Extracting Biochemical Reaction Kinetics from Time Series Data

Edmund J. Crampin[1⋆], Patrick E. McSharry[2,3], and Santiago Schnell[4,5]

[1] Bioengineering Institute, The University of Auckland,
Private Bag 92019 Auckland, New Zealand
e.crampin@auckland.ac.nz
[2] Mathematical Institute, 24–29 St Giles', Oxford, OX1 3LB, UK
[3] Department of Engineering Science, University of Oxford,
Parks Road, Oxford, OX1 3PJ, UK
[4] Centre for Mathematical Biology, Mathematical Institute,
24–29 St Giles', Oxford, OX1 3LB, UK
[5] Christ Church, Oxford, OX1 1DP, UK

Abstract. We consider the problem of inferring kinetic mechanisms for biochemical reactions from time series data. Using a priori knowledge about the structure of chemical reaction kinetics we develop global nonlinear models which use elementary reactions as a basis set, and discuss model construction using top-down and bottom-up approaches.

1 Introduction

There is a current shift in the biological sciences from reductive to systematic approaches. High-throughput experimental assays are increasingly common. The data sets generated in these experiments hold the promise of identification of the components and interactions comprising regulatory biochemical networks. At the same time, however, there is a growing requirement for the development of computational approaches suited to the analysis of these data sets, in particular when there is little prior knowledge of the chemical interactions involved.

Determining nonlinear reaction mechanisms directly from time series data seems likely, prima facie, to be a difficult problem given the difficulties encountered for parameter optimisation in biochemical pathway models [1, 2]. Several authors have tackled the simpler problem of determining the Jacobian matrix from the linear response of a chemical system near a steady state (see [3] for a review). While this provides useful information on steady state behaviour and 'connectivity' in a biochemical network, it does not reveal crucial details about the nonlinear dynamics which underlie transient behaviour and oscillations, which are of particular interest in many biological systems. In this paper we discuss an approach to inferring reaction kinetics from time series data using global nonlinear modelling techniques.

⋆ Author for correspondence.

M.Gh. Negoita et al. (Eds.): KES 2004, LNAI 3214, pp. 329–336, 2004.

2 Global Nonlinear Modelling of Time Series Data

Let us assume that time series data of length T are recorded on the concentrations of M chemical species forming a complete reaction pathway. A model $\mathbf{F} = (F_1, \ldots, F_M)$ for the underlying reaction mechanism provides a description for the rate of change of each of the concentrations $\mathbf{x}(t) = \big(x_1(t), x_2(t), \ldots, x_M(t)\big)$, and can be expressed [4] as a sum of K basis functions, ϕ_j,

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = F_i(\mathbf{x}) = \sum_{j=1}^{K} \phi_j(\mathbf{x}, \mathbf{a}_{ij}), \quad i = 1, \ldots, M \ . \tag{1}$$

A data-driven modelling approach is to select a generic form for the basis set. Neural networks are a popular choice, and have been shown to be useful in a number of applications to modelling biochemical pathways [5]. Data-driven modelling can deliver an accurate description of the data, as measured by model prediction errors. There are, however, several issues motivating an alternative approach. A mechanistic interpretation of the model, which is our aim, is rarely possible. Secondly, by imposing the structure of the laws of chemical reactions, so that the resulting model can be interpreted mechanistically, we also constrain the form which the model can take which should help in the model selection process. Thirdly, polynomial models of chemical reactions based on mass action kinetics fall into the category of pseudo-linear basis functions, which are particularly convenient for time series analysis.

2.1 Mathematical Models of Chemical Kinetics

Chemical reaction pathways are composed of a number of elementary reactions, each of which can be represented by

$$a\,\mathrm{A} + b\,\mathrm{B} \xrightarrow{\lambda} c\,\mathrm{C} + d\,\mathrm{D} \tag{2}$$

in which the chemical species A and B react to form species C and D in the proportions given by the integers a, b, c, d [6]. The molecularity of the reaction is determined by a and b, which represent the numbers of molecules of A and B, respectively, which take part in the reaction. Unimolecular reactions occur when a single molecule is transformed into one or more product molecules; bimolecular events involve the collision of two reactant molecules [7]. According to the law of mass action of chemical kinetics [6], the reaction velocity $v(t)$ is proportional to the product of the concentrations of the reactants,

