SYSTEM IDENTIFICATION CHALLENGES FROM SYSTEMS BIOLOGY

Edmund J. Crampin *

* *Bioengineering Institute, The University of Auckland, Private Bag 92019, Auckland, New Zealand*

Abstract: Systems biology is the understanding through computational modelling of the function of biological systems. New high-throughput experimental technologies can measure simultaneously the levels of expression of thousands of genes. The challenge is to extract knowledge from these data sets in order to understand the regulatory machinery of the cell. This article describes recent approaches to gene network modelling, focusing on the issues arising in the attempt to identify regulatory networks directly from high-throughput gene expression data.

1. INTRODUCTION

In recent years, new experimental approaches have been developed that allow large scale quantitative measurement of biological systems, which was not previously possible. In particular, the development of techniques to monitor which genes are actively making proteins in a cell has opened the door to network-based analysis of the cell's regulatory machinery. Systems biology is the term which has been adopted over the past decade or so to describe this new approach to understanding biological function. This review describes the current status of systems biology, focusing on the challenges thrown up in the analysis of these data.

In the past, the study of gene function has revolved around attempts to delete a gene (gene 'knock-out' experiments) in order to determine its putative biological role. This produces a 'parts list' for a biological organism, an effort which has been greatly advanced by completion of various genome sequencing projects (the Human Genome Project, completed in 2003 (Collins et al., 2003), and similar genome projects for many other organisms). However, not all of the available components in an organism's parts list are present in all its cells and, furthermore, it has been found that the same component may have different roles in different cells within the same organism. This has lead researchers to develop experimental techniques to measure which subset of an organism's genes are actively being used in different cell types at any particular point in time. This allows an even more profound study: how do the genes in an organism's genome interact to generate complex biological function.

A major challenge remains, however, in how to process and interpret the data which are produced by these technologies. The Human Genome Project determined that there are some 25 thousand genes in the human genome. While this is well short of the 140 thousand or so which were predicted in the early days of the project, the simultaneous measurement of the activity of all 25 thousand genes creates significant challenges for researchers trying to infer regulatory interactions between the genes. Predictive mathematical modelling provides a framework within which data can be used to determine the regulatory interactions between genes. Traditionally this approach has been tackled in an intensive fashion, by piecing together available information on individual gene interactions to reconstruct the network of regulatory interactions within a cell. High-throughput experiments now produce data from thousands of genes, demanding a change of emphasis to auto-

mated data analysis and modelling procedures. The task, to determine the network of regulatory interactions underlying the data, has been called reverse engineering (in analogy with the industrial practice of examining a competitor's finished product in order to learn how it performs its tasks). Below we review some of the reverse engineering techniques which have been developed for this task, in particular focusing on the system identification challenges. Before that we review the fundamentals of molecular biology, and describe the experimental approaches which have lead to these new challenges for systems biology.

1.1 Gene Expression Primer

Genes are the fundamental units of heredity, by which information is passed on from one generation to the next. A gene is a section of DNA, the molecule on which the heritable information is carried, which codes for a protein. More precisely the sequence of bases (the bridges between the two strands of the DNA double helix) along the DNA encodes the sequence of amino acids, the building blocks that make a protein. Gene *expression* is the process in which the information carried by the gene is read and used to produce a protein. This process takes two steps (see Fig. 1). Firstly a gene is *transcribed* to an intermediate molecule, mRNA, by an enzyme which reads the sequence of bases. Secondly, the mRNA molecule is *translated* into a protein (the correct sequence amino acids for the protein is assembled at a ribosome, according to the *genetic code*).

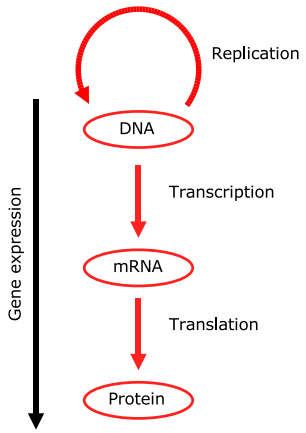

Fig. 1. The basic events in gene expression

In the double-helical structure of DNA, determined by Watson and Crick, the information content is duplicated. Information contained in the sequence of bases along one strand is the same as that in the sequence of bases along the other. This is due to the complementary base pairing principle. Each of the four possible bases, A (Adenine), T (Thymine), G (Guanine) and C (Cytosine), can only bind to one other type, to form the bridge between the two strands of the double helix. A always binds to T, G always to C. As was noted by Watson and Crick in their famous paper (Watson and Crick, 1953) this provides a template mechanism for replication of DNA when required as, for example, during cell division. This principle is also the basis for the development of high throughput molecular technologies to identify and quantify the mRNA present in the cell, and hence gene expression.

As might be expected, gene expression is very highly regulated in the cell. Not all proteins are required in all cells at any one time, and cells need to respond to changing demands and conditions, including metabolic state, growth or cell division, and so on. Most of the different steps in gene expression are regulated, and the proteins may require further *post-translational* modifications in order to become functional. However, the majority of the regulatory activity controlling gene expression is thought to take place at the transcriptional level. Transcriptional regulation is achieved predominantly by the binding of other proteins to the DNA, either to facilitate or to hinder transcription. Many proteins acting in this manner are *transcription factors*, and can be specialised to individual genes, or may act on large numbers of genes. Transcription factors which facilitate gene expression are known as transcriptional activators, while transcription factors which inhibit gene expression are known as repressors.

1.2 Modelling Gene Expression

Transcriptional regulation of gene expression involves proteins (gene products) interacting with the DNA to activate or repress transcription. Transcription factors themselves may be regulated by other proteins, or other cellular processes, such as the metabolic state of the cell. In this way, we can envisage a hierarchy of levels of regulation, which ultimately act on the DNA, shown in Fig 2.

