Progress in Biophysics & Molecular Biology 86 (2004) 77–112

Progress in Biophysics & Molecular Biology

www.elsevier.com/locate/pbiomolbio

Mathematical and computational techniques to deduce complex biochemical reaction mechanisms

E.J. Crampin[a,b,*], S. Schnell[a,c,1], P.E. McSharry[d,e,f]

[a] *Centre for Mathematical Biology, Mathematical Institute, 24–29 St. Giles', Oxford OX1 3LB, UK*
[b] *University Laboratory of Physiology, Parks Road, Oxford OX1 3PT, UK*
[c] *Christ Church, Oxford OX1 1DP, UK*
[d] *Oxford Centre for Industrial and Applied Mathematics, Mathematical Institute, 24–29 St. Giles', Oxford OX1 3LB, UK*
[e] *Department of Engineering Science, Parks Road, Oxford OX1 3PJ, UK*
[f] *Centre for the Analysis of Time Series, London School of Economics, London WC2A 2AE, UK*

Abstract

Time series data can now be routinely collected for biochemical reaction pathways, and recently, several methods have been proposed to infer reaction mechanisms for metabolic pathways and networks. In this paper we provide a survey of mathematical techniques for determining reaction mechanisms for time series data on the concentration or abundance of different reacting components, with little prior information about the pathways involved.

1. Introduction

In the biological sciences it is becoming increasingly common to collect data in high-throughput experiments on genomic, proteomic, and metabolomic scales. These data hold the promise of identifying the components and interactions which comprise large-scale regulatory biochemical networks. However, the trend towards systematic and comprehensive profiling experiments will mean increasingly large and complicated databases for analysis, demanding the development of computational approaches suited to the analysis of large-scale data sets.

*Corresponding author. Bioengineering Institute, The University of Auckland, Private Bag 92019 Auckland, New Zealand. Tel.: +64-9373-7599x88168; fax: +64-9367-7157.

E-mail address: e.crampin@auckland.ac.nz (E.J. Crampin).

[1] Current address: School of informatics and Biocomplexity Institute, Indiana University, Informatics Building, 901 East 10th Street, Bloomington, IN 47408-3912, USA.

0079-6107/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.pbiomolbio.2004.04.002

In this paper we focus on the need for algorithms and approaches to determining reaction mechanisms from time series data, which we anticipate will (soon) be automatically collected for protein interactions and metabolic pathways and networks. Our aim is to investigate techniques that allow biochemical reaction mechanisms to be inferred from time series data collected on the concentration or abundance of different reacting components of a network, with little prior information about the pathways involved. Mechanistic studies of biochemical reactions are important for several reasons: (i) an improved understanding of the functional role of different molecules can be achieved only with the knowledge of the mechanism of specific reactions and the nature of key intermediates; (ii) the control (or regulation) of different biochemical pathways can best be understood if some hypothesis for the reaction mechanism is available; (iii) kinetic modelling, which forms the basis for understanding reaction kinetics, is based on comprehensive information about the reaction mechanism. Kinetic models allow simulation of complicated pathways, and even whole-cell dynamics, which is proving to be an increasingly important predictive tool in the post-genomic era (Noble, 2002; Crampin et al., 2004).

The data required for kinetic modelling are typically time series data on the response of a biochemical system to different conditions and stimuli. However, deducing reaction mechanisms from time series data remains something of an art, and until recently, there have been few attempts to derive general approaches to this problem. In this paper we discuss several approaches to the determination of complex reaction mechanisms: both in the vicinity of a steady state and, more generally, to identify reaction network dynamics. It is not our aim to provide a survey of nonlinear biochemical kinetics. Several excellent reviews and texts are available (Gray and Scott, 1990; Scott, 1991; Epstein and Pojman, 1998; Murray, 2002, 2003). We focus instead on new computational approaches to discovering reaction mechanisms for biochemical systems, and in particular on the assumptions and methods underlying different techniques, rather than an assessment of their performance on a particular data set.

A full reaction mechanism is the set of elementary steps that specifies how a chemical reaction takes place. Elementary reaction steps are those that cannot be decomposed to reveal reaction intermediates, which might themselves be identified as separate chemical entities on a biochemically relevant timescale. One way to think of a reaction mechanism is as a network of elementary steps that connects the various reactants, intermediates, and products. A comprehensive description of the reaction network therefore determines the number of chemical species and processes, the sequence of their interactions and the rate laws governing the elementary reaction velocities. A description composed solely of elementary reaction steps aims to chart the progress of actual molecular events.

There are many experimental techniques by which data can be collected, with different approaches tied to different theoretical and modelling paradigms. One general principle is that only those reaction steps which take place on a timescale comparable with the measurement of the time series will be ‘visible’ in the data. The implication is that it is not always possible, or indeed desirable, to track each and every elementary reaction and reaction intermediate. For example, multiple substrates may bind to an enzyme to moderate its activity, but we may not be interested in the detailed kinetics of transitions between every possible state of the molecule. In particular, for small molecule or ion binding and covalent modification of proteins (e.g., phosphorylation or protonation states), where it may be difficult to distinguish experimentally between different states, the lumping of multiple states into one chemical variable is often unavoidable. Therefore, mechanistic models of biochemical reactions and networks derived from data will almost certainly be partial representations of the complete set of molecular events.

In some circumstances it is possible to formally derive simplified models from full reaction mechanisms which capture the kinetics under particular restricted conditions, or operating on certain timescales, usually fast–slow (steady-state and rapid equilibrium) approximations (see, e.g., Segel, 1972; Ermentrout, 2001; Smith and Crampin, 2004). In particular, this has been an important step in the development of representations of enzymatic reactions (Segel and Slemrod, 1989; Schnell and Maini, 2003, and outlined below). These model reduction techniques are extremely useful for the development of models of large-scale biochemical processes and networks, and for whole-cell modelling. However, in general such formal techniques may not be applicable, and ad hoc simplification is achieved from a full mechanism by focusing on the most important overall processes and reducing the number of species and steps needed to yield the behaviour of interest. A further limitation of biochemical models, sometimes referred to as the fundamental dogma of chemical kinetics, is that it is not possible to prove that a reaction mechanism is correct. Several different mechanisms may be consistent with the available data, or may even give the same mathematical representation (indistinguishability; see, e.g., Érdi and Tóth, 1989; Epstein and Pojman, 1998). We can only disprove mechanisms by showing inconsistency with data, or with theoretical requirements for a model. Indeed, developing alternative models for different reaction mechanisms to compare to data is one of the most effective ways to disprove a hypothetical reaction mechanism.

The rest of the paper is organised as follows: we begin with a review of modelling frameworks for describing reaction mechanisms and rate laws. Section 2.1 defines the notation for mass action-derived polynomial models that will form the basis for most of the subsequent discussion in this paper. Next, in Section 3, we present several techniques, which aim to determine the properties of a reaction network close to a steady state from time series data collected on the response to small perturbations. In Section 4, we discuss approaches in which network connectivity is inferred directly from time series measurements, and in Section 5, we describe the application of methods from nonlinear time series analysis to biochemical network data. Finally, we close with a discussion of the use of mathematical and computational techniques in deducing biochemical reaction mechanisms.

2. Mathematical representation of chemical kinetics

The goal of a mechanistic study of a biochemical system is to clarify the nature of reaction intermediates and their interactions (how they react with, or transform into, each other) and to determine the rates of these transformations. This can be broken down into two aspects: first, a connectivity or ‘wiring diagram’ is established and second, the individual interactions are assigned appropriate kinetic properties, or rate laws. Therefore, the mathematical models describing complex biochemical reaction mechanisms can be divided in two groups according to their structure: stoichiometric and kinetic models. The former are based on the time invariant characteristics of the reactions and the latter are based on both the stoichiometry and reaction rates.

Stoichiometric models concern the proportions of rates of change in the concentrations of the reacting species. These proportions are the result of the topological structure of the reaction mechanism, indicating which species are linked by reactions. This perspective can lead to important general results about the properties of reactions. Stoichiometric properties do not depend on the mathematical description of the rate laws and are often more easily established than the kinetic parameters of the reactions. However, stoichiometric models have a major drawback: the predictive power is limited due to the lack of regulatory information, which can only be included in the formulation of a kinetic model.

To formulate kinetic models we require knowledge of the reaction stoichiometry and detailed information about the kinetics of a reaction pathway. Consider the following stoichiometric pathway scheme:

$$\begin{tikzcd} \mathrm{S} \arrow[d, "F_s"] & & \mathrm{P} & \\ \mathrm{C}_1 \arrow[r, "F_1"] & \mathrm{C}_2 \arrow[r, "F_2"] & 2\ \mathrm{C}_3 \arrow[r, "F_3"] \arrow[u, "F_p"] & \mathrm{C}_4 \arrow[lll, bend left, "F_4"] \end{tikzcd} \tag{1}$$

In this scheme we are considering a biochemical pathway in which a product P is synthesised from a substrate S through the four intermediates C_i, $i=1,\ldots,4$. The pathway has a negative feedback on C_1 triggered by the end intermediate C_4. Based on the stoichiometry, we can write the mass balance equation for the rate of change of intermediate C_3, for example, as

$$\frac{\mathrm{d}[C_3]}{\mathrm{d}t} = 2F_2 - (F_3 + F_p), \tag{2}$$

where F_i are the reaction rates, or fluxes, and $[C_3]$ is the concentration of the intermediate species. (In this paper we will use upper case notation to represent the names of different chemical species, and lower case letters or brackets $[\cdot]$ to represent concentrations.) Here we are considering that there is no mass flow due to convection or diffusion. The algebraic expression for the reaction rates F_i will depend on the kinetics under consideration. Many biochemists employ phenomenological rate laws derived empirically. In the literature, a number of approaches to deriving kinetic functions are commonly used: mass action kinetics, the Michaelis–Menten type and allosteric kinetics, and the power-law approximation, which we will summarise in the following sections.

2.1. The law of mass action

The behaviour of a homogeneous chemical system can be described by a system of ordinary differential equations (ODEs) obtained from the reaction mechanism by the law of mass action: the rate of any given elementary reaction is proportional to the product of the concentrations of the species reacting in the elementary process (reactants). This so-called law was postulated more than a century ago to describe observations about the rate of chemical reactions and so, it is fair to say, is empirical in its origins, although it has been shown to be consistent with results in non-equilibrium thermodynamics (see, e.g., Keizer, 1987). The proportionality constant, known as the rate constant, depends on the reaction conditions (temperature, solvent, pH, etc.); biochemists generally try to hold the reaction conditions constant to avoid dealing with higher-order complexities arising from variations in these parameters.

Most reactions involve a number of simultaneous elementary steps. The rate of change of the concentration of any given species is then a sum of the rates of change due to the elementary reactions in which that species participates. An arbitrary number, m, of simultaneous elementary steps may be represented in the form

$$\sum_{i=1}^{n} v'_{i,j} X_i \rightarrow \sum_{i=1}^{n} v''_{i,j} X_i, \quad j = 1, \ldots, m, \tag{3}$$

where X_i is the ith of a total of n chemical species. Integers $v'_{i,j}$ and $v''_{i,j}$ are the stoichiometric coefficients for species i appearing as a reactant and as a product, respectively, in the jth reaction step. If no processes other than chemical reactions cause a change in the concentration of the ith species, x_i, then the net rate of production of species i will be given by

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = \sum_{j=1}^{m} F_{i,j} = F_i \quad i = 1, \ldots, n, \tag{4}$$

where $F_{i,j}$ is the rate of change of species i due to reaction j. The law of mass action implies that a reaction rate $F_{i,j}$ for the jth reaction step, given in expression (3), may be defined as

$$F_{i,j} = (v''_{i,j} - v'_{i,j}) k_j \prod_{l=1}^{n} x_l^{v'_{l,j}} \quad i = 1, \ldots, n, \quad j = 1, \ldots, m, \tag{5}$$

where k_j is a rate constant specific to step j.

Thermodynamically, all reactions are in principle reversible (although in some cases the backward reaction rate may be negligibly small). In the case of reversible reactions it is convenient to replace (3) by the equivalent scheme

$$\sum_{i=1}^{n} v'_{i,j} X_i \rightleftharpoons \sum_{i=1}^{n} v''_{i,j} X_i, \quad j = 1, \ldots, m', \tag{6}$$

where m' represents the number of reversible reactions (forward and backward steps are grouped together, thus if m is the total number of uni-directional reaction steps then $m' = m/2$). With this convention, Eq. (5) is replaced by

$$F_{i,j} = (v''_{i,j} - v'_{i,j}) \left[k_j \prod_{l=1}^{n} x_l^{v'_{l,j}} - k_{-j} \prod_{l=1}^{n} x_l^{v''_{l,j}} \right], \quad i = 1, \ldots, n, \quad j = 1, \ldots, m', \tag{7}$$

where k_j and k_{-j} are positive reaction rate constants for the forward and backward reactions of the jth reaction step, respectively. Let us consider an example. Applying the law of mass action, the governing equations of the reaction

$$X_1 + X_1 \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} X_2 \tag{8}$$

are given by

$$\frac{\mathrm{d}x_1}{\mathrm{d}t} = -2k_1 x_1^2 + 2k_{-1} x_2, \tag{9}$$

$$\frac{\mathrm{d}x_2}{\mathrm{d}t} = k_1 x_1^2 - k_{-1} x_2. \tag{10}$$

In this example, the stoichiometric coefficients take integer values, but for reactions occurring in non-ideal solutions, they can take non-integer values (Othmer, 1981). This is part of the basis of the power-law formalism, which is discussed below.