$$v = \lambda\, x_{\mathrm{A}}^{a}\, x_{\mathrm{B}}^{b} \tag{3}$$

where x_{A} and x_{B} represent the concentrations of A and B, and the rate parameter λ is the constant of proportionality. C is produced in the reaction at c-times this reaction velocity, for example. Therefore the rates of production of C and D and removal of A and B, determined from the overall reaction velocity, are

$$v(t) = -\frac{1}{a}\frac{\mathrm{d}x_\mathrm{A}}{\mathrm{d}t} = -\frac{1}{b}\frac{\mathrm{d}x_\mathrm{B}}{\mathrm{d}t} = \frac{1}{c}\frac{\mathrm{d}x_\mathrm{C}}{\mathrm{d}t} = \frac{1}{d}\frac{\mathrm{d}x_\mathrm{D}}{\mathrm{d}t} \quad , \tag{4}$$

which imposes a structural constraint on the kinetics of elementary reactions.

Reaction pathways are typically characterised by chains of elementary reactions, rather than dense networks, and so only a small subset of the total number of possible reactions between biochemical species will be present. Kinetic equations describing the net rate of change of each species for the whole pathway are be found by summing the appropriate velocity terms over the subset of elementary reactions in which each species is involved.

2.2 Pseudo-Linear Models Based on Elementary Reactions

A particularly convenient class of basis functions has the general form [4, 8]

$$F_i(\mathbf{x}, \mathbf{a}_i) = \sum_{j=1}^{K} a_{ij}\phi_j(\mathbf{x}) \quad . \tag{5}$$

The fact that a_{ij} appears linearly greatly simplifies fitting the model to data. The model parameters, $\mathbf{a}_i = \{a_{ij}\}_{j=1}^{K}$, can be determined for each species i by solving the linear system of equations $\mathbf{y}_i = \Phi \cdot \mathbf{a}_i$ in the least squares sense, where $\mathbf{y}_i = \{\mathrm{d}x_i(t_k)/\mathrm{d}t\}_{k=1}^{N}$ and $\Phi_{jl} = \phi_j\big(\mathbf{x}(t_l)\big)$ is the model design matrix. This is achieved by seeking $\mathbf{a}_i$ which minimises $\chi_i^2 = ||\mathbf{y}_i - \Phi \cdot \mathbf{a}_i||^2$. Both χ_i^2 and $||\mathbf{a}_i||$ are minimised by choosing $\mathbf{a}_i = \Phi^+\mathbf{y}_i$, where Φ^+ is the Moore-Penrose pseudo-inverse of Φ [9].

Let us suppose that the reaction system governing the concentrations of the chemical species is represented by a polynomial model structure of order p. A quadratic model, $p = 2$, may be written as

$$F_i(\mathbf{x}, \mathbf{a}_i) = a_i + \sum_{j=1}^{K} b_{ij}x_j + \sum_{j=1}^{K}\sum_{k=1}^{K} c_{ijk}x_j x_k \quad . \tag{6}$$

In the context of chemical reactions, the parameters $\mathbf{a}_i = \{a_i, b_{ij}, c_{ijk}\}$ determine the rate constants for constant flux terms (sources and sinks), linear and quadratic interactions. The model can, therefore, represent all possible unimolecular and bimolecular interactions between the species, and encompasses all possible elementary reaction velocities of the type represented in (3). However, general multivariate polynomial models have limited usefulness for modelling reaction systems, for the following reasons. Firstly, the large number of free parameters can quickly become intractable with large numbers of variables [10]. Secondly, the set of multivariate polynomial basis functions includes a large number of combinations of polynomials which cannot be interpreted in terms of reaction mechanisms (4). For example, a positive quadratic term $c_{ijk}x_j x_k$ in the kinetic equation for the ith chemical species suggests that X_i is produced in a bimolecular reaction between X_j and X_k with rate parameter $\lambda = c_{ijk}/n_i$, where n_i is the number of molecules of species i produced in the bimolecular reaction. In which case, the kinetic equations for the species j and k must also reflect this reaction, according to (4), by including terms $-(c_{ijk}/n_i)x_j x_k$. (Note

that a negative quadratic term in which x_i is not one of the variables could have no chemical meaning in the kinetic equation for X_i.)