The cellular regulatory network that controls gene expression can me modelled at several different levels of abstraction:

Comparative studies are aimed at finding patterns in gene expression data. Examples include clustering of genes into groups according to correlation of their expression profiles (co-expression) under different experimental conditions.

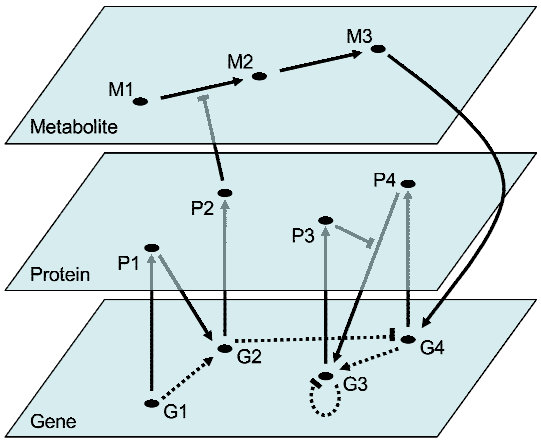

Fig. 2. Interdependence of Cellular Networks. Top layer: cell signalling and metabolic network (metabolome); middle layer: protein-protein interactions (proteome); bottom layer: gene expression (transcriptome). Arrows represent activation and bars represent inhibition. Adapted from Brazhnik et al. (2002).

Gene Regulatory Network models focus on data at the transcriptional level, to identify a regulatory 'wiring diagram', called a Gene Regulatory Network, illustrated by the dashed arrows in Fig 2. This simplified representation of cellular signalling projects the pathways mediated by protein and metabolite interactions on to the transcriptional level (dashed arrows), reducing the problem to identifying (indirect) gene-gene connectivity from gene expression data. In particular, the wiring diagram identifies connections between genes which up- and down- regulate gene expression. For example, in the figure gene 1 (G1) makes protein P1 which is a transcription factor that activates expression of G2. Therefore in the gene regulatory network, G1 activates G2. Similarly, G3 inhibits its own expression, as its protein P3 inhibits the action of activatory transcription factor P4, and so on.

Mechanistic models characterise the biophysical and biochemical details of DNA, protein and metabolite interactions, to produce a fully predictive kinetic model of gene regulation and cellular signalling. These models include all three of the interaction levels illustrated in the figure.

The majority of this review will focus on modelling gene regulatory networks from gene expression data. Next we discuss how these data are generated.

2. MEASURING GENE EXPRESSION

Gene expression is quantified by measuring the amount of mRNA corresponding each gene that is present in the cell. The two issues facing experimental measurement gene expression are firstly to determine which mRNAs are present in the cell, and secondly to count how many copies of each molecule are present. Complementary base pairing of DNA and RNA molecules can be used to solve both these problems, and the development of technologies that enable parallelisation of these manipulations has given birth to transcriptomics and systems biology.

Because of complementary base pairing, sequences of RNA (and single strands of DNA) bind ('hybridise') much more readily with molecules having the complementary sequence of bases, than with molecules with a different sequence. Even a single difference in the sequence greatly reduces their binding affinity. Thus complementary base pairing can be used to identify transcripts, by hybridisation of a sample with molecules of known sequence.

This basic process has been turned into a high throughput mechanism for measuring gene expression by parallelising the hybridisation reaction, so that the abundance of thousands of gene transcripts can be detected at once. This technology is described below.

Microarrays: The most widely adopted technology for high throughput measurement of gene expression is DNA microarrays. DNA microarrays use the complementary binding properties of mRNA to recognise the specific sequences of mRNA molecules extracted from cells. mRNA isolated from the sample is used to make fluorescently labelled molecules which hybridize with molecules immobilised onto glass slides or chips. These immobilised molecules are manufactured to have sequences corresponding to genes of interest, and are laid out in an array, where each point on the array corresponds to a particular gene sequence. When the fluorescently labelled sample is washed over the array, the intensity of the fluorescence at a particular location in the array reveals the amount of the gene transcript which is present in the sample.

This technology is most commonly used to quantify the ratio of fluorescence intensities between sample and a control, or between two samples, using red and green fluorescent labels. The sensitivity that can be achieved can be as high as a few mRNAs per cell, with relative discrimination of around two-fold concentration changes, due to the high affinity and specificity of complementary nucleotide binding.

Typically, two types of microarray experiments are conducted. Gene expression in two different samples can be compared directly on the same array (for example 'normal' against disease). Alternatively, time course data can be generated by sampling from a cell population at different times following a stimulus, and compared to gene expression in an unstimulated (control) population.

Microarrays are now produced for thousands of genes, enough to cover entire genomes, however, microarray experiments give only relative levels of gene expression. Absolute quantitative studies can be carried out with RT-PCR and SAGE technologies.

Quantitative PCR: The Polymerase Chain Reaction (PCR) allows small amounts of DNA to be amplified, using an enzymatic reaction that produces copies with the exact same sequence. The PCR technique has been adapted to provide quantitative and highly sensitive measurements of the amount of a specific mRNA present in a sample. The technology is not easily parallelised, however, and is relatively slow, therefore it is most commonly used for detailed study and validation of results.

Both PCR and Microarray technologies require molecules of the appropriate sequence for all genes of interest (required as 'primers' in the PCR reaction). Therefore, neither technique can be used to measure the expression of previously unknown or unrecognised genes. One technique which can identify and measure transcripts without prior knowledge of the set of genes of interest is SAGE.

Serial Analysis of Gene Expression (SAGE): SAGE (Velculescu et al., 1995, 2000) is an approach in which large numbers of mRNA transcripts can be counted and analysed efficiently by sequencing only short pieces, or 'tags', from a defined location of each mRNA molecule. These tags provide a signature that uniquely identifies the corresponding mRNA molecule (the length of the tags is such that there is a one-to-one mapping from tag sequences to genes). Tags extracted from a sample are assembled into a single molecule, which is then amplified and sequenced. Counting the number of appearances of each tag quantifies the abundance of the corresponding mRNA in the original sample. Because it is not necessary to know the gene corresponding to a particular tag at any point in this process, SAGE can therefore be used to identify and quantify the expression of previously unknown genes.

