is activated by FBP ( $X_3$ ). Even though our basis function does not represent feedback reactions, our method does show an indication of possible involvement of this component FBP ( $X_3$ ) in the synthesis

of pyruvate ( $X_6$ ). The same is the case with the involvement of FBP ( $X_3$ ) in the synthesis of lactate ( $X_7$ ). Our method also predicts an additional reaction  $X_1 \rightarrow X_3$  which is not originally seen in the glycolytic pathway.

## 5 Discussion

We have presented a new approach for identifying biochemical reaction mechanisms from time series data based on global non-linear 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 *Lactococcus lactis*. Biological interactions are confined to follow the laws of chemistry. We used this information to construct the model design matrix based on the principle of mass action (Section 2.1). The model design matrix, which provides a description of all feasible chemical interactions between the set of species, is the key aspect of the method. This approach rules out the identification of chemically impossible combinations. The 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 global non-linear method for model selection is linear in the coefficients to be determined. This dramatically simplifies the problem from a non-linear inference to a linear optimization problem. The iterative approach to model selection suggests both specific to general and general to specific approaches to analysis of the data. We have found that the general to specific method out-performs the specific to general approach. Specific to general appears to suffer from a divergent approach in minimizing the cost function, which could result in local minima instead of attaining a global minimizer [30].

We have introduced the use of a sensitivity-based analysis for model verification. This gives us insight into the applicability of our method over a variety of chemical reactions based on their complexity. The two hierarchical levels of sensitivities we introduced, namely total sensitivity and topological sensitivity, show two measures of algorithm performance, both in terms of the correct ‘connectivity’ of the inferred pathway, and the accuracy of the inferred model as a quantitative model of the pathway kinetics. For unimolecular reactions the topological sensitivity (Figure 4a) and the total

sensitivity (Figure 4c) were 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 was 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 with just one set of time series data, while traditional methods typically need steady state data or data from a variety of perturbations in order to decipher the mechanism. The data used in Example 2 is experimentally measured data, which proves to be an ideal candidate to test the efficiency of the method. A set of time series data of seven metabolites 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.

The time involved in the inference process for these examples is minimal. However, there are computational limitations to this approach. The number of basis functions generated increases significantly for larger sets of chemical species and so therefore does the computation time. With a larger number of basis functions, it is possible to find more than one possible interaction set for the same data set. Under this circumstance the set of basis functions can be further restricted to a subset of the complete dictionary. 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 non-linear 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.

### **Acknowledgements**

SS would like to acknowledge support from National Science Foundation (USA) Grant IIS-0513701. EJC acknowledges the University of Auckland research committee grant 3606318. PEM 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.

## References

- [1] Hutchinson, E. G., Thornton, J. M., *Protein Sci.* 1996, 5,212-220.
- [2] Kosak, S. T., Groudine, M., *Science* 2004, 306,644-647.
- [3] Enright, A. J., Iliopoulos, I., Kyrpides, N. C., Ouzounis, C. A., *Nature* 1999, 402(6757),86-90.
- [4] Ng, S. K., McLachlan, G. J., Wang, K., Ben-Tovim Jones, L. *et al.*, *Bioinformatics* 2006,doi:10.1093/bioinformatics/btl165
- [5] Neves, A. R., Ventura, R., Mansour, N., Shearman, C. *et al.*, *J. Biol. Chem.* 2002, 277,28088-28098.
- [6] Szyperski, T., *Quart. Rev. Biophys.* 1998, 31, 41-106.
- [7] Gerner, C., Vejda, S., Gelbmann, D., Bayer, E. *et al.*, *Mol. Cell Proteomics* 2002, 1,528-537.
- [8] McKenzie, J. A., Strauss, P. R., *Anal. Biochem.* 2003, 313, 9-16.
- [9] Alizadeh, A. A., Ross, D.T., Perou, C.M., van de Rijn, M., *J. Pathol.* 2001, 195, 41-52.
- [10] Goodenowe, D., in: Goodacre, R., Harrigan, G. G., (Eds.), *Metabolomic analysis with Fourier transform ion cyclotron resonance mass spectrometry.*, Kluwer Academic Publishing, Dordrecht, The Netherlands 2003, pp. 125-139.
- [11] Fresht, A. R., *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding.*, New York: W. H. Freeman and Co. 1999.
- [12] Sontag, E. D., Kiyatkin, A., Kholodenko, B.N., *Bioinformatics* 2004, 20, 1877-1886.
- [13] Eiswirth, M., Freund, A., Ross, J., *Adv. Chem. Phys.* 1991, 80,127-199.
- [14] Chevalier, T., Schreiber, I., Ross, J., *J. Phys. Chem.* 1993, 97, 6776-6787.
- [15] Vance, W., Arkin, A., Ross, J., *Proc. Natl. Acad. Sci. USA* 2001, 99, 5816-5821.
- [16] Arkin, A., Ross, J., *J. Phys. Chem.* 1995, 99,970-979.
- [17] Eisen, M. B., Spellman, P.T., Brown, P.O., Botstein, D., *Proc. Natl. Acad. Sci. USA* 1998, 95, 14863-14868.