This latter problem, at least, can be improved if polynomial basis functions are based instead on elementary reaction steps, as defined by (4). For example, if the jth basis function represents a bimolecular reaction between X_p and X_q

$$X_p + X_q \xrightarrow{\lambda_j} X_r, \quad p, q \in \{1, \ldots, M\}, \quad r \in \{1, \ldots, M; r \neq p, q\} \tag{7}$$

then an appropriate basis function, $\phi_j(x_p, x_q)$, operating on three variables $\{x_p, x_q, x_r\}$, would be the set

$$\left\{ \frac{\mathrm{d}x_p}{\mathrm{d}t} = -x_p x_q, \frac{\mathrm{d}x_q}{\mathrm{d}t} = -x_p x_q, \frac{\mathrm{d}x_r}{\mathrm{d}t} = x_p x_q \right\} \tag{8}$$

and the coefficient $a_j = \lambda_j$ for this basis function is the rate parameter, which must be positive. In this way a set of pseudo-linear basis functions can be built up to represent possible elementary reactions. To implement this approach we concatenate the multivariate problem to a single vector of length $N = M \times T$ of the entire data set. The design matrix can then be constructed in the usual way.

3 Iterative Approaches to Model Selection

We wish to minimise

$$\chi^2 = ||\mathbf{y} - \boldsymbol{\Phi} \cdot \mathbf{a}||^2 \quad \text{subject to} \quad a_j \geq 0, \ \forall j \quad \text{and} \quad \mathcal{N}(\mathbf{a}) = K \tag{9}$$

where $\mathcal{N}(\mathbf{a})$ is the number of nonzero components of $\mathbf{a}$, i.e. the number of basis functions used in the model. In general, increasing the model size K is always likely to marginally improve the prediction errors. Our expectation is that the data are generated by just a small number from amongst the set of possible interactions, and therefore we wish to identify a parsimonious model which is consistent with the data [11]. A least squares calculation for the full complement of possible elementary reactions will tend to use all available basis functions to minimise the residuals which, even for small N, is unlikely to yield useful results. Hence a model with too many parameters will not distinguish between the generative dynamics that we wish to identify and artifacts due to noise, which is known as over-fitting.

The optimal model size can be determined using a maximum likelihood approach by adding a penalty term which favours more concise models. This can be achieved using a cost function based on the Akaike Information Criterion [12] (derived by maximising the log-likelihood functions for a set of models with different numbers of parameters). Assuming independent and normally distributed errors, the maximum likelihood parameter estimates are the least squares estimates, in this case calculated using a non-negative least squares algorithm [13]. If $\mathbf{E}^{(K)} = \mathbf{y} - \boldsymbol{\Phi} \cdot \mathbf{a}$ with $\mathcal{N}(\mathbf{a}) = K$ is the vector of residuals with model size K, the cost function to be minimised over K is

$$C_{\mathrm{AIC}}(K) = \frac{1}{N} \mathbf{E}^{(K)^{\mathrm{T}}} \mathbf{E}^{(K)} + K \ . \tag{10}$$

(An alternative cost function, Rissanen's Minimum Description Length [14] based on minimising the coding length of a model and associated residual errors, could also be used here.)

We are now left with the problem of determining, for each model size K, the optimal subset of K basis functions from the pool of all possible basis functions, so that we can subsequently determine the optimal model size which minimises the cost function $C_{\mathrm{AIC}}(K)$. The non-orthogonality of the basis functions means that the optimal subset of size $K+1$ is not necessarily the optimal subset of size K plus the 'next best term'. The selection process must therefore be iterative. The approach we propose is based on finding criteria for adding and eliminating elementary reactions to sequentially expand or contract the current basis.

3.1 Simple-to-General and General-to-Specific Model Selection

Following Judd and Mees [8], let $\boldsymbol{\mu} = -\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{E}^{(K)}$ be the projection of the vector of residuals onto the model design matrix for the entire set of basis functions $\boldsymbol{\Phi}$. A sensitivity analysis of the minimisation problem (9) shows that the basis function corresponding to the largest positive element in $\boldsymbol{\mu}$ should be added to the current subset of basis functions to give the largest marginal improvement to the mean squared error. Similarly, the basis function which can be eliminated doing least damage to the residuals can be shown to be the one with the smallest coefficient a_j.