2.1 Measuring Protein Abundance

The two higher levels of the hierarchy shown in Fig 2 are more difficult to measure in a high throughput fashion. Protein molecules have complicated secondary and tertiary structure which makes them intrinsically much less easy to work with for high-throughput studies than DNA and RNA. Nevertheless, techniques are available which allow large scale analyses of the abundance of different proteins in the cell, including separation techniques (for example using the mass and electrostatic charge to separate proteins). Quantitative measurements can then be made following separation. Mass spectrometry can be used to identify proteins separated out in this way.

2.2 Gene Expression Data Sets

A large amount of data is generated in each microarray experiment, however, currently the availability of multiple repeat measurements, both on-slide repeats and repetition of entire experiments, is limited by the expense of the technology. Stringent experimental procedures are required to get meaningful, reproducible data sets. Microarray experiments involve many separate steps in the preparation of samples, of the arrays and in the subsequent image acquisition and analysis (Hauser et al., 1998). Significant microarray-to-microarray variability is therefore to be expected, and normalisation and pre-processing of the raw data is a crucial element in microarray analysis. Repeat experiments are clearly of importance for the statistical analysis of microarray experiments and for statistical significance tests on inferences drawn from the data (Speed, 2003). Recently a set of standards was adopted by the microarray community in an attempt to make experiments more easily reproducible, and results more straightforward to interpret, and to ensure that experimental data is made publicly available. This comprises a set of principles, called Minimum Information About a Microarray Experiment (MIAME) (Brazma et al., 2001), and many journals require that these be adhered to before publication.

However, there remains a significant problem for the development of system identification approaches to gene regulatory network inference, in that there are no large-scale validated data sets where the underlying network is fully known. There are some smaller data sets available, corresponding to well characterised sub-networks (Gardner et al., 2003), but often the development of reverse engineering algorithms has relied on synthetic data sets, generated using mathematical models of gene regulatory networks (Mendes et al., 2003).

2.3 Data Preprocessing

Data preprocessing done prior to the implementation of high level analysis techniques is used to arrive at the best estimate of the mRNA level in the original sample from the experimental measurement (fluorescence intensity of the microarray image, for example). Calibration, normalisation and scaling of the data, as well as log-transformation of relative gene expression levels and technique-specific analysis such as image processing of fluorescence intensities for microarray studies, are crucial if correct inferences are to be drawn from the data (Speed, 2003).

2.4 Data Requirements for Reverse Engineering Gene Networks

The above considerations aside, the data typically generated in high-throughput gene expression experiments raise significant difficulties for systems identification approaches. The principal concern is the so-called curse of dimensionality: the parameter space for a model grows exponentially with the number of genes, while typically relatively few independent experiments are done. This makes finding appropriate parameter values for large scale gene network studies a major challenge.

One key to this problem is to incorporate *a priori* knowledge about the system into the data analysis. In general, many gene-gene interactions will have already been identified, often using gene-scale (rather than genome-scale) experimental approaches. These data can be used to restrict the dimension of the parameter search space. Other global features of gene regulation can also be incorporated to deal with this problem. One such feature is that gene networks are sparsely connected (that is, there are many fewer actual than possible connections in the network). Imposing an upper limit to the number of regulatory interactions per gene (estimated at less than 10 for higher organisms) reduces the difficulty of the system identification problem.

To fully characterise network behaviour it is necessary to sample gene expression under as many different combinations of inputs and perturbations as possible. To this end gene activity can be manipulated by using a variety of molecular biology techniques, including gene deletions (knock-outs), in which a gene is removed from the network; knock-downs in which gene expression is reduced for a target gene; and over-expression, increasing the expression level of a target gene (for example using plasmids to express the mRNA). A simple strategy for rational selection of genes to perturb for gene network inference has been suggested by Tegnér et al. (2003). In order to maximize the amount of information extracted from each experiment, genes whose activity has changed the least during all previous experiments are selected. This iterative procedure can be supplemented by then selecting genes whose network connections are statistically most uncertain. Other than this, very little has been said regarding experimental design in the system identification of gene regulatory networks.

A wide variety of techniques have been used to investigate gene expression profiles from microarray studies, including principal components analysis (Holter et al., 2000, 2001; Alter et al., 2000), correspondence analysis (Fellenberg et al., 2001), and the construction of statistical models (Zhao et al., 2001). In the following section we discuss one of the most common approaches, cluster analysis, in which similar gene expression profiles are grouped together.

3. CLUSTERING GENE EXPRESSION DATA

A common starting point for the analysis of gene expression data is to use a clustering technique to group together genes with similar expression profiles (Eisen et al., 1998; Chu et al., 1998; Spellman et al., 1998; Wen et al., 1998; Iyer et al., 1999; Alon et al., 1999; Tusher et al., 2001; Yeung et al., 2001). This approach is based on the idea that genes will respond in one of only a limited number of ways, and seeks to identify these groupings. Typically, however, the different ways in which genes may respond in experiments are not known, and therefore unsupervised techniques are most commonly used. Genes are said to be co-expressed if there is strong correlation in their expression profiles (over different experimental perturbations) which may imply that they are under the same regulatory control. Co-expressed genes 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 cluster a data set the *similarity* between two data points has to be quantified. Distance metrics commonly used are Euclidean distance between expression profiles, or a distance based on the correlation coefficient (this being particularly suitable for comparisons of shape, rather than magnitude, of expression levels). Other measures that have been used include mutual information and rank correlation. Many different algorithms have been applied to cluster gene expression data. Techniques can be divided into hierarchical and non-hierarchical approaches. Non-hierarchical techniques iteratively partition the data into a predetermined number of groupings, so as to optimize some selection criterion. For example, K-means (Tavazoie et al., 1999) seeks it-

eratively to partition N data points into K groups so that the sum of K sums of squared distances from the means of each group is minimized (hence *K-means*). A drawback of this approach is that the number of clusters must be specified at the outset. Typically for a given data set the algorithm is applied in succession for different values of K, and the 'best fit' selected according to some selection criterion.

