



ELSEVIER

BioSystems 65 (2002) 1–18



www.elsevier.com/locate/biosystems

Mathematical modelling in the post-genome era: understanding genome expression and regulation— a system theoretic approach

Olaf Wolkenhauer *

*Department of Biomolecular Sciences, Department of Electrical Engineering and Electronics, Control Systems Centre, UMIST,
Manchester M60 1QD, UK*

Received 18 September 2001; accepted 26 September 2001

Abstract

This paper introduces a mathematical framework for modelling genome expression and regulation. Starting with a philosophical foundation, causation is identified as the principle of explanation of change in the realm of matter. Causation is, therefore, a relationship, not between components, but between changes of states of a system. We subsequently view genome expression (formerly known as ‘gene expression’) as a dynamic process and model aspects of it as dynamic systems using methodologies developed within the areas of systems and control theory. We begin with the possibly most abstract but general formulation in the setting of category theory. The class of models realised are state-space models, input–output models, autoregressive models or automata. We find that a number of proposed ‘gene network’ models are, therefore, included in the framework presented here. The conceptual framework that integrates all of these models defines a dynamic system as a family of expression profiles. It becomes apparent that the concept of a ‘gene’ is less appropriate when considering mathematical models of genome expression and regulation. The main claim of this paper is that we should treat (model) the organisation and regulation of genetic pathways as what they are: dynamic systems. Microarray technology allows us to generate large sets of time series data and is, therefore, discussed with regard to its use in mathematical modelling of gene expression and regulation. © 2002 Elsevier Science Ireland Ltd. All rights reserved.

Keywords: Genome expression; Systems biology; Microarray data analysis; Gene network modelling

1. Introduction

It should be made clear from the start that this contribution is dealing with ‘science fiction’. Cur-

rent technology, such as DNA microarrays, does not as yet deliver the data that would be required to identify accurate, predictive models of cellular processes realised by complex networks of chemical reactions. For a comprehensive model of genome expression, current microarray technology does not provide a sufficiently high resolution, and the activity of transcriptional factors,

* Tel./fax: +44-161-200-4672; <http://www.umist.ac.uk/csc/people/wolkenhauer.htm>.

E-mail address: o.wolkenhauer@umist.ac.uk (O. Wolkenhauer).

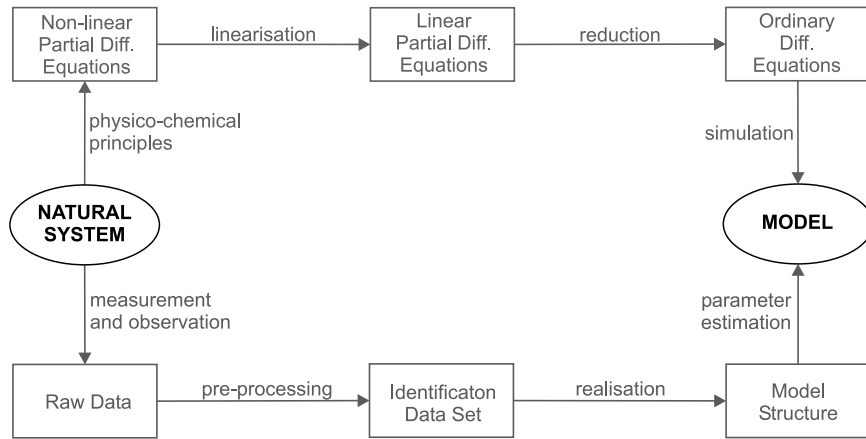


Fig. 1. Modelling vs. identification. In both cases we have the same objective—to obtain a model that is a good representation of the natural system under consideration. In any case we usually have to make a number of assumptions. While in modelling, the purpose is usually simulation, independent of measured data, we can use more complex and often more realistic non-linear models. The quality and quantity of measured data usually force us to make simplifying assumptions in identification.

for instance in regulation, can be influenced by post-translational modification and interactions that remain undetectable at the expression level. However, we begin to have measurements that allow us to discuss mathematical models of genome expression and regulation that go beyond the traditional approach using differential equations to describe observed phenomena. The estimation of unknown parameters from data is notoriously difficult and is probably the reason why mathematical biology has, in some respects, evolved into a discipline less concerned with some of the problems of vital importance to biologists.

Systems biology is an emerging field of biological research that aims at a system-level understanding of genetic or metabolic pathways by investigating interrelationships (organisation or structure) and interactions (dynamics or behaviour) of RNA transcripts, proteins, and metabolites (Kitano, 2001). Mathematical systems biology is then the application of systems and control theory to this end. Systems theory is the study of organisation and behaviour per se, control theory focuses specifically on mathematical modelling and identification of dynamic systems. While the aim of mathematical modelling is often the simulation of natural systems, in system identification, we aim to identify the parameters of a

model from measured data. Mathematical modelling of cellular systems has always been part of mathematical biology. The renewed interest in mathematical modelling in the post-genome era of the life sciences stems from the availability of technologies that allow us to make direct measurements of cellular processes. Over the coming years it should, therefore, become possible to identify parametric models from data. A recent survey and discussion of systems biology can be found in Wolkenhauer (2001a) and for a review of computational studies of gene regulatory networks we refer to Hasty et al. (2001). The relationship between modelling and identification is also illustrated in Fig. 1. While in simulation we can use more complex, for example, non-linear methodologies, in identifying models from numerical data, the quality and quantity of data often forces us to make simplifying assumptions and to be precise about the uncertainty in the data and the model.

The importance of what we now call systems biology was pointed out in 1948 by Norbert Wiener in defining the area of cybernetics (Wiener, 1948). Wiener's work was followed by intensive research into the mathematical foundations of systems and control theory. These developments have been accompanied by numerous

applications of the theory in engineering but generally failed to impress molecular biologists. In the 1960s and in the 1970s, Francois Jacob and Jacques Monod (Jacob and Monod, 1960; Monod, 1970) investigated regulatory proteins and the interactions of allosteric enzymes. They introduced a distinction between ‘structural genes’ (coding for proteins) and ‘regulatory genes’ that regulate the rate at which structural genes are transcribed. This control of the rate of synthesis of proteins gave the first indication of such processes being most appropriately viewed as dynamic systems driven by a multitude of factors and feedback regulated. Indeed, negative feedback is used in all cells and in metabolic pathways in particular. Control of such processes is achieved through regulatory enzymes that respond to effector concentrations by increase or decrease in their reaction rates.

In this paper we reconsider some mathematical models and their application to genome expression, the process by which information stored in the DNA, is transformed into the processes that maintain the cell’s function. A key aspect of this paper is the dynamic systems perspective of genome expression and regulation. This is in contrast to the common presentation in which DNA, RNA, and proteins are seen as material objects described by their spatio–temporal relationships. In genomics, the shift of focus from molecular characterisation to an understanding of functional understanding is driven by new technology such as DNA microarrays which allow, if only limited, access to a temporal analysis of genome expression and regulation. Time course experiments and time-series data from microarrays¹ provide the motivation for the present investigation.

¹Microarrays can analyze the expression levels of many thousands of DNA segments simultaneously. The arrays are made by binding to spots on a glass slide precisely measured quantities of single-stranded DNA that are transcribed into mRNA. In the experiment, they are rinsed with labeled, single stranded mRNA or cDNA mixtures from the cells of interest. Complementary DNA or RNA molecules hybridize to each other, attaching the fluorescent molecules to particular spots on the array. The brightness of the spots can be measured to yield a quantitative measure of the likely number of mRNA molecules present in the cells at the moment of sampling.