2.1.1. Enzyme kinetics and the law of mass action

Enzyme-catalysed reactions can be described by incorporating all the elementary steps of the enzyme–substrate association–dissociation, isomerisation of intermediates and formation of products. A major problem of this approach is that it produces systems of highly nonlinear differential equations with many kinetic and stoichiometric parameters. Typically these systems are stiff, have multiple timescales and are computationally demanding to solve numerically, also making fitting to data difficult.

The kinetic modelling of enzymatic reactions can be simplified considerably if the overall reaction is studied with the aid of the quasi-steady-state or equilibrium approximations. Consider the simplest enzymatic reaction, in which there is a reversible association between an enzyme E and a substrate S, yielding an intermediate enzyme–substrate complex C which irreversibly breaks down to form a product P:

$$S + E \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} C \overset{k_2}{\rightarrow} E + P. \tag{11}$$

The time evolution of these reactions is obtained by applying the law of mass action to yield the set of coupled nonlinear differential equations

$$\frac{d[S]}{dt} = k_1(-([E_0] - [C])[S] + K_S[C]), \tag{12}$$

$$\frac{d[C]}{dt} = k_1(([E_0] - [C])[S] - K_M[C]), \tag{13}$$

$$\frac{d[P]}{dt} = k_2[C] \tag{14}$$

with the conservation law

$$[E](t) = [E_0] - [C](t) \tag{15}$$

and with initial conditions ([S], [C], [P]) = ([S_0], 0, 0) at time $t = 0$. In this system $K_S = k_{-1}/k_1$ is the equilibrium dissociation constant for the enzyme–substrate complex and $K_M = (k_{-1} + k_2)/k_1$ is known as the Michaelis–Menten constant. Under the quasi-steady-state conditions (Segel, 1988; Schnell and Maini, 2000),

$$[E_0] \ll K_M + [S_0], \tag{16}$$

the substrate and product concentrations in the reaction scheme (11) are well approximated by the Michaelis–Menten equation (Boyde, 1980)

$$\frac{d[P]}{dt} = -\frac{d[S]}{dt} = \frac{v_{max}[S]}{K_M + [S]}, \tag{17}$$

where $v_{max} = k_2[E_0]$ is the maximum reaction velocity. Due to its ubiquity this equation is arguably among the most important in biochemistry. In this approximation the complex concentration is described by an algebraic expression of the form:

$$[C](t) = \frac{[E_0][S](t)}{K_\mathrm{M} + [S](t)}. \tag{18}$$

The successful application of the quasi-steady-state approximation relies on the existence of a separation of timescales between the fast and slow species, in this case the enzyme–substrate complex $[C]$ and the substrate $[S]$, respectively (Segel, 1988; Segel and Slemrod, 1989). Simplification of reaction kinetics by reducing the dimension of the system of equations on the basis of a difference in timescales is an important and commonly used tool (see Ermentrout, 2001, for a general discussion; Smith and Crampin, 2004, for the rapid equilibrium approximation for simplifying kinetic models of membrane pumps and exchangers). The identification of the appropriate timescales and ranges of validity for the simplification of enzyme kinetics has, however, been the subject of some controversy, for the single enzyme–substrate reaction in particular. Despite considerable work, the literature until quite recently provided insight only into the determination of timescales and the validity of the standard quasi-steady-state approximation presented above. In vivo, however, in some situations the amount of enzyme available will be comparable to the amount of substrate, and the above in vitro conditions (16) will not hold. In this case Schnell and Maini (2000) have identified appropriate timescales for dimension reduction and determined the corresponding parameter regime in which the approximation is valid. While the advantage of employing such quasi-steady-state or equilibrium approximations is that they reduce the dimension of the governing differential equation system, and hence the number of parameters to be determined, a disadvantage of this approach is the restriction of the applicability of the reduced model to a restricted region of the parameter domain (for a review, see Schnell and Maini, 2003).

An important aspect of enzyme kinetics in complex biochemical pathways is the effect of inhibitors and activators of the enzyme. For example, enzyme inhibition can be generalised by the scheme (mixed inhibition)

$$\begin{tikzcd}
E + S + I \arrow[r, shift left, "k_1"] \arrow[d, shift left, "k_3"] & ES + I \arrow[l, shift left, "k_{-1}"] \arrow[r, "k_2"] \arrow[d, shift left, "k_4"] & E + I + P \\
EI + S \arrow[u, shift left, "k_{-3}"] \arrow[r, shift left, "k_5"] & ESI \arrow[l, shift left, "k_{-5}"] \arrow[u, shift left, "k_{-4}"]
\end{tikzcd} \tag{19}$$

where I is an inhibitor and E, ES, EI, and ESI, represent the free enzyme, the enzyme–substrate intermediate complex, the enzyme–inhibitor complex, and the enzyme–substrate–inhibitor complex. Equilibrium constants can be defined for the inhibitor and substrate–inhibitor complexes, respectively: $K_I = k_{-3}/k_3$ and $K_{SI} = k_{-4}/k_4$. The reaction type is defined by imposing constraints on this scheme: noncompetitive inhibition requires $K_I = K_{SI}$; and inhibition is said to be uncompetitive or competitive if the formation of EI or ESI, respectively, is excluded. The rate law for mixed inhibition, within the Michaelis–Menten framework, is given by

$$\frac{\mathrm{d}[P]}{\mathrm{d}t} = -\frac{\mathrm{d}[S]}{\mathrm{d}t} = \frac{\tilde{v}_\mathrm{max}[S]}{\tilde{K}_\mathrm{M} + [S]}, \tag{20}$$

where $\tilde{K}_{\mathrm{M}}$ and $\tilde{v}_{\max}$ are referred to as the "apparent" Michaelis–Menten constant and maximum velocity, given by

$$\tilde{v}_{\max} = \frac{v_{\max}}{(1+[I]/K_{SI})} \quad \text{and} \quad \tilde{K}_{\mathrm{M}} = K_{\mathrm{M}}\frac{(1+[I]/K_I)}{(1+[I]/K_{SI})}. \tag{21}$$

This Michaelis–Menten type rate law exhibits saturation at high substrate concentrations, a hallmark behaviour of enzymatic reactions. Empirically, for many reactions the rate of product formation follows sigmoidal kinetics with the substrate concentration. One of the first to appreciate the sigmoidal behaviour of proteins was Hill (1910), studying the binding of oxygen to haemoglobin. He employed an empirical equation to describe the binding mathematically, which is of the form for enzyme–substrate reactions:

$$\frac{\mathrm{d}[P]}{\mathrm{d}t} = -\frac{\mathrm{d}[S]}{\mathrm{d}t} = \frac{v_{\max}[S]^n}{K_{\mathrm{M}}^n + [S]^n}, \tag{22}$$

where n is the Hill coefficient. While an integer Hill coefficient may in some cases have a mechanistic explanation, i.e., the number of substrate molecules that can bind simultaneously to the enzyme, sigmoidal kinetics of this form are often used to fit data with non-integer Hill coefficients. Sigmoidal kinetics can also be obtained with other reaction mechanisms for enzymes where several subunits interact with each other cooperatively, such as the Monod–Wyman–Changeux model. For further details on these, and other enzymatic reactions, the reader is directed towards the books by Segel (1975), Cornish-Bowden (1995) and Fersht (1999).

2.2. The power law approximation: S-systems

Savageau (1969, 1976) proposed the power-law, or "synergistic-system" (S-system) approximation as an alternative approach for modelling reactions following non-ideal kinetics, such as those occurring under molecular crowding within cells (Savageau, 1992). In contrast to Eqs. (4) and (7), the power-law approximation assumes that the rate of change of a state variable is equal to the difference of two products of variables raised to non-integer powers:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = \alpha_i \prod_{j=1}^{n} x_j^{g_{i,j}} - \beta_i \prod_{j=1}^{n} x_j^{h_{i,j}} \quad i = 1,\ldots,n, \tag{23}$$

where the first term represents the net production, and the second term the net removal rate for the ith species. While this expression is based less on physical principles and more on mathematical amenity, it has been motivated by linearising enzyme kinetics rate expressions in terms of the concentrations (Heinrich and Rapoport, 1974; Palsson et al., 1985) or in terms of the reaction parameters (Kedem and Caplan, 1965; Rottenberg, 1973). Following Savageau (1969), if we assume that the net reaction rate F_i for the ith chemical species can be written as a polynomial or rational function of the concentrations $x_j, j=1,\ldots,n$, then taking the logarithmic transform and truncating a Taylor series expansion about an arbitrary point $(x_1^*, x_2^*, \ldots, x_n^*)$ at linear order, we find

$$\ln F_i(\ln(x_1),\ldots,\ln(x_n)) = \ln F_i^* + \sum_{j=1}^{n}\left(\frac{\partial \ln F_i}{\partial \ln x_j}\right)_{\mathrm{x}^*}(\ln x_j - \ln x_j^*) + \text{h.o.t.}, \tag{24}$$

$$F_i(x_1, \ldots, x_n) \approx F_i^* \prod_{j=1}^{n} \frac{\exp\left\{ \left(\dfrac{\partial \ln F_i}{\partial \ln x_j} \right) \ln x_j \right\}}{\exp\left\{ \left(\dfrac{\partial \ln F_i}{\partial \ln x_j} \right) \ln x_j^* \right\}} = F_i^* \prod_{j=1}^{n} \frac{x_j^{g_{i,j}}}{x_j^{*g_{i,j}}} \tag{25}$$

which has the same form as each of the terms in Eq. (23). For example, for the net production term

$$\alpha_i = F_i^* \prod_{j=1}^{n} \frac{1}{x_j^{*g_{i,j}}} \quad \text{and} \quad g_{i,j} = \left(\frac{\partial \ln F_i}{\partial \ln x_j} \right) = \frac{x_j^*}{F_i^*} \left(\frac{\partial F_i}{\partial x_j} \right)_{\mathbf{x}^*}.$$

Although this analysis motivates the form of the power law approach, it should be pointed out that expression (23) cannot be obtained from a general mass action or Michaelis–Menten rate law as the net production and removal terms will not in general be separable under this transformation.

A number of properties of reaction kinetics can be modelled with this approximation, but it fails to describe many important biochemical effects such as saturation and sigmoidicity (Heinrich and Schuster, 1996). In the power-law rate expression (23) the parameters α_i and β_i play the role of the rate constants and the exponents $g_{i,j}$ and $h_{i,j}$ are the kinetic orders. For mass action-derived expressions, the kinetic orders are given by the positive integer stoichiometric coefficients. In the power-law formulation they are phenomenological parameters which may or may not be integer, and can take negative values to describe the effects of inhibitors, which also has the unfortunate effect of introducing singular behaviour into the rate-law expressions. For these reasons, and those outlined above, the application of the power-law approximation seems most appropriate when our aim is to find phenomenological expressions to approximate complex biochemical data, when a detailed mechanistic understanding of the kinetics is not required.

2.3. *Alternative modelling frameworks*

While there is some arbitrariness about the choice of mathematical framework in any modelling effort, different mathematical formulations are better able to capture one or another feature of the data, and represent the underlying hypotheses of a theory more or less faithfully. There is currently an active debate in the biochemical modelling literature as to the appropriate modelling formulation for biochemical network kinetics (Schnell and Turner, 2004). The traditional approach, which encompasses the law of mass action and power-law approximation, described above, has been the basis for kinetic modelling for over a century, in particular in modelling enzyme kinetics. This approach naturally extends to include partial differential equation models for spatially distributed processes, such as reaction–diffusion models (see Roussel and Roussel, this issue), which are derived from conservation equations for the net production and transport processes and to stochastic differential equations to include (extrinsic) noise-driven processes.

At the molecular level, random fluctuations are inevitable when molecules are at low numbers per cell, for example, in the regulation of gene expression where small numbers of regulatory proteins interact with DNA binding sites in the gene's promoter region. These intrinsic noise effects have been recently measured in gene expression using fluorescent probes (see, e.g., Elowitz et al., 2002; Blake et al., 2003). Low copy numbers of expressed RNAs may be significant for the regulation of downstream pathways (McAdams and Arkin, 1997). In this and other cases when there are only small numbers of molecules in the reaction volume a stochastic modelling approach is required (Gillespie, 1977; Morton-Firth and Bray, 1998; Burrage et al., 2004).

There is also growing evidence of the importance for reaction kinetics of the spatial organisation of the intracellular environment, which is far from the homogeneous, well mixed solution typical of the in vitro experiments in which rate laws and parameters are measured. There is a high level of macromolecular crowding within the cell (Medalia et al., 2002), in particular, much higher than in a typical biochemical assay, (Ellis and Minton, 2003), which has been shown to affect the rate of enzymatic reactions both experimentally (Rohwer et al., 1998) and via modelling (Berry, 2002; Schnell and Turner, 2004); however, in what follows we will restrict our discussion to mass action-derived kinetic expressions.

3. Perturbation methods

Whatever framework is chosen for the representation of a biochemical network, the tasks of determining network connectivity and parameterising network interactions are highly difficult problems in themselves. These are not unrelated tasks, and most of the techniques which have been proposed for biochemical network inference try to find both the network interactions and the interaction strengths at the same time.

By contrast, traditional approaches to measuring the rate constants of individual reactions rely on the reaction mechanism being known or, at least, easily guessed. Typically these techniques measure the rate at which equilibrium is approached after initiating the reaction, as in rapid mixing and flash photolysis (photochemical release of caged compound) experiments or, for relaxation methods, in response to a shift of the equilibrium position (Fersht, 1999). The equilibrium constant for a reaction depends on intensive parameters such as temperature and pressure, and so the equilibrium may rapidly be shifted by a step change in the parameter, for example, the temperature jump method, illustrated in Fig. 1. The rate at which the new equilibrium is approached can be used to determine the rate constant at the new temperature.