More widely used are hierarchical clustering techniques which generate a tree structure linking genes according to how closely their expression profiles are correlated. This is tackled as either a top-down problem (Alon et al., 1999), where one large cluster is successively divided into smaller groupings, or a bottom-up approach (Eisen et al., 1998) in which smaller clusters are successively combined into larger ones. The disadvantage of these methods is that it can be difficult to know which associations are significant, and which are artefactual, i.e. where to 'cut the tree'.

Problematically, different clustering algorithms will find different partitions of the same data set. So far, however, there are few indications as to which technique provides the best clustering for a specific data set or application. There is a pressing need for improved statistical analyses of the results of clustering techniques which give confidence levels for the clusters that are found.

3.1 Principal Components Analysis

Another low-level data analysis approach is to use principal components analysis (PCA) to reduce the data set to a few simple underlying modes of gene expression (Holter et al., 2000, 2001; Alter et al., 2000). Principal components analysis performs a linear transformation of the data such that the majority of the variance in the data is captured by the first few modes (principal components). This provides a way to simplify the data set and is particularly useful for oscillatory patterns of gene expression where responses with different period are distinguished easily (although Yeung and Ruzzo (2001) have commented on the detrimental effect of using PCA to simplify the data before applying a clustering algorithm).

4. REVERSE ENGINEERING GENE NETWORKS

Gene network analysis tries to identify the regulatory interactions underlying the gene expression data. Many different approaches have been applied to the problem of inferring the structure of gene networks from expression data, and to developing predictive kinetic models. Below we describe these techniques in turn, discussing their merits and data requirements, with examples.

4.1 Boolean and Logical Networks

One way to greatly simplify the mathematical representation of networks is to ignore the details of molecular interactions and focus instead on their outcomes, namely whether a gene is 'on' or 'off', according to whether its transcription level is above or below a given threshold. Regulation is then represented by logical operations (AND, NOR *etc.*) on the gene expression states, according to whether interactions activate or repress transcription. Gene expression levels are Boolean variables which are updated synchronously according to a rule table (a set of *if ... then ...* instructions) which describes the logical operations representing the regulatory interactions between genes. A Boolean network is thus quite a natural way to represent a wiring diagram for the gene network.

Although the properties of Boolean networks are much simpler than their continuous variable counterparts, they retain many of the properties of networks that are important in terms of gene function. For example, steady states are achieved as the network settles down into a stationary or a repeating pattern of logical states (oscillation), for which stability properties can be determined.

Several authors have been able to piece together logical network models for smaller scale gene networks (typically models representing a small part of a larger network) from the literature on known regulatory interactions between genes, as discovered by mutation screens and molecular studies. Mendoza et al. (1999) used a generalised formalism of the logical network to study gene interactions underlying the development of flower buds in Arabidopsis (a plant commonly used to study plant development and genetics). Knowledge of pairwise interactions for a network of 10 genes was extracted from the literature 'by hand' to build the network.

The aim of a reverse engineering approach is to infer the logical rule table *directly* from data. Techniques have been proposed which demonstrate that in principle a Boolean network can be constructed from data using no prior knowledge. Liang et al. (1998) have developed an algorithm, which they call 'REVEAL', to determine Boolean models from data using an information theoretic approach. Mutual information is used to identify the minimal set of inputs required to determine the output for each gene in the network. Look-up tables are then used to reveal the corresponding logical operations acting on the inputs, from which the network 'wiring' is determined.

The appeal of this modelling framework is its simplicity. However, much biological detail is clearly lost. There are many aspects of cellular signalling which cannot easily be described with Boolean variables. Regulation cannot be additive, nor can regulatory mechanisms such as negative feedback, well characterised in biological systems, be easily accommodated. Furthermore, in reality gene expression levels recorded in time course microarray studies, for example, seem to spend much of their time at 'intermediate' levels, rather than quickly saturating at maximal expression rate, or falling to negligible levels. Boolean networks may therefore be a good modelling strategy when the data quality is poor, and where intermediate expression levels cannot be resolved.

Some of these drawbacks can be overcome by generalisations which allow multiple logical values for each gene, asynchronous updates and multiple distinct thresholds for switching between states, but all of this comes at the expense of the simplicity of the Boolean formulation. Control of logical variables by continuously varying metabolite or cell signalling networks can be included, and various hybrid models with some logical and some continuous variables have been proposed.

Davidson et al. (2002b) have lead an intensive effort to identify and model the gene regulatory network for specification of endomesoderm (an early tissue formation event during development) in the Sea Urchin embryo (see also Davidson et al., 2002a; Bolouri and Davidson, 2002). This includes cell-to-cell signalling pathways (ligand-receptor binding) to provide spatial coupling between cells, each of which has the same underlying logical gene network. A model of the whole network was built up in a piece-by-piece fashion, using both Boolean and algebraic logical elements, as appropriate for the available data (Brown et al., 2002).

4.2 Bayesian Networks

The Bayesian approach to gene network modelling treats the expression level of each gene as a random variable and regulatory interactions as probabilistic dependencies between variables (Friedman et al., 2000; Pe'er et al., 2001; Hartemink et al., 2002; Imoto et al., 2003). Bayesian analysis reveals statistical relationships between the genes from data. These relationships can be represented as a directed graph. If a directed edge exists from gene X to gene Y then the expression of Y is found to be directly dependent on X, and so forth. More complicated dependencies are represented in a similar way. Quantitatively, the statistical relationships between genes found in the data are described as joint probability distributions on the variables (for example the probability of gene Y being expressed at a certain level given that gene X is expressed).