- [18] Crampin, E. J., Schnell, S., McSharry, P. E., *Prog. Biophys. Mol. Biol.* 2004, 286,77-112.
- [19] Moles, C. G., Mendes, P., Banga, J. R., *Genome Res.* 2003, 13,2467-2474.
- [20] Crampin, E. J., McSharry, P. E., Schnell, S., *Lecture Notes in Artificial Intelligence* 2004, 3214, 329-336.
- [21] Judd, K., Mees, A., *Physica D* 1995, 82,426-444.
- [22] Akaike, *IEEE Trans. Automat. Contr.* 1974, 19,716-723.
- [23] Schwarz, G., *Annals of Statistics* 1978, 6,461-464.
- [24] McSharry, P. E., Ellepola, J. H., von Hardenberg, J., Smith, L. A. *et al.*, *Intl. J. Heat Mass Transfer* 2002, 45,237-253.
- [25] Wildenhain, J., Crampin, E. J., *IEE Proc. Systems Biology* 2006, (*in press*).
- [26] Temkin, O. N., Zeigarnik, A. V., Bonchev, D. G., *J. Chem. Inf. Comput. Sci.* 1995, 35,729-737.
- [27] Schnell, S., Maini, P. K., *Comments Theor. Biol.* 2003, 8,169-187.
- [28] Hoefnagel, M. H. N., Hugenholtz, J., Snoep, J. L., *Mol. Biol. Rep.* 2002, 29,157-161.
- [29] Voit, E. O., Almeida, J., Marino, S., Lall, R. *et al.*, *Systems Biology* 2006, (*in press*).
- [30] Hendry, D. F., Krolzig, H. M., in: Stigum, B. P., (Eds.), *Econometrics and Philosophy of Economics*, Princeton University press, Princeton, 2003, pp. 379-422.

## Figure Legends

**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.

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

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

**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 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. Total sensitivity decreases as the complexity index crosses 60.

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

**Figure 6** Glycolytic pathway topology : (a) A simplified topology of the glycolytic pathway of the *Lactococcus lactis* as given by Hoefnagel et al [28] (b) Predicted topology by our method y the input of seven time series. The predicted topology seems to be fairly similar to the parent pathway. The time series of those components inside the dotted circle was not available for input. The reaction step in the dotted rectangle shows the predicted reaction by our method.

**Figure 7** Differential rate equations predicted by our method which corresponds to the predicted network described in Figure 6a. Note that our method has predicted the conversion of Glucose ( $X_1$ ) to FBP( $X_3$ ) which does not seem to appear in the original pathway.