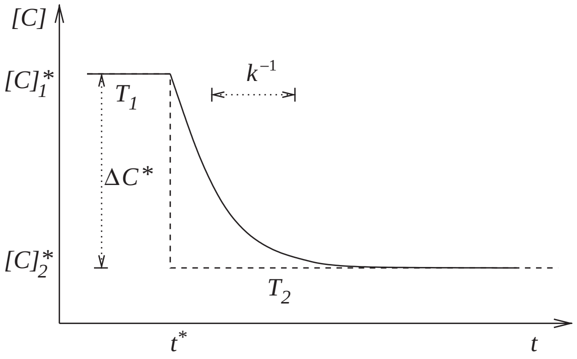

Fig. 1. Temperature jump method: exponential relaxation to new equilibrium concentration $[C]_2^*$ in response to temperature jump at time t^*. The time constant for the relaxation is related to rate constants for the reaction.

For first-order reactions, in which the concentrations appear linearly in the kinetic equations, the biochemical kinetics are exponential in time, characterised by a number of time constants: a single time constant for a single reversible reaction, which is equal to the reciprocal of the sum of the reaction rates; for n first-order reaction steps in series there are in general n time constants each of which are functions of the rate constants (so-called normal modes, in analogy with the mathematically equivalent situation of normal modes of vibration for coupled linear oscillators). These may be resolved if the rate constants for the various steps differ sufficiently. Rate constants for more complicated first-order reaction schemes can also sometimes be deduced, if the reaction mechanism is known beforehand, although in general this is much less straightforward (Bernasconi, 1976; Gutfreund, 1995; Fersht, 1999).

For more general reaction networks, with feedbacks and higher-order nonlinear interactions, the response to an arbitrary perturbation will not necessarily be a simple combination of exponential processes. If the reaction mechanism is not known then the rate constants cannot be deduced even if the kinetics are exclusively first order. However, the theory of ordinary differential equations tells us that if the perturbation made to the system is sufficiently small then the resulting kinetics will again be exponential. Therefore, instead of trying directly to determine rate constants, a great deal of information can be obtained by classifying the linear properties of the network near to equilibrium.

3.1. Linear behaviour near equilibrium

Let us assume a reaction mechanism involving n species is described by the system of equations

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = F_i(x_1, \ldots, x_n; p_1, \ldots, p_{n_p}), \quad i = 1, \ldots, n, \tag{26}$$

where x_i is the concentration of the ith chemical species and F_i is the (in general nonlinear) function describing the production and consumption of x_i. For a stationary state of the reaction

$$F_i(x_1^{\mathrm{s}}, \ldots, x_n^{\mathrm{s}}; p_1, \ldots, p_{n_p}) = 0, \quad i = 1, \ldots, n, \tag{27}$$

where the steady-state concentrations x_i^{s} are functions of the parameters p_j, $j = 1, \ldots, n_p$. Expanding in a Taylor series about the steady state, the reaction kinetics are determined by the system of equations

$$\frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_{j=1}^{n} \left(\frac{\partial F_i}{\partial x_j}\right)_{\mathbf{x}^{\mathrm{s}}} u_j + \mathrm{o}(|\mathbf{u}|), \quad i = 1, \ldots, n, \tag{28}$$

where $u_i(t) = x_i(t) - x_i^{\mathrm{s}}$ is the distance from equilibrium, and near enough to equilibrium $|\mathbf{u}|$ is small so that higher-order terms $\mathrm{o}(|\mathbf{u}|)$ are neglected. A great deal of analysis can now be brought to bear on this linear system, and techniques from linear control theory can be applied to determine the properties for an experimental system near its steady state.

The Jacobian matrix, $J_{ij} = \partial F_i/\partial x_j$ evaluated at the steady-state concentrations, contains all the information needed to classify the type and stability of the steady state and much useful detail about the nature of the reaction mechanism. For example, a positive entry $J_{ij} > 0$ indicates that near steady state an increase in x_j gives rise to production of x_i, whereas negative elements in the matrix indicate that x_i is removed as a direct result of an increase in x_j. A zero element of the matrix indicates that there is no *direct* interaction from x_j to x_i. Thus information on the network connectivity is revealed in the Jacobian. Several attempts have been made to see to what extent reaction mechanisms may be classified from knowledge of the Jacobian alone (Chevalier et al., 1993), in particular for oscillatory reactions where the relationships between the signs of the entries may be used to determine the nature of processes destabilising the steady state (Tyson, 1975) (see Section 3.3).

3.2. Determining the Jacobian matrix

Each of the methods described in this section aims to probe the biochemical network using small-amplitude perturbations to the system, such that the system response is linear and the corresponding Jacobian matrix can be determined. However, there is still considerable choice in how to go about this: in particular the choice of measurement protocol (whether the concentration of every species need be measured, or whether system properties can be determined by measuring only a subset of the constituent species) and perturbation strategy (whether all or just a subset of the species need be perturbed; whether random perturbations are applied at regular intervals or a uniform perturbation applied at irregular time intervals). Indeed, it is possible to determine the properties of the Jacobian using perturbations to system parameters, rather than the concentrations of the species themselves (using a sensitivity analysis approach, described in Section 3.2.2). Finally, some considerations for perturbation analysis of large-scale problems are discussed in Section 3.4, which are particularly relevant for the application to microarray data on gene regulatory networks.

The most direct approach to evaluating the Jacobian matrix for a biochemical network is to experimentally perturb one or more of the concentrations from steady state and monitor the response of each of the chemical species as the system relaxes. From these data it is at least theoretically a straightforward matter to calculate the Jacobian. For the linear system of equations arising from (28) this is essentially an exercise in linear regression

$$\frac{\mathrm{d}u_i}{\mathrm{d}t} = \mathbf{J}_i \cdot \mathbf{u}(t), \tag{29}$$

where the vector $\mathbf{J}_i$ is the ith row of the Jacobian matrix. Note that we must know the steady-state concentrations $\mathbf{x}^{\mathrm{s}}$ to calculate $\mathbf{u} = \mathbf{x} - \mathbf{x}^{\mathrm{s}}$ from the measured concentrations $\mathbf{x}(t)$. The rates of change can be approximated using numerical differencing: $\Delta t\, \mathrm{d}u_i/\mathrm{d}t \approx u_i(t) - u_i(t - \Delta t)$. Then in matrix form, for m time points,

$$\frac{\mathrm{d}}{\mathrm{d}t}[\mathbf{u}_0|\mathbf{u}_1|\cdots|\mathbf{u}_{m-1}] = \mathbf{J} \cdot [\mathbf{u}_0|\mathbf{u}_1|\cdots|\mathbf{u}_{m-1}], \tag{30}$$

where $\mathbf{u}_k \equiv \mathbf{u}(t_k)$ is a column vector of concentrations at time t_k. If the data set contains more data points than there are concentrations, $m > n$, as is usually the case, then this matrix inversion problem is overdetermined. The linear system can then be solved in the least squares sense to give 'best-fit' values for the entries in the Jacobian by seeking parameters $\mathbf{J}_i$, which minimise the sum of the squared residual errors:

$$\chi^2(\mathbf{J}_i) = \sum_{k=1}^{m} \left(\frac{\mathrm{d}u_i}{\mathrm{d}t}(t_k) - \mathbf{J}_i \cdot \mathbf{u}_k \right)^2 \tag{31}$$

for $i = 1,\dots,n$. Singular value decomposition (SVD) is the technique of choice here, in particular as any difficulties which might arise due to possible near-singularity of the matrix are avoided (Press et al., 1992). One further consideration is the calculation of time derivatives from concentration data by numerical differencing, which may generate significant errors in the estimates. Several possibilities exist to improve this situation, including the application of smoothing techniques to the data. Alternatively, by formally integrating the linear equations (29) the problem is cast as a nonlinear fitting procedure to a sum of exponentials, much as for the classical relaxation techniques. Then the dependence on the parameters (relaxation times and coefficients for the linear modes) is nonlinear, and so a nonlinear modelling technique (such as the Levenberg–Marquardt method, Press et al., 1992) must be employed. Hence, there is an algorithmic trade-off in choosing this alternative approach. These difficulties can in part be avoided by writing this expression as a multilinear regression problem (Díaz-Sierra et al., 1999), where, for constant sampling interval Δt,

$$\mathbf{u}(t) = \sum_{j=1}^{n} a_j \mathbf{v}^j e^{\lambda_j t} \tag{32}$$

$$= e^{\mathbf{J}\Delta t}\mathbf{u}(t - \Delta t). \tag{33}$$

Here $e^{\mathbf{J}\Delta t}$ is a matrix which has the same eigenvectors $\mathbf{v}^j$ as the Jacobian, $\mathbf{J}$, and has eigenvalues $\exp(\lambda_j \Delta t)$, where λ_j are eigenvalues of the Jacobian. Then an expression equivalent to (29) is

$$[\mathbf{u}_1|\mathbf{u}_2|\cdots|\mathbf{u}_m] = e^{\mathbf{J}\Delta t}\cdot[\mathbf{u}_0|\mathbf{u}_1|\cdots|\mathbf{u}_{m-1}]. \tag{34}$$

In the language of control theory, if $m \geqslant n$ and the matrix $[\mathbf{u}_0|\mathbf{u}_1|\cdots|\mathbf{u}_{m-1}]$ is nonsingular then the system is said to be *identifiable* (by measurement of all n of the variables), and we can determine the elements of $\mathbf{J}$. In practice this means that the initial perturbation $\mathbf{x}_0$ must excite all linear modes of the system (i.e., the perturbation must have nontrivial projection onto all eigenvectors of $e^{\mathbf{J}\Delta t}$). Also, sampling must be sufficiently fast otherwise some rows of this matrix will be (nearly) linearly dependent. Furthermore, the condition for identifiability will be more restrictive for noisy data (see Lee, 1964, for more discussion on this point).

3.2.1. A practical approach: multilinear regression

Perhaps the biggest problem with the methods outlined above is that the response of each of the biochemical species in the network must be monitored in order to determine the Jacobian matrix. This is a significant hurdle to the experimental application of the techniques, where for practical reasons it may not be possible to measure many concentrations concurrently. One possible solution suggested by Mihaliuk et al. (1999) requires a different perturbation strategy, and a multilinear regression approach to analysis of the data. If perturbations are made repeatedly to the system at time intervals Δt, as illustrated in Fig. 2, then concentrations will be related by

$$\mathbf{u}_k = e^{\mathbf{J}\Delta t}(\mathbf{u}_{k-1} + \mathbf{g}w_{k-1}), \tag{35}$$

where w_{k-1} is the amplitude of the perturbation $\mathbf{g} = (g_1, \dots, g_n)$ made at time t_{k-1}. The perturbation in this case can be to any of the biochemical species individually, for example, by adding a random amount at each time point, or to a combination of the species by adding random amounts in fixed concentration ratio, so that $\mathbf{g}$ does not change between intervals. This casts the

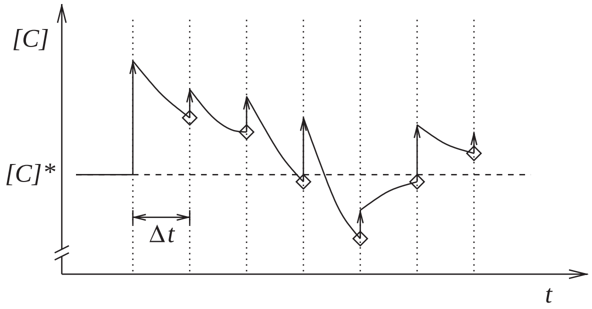

Fig. 2. Schematic illustration of periodic random amplitude perturbations and measurements: diamonds represent measurements, while vertical arrows indicate perturbations of the system in the vicinity of the steady state ($[C]^*$).

problem in a form for which linear control theory may be applied to determine the Jacobian matrix.

A measurement is made on the biochemical system directly prior to each perturbation, indicated by the diamonds in Fig. 2. This can be measurement of the concentration of any particular species, or indeed any linear combination of the set of concentrations $\mathbf{x}(t)$

$$z(t) = \mathbf{h}^{\mathrm{T}} \cdot \mathbf{x}(t) = \mathbf{h}^{\mathrm{T}} \cdot (\mathbf{u}(t) + \mathbf{x}_{\mathrm{s}}) = \tilde{z}(t) + \mathbf{h}^{\mathrm{T}} \cdot \mathbf{x}_{\mathrm{s}}. \tag{36}$$

Now $\tilde{z}_k = \tilde{z}(t_k)$ is simply a linear function of the linearised concentrations $\mathbf{u}_k$ and satisfies a multilinear autoregression expression, which can be written in terms of the *physical* measurements z_k as

$$z_k = a_1 z_{k-1} + \cdots + a_n z_{k-n} + b_0 + b_1 w_{k-1} + \cdots + b_n w_{k-n}, \tag{37}$$

where $b_0 = (1 - a_1 - \cdots - a_n)\mathbf{h}^{\mathrm{T}} \cdot \mathbf{x}_{\mathrm{s}}$. Thus the time series measurement at time t_k depends on the values at the n previous time intervals and the random disturbances only, and this dependence can be expressed as a linear combination. At least $3n+1$ data points are needed in the time series to form $2n+1$ regression equations that determine the $2n+1$ parameters, $a_1, \ldots, a_n, b_0, b_1, \ldots, b_n$, which can subsequently be calculated by inverting the matrix equation

$$\begin{pmatrix} z_{k-1} & \cdots & z_{k-n} & 1 & w_{k-1} & \cdots & w_{k-n} \\ z_{k-2} & \cdots & z_{k-(n+1)} & 1 & w_{k-2} & \cdots & w_{k-(n+1)} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ z_{k-(2n+1)} & \cdots & z_{k-3n} & 1 & w_{k-(2n+1)} & \cdots & w_{k-3n} \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_n \\ b_0 \\ b_1 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} z_k \\ z_{k-1} \\ \vdots \\ z_{k-2n} \end{pmatrix}. \tag{38}$$

As before, generally more data points than this should be collected in the time series, and the least squares minimising solution calculated using SVD for Eq. (37). These regression parameters are related to the matrix $\mathrm{e}^{\mathbf{J}\Delta t}$ and hence $\mathbf{J}$, where to do the necessary matrix inversion we require n independent repetitions with linearly independent perturbations $\mathbf{g}$, each time series requiring at least $3n+1$ data points. Therefore, $\mathcal{O}(n^2)$ measurements are required in total to determine the Jacobian for an n-species network, as might be expected. Mihaliuk et al. (1999) have demonstrated this approach by reconstructing the Jacobian from numerical simulations of the three-species Oregonator model for the Belousov–Zhabotinsky reaction (in a non-oscillatory parameter regime). The linearised system was recovered reliably when uniform additive noise was simulated, by collecting more time series data than the theoretical minimum requirement.