The strategy for this paper is as follows. In modelling genome expression and regulation a discussion of philosophical questions can be ignored but not avoided. We, therefore, start with few basic considerations of causal entailment as a basis for mathematical modelling. In Section 3, we begin with the possibly most abstract but general formulation in the setting of category theory. In its most general form, the framework captures the classical approach to describe genome expression in terms of (material) objects or components and their spatio–temporal relationships. However, as we first specialise and then generalise our approach, we put forward a dynamic systems perspective based on measured data from time course experiments. Changing the modelling paradigm from relationships among material objects to sequences of observations, we then specialise our model to obtain an input–output representation. This model is subsequently generalised to a more flexible state-space model that allows us to represent internal relationships. In the context of DNA microarray data, this approach requires some assumptions and we, therefore, consider in Section 5 a conceptual framework which defines a dynamic system as a family of expression profiles and subsequently subsumes the state-space approach and AR autoregressive (AR) models as special cases. While in engineering or forecasting the main criteria for a model is its accuracy of prediction, here the purpose of the model is to help, explain the relationships in those cellular processes that led to the observations made. In Section 6, we introduce a formal representation of the modelling process itself.

A summary of mathematical conventions and notation is given at the end of the paper.

2. The philosophical foundation

The most comprehensive and unsurpassed study of causality or, more generally, of explanation is the work of the philosopher Arthur Schopenhauer who corrected and extended the work of Immanuel Kant. For both, ‘why?’ questions are at the root of any scientific investigation

and their validity, implying a ‘because’, which is the subject of Schopenhauer (1818) dissertation. His ‘Principle of Sufficient Reason’ states that there is a reason, that is, an explanation, for any dictum or existent, whatever and thereby provides scientists with a motive. Furthermore, he argued that always and everywhere each thing exists merely by virtue of another thing. In other words, Schopenhauer highlighted the relational character of explanation, a theme we follow throughout this paper. The Principle of Sufficient Reason takes four forms depending on the class of objects to which it is applied.

1. Material objects subject to change (physico-chemical analysis).
2. Propositions or judgements bearing truth (logical reasoning).
3. Objects possessing mathematical properties (mathematical modelling).
4. Objects giving rise to actions under the influence of motives.

The first class of objects is, in our context, associated with molecular biology, i.e. the description of spatio-temporal relations among components. Causation is in this context the principle of explanation of change in the realm of matter. Causation is thus a relationship, not between things, but between changes of states of things. Causation as the change of states is analogous to the widely used approach of using ordinary or partial differential equations, outlined in Fig. 2 and further discussed in Section 4.

The principal purpose of mathematical models, is to identify sets of rules, statements about local associations or dependencies among variables. In genomics, mathematical models may be expected to not only describe associations but also to explain dependencies among genes. A ‘causal law’, which is not bound strictly to any specific philosophical perspective, is then understood as a ‘causal dependency’, a general proposition by virtue of which it is possible to infer the existence of an event from the existence of another.

Based on sets of first-order differential equations, the so-called state-space representation of

linear time-invariant dynamic systems has been very successful in modelling and simulating physical systems (Kailath, 1980). This approach, which has dominated physics, applied mathematics, and various engineering disciplines, has drawbacks that become particularly unsatisfactory when considering molecular dynamics. First, to make a formal analysis feasible we usually have to consider a system as isolated from its environment and, secondly, knowledge of independent (input) and dependent (output) variables is assumed. On the other hand, recent advances in microarray technology provide simultaneous measurements of gene activity levels across a whole genome and are thus a means to investigate the dynamics of genome expression. From such experiments we obtain a large set of time series. Although, we may have some idea of which genes are involved in a particular process, say antibiotic production in *Streptomyces* we, unfortunately, will usually not know the causal nature of dependencies in such a genetic network.

In the following sections, we are going to discuss a mathematical framework to represent causal entailment in natural systems. From the foregoing discussion, the emphasis will be very much on representation, not of the physico-chemical interactions of components in a cellular process but of the changes of states, observable through for example DNA microarrays. We start with the most abstract and hence, general representation of molecular systems.

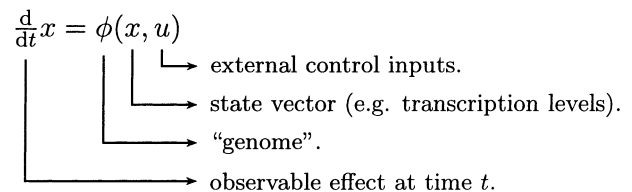


Fig. 2. Mathematical modelling using differential equations. Note that the equations by themselves do not state that changes are produced by anything, but only that they are either accompanied or followed by certain other changes. Considering $dx/dt = \phi(\cdot)$ or equivalently $dx = \phi(\cdot) dt$, is merely asserts that the change dx undergone during the time interval dt equals $\phi(\cdot) dt$. See Rosen (1985) for a full discussion.

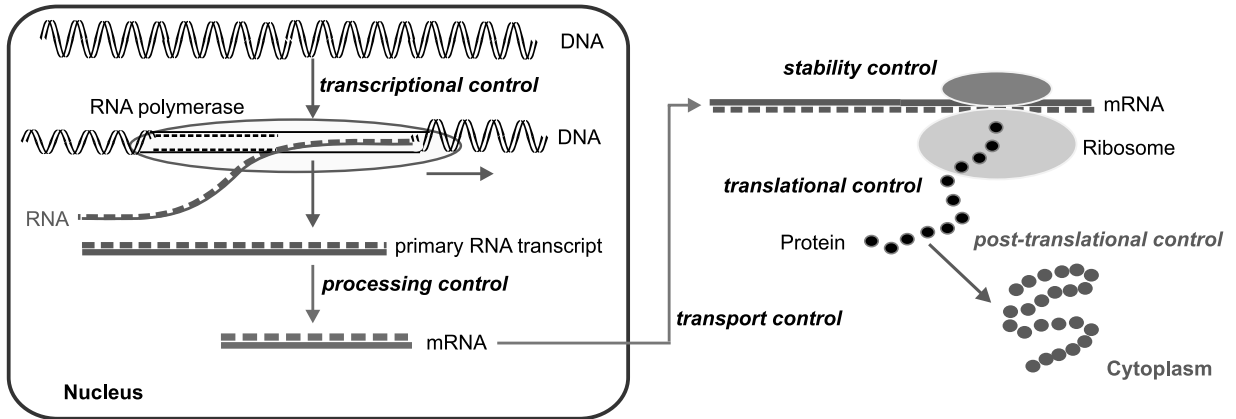


Fig. 3. Illustration of gene expression and regulation in a eukaryotic cell with a nucleus. Each step by which information, stored in the DNA, is transformed into products such as protein is a controlled dynamic process. The intermediate mRNA product is unstable and its production, therefore, requires a control mechanism. Other levels of control are also exerted to modulate the concentration of active proteins. Some of these are illustrated here. Microarrays are measuring the amount of mRNA produced from each gene. The picture, therefore, also illustrates that microarray data alone do not provide a comprehensive picture of biological regulation. However, microarrays are revolutionary as they allow us to study the activity levels for the entire repertoire of genes.

3. Mathematical modelling of genome expression

The purpose of this section is to introduce a general framework to describe the time evolution of RNA transcription or expression levels. We first introduce a model of genome expression in terms of its components as illustrated in Fig. 3. However, in later sections we are going to modify this ‘classical’ approach by focussing on temporal relations.