The Bayesian approach is not limited to pairwise or linear interactions between genes. It is robust to noisy data and can in principle be extended to handle missing data and even latent (unobserved) variables. However, it is the ability to include *prior* information, for example on known gene interactions or protein data such as transcription factor binding locations *etc.*, and indeed protein concentrations, in a rigorous manner as additional variables which is perhaps the main strength of this approach. In its simplest form the Bayesian network gives a static model of the data set, and does not describe dynamic processes such as feedback, known to characterise regulation. The formalism can be extend to dynamic Bayesian networks, a series of connected models for which probabilities span timesteps.

Alternative models for the underlying network can be ranked against each other according to the Bayesian scoring metric—the logarithm of the probability that model is correct, given the data—which measures, essentially, the economy with which the model explains the data. This is, however, a more difficult task when latent variables are included, as is the assignment of appropriate weights in the scoring metric when prior information is incorporated.

Hartemink et al. (2002) collected expression data on galactose metabolism in the yeast *Saccharomyces cerevisiae* using 52 samples from both wild type and mutant strains under variety of environmental conditions. Data was collected using DNA chips with all 6135 genes of the *S. cerevisiae* genome. A static model of the statistical dependencies between genes was built up for (hand-picked) components of the genome demonstrating the feasibility of this approach to network modelling.

4.3 Continuous Variables: Linear Network Modelling

Systems of differential equations provide a very natural modelling framework for the kinetic behaviour of gene networks. Overlooking those situations where stochastic models are required (Kepler and Elston, 2001; Arkin et al., 1998; McAdams and Arkin, 1999; Crampin et al., 2004c for review; which are unlikely to be amenable to reverse engineering approaches), the level of expression of each gene is represented as a continuous variable which changes over time according to a differential equation with 'reaction' terms describing regulatory inputs from other variables. This gives a coupled system of differential equations to solve for the network behaviour.

Near to a steady state of such a dynamical system, the kinetics are well described by a model which is linear in the variables. A general linear model for the concentration of the ith mRNA, $v_i(t)$, is given by

$$\frac{dv_i}{dt} = \sum_{j=1}^{N} A_{ij} v_j(t) - \lambda_i v_i(t) + b_i(t)$$

for $i = 1, \ldots, N$ genes. λ_i is a degradation rate for the ith mRNA and the $b_i(t)$ are known functions which represent experimental perturbations applied to the network. A_{ij} is the connectivity matrix for the network (the Jacobian). Reverse engineering aims to determine the unknown parameters A_{ij} and hence the connectivity in the linear model. Ideally sufficient data would be collected so that the connectivity matrix can be completely determined (i.e. parameters chosen to minimise the least squares discrepancy of the model from the data set). However, typically for large networks there are more parameters to be determined than there are (independent) data points. Microarray experiments rarely have more than 10 time points, for example, and the number of genes N may be an order of magnitude larger. This *underdetermined* problem does not have a unique solution but can still be solved, in the least squares sense, using singular value decomposition (SVD) to find the solution space—the family of solutions which are consistent with the data (in fact SVD algorithms will pick the least squares solution from this solution space, Press et al., 1992).

A supplementary criterion is needed to select the most appropriate solution in this case. One suggestion that has been made by Yeung et al. (2002) is that gene networks are observed to be sparse, that is, there are relatively very few regulatory interactions between genes, and so most of the entries in the connectivity matrix should be zeros. This property can be exploited to choose the solution which is consistent with the data and minimises the number of nonzero elements. These authors have demonstrated this to be a tractable approach, for data generated from model networks at least, using an algorithm from robust regression analysis to find the optimal solution. The relatively low data requirement—they show that only approximately as many data points as genes are required for sparse networks—comes at the expense of an increased computational cost, however, the approach can be used to efficiently compute connectivity matrices for networks containing thousands of genes. Further details on these techniques, and variations on these approaches, can be found in Crampin et al. (2004b).

Steady-state Perturbation Data: Two reverse engineering algorithms based on this approach have recently been described for steady state experimental data. In this scenario, the cell is allowed to settle into a new stable steady state following perturbation to one or more genes, and the steady gene expression levels recorded. Gardner et al. (2003) described an algorithm called Network Inference via multiple Regression (NIR), which solves the following steady state problem

$$0 = \sum_{j=1}^{N} A_{ij} v_{jl} + b_{il}$$

for $i = 1, \ldots, N$ genes, where b_{il} is the perturbation made to the ith gene and v_{jl} is the steady state expression level of the jth gene in the lth experiment. Gardner et al. (2003) collected data on the SOS response network in the bacterium *E. Coli*. They considered a small sub-network of 9 genes involved in sensing damage to DNA and triggering a the synthesis of a number of proteins involved in DNA repair. The NIR algorithm assumes that each of the N genes in the network has at most $K < N$ regulatory interactions, and for each gene iteratively searches for the best choice of K interactions using regression. By doing this for different K and computing the best fit over K for the data, an optimum network can be determined. This approach performed very well for the SOS response network, when network predictions were compared with known interactions between the genes.