In practice it may be most convenient to perturb the network by systematically adding random amounts of each species in turn. An alternative, but equivalent approach which again may be more practicable is to add uniform amounts of chemicals at random time intervals, rather than subjecting the system to regular perturbations. An equivalent analysis can be followed through to determine the Jacobian. In fact, to calculate the Jacobian **J** from the regression parameters, the form of the measurement **h** is not required to be known, as long as it is some linear combination of the concentrations. This is particularly useful as many measurement techniques which record concentration changes using optical absorption, conductivity or pH can be linearly related to concentration (e.g., optical absorbance is a linear function of concentration according to Beer's law; Fersht, 1999).

3.2.2. *Using sensitivities*

An alternative mechanism for probing chemical reactions near to steady state, which may be more practically applicable in many biochemical problems, is to manipulate system parameters (interpreted broadly as all those parameters on which the steady state concentrations depend), rather than concentrations of the reactants and reaction intermediates themselves. Concentration shift experiments (Eiswirth et al., 1991; Chevalier et al., 1993) are concerned with the variation of input concentrations, x_i^0, which appear as parameters of the model equations

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = F_i(\mathbf{x}, \mathbf{p}) = F_i'(\mathbf{x}, \mathbf{p}') + k_0(x_i^0 - x_i), \quad i = 1, \ldots, n. \tag{39}$$

The experimental apparatus in mind here is the continuous-flow stirred tank reactor (CSTR) in which the reactants are continuously fed into and removed from the reaction vessel. $(k_0)^{-1}$ is the residence time for the CSTR: a measure of how long a molecule spends in the reactor, on average, before it is removed in the out-flow. The objective is to find the elements of the Jacobian ($\mathbf{J}_{ij}= \partial F_i/\partial x_j$) from data obtained on the steady-state concentrations $\mathbf{x}^{\mathrm{s}}(k_0, \mathbf{x}^0, \mathbf{p}')$ when the input concentrations x_i^0 are varied. Differentiating the *steady-state* equations with respect to x_i^0, and evaluating at the steady state,

$$\sum_{k=1}^{n} \frac{\partial F_i}{\partial x_k} \frac{\partial x_k^{\mathrm{s}}}{\partial x_j^0} + \frac{\partial F_i}{\partial x_j^0} = 0, \tag{40}$$

where the second term is equal to k_0 for $i = j$ and zero for $i \neq j$, as x_i^0 only appears in the kinetic equation for the *i*th chemical species, as shown in Eq. (39). The constants $\partial x_k^{\mathrm{s}}/\partial x_j^0$ are the steady-state sensitivities with respect to the input concentrations (note that the steady-state concentration for the *i*th biochemical species, x_i^{s}, does vary with input concentration x_j^0 if species *i* and *j* share any direct or indirect interaction with each other). Then in matrix notation, where $(\partial \mathbf{F}/\partial \mathbf{x})_{ij} = \partial F_i/\partial x_j$,

$$\mathbf{J} = \frac{\partial \mathbf{F}}{\partial \mathbf{x}} = -k_0 \left(\frac{\partial \mathbf{x}^{\mathrm{s}}}{\partial \mathbf{x}^0} \right)^{-1} \tag{41}$$

and so in this case the Jacobian is proportional to the inverse matrix of sensitivities (where the constant of proportionality k_0 may or may not be known a priori). To determine the Jacobian using this approach, each species must have an input flux which can be varied, and the sensitivities $\partial x_i^s/\partial x_j^0$ of the steady state concentrations to each input flux must be measured. In practice these can be approximated by the ratio of the change in x_i^s to the change in the input flux, $\Delta x_i^s/\Delta x_j^0$.

For the CSTR apparatus, input concentrations are easily controlled and responses to these perturbations can be measured from the out-flow concentrations once steady state is re-established. However, it is not immediately obvious how this approach can be extended to study biochemical reactions in vivo. Differentiating the steady-state equations (27) for a general biochemical network with respect to parameters $\mathbf{p}$ gives a generalisation of Eq. (40) for the biochemical network:

$$\sum_{k=1}^{n} \frac{\partial F_i}{\partial x_k}\frac{\partial x_k^s}{\partial p_j} + \frac{\partial F_i}{\partial p_j} = 0 \tag{42}$$

from which, in matrix notation,

$$\frac{\partial \mathbf{F}}{\partial \mathbf{x}} = -\frac{\partial \mathbf{F}}{\partial \mathbf{p}} \cdot \left(\frac{\partial \mathbf{x}^s}{\partial \mathbf{p}}\right)^{-1}, \tag{43}$$

where a matrix multiplication is implied by the dot. Now in general, even if the sensitivities $\partial x_i^s/\partial p_j$, can be measured, the unknowns $\partial F_i/\partial p_j$ prevent us from determining the Jacobian. However, if a set of n parameters can be found, $p_j, j=1,\dots,n$, (i.e., the same number of parameters as species), each of which only affects one reaction function F_i, $i=1,\dots,n$, then $\partial F_i/\partial p_j = \mathrm{diag}(\partial \mathbf{F}/\partial \mathbf{p})$ is a diagonal matrix, and the Jacobian matrix can be determined up to an arbitrary scaling of each of the rows (Kholodenko and Sontag, 2003). Despite these n free parameters, which remain undetermined, information on the relative signs and magnitudes of the Jacobian elements within rows is nevertheless very useful, and if one of the interaction pairs is already known, or can be determined by other means, then the unknown scaling factor is determined for that row of the matrix.

Kholodenko et al. (2002) have extended this approach to exploit the modularity, or compartmentalisation, of many functioning reaction networks. Preferring to work with fractional concentration changes, $\Delta x_i/x_i$, (to be consistent with the notation of metabolic control analysis, Heinrich and Schuster, 1996; Fell, 1997), they show that the approach readily extends to the situation in which self-contained reaction subsystems (modules) interact with each other through connecting reaction intermediates. Each element of the Jacobian then corresponds to the influence of one module's activity on another, and only the concentrations of the interconnecting species need to be monitored. This makes a lot of sense, biologically, where many cellular networks have modular structure, such as signalling pathways (their example: mitogen-activated protein kinase cascades, MAPK) and gene expression networks, and the approach can readily be extended to encompass multiple connecting reaction intermediates between modules (Kholodenko and Sontag, 2003).

What is so appealing about the concentration shift experiments in the CSTR is firstly that the nontrivial entries in $\partial F_i/\partial p_j$ all take the same value, k_0, the reciprocal residence time, (and so we can recover a matrix proportional to the true Jacobian, scaled by k_0 in this case), and secondly that the perturbations Δx^0 to the input concentrations are easily quantified, so that the sensitivities can be estimated. This is not so straightforward, and indeed may not be possible in the biochemical setting, where perturbations made to modules in vivo may be indirect and hence difficult to quantify. Importantly for the applicability of these methods, Kholodenko et al. (2002) have shown that the parameter changes Δp themselves do not actually need to be known. Instead, by assuming linearity of the kinetic functions F with small changes in parameters it is possible to approximate the derivatives from measurements of the steady-state concentration changes alone. This result may turn out to be very useful in situations where a parameter can be indirectly varied, but the magnitude of the change difficult to quantify, as may be the case for in vivo studies.

3.3. Classification of reactions based on the Jacobian

Once the Jacobian has been determined, the entries J_{ij} can be used to classify steady states and determine the interactions in the network (those which persist to first order at the steady state). The development of systematic means of interpretation of these data for chemical reactions have been described by several authors (see, e.g., Chevalier et al., 1993; Epstein and Pojman, 1998; Thomas and Kaufman, 2002, and references therein). As we have noted, a lot can be said about the type of interactions in a reaction network from the signs of the entries alone. If $J_{ij} = 0$, then to first order x_j does not influence x_i: there is no reaction involving the jth species producing the ith chemical. On the other hand, if $J_{ij} > 0$ then increasing x_j causes a rise in the production rate of the ith species. In the case $i = j$ this can be interpreted as direct autocatalysis for $J_{ii} > 0$. Similarly, for $J_{ij} < 0$ the rate of production of x_i decreases with x_j, indicating an inhibitory interaction. The signs of the Jacobian elements can thus be interpreted in terms of the connectivity of the biochemical network. Trivially, we can also infer indirect interactions: if x_i acts on x_j which acts on x_k, then x_i acts indirectly on x_k, and the sign of this indirect influence is determined by the signs of the direct steps. Going back to Tyson (1975), these interactions can be interpreted in terms of positive and negative feedback loops, in which terms the reaction network can also be classified.

3.3.1. Identifying oscillatory dynamics: quenching

When the Jacobian is known for a reaction network then, naturally, the tools of linear dynamical systems can be applied. The stability of steady states can be determined: if all the eigenvalues of $\mathbf{J}$ have negative real part, then the steady state is stable. One important reason for studying the Jacobian is to identify instabilities leading to stable oscillations in the concentrations—if any eigenvalue of the matrix has a positive real part, then the steady state is unstable and oscillations may (possibly) occur (as a limit cycle appears through a Hopf bifurcation). In general, of course, the magnitudes of the Jacobian entries are required for the eigenvalue problem; however, conditions on the signs of the Jacobian entries may be found by which stability properties can be determined. These, in turn, can be classified in terms of feedback loops (Luo and Epstein, 1990; Epstein and Pojman, 1998). In many cases partial information on the Jacobian may be enough to identify instability.

An alternative technique has been proposed for studying oscillatory reactions operating near to the bifurcation point at which the limit cycle gains stability: a perturbation is found to *quench* the oscillation, i.e., which temporarily halts the oscillatory behaviour (Hynne et al., 1990; Sørensen et al., 1990). The phase of the limit cycle and amplitude of perturbation required to quench the oscillation can be related to the Jacobian matrix. However, it is not clear how suitable this approach will be for in vivo studies, and in particular, for biochemical networks with multiple regulatory and homeostatic properties.

3.4. Linear modelling for large reaction networks

Several related techniques have been proposed for analysis of large data sets, in particular those arising from microarray gene expression studies, for which particular problems and issues arise. Microarray data sets are routinely generated for large numbers of dependent variables (microarrays and DNA chips can measure the gene expression levels for thousands of genes simultaneously, enough to cover the entire genome of some organisms; see, e.g., Baldi et al., 2002). Linear modelling of normalised gene expression levels has been suggested (amongst a plethora of other techniques including Boolean networks (Liang et al., 1998) and Bayesian networks (Friedman et al., 2000; Hartemink et al., 2002) for data analysis with the aim of revealing the connectivity in the underlying gene regulatory network. Typically the assumption made is that gene transcripts (mRNAs), which reflect the expression levels of the genes, interact with each other in a linear manner

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = \sum_{j=1}^{n} A_{ij} x_j(t) + b_i(t), \tag{44}$$

where b_i are perturbations made to explore the network and other external factors. For a large network the task of determining the elements A_{ij} from time series data $x_i(t_k)$ with stimuli b_i is typically *underdetermined*. Therefore the inverse problem will not have a unique solution. However, SVD is still the method to use to find a family of solutions (the particular solution picked out by the SVD algorithm will be the one which is smallest in the least squares sense). Yeung et al. (2002) have suggested a further constraint to pick a biologically optimal solution from those found by SVD to be consistent with the data, which is to maximise the sparsity of the resulting connectivity matrix. That is, to find the solution consistent with SVD which has the largest number of zero entries in A. This constraint stems from the observation that large biochemical networks, metabolic and genetic, are characterised by having a much lower number of connections between nodes than would a fully connected network (Jeong et al., 2000, 2001; Wagner and Fell, 2001). In order to minimise the number of nonzero elements the authors have borrowed a technique from robust statistics. The solution determined in this way was found to match up well to data generated from sparsely connected gene network models. Importantly, they show that many fewer measurements (from different perturbations) are needed to find this minimising solution than there are genes in the network, whereas to determine the sparse network using SVD alone, generally as many data sets as genes would be required.

4. Network connectivity

The linear modelling approaches we have described are potentially very useful for providing information on network structure. Jacobian entries reveal interactions between species in the network and, importantly, which species do not interact with each other. Several alternative approaches have been put forward, which aim to directly establish the connectivity of biochemical networks without first examining linear properties near the steady state. Rather, these approaches seek to deduce connectivity from observations on the response to perturbations of *arbitrary* amplitude made at different locations in the network. Inferences about network connectivity can be made from the order and magnitude of the responses of different species to stimuli at different points in the network.