The first step in our approach is to view genome expression as a system defined by a set or sets of (material or abstract) objects and relations defined on these sets. We denote these sets with uppercase letters such as A and B with elements $a \in A$ or $A = \{a\}$. The mapping $f: A \rightarrow B$ describes how pieces of DNA,² represented by the objects $a \in A$ are associated with the RNA transcript for

which they code. A mathematical representation of such knowledge employs the following standard notation:

$$f: A \rightarrow B$$

$$a \mapsto b = f(a) \quad (1)$$

Modelling transcription, a particular aspect of a genetic pathway or ‘genetic-network’, is then the process of selecting subsets of A , B and identifying the map f . The exponential $B^A = \{f: A \rightarrow B\}$ denotes the set of all maps from A to B and, therefore, represents all transcription processes which are realisable by the cell. The transcription mapping f can be subject to changes, either disturbances such as mutations or deliberate changes to the cell’s operating conditions. Both cases are regulated by the cell. We next introduce the translation step of transcripts in B to a set of proteins in C . More specifically, we are going to assume that the translation map $g: B \rightarrow C$ describes the synthesis of all those components which are essential for transcription to take place, e.g. repressors or RNA polymerase. Considering the elements of C as realising the map f , we can identify the co-domain of g with B^A , leading to the following illustration of the transcription–translation process:

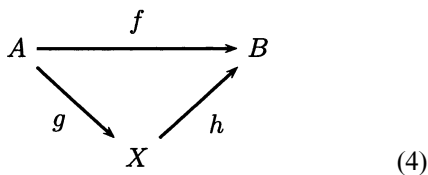
² We deliberately avoid the association of transcripts with genes and instead refer to pieces of DNA, i.e. subsets or sections of the genome. The reason is that genes code for more than one product and hence $f: A \rightarrow B$ is a relation. Our formulation allows us therefore, to treat the transcription mapping f as one-to-one (injective). The term, ‘gene’ appears inappropriate here and instead of looking for a unit that is the causal agency of proteins, in this paper we are going to replace a material object by a sequence of measurements which generates it.

$$y_t = \sum_{k=0}^{t-1} D_{t-k} u_k, \quad t \in \mathbb{Z}_+ \quad (3)$$

where $D_t \in \mathbb{R}^{p \times m}$ denote the coefficient matrices which characterise the process. For each t , Eq. (3) specifies a set of q equations in $q \times m$ unknowns of the matrix D_t . With reference to our more abstract model Eq. (1) above, we find that Eqs. (1) and (3) are isomorphic, i.e. there exist a one-to-one correspondence:

$$f \cong \{D_1, D_2, D_3, \dots\}.$$

The measurements of input–output data (a, b) describe the system in an external sense. The output of a system, in general, depends on both the present input u_t and the past history of the system. To allow us to present inner relations we say, therefore, that the present output depends on the state of the system, and define the (present) state of the system as that part of the present and past history which is relevant to the determination of present and future outputs. A state is defined subsequently by a set of internal or state variables which must not necessarily be directly observable (see Casti (1992) and Kailath (1980) for further details). The problem of explaining the internal dependencies, which generate the observed behaviour, using a mathematical model called the realisation. This concept is a straightforward extension of the input–output map $f: A \rightarrow B$ by adding a set of states and two new maps, g and h connecting this state-space with the input and output space⁴:



⁴ Although every attempt was made to have unambiguous, consistent notation throughout the paper, note that the output map h is unrelated to the evaluation map discussed earlier and is also unrelated to the impulse response function of dynamic systems, discussed further below.

As before, $a = [u_0, u_1, \dots]$, $a \in A$, and $u_t \in U$ are measurements of the input variables at instances of time $t \in \mathbb{Z}_+$. Similar $b = [y_1, y_2, \dots]$, $b \in B$ and $y_k \in Y$. The measured data we have available for the identification of system parameters are $\{(u_t, y_{t+1})\}$. In the diagram (4), X denotes the state-space and the map g is assumed to be surjective (onto X), i.e. more than one input sequence $a \in A$ can map into the same state $x \in X$. The output map h is one-to-one (injective); i.e. any $x \in X$ maps to exactly one output sequence $b \in B$.

For the representation (4) to give rise to the observations $\{(u_t, y_{t+1})\}$, we must be able to construct the state-space X and the maps

$$\begin{aligned} \phi: X \times U &\rightarrow X \quad \text{such that} \quad x_{t+1} = \phi(x_t, u_t) \quad (5) \\ h: X &\rightarrow Y \quad y_t = h(x_t) \end{aligned}$$

In other words, given the present state $x_t \in X$ and input $u_t \in U$ the (non-linear) map ϕ determines the next state and for every state x , the output map h determines an output y_t . It is usually assumed that $X = \mathbb{R}^n$ and thereby any state can be represented as a point⁵ in X . Note that the concept of state is a general notion, defining a set of n state-variables such that the knowledge of these variables at some initial point in time $t = t_0$ together with the knowledge of the input for $t \geq t_0$ completely determines the behaviour of the system for any time $t \geq t_0$. State variables need not be physically measurable or observable quantities.

State-space equations Eq. (5) form the basis for two well established conceptual frameworks: automata theory and control theory. An automaton is a discrete–time system with finite input and output sets U and Y , respectively. In this context, ϕ is referred to as the next-state function. If at any time t the system is in state x_t and receives input u_t , then at time $t + 1$ the system will be in state $\phi(x_t, u_t)$. We say the automata is finite if X is a finite set.⁶ Automata theory has been used to

⁵ A dynamical system is finite dimensional if X is a finite dimensional linear space; it is finite state if X is a finite set. It X, U , and Y are finite sets and the system is discrete time, it is known as a (finite) automation.

⁶ The state of a linear dynamic system, continuous-time or discrete-time evolves in \mathbb{R}^n , whereas the state of an automaton resides in a finite set of ‘symbols’.

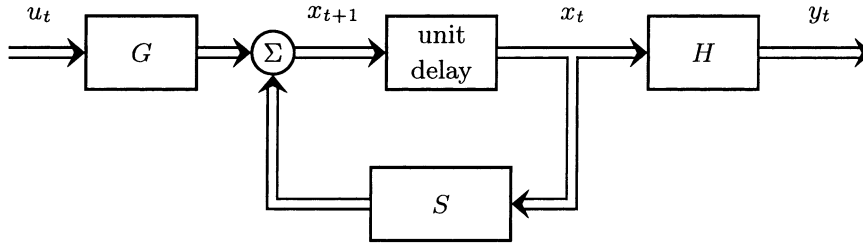


Fig. 4. Matrix block diagram of the general linear discrete-time dynamic system.

model numerous systems including ‘gene networks’ (Thieffry, 1999; Hatzimanikatis et al., 1999). A drawback to validate such models with data, for example from microarrays, is the finiteness of spaces in which the inputs and outputs take their values. Measured values are first quantized and then discretized but with typically weak signals and few samples, the loss of information can be unacceptable.

We now consider the (linear) dynamic systems approach which has been a central theme in control theory and for which a wealth of theory and experience in applications exists. For linear, time-invariant systems, the state-equation and output map Eq. (5) may be simplified to

$$x_{t+1} = Sx_t + Gu_t \quad (6)$$

$$y_t = Hx_t \quad (7)$$

where x_t denotes the state-vector of size n and, therefore, S is the $n \times n$ state or system matrix, G is $n \times m$ and is called input matrix while H is the $p \times n$ output matrix. These matrices determine how inputs are coded into states and states are decoded into outputs. As can be readily seen in Fig. 4, most important is the system matrix S , which determines the internal dynamics and, therefore, how present and past are related. While the discrete-time representation is based on first-order difference equations, the continuous-time equivalent of Eq. (6) describes the system dynamics by a first-order vector differential equation.

$$\frac{d}{dt} x = Sx + Gu \quad (8)$$

$$y = Hx$$

The difference between discrete-time and continuous-time is not relevant for our discussion and we only note that Eq. (8) suits our discussion about the role of change in modelling causal entailment (cf. Section 2). With the assumptions⁷ implicit in Eq. (6), the equivalence to the input-output model Eq. (3) is established by

$$D_t = HS^{t-1}G \quad \forall t \in \mathbb{Z}_+$$

In Fig. 4, a general linear discrete-time dynamic system like Eq. (6) is illustrated by a block diagram. To illustrate state-space modelling we consider an artificial noise-free continuous-time system with two inputs and two outputs. Experiments are used to determine the response of the system to impulses. The impulse response plays an important role in that it allows us to fully characterise the system through the convolution integral and thereby defines the mapping $f: U \rightarrow Y$:

$$y_t = \int_0^t h_{t-\tau} u_\tau d\tau \quad \text{and in discrete-time} \\ y_t = \sum_{k=0}^{t-1} h_{t-k} u_k \quad (9)$$

In Eq. (9), h_t denotes the impulse response for a single-input, single-output system. For a multi-variable discrete-time system with m inputs we have:

⁷ Mathematical assumptions and their relation of phenomena in the natural system under consideration should consider issues such as observability, reachability initial conditions, etc. For a detailed account of such questions and examples of state-space modelling in the engineering sciences, the reader is referred to any book on linear systems and control theory. See for example (Kailath, 1980; Wolkenhauer, 2001a,b). A classic text for system identification is (Ljung, 1998).