This suggests two approaches to model construction. To expand from the optimal model of size K to $K+1$ these two criteria are applied to alternately add then remove a basis function until there is no further change (when the term to be added to the subset is the same as the term to be removed). A 'simple-to-general' algorithm uses this approach to successively expand the model, starting with a single basis function, in order to find the model which minimises $C_{\mathrm{AIC}}(K)$ (for more details of the algorithm see [8, 15]). Alternatively, a 'general-to-specific' approach [16] can be developed where the initial set contains all basis functions and the same criteria are applied alternately to eject then add a basis function until the same basis function is chosen and is removed from the subset, reducing the model size by one (P. E. McSharry, unpublished).

3.2 Example Biochemical Pathways

We demonstrate our approach with two very simple yet realistic types of biochemical networks (which have also been used to test other network identification methods [17]). The first type consists of chains of unimolecular reactions and the second is an enzymatic reaction involving a bimolecular reaction step. In each case, a full set of basis functions was used including source and sink terms, unimolecular and bimolecular reactions.

Unimolecular Reaction Pathways: We tested both model construction approaches on two unimolecular (linear) pathways: one with a reversible step, which has five elementary reactions in the pathway

$$\xrightarrow{\lambda_0} X_1 \underset{\lambda_2}{\overset{\lambda_1}{\rightleftharpoons}} X_2 \xrightarrow{\lambda_3} X_3 \xrightarrow{\lambda_4} \qquad \text{(a)}$$

and one a branched pathway, which has six elementary reactions

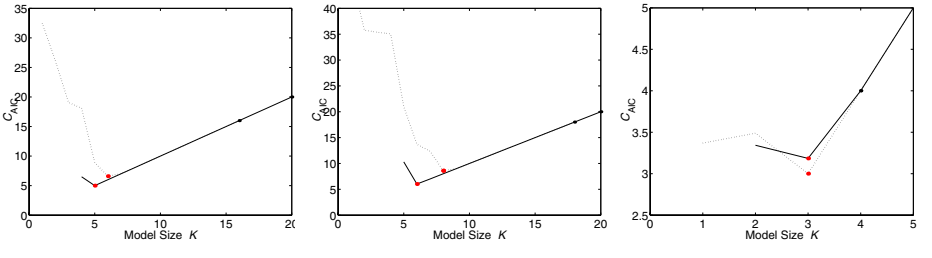
$$\begin{array}{l} \xrightarrow{\lambda_0} X_1 \xrightarrow{\lambda_1} X_2 \xrightarrow{\lambda_2} X_3 \xrightarrow{\lambda_3} \\ \qquad\qquad\quad \lambda_4 \uparrow \\ \qquad\quad \xrightarrow{\lambda_5} X_4 \end{array} \qquad \text{(b)}$$

Enzyme-Substrate Reaction: A simple model for the enzyme-substrate reaction proposed by Fersht [7], which gives similar kinetics to the standard Michaelis-Menten reaction, is given by

$$X_1 + X_2 \xrightarrow{\lambda_1} X_3 \xrightarrow{\lambda_2} X_2\ (+X_4) \qquad \text{(c)}$$

where the enzyme X_2 is recovered from the enzyme-substrate complex X_3 so that the overall reaction is transformation of substrate X_1 into product X_4.

We ran simulations of the ODE kinetic models for each of these three pathways, collecting around 20 data points on each species, for a single time series 'experiment' which starts from initial conditions perturbed from the steady state. These data were then used to attempt to extract the reaction mechanism using the two algorithms discussed above, which were written in Matlab. The results are shown in Fig. (1), where the cost function is plotted against model size K.



(a) (b) (c)

Fig. 1. Cost function C_{AIC} against model size K for (a) linear pathway (with $\lambda_0 = 1.0, \lambda_1 = 2.0, \lambda_2 = 1.5, \lambda_3 = 0.5$ and $\lambda_4 = 1.0$), (b) branched pathway (with $\lambda_0 = 1.0, \lambda_1 = 2.0, \lambda_2 = 8.0, \lambda_3 = 3.0, \lambda_4 = 4.0$ and $\lambda_5 = 0.5$) and (c) enzyme-substrate bimolecular reaction ($\lambda_1 = 2.0, \lambda_2 = 1.0$); showing general-to-specific (solid line) and simple-to-general (dotted line) approaches

4 Discussion

As can be seen from the Figure, the general-to-specific algorithm out performs the simple-to-general approach. For pathways (a) and (b) the general-to-specific algorithm identified the appropriate model size and the correct set of elementary reactions in each case, while the simple-to-general approach found models minimising the cost function which were not identical to the generative elementary reactions. For the substrate-enzyme reaction both approaches added an extra elementary reaction which was not part of the generative kinetics, but both nevertheless captured the main features of the pathway. Hendry and Krolzig [16] have pointed out that the simple-to-general approach to model building is a divergent branching process, and likely to be susceptible to finding routes to local minima in the cost function. For this reason it seems natural to favour the general-to-specific approach.