A qualitative form of impulse response analysis can be used to gather information on the connectivity of a biochemical network, which can be pieced together to determine network connectivity, and in some cases the entire wiring diagram. If the concentration of one (or more) species of a network at steady state is increased by some arbitrary amount, unlike the previous methods for determining the Jacobian which relied on small amplitude perturbations, the responses in the concentrations of the other species will reveal qualitative properties of the network. Vance et al. (2002) suggest several observations that can be made on such time series data to reveal network connectivity. As the concentrations rise and fall following the initial disturbance (or impulse), the order of the appearance of peaks and troughs in the time courses for the different species reveals information about their ordering in a pathway. For example, an unbranched chain of reactions perturbed at one end will show (dissipative) propagation of the pulse along the chain. In this trivial case the order of species is revealed in a straightforward manner. The connectivity of the following simple branched pathway:

$$\begin{tikzcd}
X_1 \arrow[r, "k_1"] & X_2 \arrow[r, "k_2"] & X_3 \arrow[r, "k_3"] & {} \\
{} \arrow[r, "k_5"] & X_4 \arrow[u, "k_4"'] & &
\end{tikzcd} \tag{45}$$

can be determined by pulses applied to components X_1 and X_4, as illustrated in Fig. 3.

For more complicated networks, however, the responses will be considerably harder to interpret and the network much more difficult to reconstruct. Qualitatively, at least, the time at which an extremum in concentration appears will increase with 'distance' from the perturbed species. Vance et al. (2002) argue that by perturbing the different components of a network in turn, sufficient information can often be gathered about the causal order of the responses of the

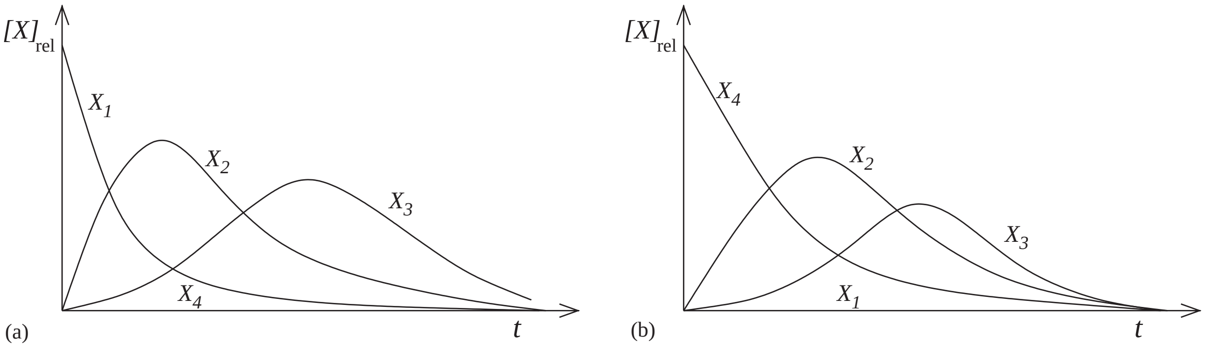

Fig. 3. Relative concentration time series following impulse changes applied to (a) X_1 and (b) X_4 for the reaction pathway (45). Nonzero gradients in the initial responses of X_2 suggests direct connections from both of these species (see text for details). The response of X_3 follows X_2 in both experiments. These data can therefore be interpreted as identifying the convergent branched pathway (45), where all connections other than the flux into X_4 have been revealed.

other concentrations that in large part the network connectivity can be determined. Further information can also, in theory, be obtained from the responses; for example, those species in direct reaction with the perturbed chemical will have nonzero initial slope, whereas the time series data for other species will initially have zero gradient after the impulse (although reaction steps on fast timescales will clearly not be resolved in this way). For open systems the pulse propagates through the network dissipatively and inevitably the best information is recovered on those species closest in a pathway to the point of disturbance.

Vance et al. (2002) motivate this approach via numerical simulation of several example hypothetical networks, with some of the properties of real biochemical networks including branching, feedback, and regulatory interactions. There is, however, no rigorous analysis of the method, for example, of how many species may need to be perturbed or to consider which types of network may or may not be amenable to such an approach, but the method has been demonstrated as viable in an experimental study of a part of the glycolysis pathway in vitro. Torralba et al. (2003) measured concentration changes in the CSTR following impulse changes to the concentrations of different reaction intermediates using capillary electrophoresis. From this time course data alone they were able to determine salient features in the reaction pathway, including irreversibility of the PFK reaction, and (positive and negative) regulatory effects of NADH at the appropriate juncture in the pathway.

4.1. *A correlation-based approach*

This type of qualitative approach can be put into a more quantitative framework using methods which analyse correlations in time series data. Correlation-based approaches to network identification have received a lot of attention recently for analysing gene expression data sets from microarray experiments (Eisen et al., 1998). A common starting point for the analysis of gene expression data is to group together genes with similar expression profiles using a data clustering approach. Genes with similar expression profiles may be involved in similar functions within the cell, and the association of new genes with genes of known function suggests new targets for study.

In order to perform a clustering analysis the *similarity* between two time series must be quantified. A measure based on the correlation coefficient is one possibility, which is particularly suitable for comparisons of shape, rather than magnitude, of the time series profiles. The linear correlation coefficient r_{ij} between two time series $x_i(t)$ and $x_j(t)$ is defined by

$$r_{ij} = \frac{1}{m}\sum_{k=1}^{m}\left(\frac{x_i(t_k)-\bar{x}_i}{\sigma_i}\right)\left(\frac{x_j(t_k)-\bar{x}_j}{\sigma_j}\right), \tag{46}$$

where $\bar{x}_i$ is the mean value of the time series $x_i(t)$ and σ_i is the sample standard deviation

$$\sigma_i = \sqrt{\frac{1}{m}\sum\nolimits_{k=1}^{m}(x_i(t_k)-\bar{x}_i)^2}. \tag{47}$$

The correlation coefficient is symmetric, $r_{ij} = r_{ji}$, and takes values between -1 and 1. When all of the points $(x_i(t_k),\ x_j(t_k))$, for $k = 1,\ldots,m$, lie on a straight line with a positive gradient then $r_{ij} = 1$, independently of the magnitude of the gradient, and the time series are said to be completely positively correlated (illustrated in Fig. 4). Similarly, time series for which $r_{ij} = -1$ are completely negatively correlated, when all of the points lie on a straight line with any negative slope. When

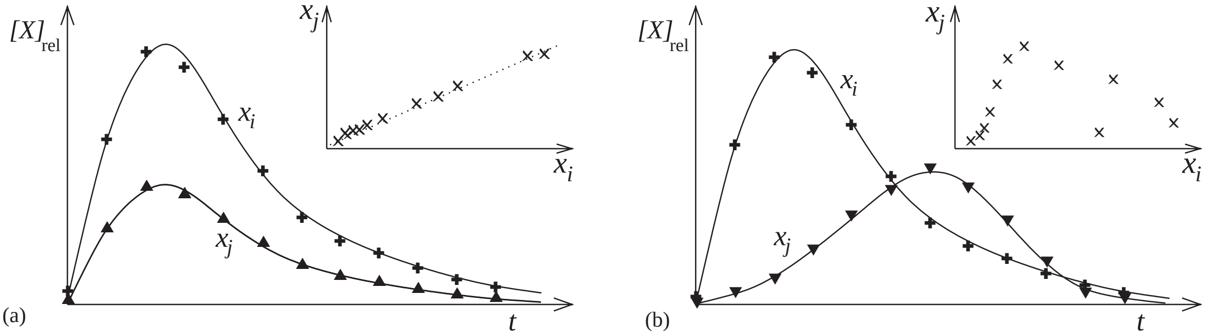

Fig. 4. Illustration of linear correlation between time series data sets. In (a) the time series for x_i and x_j have similar shape, and are thus strongly (positively) correlated with a correlation coefficient close to 1 (as shown in the inset, where the points $x_j(t_k)$ plotted against $x_i(t_k)$ for $k = 1,\ldots,m$ fall on a straight line with positive slope). In (b) the two data sets have different shape, and are not strongly correlated (inset; the points x_j against x_i do not fall on a straight line). In this case the correlation coefficient measures how well a straight line can be fitted to this cloud of points.

$r_{ij} = 0$ the data sets are completely uncorrelated and no preferred linear relation between the two time series can be found.

For a data set comprising time series profiles for n species, $x_i(t)$ for $i = 1,\ldots,n$, the correlation matrix of the $n(n-1)/2$ independent pairwise correlation coefficients can be used to cluster the data set into groups of species within which correlations between species are high, when compared to pairwise correlations between different groups. These groupings can most easily be discerned by calculating a matrix of pairwise *distances*, $d_{ij} = \sqrt{2(1-r_{ij})}$, from the correlation matrix, whereby $d_{ij} = 0$ for two species which are completely positively correlated and increases as the pairwise correlation r_{ij} decreases. This distance matrix can subsequently be analysed to find clusters in the data, for example, using hierarchical clustering techniques, or the K-means clustering algorithm (Chipman et al., 2003).

4.2. Correlation metric construction

Correlations in time series data for biochemical networks can similarly be used to reveal dependencies between variables, and to infer connectivity between species. As discussed above, the influence of one species on another takes some finite amount of time to propagate through the network, and the ordering of responses to impulse stimuli reveals information about network connectivity. In particular, this will be evident in the time series if the time interval Δt between concentration measurements is smaller than the characteristic response timescales for the network. Therefore, two time series which have a low correlation coefficient may in fact be strongly correlated if a *time lag* is allowed for between the data points for the two species, as illustrated in Fig. 5.

By calculating correlations with different time lags between time series, interactions between species can therefore be discerned. The response to more direct, proximal interactions might be expected to be more rapid, and to give a stronger correlation at the appropriate time lag, than to an effect from further away in the network. As proposed by Arkin and Ross (1995), a time-lagged correlation matrix can be computed from the data at a series of delays $\tau\Delta t$, where

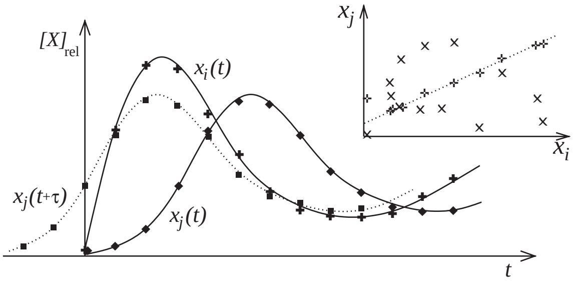

Fig. 5. Illustration of time-lagged correlation analysis. Time series for species x_i and x_j (solid lines) show little sign of linear correlation (inset; × 's); however, comparison of $x_i(t)$ and time-lagged data $x_j(t+\tau)$ (dotted line, time lag $3\Delta t$) shows strong positive correlation (inset; + 's, correlation indicated by the dotted straight line). This suggests an interaction where X_j is activated by X_i.

$\tau = 0,1,2,\ldots,m-1$ for a time series of length m, by generalising the linear correlation coefficient (46)

$$r_{ij}(\tau) = \frac{1}{m-\tau}\sum_{k=1}^{m-\tau}\left(\frac{x_i(t_k)-\bar{x}_i}{\sigma_i}\right)\left(\frac{x_j(t_{k+\tau})-\bar{x}_j}{\sigma_j}\right), \tag{48}$$

where $t_{k+\tau} = t_k + \tau\Delta t$. For each pair i, j the largest magnitude correlation value as τ is varied gives an indication of the strength of interaction between the species. In fact the time-lagged correlation matrix also has a symmetry: $r_{ij}(\tau) = r_{ji}(-\tau)$, and therefore, we need only consider positive time lags for each pair: $\max_{\tau \geqslant 0}|r_{ij}(\tau)|$ is the maximum absolute correlation value (positive or negative) for the influence of x_i on x_j, while $\max_{\tau \geqslant 0}|r_{ji}(\tau)| \equiv \max_{\tau < 0}|r_{ij}(\tau)|$ is the strongest effect of the jth on the ith species.[2] The sign of the maximal correlation in each case, whether a positive or negative correlation, determines if the interaction is activating or inhibitory, and the magnitude suggests whether the interaction is more or less direct.

To use this approach, termed correlation metric construction (CMC) by Arkin and Ross, a biochemical network near steady state is subjected to a series of random, large amplitude concentration changes, and time series data is measured for as many species as possible at time intervals Δt shorter than the relaxation timescales for the network. These data are used to calculate pairwise correlation coefficients at all positive time lags $\tau = 0,1,\ldots,m-1$, (although in practice the values for large time lags are less useful as they are calculated with very few data points). The extremal correlation values can be used, in the same way as above, to construct a distance matrix, $d_{ij} = \sqrt{2(1-\max_{\tau \geqslant 0}|r_{ij}(\tau)|)}$, for which smaller distances are associated with mechanistic connections between species. The n diagonal elements d_{ii} are zeros, trivially, as any time series completely correlates with itself at zero lag; however, each of the $n(n-1)$ independent entries reflects the strength of the directed interaction, $x_i \to x_j$.

Much of the information contained in the distance matrix can be extracted and interpreted graphically. Arkin and Ross have suggested an analysis which finds a particular projection of the distances on to the plane in which the stronger connections (shorter distances) between species are represented as a connected graph. Weaker interactions (longer distances) are collapsed, and can

[2] This is equivalent to taking the maximal absolute correlation values at positive lag and at negative lag for each i, j pair.

be ignored. A technique called multidimensional scaling (MDS) finds the optimum projection, for a given distance matrix, to give the best separation of the data. Graphical interpretation of the projected data reveals information about the ordering of species in pathways and some clues as to the type of interaction, for example, species which are strongly localised may form a subsystem which is weakly coupled to the rest of the network, or may represent interconversion of reaction intermediates at quasi-steady state. The MDS analysis allows a reduction of the highly complicated correlation matrix to a more easily visualised and interpreted representation (Samoilov et al., 2001).