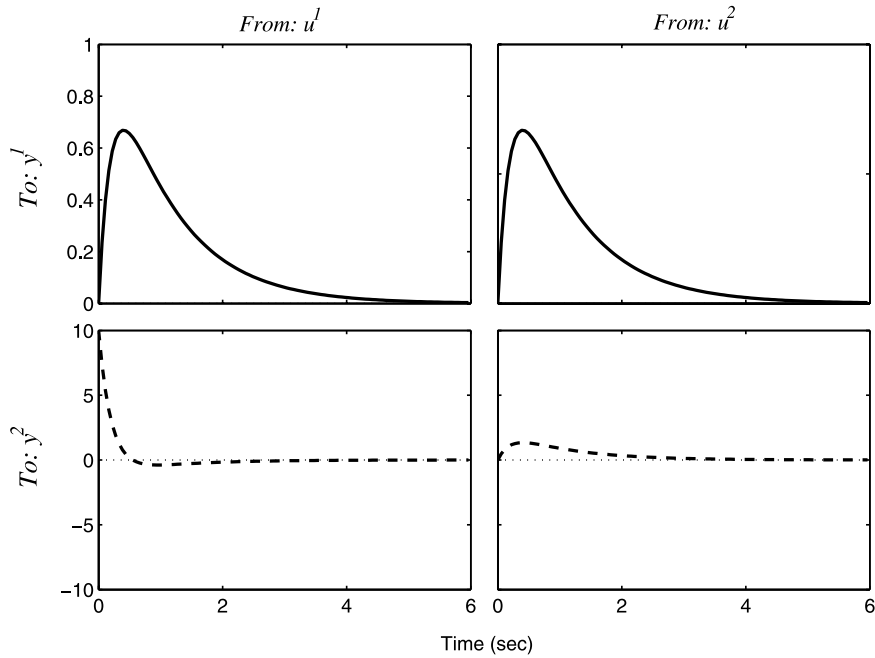


Fig. 5. Impulse responses of a two-input, two-output system.

$$y_t^j = \sum_{r=1}^m \sum_{k=0}^{t-1} h_k^{rj} u_k^r$$

where h^{rj} denotes the impulse response from input r to output j . In Fig. 5, four such impulse responses are shown for a two-input, two-output system. Accompanied with some mathematical assumptions we find also that for our model Eq. (6) the relation $h_{t-r} = HS^{t-k-1}G$ holds such that

$$y_t = HS^t x^0 + \sum_{k=0}^{t-1} HS^{t-k-1} G u_k \quad t = 1, 2, 3, \dots,$$

where S^r is the r -fold matrix product $S \times S \dots$ (r -times) and x^0 is the initial state at time $t = 0$. In Fig. 5 we have four plots showing us the impulse response patterns observed for the two output variables. A possible realisation for this system is given by the state-equations Eq. (8)

$$\frac{d}{dt} \begin{bmatrix} x^1 \\ x^2 \\ x^3 \\ x^4 \end{bmatrix} = \overbrace{\begin{bmatrix} -6 & -2.5 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 0 & -6 & -1.25 \\ 0 & 0 & 4 & 0 \end{bmatrix}}^s \begin{bmatrix} x^1 \\ x^2 \\ x^3 \\ x^4 \end{bmatrix}$$

$$+ \overbrace{\begin{bmatrix} 4 & 0 \\ 0 & 0 \\ 0 & 2 \\ 0 & 0 \end{bmatrix}}^G \begin{bmatrix} u^1 \\ u^2 \end{bmatrix} \quad (10)$$

$$\begin{bmatrix} y^1 \\ y^2 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 0.625 & 0 & 0.625 \\ 2.5 & 0.625 & 0 & 1.25 \end{bmatrix}}_H \begin{bmatrix} x^1 \\ x^2 \\ x^3 \\ x^4 \end{bmatrix} \quad (11)$$

In these equations, the superscripts in x^i or u^r denote the i th state or r th input, respectively. With four internal states, we notice that the two inputs only act on the first and third state. While the first output is determined by a combination of state two and four, the second output is determined by a linear combination of three states. Once a mathematical model is established, we can simulate responses to arbitrary input pattern. Such simulations can be matched with experimental observations and thereby help in experimental design. In Fig. 6, three plots show the response to a unit step on input 1, input 2 and then a simultaneous positive/negative input step at both inputs.

The model (4) is clearly more flexible, allowing us to answer questions about the influence inputs have on the system and what the observed outputs tell us about the system. The theory of state-space modelling provides us also with a statistical theory to account for the observed randomness as well as correlations in the observations. We do not detail this extension but point out that the theory for stochastic linear systems is well developed and documented in virtually every engineering textbook on system modelling and

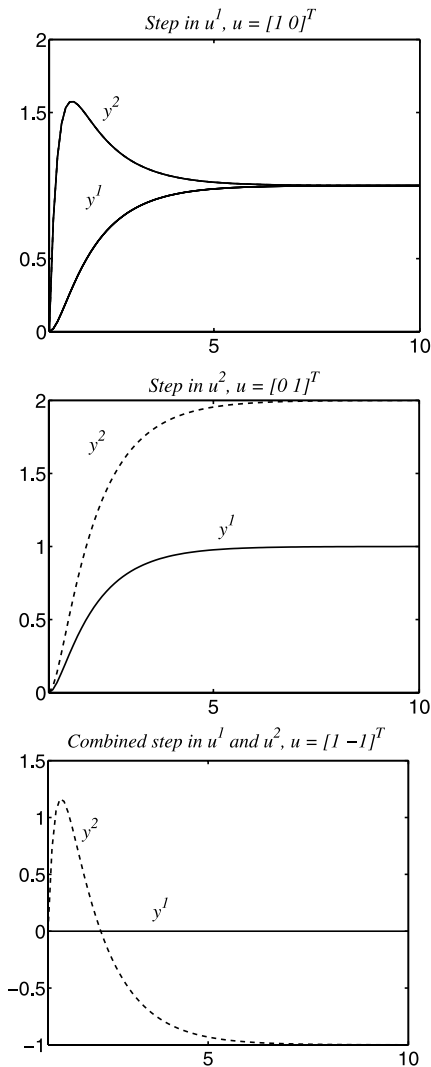


Fig. 6. Step response pattern for a two-input, two-output system (time in seconds).

identification. As indicated above, it is also particularly well suited for multivariable systems. Of utmost importance in state-space modelling is the formation of the state-space X , i.e. how an input is encoded into an element of X . The mathematical theory for this is well developed and allows numerous convenient realisations. The more challenging task is, therefore, to attach biologically meaningful interpretation to the states. A drawback of state-space modelling is further that it requires us to have knowledge of the set of independent input variables and one or more dependent (output) variables. DNA microarray experiments provide us, however, with a large set of expression profiles and to identify dependencies among variables (e.g. RNA transcripts) is often the main purpose of such experiments. We would, therefore, like to start with a conceptual framework for mathematical modelling that defines a dynamic system as a family of time-series. Such a mathematical framework has been developed by Jan Willems (Willems, 1991) and is introduced in Section 6. As before, this approach is not an alternative but rather closely related to our previous discussion. The state-space model Eq. (6) is obtained as a special case and popular regression models can also be obtained. Most ideas on gene network modelling that have been published in the context of microarray data describe AR models and the conceptual framework outlined in this paper provides a common mathematical basis.

