Models and computational strategies linking physiological response to molecular networks from large-scale data

BY FERNANDO ORTEGA†, KATRIN SAMEITH†, NIL TURAN†, RUSSELL COMPTON, VICTOR TREVINO, MARINA VANNUCCI and FRANCESCO FALCIANI*

1 School of Biosciences and IBR, University of Birmingham, Birmingham B15 2TT, UK
2 Department of Statistics, Rice University, Houston, TX 77251-1892, USA

An important area of research in systems biology involves the analysis and integration of genome-wide functional datasets. In this context, a major goal is the identification of a putative molecular network controlling physiological response from experimental data. With very fragmentary mechanistic information, this is a challenging task. A number of methods have been developed, each one with the potential to address an aspect of the problem. Here, we review some of the most widely used methodologies and report new results in support of the usefulness of modularization and other modelling techniques in identifying components of the molecular networks that are predictive of physiological response. We also discuss how system identification in biology could be approached, using a combination of methodologies that aim to reconstruct the relationship between molecular pathways and physiology at different levels of the organizational complexity of the molecular network.

Keywords: functional modules; statistical modelling; network inference; systems biology

1. Introduction

The development of functional genomics technologies has provided researchers with tools to acquire information on the molecular state of tens of thousands of single nucleotide polymorphisms (SNPs), mRNAs, proteins and metabolites (Hobman et al. 2007). In addition to the ability to characterize the phenotypic state of a cell or tissue using a battery of physiological tests, this has created a formidable challenge for the development of comprehensive computational models that integrate molecular and physiological information.

* Author for correspondence (f.falciani@bham.ac.uk).
† These authors contributed equally to the study.

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsta.2008.0085 or via http://journals.royalsociety.org.

One contribution of 12 to a Theme Issue ‘The virtual physiological human: building a framework for computational biomedicine’.
Mathematical models representative of a broad range of biological systems have been developed (Gavaghan et al. 2006). Some of them have relevance in a clinical setting, elucidating the molecular mechanisms of pathology or suggesting useful surrogate markers. There is, however, a huge gap between the large amounts of data produced by functional genomics technologies (including the characterization of the complex physiological response of an organism) and the ability to derive computational mechanistic models based on realistic parameters and constraints. In order to address this issue, a number of methodologies have been developed to infer mathematical models of molecular networks from experimental data. Given the overall lack of information on the dynamics of the interactions for the majority of the biological components, network inference methods tend to be based on information theory or probabilistic methods. Moreover, the development of mathematical models representative of the whole complexity of a cell is not likely to be achievable in the near future, at least for complex eukaryotic systems. The extremely large and yet undetermined number of components, the limited understanding of their function and the complexity of their interactions make the task nearly impossible.

In this paper, we discuss the prospective that system identification in biology should be approached using a combination of methodologies that aim to reconstruct the relationship between molecular pathways and physiology at different levels of the organizational complexity of the molecular network. In this context, we initially discuss the hypothesis of the existence of a higher-level modular organization of molecular networks defined by relatively independent subnetworks of connected components that perform a specific function. We provide examples of how these modules can be identified from experimental data. We further report a series of observations in support of the usefulness of modularization and other modelling techniques in identifying components of the molecular networks that are predictive of physiological response. We show that these techniques have the potential to identify the components of regulatory circuits that are directly associated to the expression of a phenotype. However, the development of mathematical models linking the individual components of the system in an organized network structure requires a different type of methodology, generally called reverse-engineering techniques. These have been demonstrated to be effective but still they have severe limitations when applied in a biological context.

A common methodological challenge in all these approaches is that, owing to the scarcity of the experimental data, many equally possible models can be inferred. For this reason, we discuss the use of computational procedures that are able to search for ‘good’ biologically meaningful models. Some of these methodologies are based on the inclusion of biological knowledge in the form of general rule-based principles or specific information supported by experiments.

We conclude by proposing a strategy based on the combination of these methodologies that may represent different degrees of certainty in the mechanisms underlying system behaviour.

2. System identification in biology

A computational model can be defined as a mathematical representation of a system that allows simulation of its behaviour. Models can be built entirely
on first principles. These are conventionally called white-box models and they need to incorporate a great deal of knowledge on the mechanisms behind the behaviour of the system itself. This approach is quite well established in engineering, physics, biology, medicine and many other fields of research (Gavaghan et al. 2006).