There are, however, some limitations to this approach. The implementation of non-negative least squares, to ensure positive coefficients, appears to be very slow for larger sets of chemical species ($M \sim 10$) where there is a very large number of possible elementary reactions. It is therefore desirable to limit the size of the basis set using whatever a priori knowledge of the reaction pathway may be available. The requirement for time series data on every species, and the restriction to elementary reactions of the form described above also limit applicability. This latter restriction can be relaxed, to include Michaelis-Menten-type expressions for reaction rates for example, at the expense of the pseudo-linearity of the model. A crucial component of the model building process is experimental design. Data on the response of the system to many different perturbations, which excite different modes of behaviour, are needed to characterise the full 'model space'. While multiple data sets can straightforwardly be incorporated using this framework, the issue of what data should optimally be collected has not yet been addressed.

Biological networks and reaction pathways are in general sparsely connected, and so we may be justified in assuming that parsimonious models of the data are most likely to establish the correct generative mechanisms. As our understanding of biochemical pathways and networks improves we may be able to derive better criteria, based specifically on the currently emerging details on the structure and topology of individual biological pathways themselves.

Acknowledgements

This work was supported by a New Zealand Institute of Mathematics and its Applications (NZIMA) fellowship to EJC, the Wellcome Trust through Research Training Fellowships in Mathematical Biology to EJC and SS and a Royal Academy of Engineering Research Fellowship to PEM, who also acknowledges the support of the Engineering and Physical Sciences Research Council, UK.

References

[1] P. Mendes and D. B. Kell. Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics*, 14:869–883, 1998.

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

[3] E. J. Crampin, S. Schnell, and P. E. McSharry. Mathematical and computational techniques to deduce complex biochemical reaction mechanisms. Prog. Biophys. Mol. Biol. (in press), 2004.

[4] H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis*. Cambridge University Press, Cambridge, 1997.

[5] J. S. Almeida. Predictive non-linear modeling of complex data by artificial neural networks. *Curr. Opin. Biotech.*, 13:72–76, 2002.

[6] P. Érdi and J. Tóth. *Mathematical Models of Chemical Reactions: Theory and Applications of Deterministic and Stochastic Models*. Princeton University Press, Princeton, 1989.

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

[8] K. Judd and A. Mees. On selecting models for nonlinear time series. *Physica D*, 82:426–444, 1995.

[9] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C*. Cambridge University Press, Cambridge, 2nd edition, 1992.

[10] J. D. Farmer and J. J. Sidorowich. Predicting chaotic time series. *Phys. Rev. Lett.*, 59(8):845–848, 1987.

[11] P. E. McSharry and L. A. Smith. Consistent nonlinear dynamics: identifying model inadequacy. *Physica D*, 192:1–22, 2004.

[12] H. Akaike. A new look at the statistical identification model. *IEEE Trans. Automat. Contr.*, 19:716–723, 1974.

[13] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems*. Number 15 in Classics in Applied Mathematics. SIAM, 1995.

[14] J. Rissanen. Consistent order estimates of autoregressive processes by shortest description of data. In O.L.R. Jacobs and et al., editors, *Analysis and Optimisation of Stochastic Systems*. Academic Press, New York, 1980.

[15] P. E. McSharry, J. H. Ellepola, J. von Hardenberg, L. A. Smith, D. B. R. Kenning, and K. Judd. Spatio-temporal analysis of nucleate pool boiling: identification of nucleation sites using non-orthogonal empirical functions. *Int. J. Heat Mass Transfer*, 45:237–253, 2002.

[16] D. F. Hendry and H. M. Krolzig. New developments in automatic general-to-specific modelling. In B. P. Stigum, editor, *Econometrics and the Philosophy of Economics*. Princeton University Press, Princeton, 2003.

[17] W. Vance, A. Arkin, and J. Ross. Determination of causal connectivities of species in reaction networks. *Proc. Natl. Acad. Sci. U. S. A.*, 99:5816–5821, 2002.