5. Dynamic modelling of microarray data

In this section, we show that a number of proposed gene-network models are special cases of a more general modelling framework. The purpose of such models is usually to model specific aspects of genome expression and regulation or to identify interactions from microarray data. For a good introduction to the subject of genome expression and regulation we refer to the book by Ptashne (1992). An interesting survey and review of computational studies of gene regulatory networks can be found in Hasty et al. (2001).

In the theory of dynamic systems we generally have to make a decision whether to regard the

process as a deterministic⁸ non-linear system but with a negligible stochastic component or to assume that the nonlinearity to be only a small perturbation of an essentially linear stochastic process. Genuine non-linear stochastic processes have not yet been shown to be applicable for practical time-series analysis. Although natural phenomena are never truly linear, for a very large number of them linear (stochastic) modelling is often the only feasible option. Complex behaviour can often be captured by choosing the order of the model high enough and a time-varying system can be identified using recursive updating of the model. This would correspond to a local linearization of the non-linear system. For gene networks a number of models have been proposed, including finite-state machines fitting the definition of an automaton, given above (D’haeseleer et al., 2000; Savageau, 1998; Wessels et al., 2001 and references therein). Typically for such ‘Boolean model’ a gene is considered switched ‘on’ if it is being copied into RNA, and ‘off’ if not. Relative transcription levels and changes are more difficult to capture with this methodology and one would expect that the more intricate aspects of gene regulation are dependent not on whether a gene is transcribed but rather the level of transcription and relative degree in relation to other transcripts.

For reasons outlined above, we, therefore, consider a network structure in which the effect (expression level) of a gene is modelled as a linear combination (weighted sum) of other expression levels (Wessels et al., 2001):

$$x_{t+1}^i = \gamma^i \cdot \phi \left(\sum_{j=1}^n s^{j,i} x_t^j + \sum_{r=1}^m g^{j,i} u_t^r + b^i \right) - \underbrace{\lambda_i x_t^i}_{\text{degradation}} \quad (12)$$

where, x_t^i is the expression level of the i th EST at time t ; u_t^r the r th external input at time instance t ; $\phi(\cdot)$ is called activation function, $\phi(z) = 1/(1 +$

$\exp(-z))$; γ^i denotes the rate constant of gene i ; u_t^r denotes the r th external input at time instance t ; $g^{r,i}$ describes the influence of the r th external input on gene i ; b_i is the basal expression of gene i ; λ_i is the degradation constant of the i th gene expression product.

This can be further simplified to

$$x_{t+1}^i = \sum_{j=1}^n s^{j,i} x_t^j + \sum_{r=1}^m g^{j,i} u_t^r$$

Ignoring any inputs to the network we obtain an even simpler AR model (Van Someren et al., 2000):

$$x_{t+1}^i = \sum_{j=1}^n s^{j,i} x_t^j \quad (13)$$

where x_t^j is the expression level of gene j at time t and weight $s^{j,i}$ indicates the influence of gene j on gene i . If $s^{j,i} > 0$ the control action of gene j activates gene i while for $s^{j,i} < 0$ gene j has an inhibiting effect. With this interpretation the matrix S is called ‘gene regulation matrix’, ‘weighting matrix’ or ‘interaction matrix’. The total number of genes considered is n . Parameter estimation for such models is discussed in Section 6. Comparing Eqs. (12) and (13) with the state difference-equation Eq. (6), we find that the linear gene network model is in fact a special case of the discrete-time state-space model, where $S = [s^{i,j}]$, $G = [g^{i,j}]$. If we consider mRNA transcription levels as measurements of state variables, and let the gene products be denoted by y , our state-space model Eq. (7) describes the more realistic situation that not all transcripts are necessarily translated into protein. This situation was in fact already established in our discussion of the abstract representation 4 and illustrated by the example Eqs. (10) and (11). A state $x \in X$, therefore, represents an encoding of the input u in the most compact form that is consistent with the production of the output (protein) y from the input via the input/output map f . More specifically, each state is an equivalence class of inputs, where we regard two inputs u and u' as equivalent if they generate the same output under f , i.e. $u \approx u'$ if and only if $f(u) = f(u')$. Thus, the encoding of $u \mapsto x = [u]_f$ represents the memory, the way the system remembers the input u .

⁸ A system is said to be deterministic if its state and output at any time t can be determined with certainty from a complete knowledge of its state at some initial time t_0 and its input over the time interval $[t_0, t)$. Conversely, a system is stochastic (or nondeterministic) if such knowledge of state and input suffices only to provide a statistical description of the state and output at time t .

Before we discuss how network models can be identified from data, we have a look at the data themselves. Here we focus on time-course experiments using DNA microarrays and which produce a vector-valued time-series

$$\mathcal{W} = [w_1, \dots, w_N], \quad w_t \in W = \mathbb{R}^n. \quad (14)$$

At each of the $t = 1, \dots, N$ consecutive measurements, the vector w_t represents the expression levels of n EST.⁹ The vector-valued time-series can also be represented as a matrix, called the genome expression matrix \mathcal{W} . A row of the matrix \mathcal{W} is referred to as an expression profile. The matrix \mathcal{W} is not obtained directly from the microarrays but follows from a number of ‘pre-processing’ steps.

1. Treatment of outliers and missing data.
2. Averaging over array replicates.
3. Normalisation against some chosen reference.
4. Addressing the dimensionality problem.
 - (a) Smoothing or re-sampling.
 - (b) Selection of ‘informative’ signals.
5. Followed by multivariate data analysis,
 - (a) classification;
 - (b) network modelling.

In terms of column vectors, w_t , of expression matrix $\mathcal{W} = [w_1, \dots, w_N] \in \mathbb{R}^{n \times N}$, the linear model defined by Eq. (13) can be re-written as:

$$w_{t+1} = S w_t \quad t = 1, \dots, N-1 \quad (15)$$

In general, the system of Eq. (15) will be under-determined which means that there exist multiple solutions of S that fit the data in \mathcal{W} equally well. An advantage of the simple linear regression model Eq. (15) is that we can visualise the interactions of a relatively large number of variables. One possible representation is illustrated in Fig. 7.

We refer to Wessels et al. (2001) for a more comprehensive comparative study of network models and mention only two possibilities that arise from the linear model Eq. (15). While in

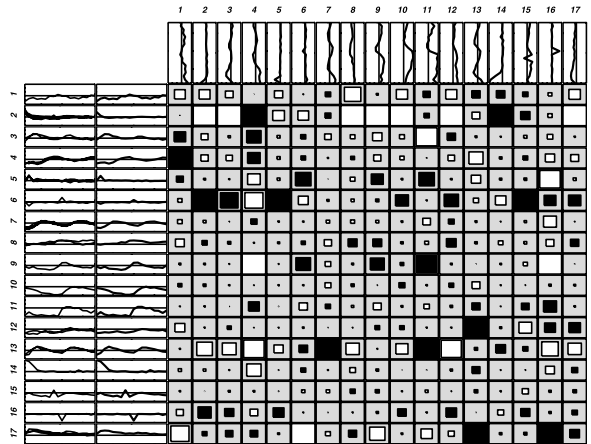


Fig. 7. Using a simple AR model, we can visualise the interactions of a relatively large number of variables. While the colour (white or black) indicates direction of interaction, i.e. whether the interaction is positive or negative, the size of the square indicates the strength of interaction (Van Someren et al., 2000).