However, a limitation of this approach is that the number of connections K per gene in the network must be guessed at, and each gene is assumed initially to have this number of regulatory interactions (although some may have near-zero strength). This is particularly important as evidence suggests that transcriptional networks are not well characterised by an average connectivity, but show a power-law distribution in the number of connections per gene (Featherstone and Broadie, 2002). Recently we have considered a similar reverse engineering approach, which deals with these two distinct issues (Wildenhain and Crampin). Firstly, the algorithms we have proposed do not impose a fixed number connections per gene, but allow this number to vary, as determined by a formal criterion such as the Akaike Information Criterion (Akaike, 1974), or Minimum Description Length (Rissanen, 1980). Secondly, the selection of connections for each gene from the N possible gene-gene interactions uses an iterative model selection scheme by Judd and Mees (1995), and described in Crampin et al. (2004a). We recently compared the performance of two algorithms based on these ideas, one which constructs a network by building from an ini-

tially unconnected set of genes, while the other starts with a fully connected network and removes connections until an optimal solution is found. These algorithms were found to perform well on simulated gene expression data sets, in particular when the underlying network was constructed to have the power-law distribution of connections.

These methods concentrate on identifying the connectivity in the network, rather than the more difficult task of characterising the nonlinear dynamics of the regulatory interactions. The linearization step which is used to reduce the problem to the determination of the connectivity matrix is only valid near to a steady state of the system, and so relies on relaxation data for small perturbations from the steady state. Methods to determine nonlinear aspects of the regulatory mechanisms are discussed below.

Kinetic Models with Nonlinear Response Functions: For well characterised networks in which the relevant genes have been identified and the wiring diagram determined (perhaps through one of reverse engineering schemes discussed above), biophysically realistic kinetic functions can be assumed for the regulatory interactions. This allows quantitative prediction of transcription rates *etc.* in response to perturbations of the network. Kinetic parameters can be determined for individual reactions (for example the binding of transcription factors, degradation rates of mRNAs, *etc.*) using timecourse data from microarrays or other experimental approaches. Ronen et al. (2002) made *in vivo* measurements on the DNA repair system in the bacterium *E. coli* using intensity of a green fluorescent protein (GFP) linked to genes of interest, to determine their expression levels (Kalir et al., 2001). Their network has a simple known architecture: a 'single input module' in which a single master repressor gene (LexA) regulates approximately 30 targets (of which they measured the response of 8). They were able to fit their data to a model with sigmoidal response functions, determining parameters by least squares fitting (using SVD).

This method can easily be generalised to more complicated transcription factor-gene interaction functions, allowing positive and negative feedbacks between gene products and gene transcription, and to allow the regulatory input of several genes. Detailed kinetic models can incorporate kinetic descriptions of promoter substructure, detailing interactions of transcription factors, with appropriate functions for the rate of transcriptional output for given binding state. Reverse engineering methods can be devised for nonlinear models such as these (see Crampin et al., 2004a,b for more details). However, there is a trade-off between model complexity and tractability. The derivation of functional forms for regulatory interactions is not a simple process, and as the parameter space expands, more and better high resolution timeseries data on the transcription of each gene under a wide variety of physiological conditions and gene perturbations is needed to parameterise the model.

5. MODELLING SPATIALLY DISTRIBUTED NETWORKS

During many developmental processes, interactions between cells and across tissues are critical in determining patterns of differentiation into distinct tissue types. While, of course, each nucleus carries the same DNA, the profile of gene expression in different cell types (the *state* of the network) will be distinct. In the course of development, cells from an initially homogeneous tissue may receive different signals and adopt different patterns of gene expression, leading to different cell fates. Models of developmental processes must therefore incorporate mechanisms for signalling between cells, transport of signalling molecules through the tissue, as well as signalling networks within cells.

5.1 Spatial Modelling using Differential Equations

Several approaches have been used to model and study gene regulation in spatially extended systems using continuous variable gene network models. In particular, two studies have focused on different stages of the early development of the fruit fly, *Drosophila melanogaster*, which have necessitated very different modelling approaches. During early development in *Drosophila* groups of genes called gap, pair rule and segment polarity are successively expressed in periodic bands along the embryo. This pattern in the expression of groups of genes is a result of an underlying regulatory network, with initial input from genes expressed by the mother. The symmetry of the banded patterns lends itself to a spatially one-dimensional analysis.

A network model for segment polarity gene expression, which occurs after the expression of pair rule genes has been constructed by von Dassow et al. (2000). From a survey of the literature, they were able to write down all of the known gene interactions which have been discovered by saturation mutagenesis studies, and chose suitable functional forms for receptor-ligand binding, transcription rates and so forth. Parameter values for most of these processes have not been experimentally measured. Surprisingly, however, an exploration of the model with randomly chosen param-

eter sets found that the experimentally observed pattern of gene expression was recovered in a relatively large region of parameter space. From this evidence, the authors suggested that *robustness* to parameter variation is a property of gene regulatory networks. This would of course be extremely useful for reverse engineering approaches, as it would suggest that network behaviour is determined primarily from is structure, and the details of the kinetics relegated to a less important role (Albert and Othmer, 2003). Clearly robustness in biological networks also has obvious biological and evolutionary implications.

A computational optimization approach has been applied to the highly specific situation arising earlier in *Drosophila* development, during which time gap and pair rule expression patterns are determined from the maternal gene expression (Mjolsness et al., 1991; Reinitz et al., 1995; Sharp and Reinitz, 1998). At this stage the embryo consists of a syncytium, in which multiple nuclei exist in an extended cell-like compartment, and are not separated by cell membranes. For this unique situation the transport of gene products can be modelled by solely by diffusion.

'Gene Circuit' Approach: In this case Reinitz et al. (1995) have demonstrated that a reverse-engineering approach is possible, using digitized images of immunofluorescence staining of the gene products. Local fluorescence intensity was assumed to be proportional to protein concentration. As all of the genes involved in this developmental stage have been identified the authors have built up a database of gene expression images from wildtype and mutant embryos for each gene known to be involved in the network and at each discernible developmental stage.