Perhaps the most attractive feature of the correlation-based approaches described in this section is that information can be extracted from a data set with relatively little a priori knowledge of the underlying mechanisms. To some degree, at least, they can deal with the effects of unobserved species on the network inference problem. This is because correlation between x_i and x_j will still be observed in the data even if their interaction is mediated via some intermediate, or intermediates, which are not measured. For example, in an enzymatic reaction it is not necessary to monitor the concentration of the enzyme–substrate complex to reveal correlations between measured substrate and product concentration time series data. This is enormously helpful in biochemical networks, where most reactions between metabolites are enzymatic, as the metabolites are by-and-large small molecular species which can more straightforwardly be measured. In an experimental study in the CSTR, also focusing on the first few steps of the glycolysis pathway, Arkin et al. (1997) have demonstrated that useful information can be extracted using the correlation-based CMC technique. However, a good deal of chemical knowledge about plausible interconversions for the metabolites was required to infer most of the interactions in the pathway. It is less clear how far it would be possible to go with less well-characterised chemical components, or for more complicated networks.

5. Nonlinear reaction models from biochemical time series data

Depending on the timescales of the biochemical reactions and the degree of experimental manipulation to which the system can be subjected, it may be possible to observe sufficient data to obtain a representation of the full underlying mechanism. The use of global nonlinear models provides a method for establishing a mathematical description of the biochemical pathway which is valid over the full range of the data, i.e., not solely in the vicinity of a steady state, for example. Typically for global nonlinear models the mathematical functions to be used in the model are chosen empirically. Biochemical reactions are constrained by the fundamental laws of chemistry and physics, and so this is naturally a good starting point for mathematical representation of the reaction kinetics. Employing fundamental laws to specify models may increase the likelihood of obtaining an accurate mathematical representation of the underlying reaction mechanism, by restricting the model class (the number and type of kinetic functions) that may be used to describe the biochemical reactions.

The techniques described below have not, to our knowledge, routinely been applied to biochemical reaction data; however, the anticipated increase in the availability of comprehensive data sets will allow the development and refinement of this approach to complex biochemical pathway and network data. In the following sections, we present a general mathematical framework for describing the mechanism underlying a system of biochemical reactions from time series data using global nonlinear models. The advantages and disadvantages of a number of different types of basis functions available for constructing global nonlinear models are described in Section 5.2, along with a very general approach to parameter estimation: genetic algorithms. The amount of experimental data available and the complexity of the biochemical reactions will be the deciding factor when selecting an appropriate model. Techniques for choosing model complexity and model construction are reviewed in Section 5.3. We start with a discussion of uncertainty in experimental data.

5.1. Uncertainty in experimental measurements

The analysis of experimental time series data is complicated by uncertainty due to measurement errors, noise, artefacts, and missing data. The construction of linear models from time series with additive measurement errors (usually assumed to be normally distributed) has received a lot of attention and there is a large body of research on this and related problems (see, e.g., Chatfield, 1989). In contrast, the use of data-driven nonlinear models, which are likely to be required for all but the simplest biochemical mechanism, is a relatively new field of research (see Kantz and Schreiber, 1997, and references therein). The construction of nonlinear models from noisy time series data is complicated by the fact that the noise is propagated through the model giving rise to non-normal distributions which depend on the nonlinearities (McSharry and Smith, 1999).

A further complication that can hinder time series analysis is whether or not the underlying biochemical mechanism was *stationary* throughout the period when the data was recorded. External influences that are assumed to be held constant during the experiment may, in fact, vary slowly. An obvious candidate for such problems is temperature: minute changes in temperature may perturb the reaction mechanism, thereby giving rise to biased parameter estimates in a model where temperature is assumed to be constant. It is therefore extremely important that the experimental conditions are carefully monitored. This is often referred to as *dynamical* uncertainty and includes uncertainty in the parameters and the structural equations used to describe the reaction mechanism. In addition, a good understanding of all the sources of inaccuracy inherent in the experimental apparatus and measurements, *observational* uncertainty, is needed. The modelling framework should take this observational uncertainty into account as it will influence the parameter estimates and the predictive accuracy of the resulting model (McSharry and Smith, 2004).

5.2. Specification of global nonlinear models

The goal of nonlinear time series modelling is to establish equations which describe the biochemical kinetics underlying the data. A model is composed of a number of so-called basis functions selected from a pool of “possible terms” for the model. The types of nonlinear basis functions which we will consider here are primarily motivated by the desire to produce models which represent the kinetics in a meaningful way, i.e., which represent plausible interactions between the various chemical species. Alternative approaches may provide accurate representations of the data set, but cannot subsequently be interpreted in terms of reaction mechanism. We will briefly mention one such approach, artificial neural networks (ANNs), which is becoming increasingly common in the literature and which can also be described in the same general framework which we develop here.

A model $\mathbf{F}$ for the underlying reaction mechanism, composed of the functions F_i, provides a description for the rate of production of each species in terms of the concentrations $\mathbf{x} = \mathbf{x}(t)$,

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = F_i(\mathbf{x}, \mathbf{a}_i), \quad i = 1, \ldots, n. \tag{49}$$

In general the net production rate F_i can be expressed as a weighted sum of K basis functions, Φ_j, which need not be orthogonal functions, representing contributions from different elementary processes

$$F_i(\mathbf{x}, \mathbf{a}_i) = \sum_{j=1}^{K} a_{ij} \Phi_j(\mathbf{x}, \mathbf{b}_j), \tag{50}$$

where the parameters $\mathbf{b}_j$ are particular to the specific elementary process. Below, we will discuss how to pick K, the number of basis functions; the choice reflecting the model complexity, i.e., the number of elementary processes making up the network.

There are many choices available for the set of basis functions Φ_j used to provide global approximations to the reaction mechanism (Kantz and Schreiber, 1997). In time series analysis when little is known about the underlying mechanisms generating the data, the basis functions are often chosen for mathematical convenience, or because of their ability to capture certain mathematical features of a data set. Radial basis functions, for example, have proven to be very useful in the case of spatio-temporal time series analysis (McSharry et al., 2002). As was discussed in Section 2, in biochemical kinetics there are good reasons for selecting particular model formalisms which are proposed as representations of the underlying chemical reactions. Polynomial models capture the structure of the equations arising from mass action kinetics, for example. This mechanistic approach is not always appropriate. It will not always be the case, in particular for high-throughput data sources, that the data are of sufficient quality to be able to extract the underlying reaction mechanism from the data set. In this case, the aim of nonlinear modelling is somewhat different; to extract certain nonlinear features, or simply to represent the data in a mathematically compact form, which will allow further comparisons and correlation studies to be done. In this case it may be more appropriate to take a generic approach, using, for example, artificial neural networks, reviewed in Section 5.2.3. The parameters for models of this latter type can be determined using a general approach to nonlinear parameter estimation, genetic algorithms, which is reviewed in Section 5.2.4.

5.2.1. Polynomial models

First, let us consider a more restrictive class of global nonlinear models, which is frequently used in nonlinear time series analysis and is highly suited to biochemical reaction networks, given by

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = F_i(\mathbf{x}, \mathbf{a}_i) = \sum_{j=1}^{K} a_{ij} \phi_j(\mathbf{x}), \tag{51}$$

where the basis functions ϕ_j may be powers and cross products of the components of the concentration vector $\mathbf{x}$. An attractive feature of this particular class of global model is that the parameters enter linearly. This means, as for the Jacobian determination in Section 3.2, that the model may be fitted to the data using least squares and singular value decomposition. (In fact, the linear fitting for the Jacobian matrix is simply the case where $\phi_j(\mathbf{x}) = x_j$.) The solution for the model parameters $\mathbf{a}_i = \{a_{ij}\}_{j=1}^{K}$, determined by the linear system of equations (51), is achieved by seeking parameters $\mathbf{a}_i$ which minimise the sum of the squared residuals

$$\chi^2(\mathbf{a}_i) = \sum_{k=1}^{m} \left(\frac{\mathrm{d}x_i}{\mathrm{d}t}(t_k) - \mathbf{a}_i \cdot \boldsymbol{\phi}(\mathbf{x}_k) \right)^2, \tag{52}$$

where, as before, both χ^2 and $||\mathbf{a}_i||$ are minimised using SVD.

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

$$F_i(\mathbf{x}, \mathbf{a}_i) = a_i + \sum_{j=1}^{K} b_{ij}x_j + \sum_{j=1}^{K}\sum_{k=1}^{K} c_{ijk}x_jx_k, \tag{53}$$

where the parameters determine the rate constants for linear (first-order) and quadratic (second-order) interactions, and constant flux terms (sources and sinks), and thus can represent all possible unimolecular and bimolecular interactions between the species. This approach can easily be extended to include cubic and higher order polynomials if necessary, and therefore encompasses all possible elementary interactions of the type represented in Eq. (3).

5.2.2. Polynomial basis functions based on elementary reaction steps

Although easily visualised and known to converge (via the Weierstrass approximation theorem) multivariate polynomial models have limited usefulness. The large number of free parameters to be fitted requires a large number of computations and the order of the convergence is not known (Casdagli, 1989). In fact, for an n-dimensional data set the number of parameters required for a polynomial of order p is $(p+n)!/p!n!$ (Farmer and Sidorowich, 1987); therefore, using global polynomial models quickly becomes intractable with large numbers of variables.

One possible resolution to this problem is to further restrict the basis functions to represent elementary reactions, rather than just interactions. For example, if the jth basis function represents a bimolecular reaction between species x_p and x_q of the form

$$x_p + x_q \xrightarrow{a_j} x_r, \tag{54}$$

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

$$\left\{ \frac{\mathrm{d}x_p}{\mathrm{d}t}, \frac{\mathrm{d}x_q}{\mathrm{d}t}, \frac{\mathrm{d}x_r}{\mathrm{d}t} \right\} = \phi_j(x_p, x_q) = \{-x_px_q, -x_px_q, x_px_q\} \tag{55}$$

and the rate parameter a_j can be estimated by fitting to the data set, thus retaining the advantage of linearity in the unknown parameters.

In this way a set of basis functions can be built up to represent plausible biochemical reactions between the different species, using the law of mass action formulation for the kinetics of bimolecular interactions. Similar expressions are easily found for unimolecular and trimolecular reactions. This is an example of restriction of the model basis by imposing the formalism of a fundamental chemical law on to the model structure. This approach is currently being developed by us, and will be reported in detail elsewhere.

5.2.3. Artificial neural networks

ANNs are extremely flexible and are capable of approximating very complicated functions (Casdagli, 1989; Bishop, 1995). For this reason, they may be useful in approximating time series when the underlying mechanism is not known or is too complex to be easily represented (if reaction intermediates have not been measured, for example, in enzymatic reaction pathways), or if the data set is of poor quality so that only patterns and correlations in the data are sought. Here we will restrict our discussion to the use of neural networks in time series analysis, but we point out that neural networks have other applications in biochemical and biotechnological data processing, for example, in discriminant analysis, classification tasks, pattern recognition and other machine learning applications (Almeida, 2002).

Feed-forward networks with one hidden layer have been used successfully for time series modelling. In the case of a biochemical system with n species, this model consists of a layer of n input units, one hidden layer of K neurons, and one layer of n output units (Fig. 6). The values at the hidden layer v_j, $j=1,\ldots,K$, are given by an activation function $g(\cdot)$ acting on the weighted input values. If a bias term b_j is also included at the jth neuron then the values are $v_j = g\left(\sum_{i=1}^{n} w_{ij}x_i + b_j\right)$ and for a sigmoidal activation function, $g(z) = 1/(1+\mathrm{e}^{z})$, the output values are given by

$$F_i(\mathbf{x}, \mathbf{a}_i) = \sum_{j=1}^{K} \frac{w'_{ji}}{1+\exp(\mathbf{w}_j \cdot \mathbf{x} + b_j)}, \tag{56}$$

where $\mathbf{w}_j = \{w_{ij}\}_{i=1}^{n}$ is a row vector of weights for the n input connections to the jth hidden node and w'_{ij} are the output weights, respectively. To fit the ANN model to data the parameters w_{ij} and w'_{ij} and b_j have to be determined by nonlinear minimisation techniques. Viewed in this way, the ANN is simply a global nonlinear function with K basis functions $\phi_j(\mathbf{x}) = 1/(1+\mathrm{e}^{\mathbf{w}_j \cdot \mathbf{x}+b_j})$ (Kantz and Schreiber, 1997).

A significant disadvantage of ANNs comes in the nonlinear minimisation step, since this is generally fraught with many local minima in parameter space. Therefore, one can never be certain whether or not a true minimum has been reached. Difficulty in obtaining a minimum generally means that a large amount of computational effort goes into this part of the procedure. Techniques include the iterative error back-propagation technique, or any other general nonlinear

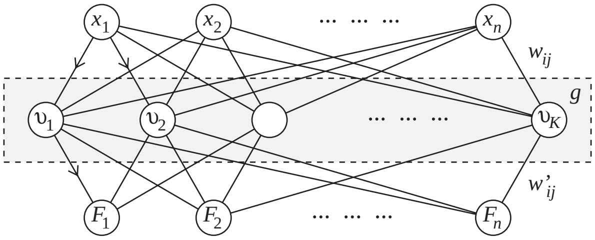

Fig. 6. Schematic of a feed-forward artificial neural network connecting n input nodes, x_i, to n output nodes, F_i, via one "hidden" layer of K neurons, v_i (shaded box). See text for details.

optimisation approach such as conjugate gradients, simulated annealing (Press et al., 1992) or genetic algorithms (GAs) (Wahde and Hertz, 2000), described below. Furthermore, while neural networks may provide an accurate approximation of the recorded time series, it is difficult to extract deeper insight about the underlying mechanism from the fitted neural network. This is the common criticism of machine learning techniques; although good predictive results may be obtained, an improved mechanistic understanding of the underlying processes does not necessarily follow. For this reason simpler and more mechanistically motivated models may be better suited to interpreting biochemical reaction data.