(van Someren et al., 2000) matrix S describes relationships between ‘genes’, in (Holter et al., 2001) this weighting matrix is used to describe the time evolution and dependencies of the characteristic modes of \mathcal{W} . These characteristic models are obtained from the singular value decomposition (SVD) of \mathcal{W} (Alter et al., 2000; Holter et al., 2001). Other authors have developed automata models for gene networks (Liang et al., 1998; Thieffry, 1999). Considering the often small number of samples and measurement uncertainty, a linear regression model has the advantage that we can be precise about uncertainty using confidence intervals. These intervals are obtained either from statistical considerations (Ljung, 1998; Draper and Smith, 1998) or, with insufficient data to estimate distributions, we may choose deterministic bounds (Walter, 1990). The inference of gene networks from microarray data is particularly difficult due to the large number of genes involved in a typical network and the few time points for which measurements can be obtained.

We note that any consideration of noisy data has been absent from our discussion so far. State-space models can be extended to include noise added to input as well as output measurements. Subsequently the whole process of identifying

⁹ Expressed sequences tags (ESTs) are cDNAs amplified by PCR that represent part of gene. An EST is, therefore, a DNA sequence that is transcribed into mRNA and can be placed in spots composing the microarray. Here we shall use the term EST instead of and to avoid the term ‘gene’, which may be composed from a number of separate DNA segments.

parametric models from time-series data can be formulated in terms of random variables, stochastic processes and expectations thereof (Wolkenhauer, 2001a,b). Although the mathematical formulations are well developed and elegant, the semantics of such models are not free from problems. We shall confine ourselves to one example, which may have an interesting biological interpretation. An intrinsic feature of a time-series is that adjacent observations are dependent and it is the purpose of mathematical models to describe this dependency. Based on an idea by Yule (Box et al., 1994), for a time series in which successive values are highly dependent, observations are regarded as generated from a series of independent random ‘shocks’ u_t . The shocks are drawn randomly from a fixed distribution, usually assumed to follow the Gaussian probability law. Such a sequence of random variables u_t, u_{t-1}, \dots is called a white noise stochastic process. The white noise process is not what is observed, say in expression profiles, but is first transformed by a linear filter which in fact generates a weighted sum of random shocks:

$$w_t = \mu + u_t + \Psi_1 u_{t-1} + \Psi_2 u_{t-2} + \dots \quad (16)$$

where μ determines the ‘level’ of the process. To obtain the popular AR models from this representation, let $\tilde{w}_t = w_t - \mu$, then

$$\tilde{w}_t = \phi_1 \tilde{w}_{t-1} + \phi_2 \tilde{w}_{t-2} + \dots + \phi_l \tilde{w}_{t-l} + u_t \quad (17)$$

which is a special case of the linear filter Eq. (16), obtained by.

1. Eliminating \tilde{w}_{t-1} on the right-hand side of Eq. (17) by substituting

$$\tilde{w}_{t-1} = \phi_1 \tilde{w}_{t-2} + \phi_2 \tilde{w}_{t-3} + \dots + \phi_l \tilde{w}_{t-l-1} + u_{t-1}$$

2. Repeating the previous step for $\tilde{w}_{t-2}, \tilde{w}_{t-3}$ and so forth, leading to an infinite series in the u ’s.

A first-order AR model, denoted AR(1), is then defined as

$$\tilde{w}_t = \phi_1 \tilde{w}_{t-1} + u_t = u_t + \phi_1 u_{t-1} + \phi_1^2 u_{t-2} + \dots$$

The interpretation of this approach to time-series analysis is then that ‘order’ (in form of the time-series \tilde{w}_t) emerges from ‘randomness’ (the white noise process).

Choosing a suitable modelling framework (linear vs. non-linear, stochastic vs. deterministic, state-space vs. input–output vs. AR models) is, therefore, an important issue. With uncertainty in modelling being a certainty, semantic aspects of models, embedded in philosophical or epistemological questions, should be guided by biological considerations, and less by mathematical convenience and computational costs. While in engineering the interpretation of models usually does not matter, in modelling cellular systems it is a central problem. The next section is, therefore, to consider a more general and comprehensive framework of mathematical modelling.

6. The behavioural approach to dynamic systems

Studying interactions of RNA transcripts we usually do not know the signal flow as required for the previous model in order to describe inputs and outputs. The model introduced in this section, referred to as the behavioural approach and first developed by Willems (1991), defines a dynamic system as a subset of time-trajectories. Previously introduced models emerge from this description as special cases.

Our previous discussion concluded that the internal structure of a mathematical model is to provide an explanation of observed phenomena but that it is also a matter of choice. This suggests that the modelling process itself is of interest and we shall here provide a more formal framework in which to discuss this process. A mathematical model is a pair (O, \mathcal{B}) with O the universe of outcomes and \mathcal{B} the behaviour. The phenomenon under consideration produces elements in the set O . A model recognises a certain subset \mathcal{B} of O and thereby describes among all possible outcomes which are possible and which are not. As such, a natural system, characterised by the behaviour \mathcal{B} , is represented by the constraints it imposes on the environment O . For example, if we are to model mRNA abundance on a single spot on a microarray, we may choose $O = \mathbb{R}$ but since there can only be a finite amount of RNA in the cell, $\mathcal{B} = (0, b]$ where b stands for some upper bound.

Our previously discussed dynamic systems are mathematical models for phenomena that evolve in time and hence are formally represented by the triple (T, W, \mathcal{B}) , with time set¹⁰ T , signal space W and behaviour $\mathcal{B} \subseteq W^T$, where W^T denotes the set of all mappings from T to W . In microarray time course experiments, the expression matrix Eq. (14) is obtained from $t \mapsto w_t \in W$, $W = \mathbb{R}^n$. Model structures, such as AR- and state-space representations, are obtained in the behavioural framework from behavioural equations such as

$$\Phi_l w_{t+2} + \Phi_{l-1} w_{t+1} + \dots + \Phi_0 w_t = 0$$

which defines an AR model of order l . The parameterisation is specified by $\Phi_0, \Phi_1, \dots, \Phi_l \in \mathbb{R}^g \times n$. Equivalently, we have the state-space representation

$$x_{t+1} = Sx_t + Gu_t$$

$$w_t = Hx_t$$

with input u , x the state, and w the observed time-series.

In order to capture the modelling process itself, let \mathcal{M} be a family of models such that each element $M \in \mathcal{M}$ denotes a mathematical model (O, \mathcal{B}) . A parameterisation (P, ρ) of \mathcal{M} consists of a parameter space P and a surjective map $\rho: P \rightarrow \mathcal{M}$. The fact that ρ is a surjection but not necessarily injective (one-to-one) implies that more than one parameter (set) $p \in P$ can lead to the same model $M \in \mathcal{M}$. Through experimentation we obtain measurements that constitute a data set $D \subset O$. Hence, each element of D is a subset of O . The modelling process is then described as follows: Given the model class \mathcal{M} and a data set $D \in \mathcal{D}$, which model(s) are an appropriate representation of the phenomena that produced the data set? We, therefore, define the modelling procedure by the map $P: D \rightarrow 2^{\mathcal{M}}$, from the data class D to the collection of all subsets of \mathcal{M} (denoted $2^{\mathcal{M}}$). We call $D \subseteq 2^O$ a data class. Following Willems' behavioural framework, a model (O, \mathcal{B}) is called unfalsified by the data $D \subseteq O$ if $D \subseteq \mathcal{B}$. We will call a model (O, \mathcal{B}_1) more powerful than

(O, \mathcal{B}_2) if $\mathcal{B}_1 \subseteq \mathcal{B}_2$. In other words, a model is more powerful if it is more restrictive in respect to possible outcomes. Starting from these most general definitions, Willems and others have developed concepts to identify a most powerful unfalsified model (MPUM) from data. This includes the AR and state-space model structures introduced in previous sections. The mathematics is somewhat abstract and as these ideas are relatively recent, its practical use in system identification is yet to be confirmed. To complete the picture, we relate the abstract mathematical model $M = (O, \mathcal{B})$ to our linear time-invariant dynamical systems discussed in the previous section. The universe of discourse in this case equals $(\mathbb{R}^n)^{\mathbb{Z}}$, and the data set $D \subseteq (\mathbb{R}^n)^{\mathbb{Z}}$, is any family of n -vector time-series Eq. (14), $w_t: \mathbb{Z} \rightarrow \mathbb{R}^n$, $t = 1, 2, \dots, N$.