However, there is the possibility that a model needs to be built in the absence of mechanistic information. Based on observed responses to experimental interventions to a biological system (e.g. biochemical, genetic or environmental perturbations), mathematical relationships among its components may be inferred. This approach is called system identification. In engineering, this term comprises algorithms and methods that are used to build mathematical models from measured data. Models can either be entirely based on inference from experimental data (black-box models) or contain some degree of mechanistic knowledge (grey-box or semi-physical models; Ljung 1999). In engineering, the field of system identification is well developed and, certainly, many of the available methodologies can be applied to the analysis of complex biological systems. However, from a systems theory point of view, the identification of a nonlinear system is a major challenge for which ideal solutions do not exist in engineering. Moreover, the inherent complexity of living systems brings many other challenges that are unique to the application of this discipline to biology. Many of these are a direct consequence of our incomplete knowledge of the rules governing biological systems and the nature of the experimental data that can be generated. For example, although the sequence of the human genome has been determined, the exact number of genes, proteins and metabolites in a cell is still a matter of debate. On the other hand, functional genomics technologies allow thousands of genotypes, mRNA expressions, proteins and metabolites to be measured in single experiments. In addition, several physiological measurements can be acquired to further characterize an experimental system.

Clearly, a black-box modelling approach is suitable in this context. However, any attempt to use large datasets for system identification purposes needs to involve some strategy for reducing the complexity of the problem.

An interesting property of complex systems that may help in achieving this initial task is ‘modularity’. In §3, we review a number of definitions of modularity that may apply to biological systems and discuss methodologies for functional module discovery based on experimental data.

3. Modularity of biological networks

(a) Functional modularity in biological systems

In §2, we have argued that biological complexity is a major challenge in system identification. Therefore, simplifying the problem by identifying subsets of cellular components that can be analysed independently of the rest of the system seems an intuitively reasonable strategy. However, a crucial question is: are molecular networks modular in nature?

In addressing this question, it is necessary to refer to a definition of modularity. According to the Oxford English Dictionary, a module is ‘each of a set of parts or units that can be used to construct a more complex structure’. This terminology emphasizes the repetitive character of modules and hence
provides a structural definition. In order to adapt this definition to a biological setting, we need to involve the concept of function. Hartwell et al. (1999) first defined a functional module as being a ‘discrete unit whose function is separable from those of other modules’. A transcriptional module, for example, may be considered as a ‘set of genes that are regulated in concert by a shared regulation program that governs their behaviour’ (Segal et al. 2003a,b). Protein complexes or metabolic pathways, on the other hand, may be defined in straightforward structural terms as ‘highly interconnected, or dense, regions of a larger protein–protein interaction network’ (Bader & Hogue 2003; Lubovac et al. 2006). All these definitions are based on the independence of molecular modules. Independence, however, does not require intermodular connections to be weak or unimportant (Watson & Pollack 2005). Connections between functional modules may be critical to the behaviour of the system and represent the integration of multiple sources of information (Hartwell et al. 1999; Petti & Church 2005). Moreover, independence does not imply that modules are non-overlapping, since some genes, proteins and metabolites may be associated to different functions by being part of alternative molecular pathways.

(b) Approaches to discover functional modules

Section 3a addressed the issue of defining modularity in biological systems. Different definitions of modularity represent features that may be used to detect functionally related subsets of genes and proteins.

Several databases reporting collections of functional modules have been developed. The most widely used so far have been defined by experts in the field with the objective to identify gene subsets with a given function and cellular localization. Well-known examples of these are ‘The Gene Ontology’ (GO; Gene Ontology Consortium 2000) and the ‘Kyoto Encyclopedia of Genes and Genomes’ (KEGG; Ogata et al. 1999).

The GO project was initially developed to address the need of consistent definitions of the function of genes and their products in an organism. GO achieves this by using a controlled vocabulary of three hierarchically structured ontologies (molecular function, biological process and cellular component).

KEGG is a collection of pathways representing portions of a larger network of signalling and metabolic reactions whose boundaries are defined by experts in the field.

While GO and KEGG can be seen as initial attempts to define modules of functionally related genes, the majority of information on the interaction between genes, proteins and metabolites is stored in many publicly available databases in the form of unstructured pairwise interactions. The information contained in these databases coupled with functional genomics datasets has been the basis for many of the methodologies developed for a more systematic identification of functional modules.

Here, we describe a number of methodologies representing different definitions of modularity. The first methodology we describe has been developed by Ideker et al. (2002) and is based on a definition of modularity, which involves the notion of aggregation of genes, whose expression is associated to a given experimental condition on a protein–protein interaction network. In this context, functional modules represent putative multiprotein complexes that are regulated in a
specific experimental condition. Using a different approach, Segal et al. (2003a) have developed a method to identify active transcriptional modules in a large network of hypothetical transcription factor–target interactions. Conceptually, the methodology is based on similar principles. In this case, however, connected subgraphs are evaluated using a Bayesian score (BS) calculated from expression-profiling data.