5.2.4. Genetic algorithms

Nonlinear modelling and prediction often involves a search for parameters which minimise residual errors, χ^2, or some more general cost function (Section 5.3.1). GAs provide a very flexible approach to nonlinear optimisation (Mitchell, 1996), which can be applied to time series analysis. GAs mimic aspects of biological evolution to find optimal solutions, namely (i) random mutation and crossover of heritable information and (ii) selection on the basis of fitness between generations. Trial parameter sets are encoded as data strings ('chromosomes') which are scored according to how well the corresponding model satisfies the optimisation task. Better minimisers are given higher fitness scores. A population of strings, encoding different parameter sets, is made to evolve over a number of generations by random mutation and crossover between strings at each generation. Optimal solutions are sought if the probability of survival to the next generation is made dependent on its fitness score. There are as many implementations of GAs as there are practitioners, however, the following basic steps present the general idea:

1. Generate a population of P individuals, each encoding a string of uniformly distributed random numbers assigned to each model coefficient. The random numbers are chosen to be within the relevant search space for each parameter.
2. For each individual in the population, $j = 1,\ldots,P$, evaluate the fitness f_j; a measure of how well the encoded model predicts the data. The algorithm is terminated if a predetermined stopping criterion on the fitness is met, for example, $\max_{j=1}^{P}\{f_j\} \geqslant f^*$. The individual with maximal fitness is returned as an optimising model within the specified criterion.

3a. *Selection*: by sampling with replacement, select a pool of P individuals with probabilities determined by fitness scores (generally the highest scoring individual from the previous round is represented at least once with probability 1).

3b. *Crossover and mutation*: from this pool generate new individuals by mixing parameters from two parents to form new individuals. A (small) degree of normally distributed, zero mean, random variation is introduced to the parameters.

4. The P resulting individuals form the next generation. Go to (2) to calculate fitness.

The success of the algorithm in converging to an optimal solution can greatly depend on the size of the population P and the choice of fitness function, stopping criterion, crossover, and mutation schedules. The underlying model basis can be very general and can even increase in complexity during a computation, for example, by including additional parameters once a certain fitness has been achieved, as a method of exploring very complicated models. Wahde and Hertz (2000) used a GA to optimise a neural network model (56) for clustering gene expression time series data, and were able to extract some qualitative features of the data set, suggesting possible interactions between the sets of genes. Working within the S-system framework given by Eq. (23), Kikuchi et al. (2003) discuss a refinement of the algorithm where a penalty is included in the fitness function so that individuals encoding models with more terms have reduced fitness. This and other techniques which attempt to prevent over-fitting the data are discussed below.

The great flexibility of GAs is that very general optimisation criteria can be implemented. Ross and co-workers have used GAs to parameterise kinetic models by optimising the efficiency of matching pathway flux to substrate demand (Gilman and Ross, 1995), and to predict the elementary steps and rate parameters for an oscillatory reaction based on fitting the period and waveform of the oscillation (Tsuchiya and Ross, 2001). An extension of genetic algorithms is to evolve programmes, rather than character strings (Koza et al., 2001; Ando et al., 2002). An individual programme describes a tree structure representing a model of the biochemical network, with mathematical operations on the nodes (chemical species) representing interactions. Thus not only can the parameters be optimised using the algorithm, but so can the underlying model structure.

5.3. Model construction

Once a model basis has been chosen, there remains the issue of how complex the model should be made when attempting to fit the experimental data. Each of the models discussed above can be made more complex by increasing the number of basis functions, K; the order of the polynomials; the number of nodes in the ANN hidden layer; and so on. Overly complex models may provide excellent approximations to the data set used for constructing the model but are unlikely to perform as well when tested on new data. An understanding of the relationships between the model and observational uncertainty (noise) is required both for parameter estimation, finding parameters for which the model best fits the data, and the subsequent assessment of the resulting model, how well it predicts the data. The following section reviews the cost function approach to identifying a model which provides an adequate fit to the data, yet is not overly complex so as to fit the noise in the data set. Methods for obtaining parameter estimates and outlines to some approaches for evaluating model accuracy are discussed in Section 5.3.2.

5.3.1. Model complexity and cost functions

As a general principle, Occam's Razor suggests that within a class of models we wish to establish the simplest model compatible with the observations. This seems particularly desirable when the model structure is to be interpreted to identify mechanistic processes underlying the data. Increasing the complexity of a model naturally gives more freedom to provide a better fit to the data, reducing the sum of the squared residuals (χ^2). Unfortunately, a model with too many parameters will not distinguish between the generative dynamics that we wish to identify and artefacts due to intrinsic fluctuations and noise. This problem is known as *over-fitting* and is easily identified by examining differences in the prediction errors between the data set used for model construction (in-sample error) and a test data set (out-of-sample error). The residual, or prediction error E_i, denotes the discrepancy between the observed data and the model prediction. If the in-sample prediction accuracy is much greater than the out-of-sample prediction accuracy then over-fitting is strongly implicated.

An intuitive approach for restricting model complexity is to add a penalty term which discriminates against models with too many parameters. Let us suppose that the model being proposed to describe the reaction kinetics is given by the polynomial basis, Eq. (53), where **a** denotes the set of model parameters. The maximum likelihood principle states that the model most likely to have generated the data $\mathbf{X} = \{\mathbf{x}(t_k)\}_{k=0}^{m}$ is the one specified by the parameters which maximise the conditional probability $p(\mathbf{X}|\mathbf{a})$ of observing the data **X** given that the underlying dynamics are described by $\mathbf{F}(\mathbf{x},\mathbf{a})$. This is typically achieved by maximising the log-likelihood function

$$L(\mathbf{X}|\mathbf{a}) = \log p(\mathbf{X}|\mathbf{a}). \tag{57}$$

By increasing the model complexity (increasing K), it is usually possible to increase the value of $L(\mathbf{X}|\mathbf{a})$. One method for taking account of the model complexity is to add a penalty term γ to the *negative* log-likelihood function, yielding a *cost function* to be minimised:

$$C(\mathbf{a}, K) = -L + \gamma(K, m), \tag{58}$$

where $\gamma(K, m)$ can be an increasing function of K, the number of terms (and hence parameters) in the model, and of the number of data points, m.

While any increasing function of K may be sufficient to yield a cost function which has a minimum for some value of $K = K_{\max}$, several suggestions have been made towards determining the appropriate penalty term to apply. By maximising the log-likelihood functions for a set of models with different numbers of parameters, Akaike (1974) has shown the optimal model to be the one that minimises the simple cost function

$$C_{\mathrm{AIC}}(\mathbf{a}, K) = -L + K. \tag{59}$$

Alternatively, Schwarz (1978) has suggested a Bayesian information-based criterion, valid as the number of data points m tends to infinity. This same cost function was obtained by Rissanen (1980) by assuming that the optimal model minimises the description length of an encoding of the data, which includes (i) a description of the model and its parameters and (ii) the residual errors from using the model to predict the data. Given a relatively accurate model of the data, fewer bits of information are required to encode the residual errors than the model itself. This reduction in bits comes at a cost, namely the length of the encoding of the model. The residuals are encoded with an average length $-L$. For m data points, the number of relevant bits in each parameter typically scale as $\sqrt{m}$, giving a minimum description length criterion

$$C_{\mathrm{MDL}}(\mathbf{a}, K) = -L + \frac{K}{2}\log m. \tag{60}$$

These different cost functions highlight that just as there is no universal 'optimal model', neither is there a universally applicable cost function. For biological networks and reaction pathways, which are in general sparsely connected, we may be justified in assuming that parsimonious models of the data are most likely to establish the correct generative mechanisms. As our knowledge and understanding of biochemical pathways and networks grows (Thieffry et al., 1998; Jeong et al., 2001; Ravasz et al., 2002; Milo et al., 2002), in particular details on scaling properties and network connectivity distributions (Jeong et al., 2000; Wagner and Fell, 2001), we may be able to derive better criteria based specifically on emerging details of the structure and topology of biological networks themselves. In the meantime, a minimum description length-based criterion remains a useful guiding principle for describing biological time series data.

5.3.2. *Parameter estimation by maximum likelihood*

For models with independently and normally distributed (IND) prediction errors (residuals), the maximum likelihood function L is essentially the sum of the squared residuals, and maximum likelihood parameter estimates are given by least squares estimates. If the measurement errors are drawn independently from a normal distribution with standard deviation σ, the probability of the data and prediction errors E_i, given the model with parameters $\mathbf{a}$, is

$$p(\mathbf{a}) \propto \exp\left[-\frac{\sum_{i=1}^{m} E_i^2}{2\sigma^2}\right] \tag{61}$$

and so the log-likelihood function (57) is maximised when p is maximised, i.e., when the least squares cost function

$$C_{\mathrm{LS}}(\mathbf{a}) = \chi^2(\mathbf{a}) = \sum_{i=1}^{m} E_i^2 \tag{62}$$

is minimised (for a given model size K). Unfortunately, this is not strictly the case for nonlinear models where the structure of the model affects the distribution of prediction errors, which generally will not be IND. In this case an alternative cost function obtained from the maximum likelihood principle has been suggested by McSharry and Smith (1999).

5.3.3. *Simple-to-general and general-to-specific modelling*

Finally, we address some issues about how to proceed with converging on the model complexity which minimises the cost function. There are two natural approaches for establishing the size of a model for describing an observed time series, i.e., the number of terms which minimises the cost function. One approach, *simple to general*, involves starting with a simple model and adding one term at each step until a minimum in the cost function is located. The issue is to determine which basis function to add at each step, for which several algorithmic approaches have been suggested (see, e.g., Judd and Mees, 1995). However, Hendry and Krolzig (2003) present a list of problems with simple to general modelling, in particular, they point out that it is a divergent branching process, which is likely to generate multiple local minima.

An alternative approach, *general to specific*, starts with a general description (including interactions between all biochemical species, using the highest-order reaction steps that are biochemically meaningful) and then successively eliminates terms until a minimum in the cost function is identified. The general to specific approach uses regression models based on the principles introduced by Hendry (1995), whereby statistically insignificant variables are eliminated at each step as the model complexity is reduced. A number of different reduction paths can be searched to prevent the algorithm from getting stuck in a sequence that inadvertently eliminates a variable that matters, while retaining other variables as proxies. In this way the global minimum in the cost function can be approached more reliably.

6. Discussion

Experimentally recorded time series of the concentrations of chemical species offer a glimpse of the mechanism underlying a system of biochemical reactions. The aim of each of the techniques we have described is to extract some mechanistic information about the reaction network from these data. The data which should be collected for optimal application of a particular technique, i.e. the experimental design, reflects what the technique aims to determine. The relaxation of a biochemical system in response to small perturbations can be used to determine the properties of the system near to a steady state, encapsulated in the Jacobian matrix, using the techniques discussed in Section 3.2. Data collected for large amplitude impulse responses were shown in Section 4 to provide information on the connectivity of a complex reaction network. By contrast, for global nonlinear models the best reconstruction can be achieved for data sets which explore as much of the potential behaviour of the system, the "model space", as possible.

One topic on which we have had little to say thus far is model validation. Generally, consistency with the data is not sufficient reason to accept a particular model, neither does it provide a rational for selecting one consistent model over another. Numerous competing models cannot easily be tested experimentally and the main criteria for their acceptance is often biological plausibility, and consistency with 'known facts'. For the time series analysis techniques which we have described in the previous section, the commonly adopted approach, which may be very useful in this regard, is known as forecasting. While models are often constructed in an attempt to gain a better understanding of the underlying processes, they are usually assessed through their ability to reproduce empirical observations. Once the model parameters have been estimated using the maximum likelihood principle, a model can be evaluated by assessing the distribution of out-of-sample prediction errors (errors for data not used in the model construction) as a measure of the quality of a particular model, as is done to guard against over-fitting to a given data set. This data-driven approach to model evaluation may prove particularly useful for large and complicated data sets where 'known facts' are unreliable, or few and far between.

It would be fair to ask what constitutes a biochemical pathway, particularly when studying systems in vivo, and in particular for those methods which require all species in a pathway to be monitored. Reaction mechanisms of complex biochemical networks can often be divided into different functional subunits. This modularity can be exploited in the investigation of reaction mechanisms, as in the parametric sensitivity analysis described in Section 3.2.2 for the steady state of a reaction system, providing an alternative to using temporal kinetic data. Indeed this approach seems particularly promising for in vivo systems, where smaller reaction pathways may be harder to 'dissect out' for detailed study.

As yet, the application of mathematical and computational techniques such as those described in this review has not been routine. Many of these approaches are in the early stages of their development, and further refinements can be expected for their application to particular sources of data. A further reason for this may be that sophisticated analysis and modelling algorithms are not routinely included in computer packages designed for biochemical data analysis. We believe that the development of high-quality high-throughput approaches to data generation will necessitate the uptake and continued development of methods such as those we have described, as has been the case for statistical clustering and data analysis for gene microarray data sets.

Acknowledgements

This work was supported by the Wellcome Trust (London) through Research Training Fellowships in Mathematical Biology to EJC and SS. PEM is a Royal Academy of Engineering Research Fellow and acknowledges the support of the Engineering and Physical Sciences Research Council, UK. EJC gratefully acknowledges the hospitality of Prof Peter Hunter during his visit to the Bioengineering Institute, University of Auckland, New Zealand.

References

Akaike, H., 1974. A new look at the statistical idenification model. IEEE Trans. Automat. Contr. 19, 716–723.

Almeida, J.S., 2002. Predictive non-linear modeling of complex data by artificial neural networks. Curr. Opin. Biotechnol. 13, 72–76.

Ando, S., Sakamoto, E., Iba, H., 2002. Evolutionary modeling and inference of gene network. Inf. Sci. 145, 237–259.

Arkin, A., Ross, J., 1995. Statistical construction of chemical-reaction mechanisms from measured time-series. J. Phys. Chem. 99, 970–979.