To decide upon a model, during its identification from data, Willems introduced two measures, of complexity and misfit. The complexity c is a mapping $c: \mathcal{M} \rightarrow L_c$ with L_c the complexity level set, e.g. $L_c = [0, 1]$. The misfit ε is a mapping $\varepsilon: D \times \mathcal{M} \rightarrow L_m$ with L_m the misfit level space, e.g. $L_m = [0, 1]$. While $c(M)$ quantifies the power of the model, $\varepsilon(D, M)$ indicates how far the model M fails to explain measurements D . In modelling we have principally two choices to proceed,

1. to fix the maximal admissible complexity; or
2. to fix the maximal tolerated misfit.

The first methodology is usually followed in system identification; a parameterised class of models is fixed and then parameters are chosen which minimise some criterion. Such parameter estimation method defines the number of free parameters as the maximal admissible complexity. All common systems identification methods leading to AR, ARMA, ARMAX models using a statistical set-up are based on the minimisation of complexity.

Viewing observations w_t as vectors, i.e. points in \mathbb{R}^n , the estimation of parameters consists of fitting a linear subspace to a set of observed vectors w_t , $t = 1, 2, \dots, N$. The most common approach to this subspace fitting problem is regression (Wolkenhauer, 2001a,b). It is customary to assume that an observation w_t is composed of two components, regressor $u_t \in \mathbb{R}^m$ and regressand

¹⁰ $T = \mathbb{R}$ for continuous-time and $T = \mathbb{Z}$ for discrete-time systems.

$y_k \in \mathbb{R}^q$. This notation includes the case of identifying matrix S in the network model Eq. (15). We ignore here the (important) question of how to decide which EST expression profiles are to be considered as independent (or ‘inputs’) and which as dependent (or ‘outputs’). We look for a matrix $L \in \mathbb{R}^n$ such that the graph of $y = L u$ fits the data in a suitable sense. In regression, an optimal matrix, denoted L^* , is obtained from minimising the criterion

$$\frac{1}{N} \sum_{k=1}^N \|y_k - L u_k\|^2$$

Using the Euclidean norm (evaluating the distance between y_k and $L u_k$) and assuming that $1/N \sum_{k=1}^N u_k u_k^T$ is invertible, L^* is obtained as

$$L^* = \left(\frac{1}{N} \sum_{k=1}^N y_k u_k^T \right) \left(\frac{1}{N} \sum_{k=1}^N u_k u_k^T \right)^{-1}$$

The complexity of the model, its order, is fixed in this setting. The graph of $y = L u$, representing the model, is an m -dimensional subspace of \mathbb{R}^n where $n = m + q$. This subspace minimises the misfit

$$\varepsilon(M, D) = \sqrt{\frac{1}{N} \sum_{k=1}^N \|e_k\|^2} \quad \text{with } e_k = y_k - L u_k$$

The matrix L^* is completely specified by the fact that the error r_k is uncorrelated by the data, i.e. $1/N \sum_{k=1}^N e_k u_k^T = 0$. In other words, it is assumed that the u_k 's have been observed without error, but that the measurements have reproduced the model outcomes $L u_k$ only up to the error $e_k = y_k - L u_k$. The model is identified by minimising the average squared error.

In microarray data analysis we usually have more ESTs than samples and hence $n \gg N$. For network modelling it is necessary to reduce n to a reasonably small figure. This can be done in numerous ways. Testing profiles to whether they reach a certain threshold throughout the experiment is often used but is rather inefficient for time-series data. Time introduces an order to the columns of the expression matrix and, therefore, provides additional information in the analysis: weak signals may still display a deterministic pattern over time. Many expression profiles will be non-informative, i.e. the signal can be considered

as a white-noise process and statistical tests can be employed to test for stationarity and randomness in order to ‘filter’ the expression matrix. Although this procedure considers each expression profile individually, the number of samples N can be too small for the tests to be reliable. Clustering algorithms can be employed to identify very similar expression profiles and replace them by a prototypical profile (van Someren et al., 2000). Yet another method to reduce the dimensionality is subspace projection by means of factorisations of the expression matrix \mathcal{W} (Alter et al., 2000; Holter et al., 2000, 2001).

7. Conclusions

Studying natural systems we have two mutually dependent sources of data and information: measurements and observation. With the aim of describing causal entailment in natural systems, biologists most commonly describe their observations, as part of an empirical analysis, using natural language, diagrams, and pictures. On the other hand, the engineering sciences have used inferential entailment in formal systems to combine measurements with mathematical models and have thereby presented their knowledge in the form of axioms, equations, and diagrams. The nature and outcome of these two forms of knowledge representation are in fact orthogonal: while biologists successfully describe very complex systems (qualitatively), engineers have only managed to describe simple systems but quantitatively. Applying mathematical models to biological systems we often find that although our inference is precise, conclusions tend to be inaccurate. In contrast, accurate biological knowledge is obtained by imprecise reasoning. It is then the combination of both forms of explanation (Schopenhauer's second and third form of the principle of sufficient reason) that is most promising for research in genomics.

Systems and control theory is a way of thinking, not a collection of facts about the real-world. Genomics, in particular the study of genome expression and regulation, deals with complex interactive phenomena. It is impossible to study quantitative relationships between relevant vari-

ables without reference to the context; nor is it possible to perform experiments or make direct observations that would isolate such relationships accurately. By constructing models for time-series we may hope to gain indirect access to the desired quantitative biological relationships, represented by the structure and parameters of the models. Current microarray technology does not permit predictive models but it is the modelling process itself that could help in the biologist's quest. Biologists describe cellular processes in terms of material objects and their spatio-temporal relationships. To represent such knowledge they use diagrams, pictures, and natural language in analysis and inference. Such 'Lego-style' modelling may not be the optimal approach.

Instead of trying to identify 'genes' as causal agents for some function, role, change in phenotype or the cellular response of proteins, we should identify these observations with sequences of events. In other words, instead of looking for a 'gene' (whatever that may mean) that is the reason, explanation or cause of some phenomenon we should seek an explanation in the dynamics (sequences of events ordered by time) that led to it. For obvious reasons and its success, molecular biology has focussed on the physico-chemical characterisation of 'parts and components' but with the emergence of new technology a shift of focus to an understanding of functional activity is taking place. DNA microarrays provide us with time-series data and the difficulties in a) finding a mathematical definition of co-expression and b) the difficulties to match 'genes' in clusters of expression profiles with information of functional classes in databases, suggest that we should focus our attention to the (dynamic) processes that lead to observable changes rather than segments of DNA. Decisions in pathways are unlikely to be conditioned on whether a gene is 'on' or 'off' but instead it will be the level/degree, a relative measure of change (in time), i.e. a process or sequence of events that determines what happens.

It is unlikely that we will soon be able to identify accurate parametric models of 'gene networks' from microarray data and systems biology should be seen as providing a conceptual framework, as 'a way of thinking'. This way of thinking

and simulations could help the biologist in experimental design, deciding which variables to measure, at what point in time to produce a microarray and what expression pattern to look for. The discussion of mathematical models of genome expression and regulation also showed that the concept of a gene is inadequate: genome expression is not just dependent on the properties of molecules but is most appropriately studied in terms of their relationships, i.e. interactions and dynamic changes of their observable effects.