Figure 1:

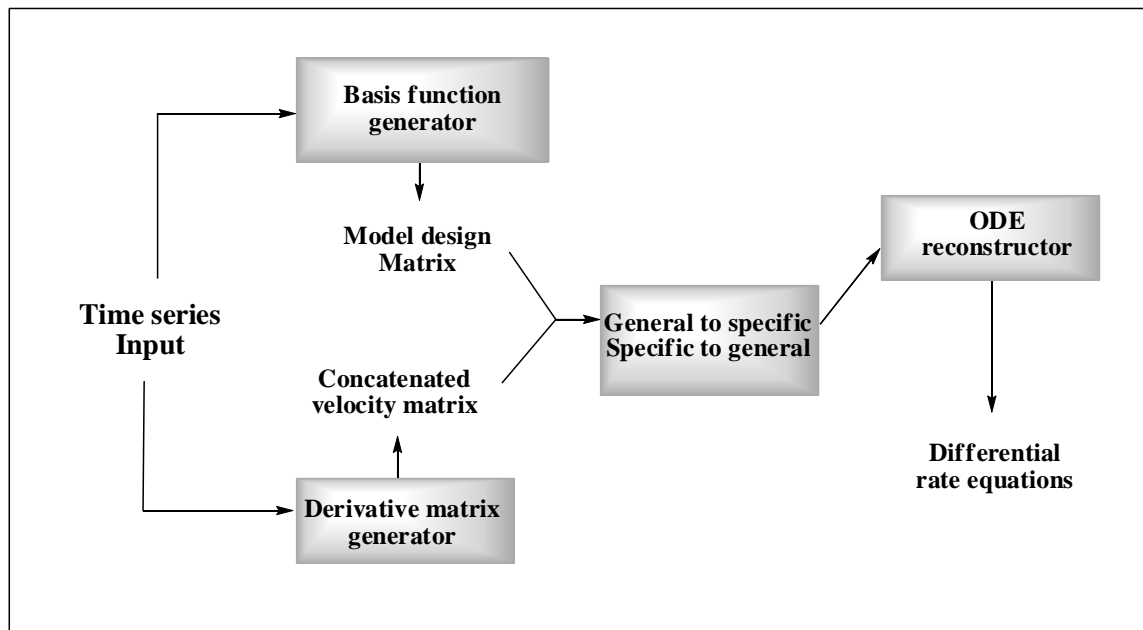


Figure 2:

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

Figure 3:

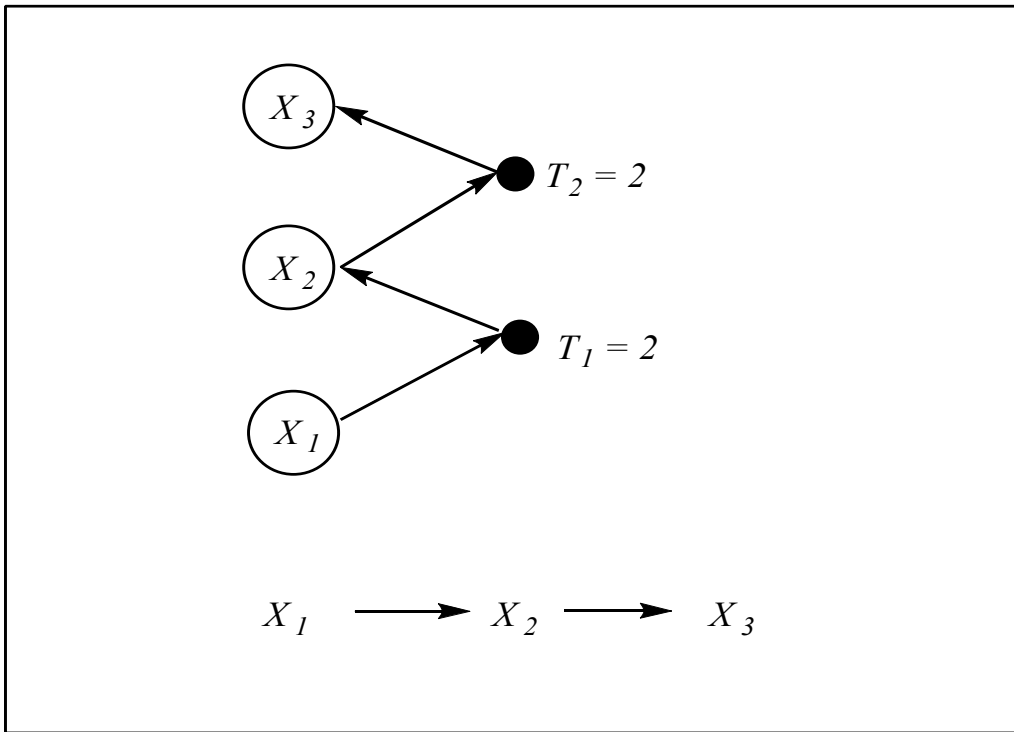


Figure 4 a,b,c:

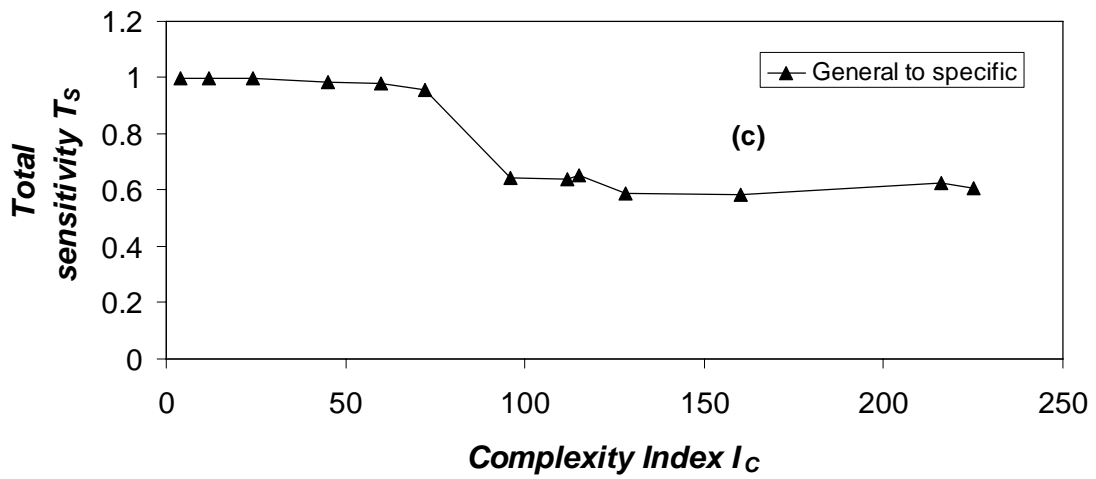
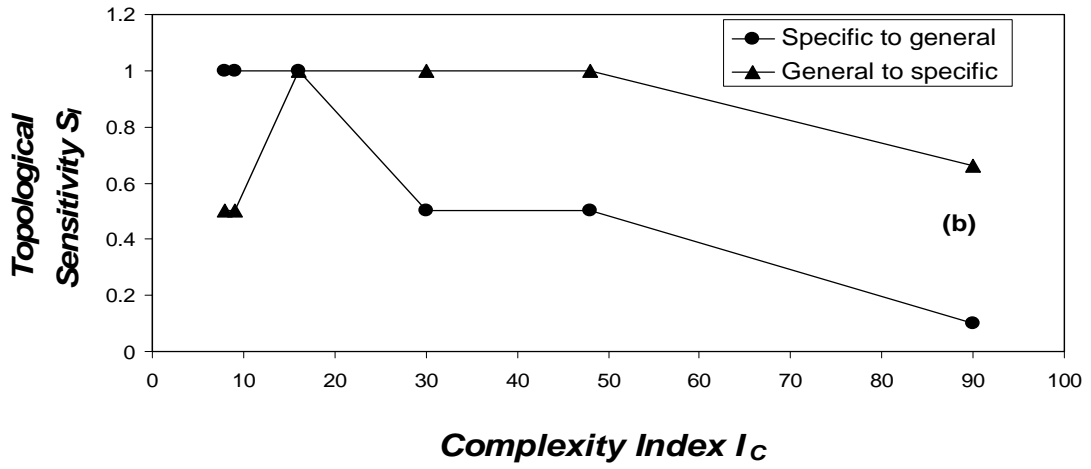
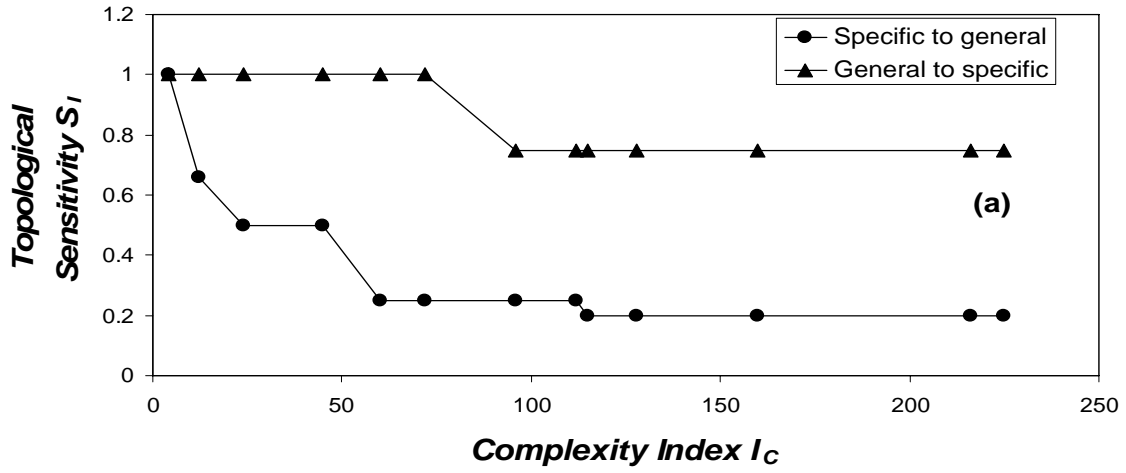