For the concentration of the ith gene product, $v_i(x,t)$, nonlinear optimization in the form of a simulated annealing algorithm (Press et al., 1992) was used to determine best fit parameters for a model of the form

$$\tau_i \frac{\partial v_i}{\partial t} = g\left(\sum_j A_{ij} v_j + m_i v_m\right) + D_i \frac{\partial^2 v_i}{\partial x^2} - \lambda_i v_i$$

where $g(\cdot)$ is a sigmoidal response function, A_{ij} is the matrix of regulatory connections between genes, known beforehand, λ_i is a decay rate and D_i the diffusion coefficient for the ith gene product and $m_i v_m$ describes the effect of underlying (maternal effect) protein gradient, $v_m(x)$, on gene i. While this model represents a much simplified view of the biophysical processes at work in the syncytium, the model could reproduce the temporal sequence of spatial gene expression patterns fir the different genes, and predict the effects of experimental interventions and mutations.

6. DISCUSSION

Gene network modelling, and in particular the development of reverse engineering techniques to automate the modelling of data, are at an early stage in their development. For reverse engineering gene networks, several approaches have been suggested and applied to specific problems, but they have yet to demonstrate general utility and applicability.

From the perspective of system identification, there are a number of pressing issues that need to be addressed in order to improve the practicality of applying reverse engineering approaches to genome scale data sets. In particular, the development of methods to incorporate prior knowledge about the biological system is required, with different degrees of confidence assigned, and to combine data from different sources, for example microarray gene expression data with proteomic data sets. Also very important is the development of techniques to reduce the dimensionality of network models. One possible approach is to apply clustering techniques to preprocess data, in order to identify a network of regulatory interactions acting on co-expressed genes. However, most problematic is the current lack of validated data sets that can be used to evaluate different network identification techniques, which makes direct comparison of the strengths and weaknesses of different approaches difficult. Also, although an accepted standard for microarray data sets has been adopted, a similar standard has only recently been proposed for biological models (MIRIAM; Minimum Information Requested In the Annotation of biochemical Models, Novère et al., 2005).

The developments in high-throughput gene expression measurement technologies pave the way for a revolution in understanding of development and disease. The ability to measure expression levels for multiple genes concurrently is driving a shift away from the study of single genes to the view that gene function can only be understood in the context of the gene network, and ultimately the many interacting regulatory networks—genetic, signalling, metabolic, *etc.*—which are the functional environment of the genes. The ability to generate predictive models describing quantitatively the interactions between genes and gene products is crucial to enabling this research.

REFERENCES

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

R. Albert and H. G. Othmer. The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *drosophila melanogaster*. *J. theor. Biol.*, 223: 1–18, 2003.

U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, 96:6745–6750, 1999.

O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA*, 97:10101–10106, 2000.

A. Arkin, J. Ross, and H. H. McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells. *Genetics*, 149:1633–1648, 1998.

H. Bolouri and E. H. Davidson. Modeling DNA sequence-based *cis*-regulatory gene networks. *Dev. Biol.*, 246:2–13, 2002.

P. Brazhnik, A. de la Fuente, and P. Mendes. Gene networks: how to put the function in genomics. *Trends Biotechnol.*, 20:467–472, 2002.

A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C.A. Ball, H.C. Causton, T. Gaasterland, P. Glenisson, F.C.P. Holstege, I.F. Kim, V. Markowitz, J.C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genet.*, 29:365–371, 2001.

C. T. Brown, A. G. Rust, P. J. C. Clarke, Z. Pan, M. J. Schilstra, T. De Buysscher, G. Griffin, B. J. Wold, R. A. Cameron, E. H. Davidson, and H. Bolouri. New computational approaches for analysis of *cis*-regulatory networks. *Dev. Biol.*, 246:86–102, 2002.

S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. O. Brown, and I. Herskowitz. The transcriptional program of sporulation in budding yeast. *Science*, 282:699–705, 1998.

F. S. Collins, E. D. Green, A. E. Guttmacher, and M. S. Guyer. A vision for the future of genomics research. *Nature*, 422:835–847, 2003.

E.J. Crampin, P.E. McSharry, and S. Schnell. Extracting biochemical reaction kinetics from time series data. *Lecture Notes in A.I.*, 3214: 329–336, 2004a.

E.J. Crampin, S. Schnell, and P.E. McSharry. Mathematical and computational techniques to deduce complex biochemical reaction mechanisms. *Prog. Biophys. Mol. Biol.*, 86:77–112, 2004b.

E.J. Crampin, N.P. Smith, and P.J. Hunter. Multi-scale modelling and the IUPS Physiome Project. *J. Mol. Histol.*, 35:707–714, 2004c.

E. H. Davidson, J. P. Rast, P. Oliveri, A. Ransick, C. Calestani, C. H. Yuh, T. Minokawa, G. Amore, V. Hinman, C. Arenas-Mena, O. Otim, C. T. Brown, C. B. Livi, P. Y. Lee, R. Revilla, A. G. Rust, Z. J. Pan, M. J. Schilstra, P. J. C. Clarke, M. I. Arnone, L. Rowen, R. A. Cameron, D. R. McClay, L. Hood, and H. Bolouri. A genomic regulatory network for development. *Science*, 295:1669–1678, 2002a.

E. H. Davidson, J. P. Rast, P. Oliveri, A. Ransick, C. Calestani, C. H. Yuh, T. Minokawa, G. Amore, V. Hinman, C. Arenas-Mena, O. Otim, C. T. Brown, C. B. Livi, P. Y. Lee, R. Revilla, M. J. Schilstra, P. J. C. Clarke, A. G. Rust, Z. J. Pan, M. I. Arnone, L. Rowen, R. A. Cameron, D. R. McClay, L. Hood, and H. Bolouri. A provisional regulatory gene network for specification of endomesoderm in the sea urchin embryo. *Dev. Biol.*, 246:162–190, 2002b.

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

D. E. Featherstone and K. Broadie. Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network. *Bioessays*, 24:267–274, 2002.