We reviewed several popular methodologies to modelling dynamic systems: the input-output model, state-space model, AR models, automata; all of which are subsumed in a common mathematical framework. The purpose of such a general and unifying framework is to enable better comparisons among techniques and to aid the user in his/her choice of a suitable methodology. Given the uncertainties in data the models should not be considered as definite descriptions of cellular processes but rather as one approach that helps us to understand the behaviour and devise experiments to test our hypotheses. The biggest challenge for the application of these models to microarray data is the problem of dimensionality: the large number of genes combined with usually a small number of measurements in time-course experiments.

8. Mathematical conventions and notation

For most parts of the document, standard mathematical notation is used, i.e. \mathbb{R} for the real line, \mathbb{Z} for integers and \mathbb{Z}_+ for non-negative integers. \mathbb{R}^r denotes the Cartesian product $\mathbb{R} \times \mathbb{R} \times \dots \times \mathbb{R}$ r times. Curly brackets $\{a, b, \dots\}$ denote an unordered list or set of elements while $[a, b, \dots]$ denotes an ordered sequence or vector. When we want to show that a is an element of set A , we write $a \in A$. 2^A denotes the power set, the set of all subsets of A . Other symbols commonly used are \therefore , the: meaning 'for which' or 'given'; \forall ('for all'); \mapsto ('maps to'). A map (or function) f from A to B is denoted $f: A \rightarrow B$, A is referred to as the domain and B as the co-domain of f . The notation $a \mapsto b$ means that b corresponds to a under

the map $A \rightarrow B$. A relation between sets U_1, U_2, \dots, U_r is written $R(U_1, U_2, \dots, U_r)$ and is a subset of the Cartesian product U_1, U_2, \dots, U_r . A binary relation $R(U, U)$ on U is called an equivalence relation on U if it is reflexive ($R(u, u)$ holds true), symmetric ($R(u, u') = R(u', u)$), and transitive (if $R(u, u')$ holds and $R(u', u'')$ holds, then also $R(u, u'')$ holds). Then for every $u \in U$, the set $[u]_R = \{u' : R(u, u') \text{ holds true}\}$ is called the equivalence class defined by u for the relation $R(U, U)$. A relation is more general than a function (mapping or map) in that it is multi-valued map, for $R(A, B)$ any $a \in A$ can map into more than one b . A mapping of A to B is a subset of the Cartesian product $A \times B$ in which each element of A occurs at most once as a first element in the ordered pairs of the subset (graph) $\{(a, b = f(a))\}$. $t \in T$ denotes time and a dynamical system is continuous-time if $T = \mathbb{R}$ and discrete-time if $T = \mathbb{Z}$. In many applications of system theory the distinction between continuous-time and discrete-time systems is not critical; the choice is often governed by mathematical convenience or ease of presentation. We do not use different symbols or fonts to distinguish matrices, vectors and single-valued variables but while y_t denotes in general a vector, y_t^i can denote an element of y_t . For a vector y_t , y_t^T denotes its transpose. Transposing a n by N (denoted $n \times N$) matrix V , we obtain $N \times n$ matrix V^T with rows and columns interchanged.

Acknowledgements

This paper was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant GR/N21871 ('Genetic Systems').

References

- Alter, O., Brown, P.O., Botstein, D., 2000. Singular value decomposition for genome-wide expression data processing and modelling. *Proc. Natl. Acad. Sci.* 97 (18), 10101–10106.
- Box, G.E.P., Jenkins, G.M., Reinsel, G.C., 1994. *Time Series Analysis*, third ed. Prentice Hall, New Jersey.
- Casti, J.L., 1992. *Reality Rules*. Wiley-Interscience, Chichester, UK.
- D'haeseleer, P., Liang, S., Somogyi, R., 2000. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16 (8), 707–726.
- Draper, N.R., Smith, H., 1998. *Applied Regression Analysis*. Wiley, New York.
- Hasty, J., McMillen, D., Isaacs, F., Collins, J.J., 2001. Computational studies of gene regulatory networks: in numero molecular biology. *Nat. Rev. Genet.* 2 (4), 268–279.
- Hatzimanikatis, V., Choe, L.H., Lee, K.H., 1999. Proteomics: theoretical and experimental considerations. *Biotechnol. Prog.* 15, 312–318.
- Holter, N.S., Maritan, A., Cieplak, M., Fedoroff, N.V., Banavar, J.R., 2001. Dynamic modelling of gene expression data. *Proc. Natl. Acad. Sci.* 98 (4), 1693–1698.
- Holter, N.S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J.R., Fedoroff, N.V., 2000. Fundamental patterns underlying gene expression profiles: simplicity to complexity. *Proc. Natl. Acad. Sci.* 97 (15), 8409–8414.
- Jacob, F., Monod, J., 1960. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* 3, 318–356.
- Kailath, T., 1980. *Linear Systems*. Prentice Hall, Englewood Cliffs, NJ.
- Kitano, H. (Ed.), 2001. *Foundations of Systems Biology*. The MIT Press.
- Lawvere, S., Schanuel, S., 1997. *Conceptual Mathematics*. Cambridge University Press.
- Liang, S., Fuhrman, S., Somogyi, R., 1998. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. *Pacific Symposium on Biocomputing* 3, pp. 18–29.
- Ljung, L., 1998. *System Identification: Theory for the user*, second ed. Prentice Hall, Upper Saddle River, NJ.
- Monod, J., 1970. *Le hasard et la nécessité*. (Engl. 'Chance and Necessity: An Essay on the Natural Philosophy of Modern Biology'). Seuil, Paris.
- Ptashne, M., 1992. *A Genetic Switch: The λ and Higher Organisms*. Blackwell Science, Oxford.
- Rosen, R., 1958. The representation of biological systems from the standpoint of the theory of categories. *Bull. Math. Biophys.* 20, 317–341.
- Rosen, R., 1985. *Anticipatory Systems*. Pergamon Press, New York.
- Savageau, M.A., 1998. Rules for the Evolution of Gene Circuitry. *Pacific Symposium on Biocomputing* 3, pp. 54–65.
- Schopenhauer, A., 1818. *On the Fourfold Root of the Principle of Sufficient Reason*. Open Court Publishing, Peru, IL (Reprinted 1974).
- Thieffry, D., 1999. From global expression data to gene networks. *Bioessays* 21 (11), 895–899.
- Van Someren, E.P., Wessels, L.F.A., Reinders, M.J.T., 2000. Linear modelling of genetic networks from experimental data. *Proceedings of the Eighth International Conference on Intelligent Systems in Molecular Biology. ISMB'00*, pp. 355–360.

- Walter, E., 1990. Estimation of parameter bounds from bounded-error data: a survey. *Math. Comput. Simul.* 32, 119–468.
- Wessels, L.F.A., Van Someren, E.P., Reinders, M.J.T., 2001. A comparison of genetic network models. *Pacific Symposium on Biocomputing. PSB'01*, 6, pp.508–519.
- Wiener, N., 1948. *Cybernetics: Control and Communication in the Animal and the Machines*. MIT Press, Cambridge, MA.
- Willems, J.C., 1991. Paradigms and puzzles in the theory of dynamical systems. *IEEE Trans. Autom. Contr.* 36 (3), 259–294.
- Wolkenhauer, O., 2001a. Systems Biology: the reincarnation of systems theory applied in biology. *Briefings Bioinformat.* 2 (3), 258–270.
- Wolkenhauer, O., 2001b. *Data Engineering: Fuzzy Mathematics in Systems and Data Analysis*. Wiley, New York.