Figure 5 a,b:

Model found using genspec1 (a)

$x[3] + 0 x[0] \rightarrow 1 x[2] + 0 x[0], k= 1.1956,$   
using basis function (49)

$1 x[1] + 1 x[2] \rightarrow 1 x[3] + 0 x[0], k= 1.9904,$   
using basis function (69)

$\text{dXdt}(1) = - 1.9904 * X(1)^1 * X(2)^1$

$\text{dXdt}(2) = + 1.1956 * X(3)^1 - 1.9904 * X(1)^1 * X(2)^1$

$\text{dXdt}(3) = - 1.1956 * X(3)^1 + 1.9904 * X(1)^1 * X(2)^1$

Model found using genspec1 (b)

$1 x[3] + 0 x[0] \rightarrow 1 x[2] + 1 x[4], k= 1.1983,$   
using basis function (106)

$1 x[1] + 1 x[2] \rightarrow 1 x[3] + 0 x[0], k= 1.9920,$   
using basis function (179)

$\text{dXdt}(1) = - 1.9920 * X(1)^1 * X(2)^1$

$\text{dXdt}(2) = + 1.1983 * X(3)^1 - 1.9920 * X(1)^1 * X(2)^1$

$\text{dXdt}(3) = - 1.1983 * X(3)^1 + 1.9920 * X(1)^1 * X(2)^1$

$\text{dXdt}(4) = + 1.1983 * X(3)^1$

Figure 6:

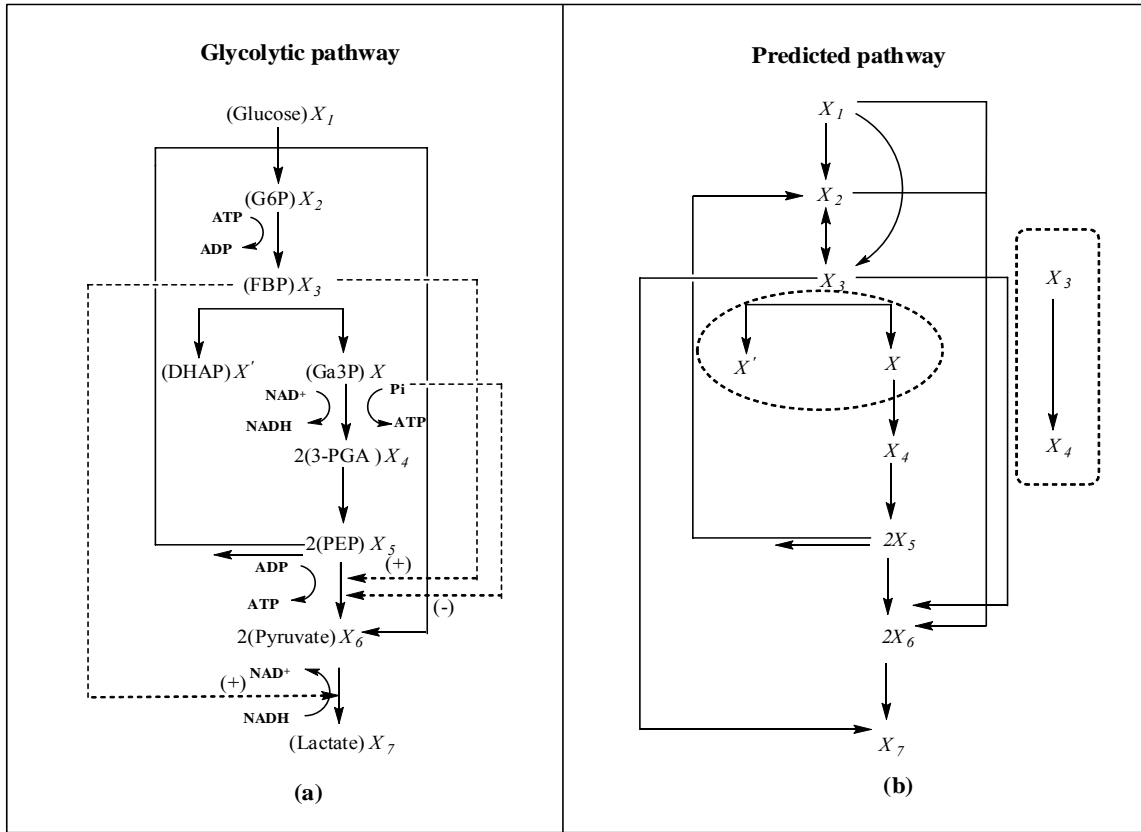


Figure 7:

$$dxdt(1) = - 0.2062*x(1)^1$$

$$dxdt(2) = + 0.0291*x(1)^1 - 2.2909*x(2)^1 - 0.8199*x(2)^2 \\ + 0.1229*x(5)^2$$

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

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

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

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

$$dxdt(7) = + 0.2454*x(3)^1 + 0.5067*x(6)^1 - 0.3543*x(7)^1$$

**Table 1 Complexity index  $I_C$  for some unimolecular and bimolecular reactions calculated based on 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**

Unimolecular reactions	$m$	$z$	$T_i$	$I_C = m z \sum_{i=1}^m T_i$
$X_1 \rightarrow X_2$	1	2	3	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
$\rightarrow X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow$ $\quad \uparrow$ $\quad \rightarrow X_4$	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	$m$	$z$	$T_i$	$I_C = m z \sum_{i=1}^m T_i$
$X_1 + X_2 \rightarrow$	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