K. Fellenberg, N. C. Hauser, B. Brors, A. Neutzner, J. D. Hoheisel, and M. Vingron. Correspondence analysis applied to microarray data. *Proc. Natl. Acad. Sci. USA*, 98(19): 10781–10786, 2001.

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

T. S. Gardner, D. di Bernardo, D. Lorenz, and J. J. Collins. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 301(5629):102–105, 2003.

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

N. C. Hauser, M. Vingron, M. Scheideler, B. Krems, K. Hellmuth, K. D. Entian, and J. D. Hoheisel. Transcriptional profiling on all open reading frames of Saccharomyces cerevisiae. *Yeast*, 14(13):1209–1221, 1998.

N. S. Holter, A. Maritan, M. Cieplak, N. V. Fedoroff, and J. R. Banavar. Dynamic modeling of gene expression data. *Proc. Natl. Acad. Sci.*

USA, 98:1693–1698, 2001.

N. S. Holter, M. Mitra, A. Maritan, M. Cieplak, J. R. Banavar, and N. V. Fedoroff. Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *Proc. Natl. Acad. Sci. USA*, 97:8409–8414, 2000.

S. Imoto, S. Kim, T. Goto, S. Miyano, S. Aburatani, K. Tashiro, , and S. Kuhara. Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. *J. Bioinform. Comput. Biol.*, 1:231–252, 2003.

V. R. Iyer, M. B. Eisen, D. T. Ross, G. Schuler, T. Moore, J. C. F. Lee, J. M. Trent, L. M. Staudt, J. Hudson, M. S. Boguski, D. Lashkari, D. Shalon, D. Botstein, and P. O. Brown. The transcriptional program in the response of human fibroblasts to serum. *Science*, 283:83–87, 1999.

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

S. Kalir, J. McClure, K. Pabbaraju, C. Southward, M. Ronen, S. Leibler, M. G. Surette, and U. Alon. Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria. *Science*, 292:2080–2083, 2001.

T. B. Kepler and T. C. Elston. Stochasticity in transcriptional regulation: Origins, consequences, and mathematical representations. *Biophys. J.*, 81:3116–3136, 2001.

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

H. H. McAdams and A. Arkin. It's a noisy business! Genetic regulation at the nanomolar scale. *Trends Genet.*, 15:65–69, 1999.

P. Mendes, W. Sha, and K. Ye. Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics*, 19 Suppl. 2:ii122–ii129, 2003.

L. Mendoza, D. Thieffry, and E. R. Alvarez-Buylla. Genetic control of flower morphogenesis in Arabidopsis thaliana: a logical analysis. *Bioinformatics*, 15:593–606, 1999.

E. Mjolsness, D. H. Sharp, and J. Reinitz. A connectionist model of development. *J. Theor. Biol.*, 152:429–453, 1991.

N. Le Novère, A. Finney, M. Hucka, U. Bhalla, F. Campagne, J. Collado-Vides, E.J. Crampin, M. Halstead, E. Klipp, P. Mendes, P. Nielsen, H. Sauro, B. Shapiro, J.L. Snoep, H.D. Spence, and B.L. Wanner. Minimum information requested in the annotation of biological models (MIRIAM). *Nature Biotechnol.*, 23:1509–1515, 2005.

D. Pe'er, A. Regev, G. Elidan, and N. Friedman. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, 17(Suppl. 1):S215–S224, 2001.

W. H. Press, W. T. Vetterling, S. A. Teukolsky, and B. P. Flannery. *Numerical Recipes in Fortran: The Art of Scientific Computing*. CUP, 2nd edition, 1992.

J. Reinitz, E. Mjolsness, and D. H. Sharp. Model for cooperative control of positional information in Drosophila by bicoid and maternal hunchback. *J. Exp. Zool.*, 271:47–56, 1995.

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

M. Ronen, R. Rosenberg, B. I. Shraiman, and U. Alon. Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics. *Proc. Natl. Acad. Sci. U. S. A.*, 99:10555–10560, 2002.

D. H. Sharp and J. Reinitz. Prediction of mutant expression patterns using gene circuits. *Biosystems*, 47:79–90, 1998.

T. P. Speed. *Statistical Analysis of Gene Expression Microarray Data*. Chapman & Hall, 2003.

P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, 9:3273–3297, 1998.

S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. *Nature Genet.*, 22:281–285, 1999.

J. Tegnér, M. K. S. Yeung, J. Hasty, and J. J. Collins. Reverse engineering gene networks: Integrating genetic perturbations with dynamic modeling. *PNAS*, 100:5944–5949, 2003.

V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, 98:5116–5121, 2001.

V.E. Velculescu, B. Vogelstein, and K.W. Kinzler. Analysing uncharted transcriptomes with SAGE. *Trends Genet.*, 16(10):423–425, 2000.

V.E. Velculescu, L. Zhang, B. Vogelstein, and K.W. Kinzler. Serial analysis of gene expression. *Science*, 270:484–487, 1995.

G. von Dassow, E. Meir, E. M. Munro, and G. M. Odell. The segment polarity network is a robust developmental module. *Nature*, 406:188–192, 2000.

J. D. Watson and F. H. Crick. Molecular structure of nucleic acids. *Nature*, 171:737–738, 1953.

X. L. Wen, S. Fuhrman, G. S. Michaels, D. B. Carr, S. Smith, J. L. Barker, and R. Somogyi. Large-scale temporal gene expression mapping of central nervous system development. *Proc.*

Natl. Acad. Sci. USA, 95:334–339, 1998.

J. Wildenhain and E.J. Crampin. Reconstructing gene regulatory networks: from random to scale-free connectivity. Submitted to IEE Systems Biology.

K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17:977–987, 2001.

K. Y. Yeung and W. L. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17:763–774, 2001.

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

L. P. Zhao, R. Prentice, and L. Breeden. Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proc. Natl. Acad. Sci. USA*, 98(10):5631–5636, 2001.