Arkin, A., Shen, P.D., Ross, J., 1997. A test case of correlation metric construction of a reaction pathway from measurements. Science 277, 1275–1279.

Baldi, P., Hatfield, G.W., Hatfield, W.G., 2002. DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modelling. Cambridge University Press, Cambridge.

Bernasconi, C.F., 1976. Relaxation Kinetics. Academic Press, London.

Berry, H., 2002. Monte Carlo simulations of enzyme reactions in two dimensions: fractal kinetics and spatial segregation. Biophys. J. 83, 1891–1901.

Bishop, C., 1995. Neural Networks for Pattern Recognition. Clarendon Press, Oxford.

Blake, W.J., Kærn, M., Cantor, C.R., Collins, J.J., 2003. Noise in eukaryotic gene expression. Nature 422, 633–637.

Boyde, T.R.C., 1980. Foundation Stones of Biochemistry. Voile et Aviron, Hong Kong.

Burrage, K., Tian, T., Burrage, P., 2004. A multi-scaled approach for simulating chemical reaction systems. Prog. Biophys. Mol. Biol. 85, 217–234.

Casdagli, M., 1989. Nonlinear prediction of chaotic time series. Physica D 35, 335–356.

Chatfield, C., 1989. The Analysis of Time Series, 4th Edition. Chapman & Hall, London.

Chevalier, T., Schreiber, I., Ross, J., 1993. Toward a systematic determination of complex-reaction mechanisms. J. Phys. Chem. 97, 6776–6787.

Chipman, H., Hastie, T.J., Tibshirani, R., 2003. Clustering microarray data. In: Speed, T.P. (Ed.), Statistical Analysis of Gene Expression Microarray Data. Chapman & Hall, Boca Raton, FL.

Cornish-Bowden, A., 1995. Fundamentals of Enzyme Kinetics. Portland Press, London.

Crampin, E.J., Halstead, M., Hunter, P., Nielsen, P., Noble, D., Smith, N., Tawhai, M., 2004. Computational physiology and the physiome project. Exp. Physiol. 89, 1–26.

Díaz-Sierra, R., Lozano, J.B., Fairén, V., 1999. Deduction of chemical mechanisms from the linear response around steady state. J. Phys. Chem. A 103, 337–343.

Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 95, 14863–14868.

Eiswirth, M., Freund, A., Ross, J., 1991. Mechanistic classification of chemical oscillators and the role of species. Adv. Chem. Phys. 80, 127–199.

Ellis, R.J., Minton, A.P., 2003. Cell biology—join the crowd. Nature 425, 27–28.

Elowitz, M.B., Levine, A.J., Siggia, E.D., Swain, P.S., 2002. Stochastic gene expression in a single cell. Science 297, 1183–1186.

Epstein, I.R., Pojman, J.A., 1998. An Introduction to Nonlinear Chemical Dynamics: Oscillations, Waves, Patterns, and Chaos. Oxford University Press, New York, Oxford.

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

Ermentrout, B., 2001. Simplifying and reducing complex models. In: Bower, J.M., Bolouri, H. (Eds.), Computational Modeling of Genetic and Biochemical Networks. MIT Press, Cambridge, MA, pp. 307–323 (Chapter 11).

Farmer, J.D., Sidorowich, J.J., 1987. Predicting chaotic time series. Phys. Rev. Lett. 59 (8), 845–848.

Fell, D.F., 1997. Understanding the Control of Metabolism. Portland Press, London.

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

Friedman, N., Linial, M., Nachman, I., Pe'er, D., 2000. Using Bayesian networks to analyze expression data. J. Comput. Biol. 7, 601–620.

Gillespie, D.T., 1977. Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. 81, 2340–2361.

Gilman, A., Ross, J., 1995. Genetic-algorithm selection of a regulatory structure that directs flux in a simple metabolic model. Biophys. J. 69, 1321–1333.

Gray, P., Scott, S.K., 1990. Chemical Oscillations and Instabilities: Non-linear Chemical Kinetics. Clarendon Press, Oxford.

Gutfreund, H., 1995. Kinetics for the Life Sciences. Receptors, Transmitters and Catalysts. Cambridge University Press, Cambridge.

Hartemink, A.J., Gifford, D.K., Jaakkola, T.S., Young, R.A., 2002. Bayesian methods for elucidating genetic regulatory networks. IEEE Intell. Syst. 17, 37–43.

Heinrich, R., Rapoport, T.A., 1974. A linear steady-state treatment of enzymatic chains. Critique of the crossover theorem and a general procedure to identify interaction sites with an effector. Eur. J. Biochem. 42, 97–105.

Heinrich, R., Schuster, S., 1996. The Regulation of Cellular Systems. Chapman & Hall, New York.

Hendry, D.F., 1995. Dynamic Econometrics. Oxford University Press, Oxford.

Hendry, D.F., Krolzig, H.M., 2003. New developments in automatic general-to-specific modelling. In: Stigum, B.P. (Ed.), Econometrics and the Philosophy of Economics. Princeton University Press, Princeton.

Hill, A.V., 1910. The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. J. Physiol. (Lond.) 40, iv–vii.

Hynne, F., Sorensen, P.G., Nielsen, K., 1990. Quenching of chemical oscillations: general theory. J. Chem. Phys. 92, 1747–1757.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., Barabási, A.-L., 2000. The large-scale organization of metabolic networks. Nature 407, 651–654.

Jeong, H., Mason, S.P., Barabási, A.-L., Oltvai, Z.N., 2001. Lethality and centrality in protein networks. Nature 411, 41–42.

Judd, K., Mees, A., 1995. On selecting models for nonlinear time series. Physica D 82, 426–444.

Kantz, H., Schreiber, T., 1997. Nonlinear Time Series Analysis. Cambridge University Press, Cambridge.

Kedem, O., Caplan, S.R., 1965. Degree of coupling and its relation to efficiency of energy conversion. Trans. Faraday Soc. 61, 1897–1911.

Keizer, J., 1987. Statistical Thermodynamics of Nonequilibrium Processes. Springer, New York.

Kholodenko, B.N., Sontag, E.D., 2003. Determination of functional network structure from local parameter dependence data, preprint: arXiv:physics/0205003.

Kholodenko, B.N., Kiyatkin, A., Bruggeman, F.J., Sontag, E., Westerhoff, H.V., Hoek, J.B., 2002. Untangling the wires: a strategy to trace functional interactions in signaling and gene networks. Proc. Natl. Acad. Sci. USA 99, 12841–12846.

Kikuchi, S., Tominaga, D., Arita, M., Takahashi, K., Tomita, M., 2003. Dynamic modeling of genetic networks using genetic algorithm and S-system. Bioinformatics 19, 643–650.

Koza, J.R., Mydlowec, W., Lanza, G., Yu, J., Keane, M.A., 2001. Automated reverse engineering of metabolic pathways from observed data by means of genetic programming. In: Kitano, H. (Ed.), Foundations of Systems Biology. MIT Press, Cambridge, MA.

Lee, R.C.K., 1964. Optimal Estimation, Identification and Control. MIT Press, Cambridge, MA.

Liang, S., Fuhrman, S., Somogyi, R., 1998. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. In: Pacific Symposium on Biocomputing, Vol. 3. pp. 18–29.

Luo, Y., Epstein, I.R., 1990. Feedback analysis of mechanisms for chemical oscillators. Adv. Chem. Phys. 79, 269–299.

McAdams, H.H., Arkin, A.P., 1997. Stochastic mechanisms in gene expression. Proc. Natl. Acad. Sci. USA 94, 814–819.

McSharry, P.E., Smith, L.A., 1999. Better nonlinear models from noisy data: attractors with maximum likelihood. Phys. Rev. Lett. 83, 4285–4288.

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

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

Medalia, O., Weber, I., Frangakis, A.S., Nicastro, D., Gerisch, G., Baumeister, W., 2002. Macromolecular architecture in eukaryotic cells visualized by cryoelectron tomography. Science 298, 1209–1223.

Mihaliuk, E., Skødt, H., Hynne, F., Sørensen, P.G., Showalter, K., 1999. Normal modes for chemical reactions from time series analysis. J. Phys. Chem. A 103, 8246–8251.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U., 2002. Network motifs: simple building blocks of complex networks. Science 298, 824–827.

Mitchell, M., 1996. An Introduction to Genetic Algorithms. MIT Press, Cambridge, MA.

Morton-Firth, C.J., Bray, D., 1998. Predicting temporal fluctuations in an intracellular signalling pathway. J. Theor. Biol. 192, 117–128.

Murray, J.D., 2002. Mathematical biology. I. An Introduction, 3rd Edition. Springer, New York.

Murray, J.D., 2003. Mathematical biology. II. Spatial Models and Biomedical Applications, 3rd Edition. Springer, New York.

Noble, D., 2002. The rise of computational biology. Nat. Rev. Mol. Cell. Biol. 3, 459–463.

Othmer, H.G., 1981. The interaction of structure and dynamics in chemical reaction networks. In: Ebert, K.H., Deuflhard, P., Jaeger, W. (Eds.), Modelling of Chemical Reaction Systems. Springer, Berlin, pp. 2–19.

Palsson, B.O., Palsson, H., Lightfoot, E.N., 1985. Mathematical modelling of dynamics and control in metabolic networks. III. Linear reaction sequences. J. Theor. Biol. 113, 231–259.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 1992. Numerical Recipes in C, 2nd Edition. Cambridge University Press, Cambridge.

Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., Barbási, A.-L., 2002. Hierarchical organization of modularity in metabolic networks. Science 297, 1551–1555.

Rissanen, J., 1980. Consistent order estimates of autoregressive processes by shortest description of data. In: Jacobs, O., et al. (Ed.), Analysis and Optimisation of Stochastic Systems. Academic Press, New York.

Rohwer, J.M., Postma, P.W., Kholodenko, B.N., Westerhoff, H.V., 1998. Implications of macromolecular crowding for signal transduction and metabolite channeling. Proc. Natl. Acad. Sci. USA 95, 10547–10552.

Rottenberg, H., 1973. The mechanism of energy-dependent ion transport in mitochondria. J. Membr. Biol. 11, 117–137.

Samoilov, M., Arkin, A., Ross, J., 2001. On the deduction of chemical reaction pathways from measurements of time series of concentrations. Chaos 11, 108–114.

Savageau, M.A., 1969. Biochemical systems analysis. I. Some mathematical properties of the rate law for the component enzymatic reactions. J. Theor. Biol. 25, 365–369.

Savageau, M.A., 1976. Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology. Addison-Wesley, Reading, MA.

Savageau, M.A., 1992. A critique of the enzymologist's test tube. In: Bittar, E.E. (Ed.), Fundamentals of Medical Cell Biology, Vol. 3a. Academic Press, New York, pp. 45–108.

Schnell, S., Maini, P.K., 2000. Enzyme kinetics at high enzyme concentration. Bull. Math. Biol. 62, 483–499.

Schnell, S., Maini, P.K., 2003. A century of enzyme kinetics. Reliability of the K_m and $v_{\max}$ estimates. Comments Theor. Biol. 8, 169–187.

Schnell, S., Turner, T.E., 2004. Reaction kinetics in intracellular environments with macromolecular crowding: simulations and rate laws. Prog. Biophys. Mol. Biol. 85, 235–260.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6, 461–464.

Scott, S.K., 1991. Chemical Chaos. Clarendon Press, Oxford.

Segel, I.H., 1975. Enzyme Kinetics: Behavior and Analysis of Rapid Equilibrium and Steady-State Enzyme Systems. Wiley, New York.

Segel, L.A., 1972. Simplification and scaling. SIAM Rev. 14, 547–571.

Segel, L.A., 1988. On the validity of the steady state assumption of enzyme kinetics. Bull. Math. Biol. 50 (6), 579–593.

Segel, L.A., Slemrod, M., 1989. The quasi-steady-state assumption: a case study in perturbation. SIAM Rev. 31, 446–477.

Smith, N.P., Crampin, E.J., 2004. Development of models of active ion transport for whole-cell modelling: cardiac sodium-potassium pump as a case study. Prog. Biophys. Mol. Biol. 85, 387–405.

Sørensen, P.G., Hynne, F., Nielsen, K., 1990. Characteristic modes of oscillatory chemical-reactions. J. Chem. Phys. 92, 4778–4785.

Thieffry, D., Huerta, A.M., Pérez-Rueda, E., Collado-Vides, J., 1998. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*. Bioessays 20, 433–440.

Thomas, R., Kaufman, M., 2002. Conceptual tools for the integration of data. CR Biologies 325, 505–514.

Torralba, A.S., Yu, K., Shen, P.D., Oefner, P.J., Ross, J., 2003. Experimental test of a method for determining causal connectivities of species in reactions. Proc. Natl. Acad. Sci. USA 100, 1494–1498.

Tsuchiya, M., Ross, J., 2001. Application of genetic algorithm to chemical kinetics: systematic determination of reaction mechanism and rate coefficients for a complex reaction network. J. Phys. Chem. A 105, 4052–4058.

Tyson, J.J., 1975. Classification of instabilities in chemical reaction systems. J. Chem. Phys. 62, 1010–1015.

Vance, W., Arkin, A., Ross, J., 2002. Determination of causal connectivities of species in reaction networks. Proc. Natl. Acad. Sci. USA 99, 5816–5821.

Wagner, A., Fell, D.A., 2001. The small world inside large metabolic networks. Proc. R. Soc. Lond. B 268, 1803–1810.

Wahde, M., Hertz, J., 2000. Coarse-grained reverse engineering of genetic regulatory networks. Biosystems 55, 129–136.

Yeung, M.K.S., Tegner, J., Collins, J.J., 2002. Reverse engineering gene networks using singular value decomposition and robust regression. Proc. Natl. Acad. Sci. USA 99, 6163–6168.