An example of the combined application of modularization principles and network inference technique (discussed in more detail in §§5 and 6) is ‘Monet’ (Lee & Lee 2005). The procedure begins by selecting a number of seed genes that have the property of being differentially regulated between two or more biologically relevant conditions. Subsequently, additional genes are grouped with a seed on the basis of a similarity score derived from the combination of functional similarity and correlation of their gene expression profiles. Once clusters of genes have been defined, their internal network structure is inferred using a Bayesian network (BN) approach whereas connection between modules is derived from commonalities in their gene content.

A different approach to identify functional modules is based on defining subgraphs with a specific topology, which are found in molecular networks with a significantly higher frequency than expected by chance (Milo et al. 2002). Motifs have been found to represent functional modules such as the cell cycle or different stress responses (Wolf & Arkiny 2003; Alon 2007).

Moreover, functional modules can be thought of as communities, which are groups of vertices in a graph sharing common properties or playing similar roles (Fortunato & Castellano 2007). Guimera & Amaral (2005) have successfully applied this approach to identify functional modules in metabolic networks.

(c) Are functional modules real?

Based on these findings, can we conclude that functional modules are, indeed, part of cellular system properties? So far, there is strong support. The modular nature of protein–protein interactions, gene regulatory and metabolic networks has been presented in a number of studies (Ravasz et al. 2002; Rives & Galitski 2003; Segal et al. 2003a). The studies discussed previously in §3b provide additional support for the existence of modularity.

However, information derived from large-scale experimental techniques, which underlay the inference of functional modules, is not always true (Qi & Ge 2006). Protein–DNA interactions, for example, are inferred by detecting physical complexes that do not necessarily imply repression or activation. As these artefactual interactions may not be randomly distributed, they could induce the detection of ‘pseudo’-functional modules.

Ultimately, regardless of the approach, a robust criterion for assessing the relevance of functional modules is to prove that their activity is associated to physiological response or a pathological condition.

(d) The activity of functional modules is predictive of physiological response

In order to illustrate that functional modules can be predictive of physiological response, we have used a simple procedure integrating gene expression information with KEGG pathways. In this approach, a functional module is defined within a KEGG pathway by the genes that show a highly significant
correlation across the samples of the expression-profiling dataset. The overall activity of the module is then computed as the average expression profile of its gene components.

The question we have addressed is whether a combination of functional modules can be predictive of the genetic rearrangements associated to different forms of acute lymphoblastic leukaemia (ALL). We have defined KEGG functional modules using a dataset representing the expression profiles of 233 patients affected by five forms of leukaemia that are associated to different chromosomal translocations or abnormalities (Yeoh et al. 2002).

Once the overall activity of the functional modules has been computed, these have been used as variables in a statistical modelling procedure (Trevino & Falciani 2006), with the objective to determine whether their activity may be associated to chromosomal rearrangement. Figure 1 shows that the activity of 16 modules is sufficient to distinguish disease type (represented graphically using hierarchical clustering). The visual inspection of the heatmap representing the expression levels of these modules in the different classes of patients is sufficient to identify clear disease-specific patterns. Other prediction methods based on the overall activity of a functional module have been developed. These are based on either principal component analysis (Bair et al. 2006) or more complex scores (van Vliet et al. 2007).

Figure 1. Functional modules predictive of ALL subtypes. A heatmap representing the expression profile of the average profile of genes included in the functional modules identified as predictive of ALL subtypes. Red refers to the expression higher than the median value of the distribution of signal, black is the median value and green represents the expression below the median value. Rows (modules) and columns (clinical samples) are arranged by cluster analysis.
4. Linking cellular and molecular components to phenotypic response using statistical modelling

(a) A study using simulated data

In §3d, we have seen how modularization can help to identify groups of functionally associated genes whose coordinate expression is associated to phenotypic response. This methodology helps the development of a model starting from complex datasets. It does that by providing a high-level view of what the important cellular processes may be associated to phenotypic responses. Individual components of these modules may then be further analysed and be the basis for the development of a black-box model.

An alternative approach for the identification of informative components of a biological system is the use of supervised classification or regression methodologies, coupled to multivariate variable selection procedures. These approaches are based on a procedure that identifies optimal subsets of gene expression profiles that maximize the prediction of a phenotypic variable using a statistical model. In our group, we have developed variable selection methods based on genetic algorithms and Bayesian variable selection for the selection of predictive variable subsets. We have shown that these can be more effective than univariate variable selection methods that do not take into account the interaction between different genes (Sha et al. 2004; Trevino & Falciani 2006). The effectiveness of these modelling techniques has been assessed using statistically simulated datasets and applied to experimental data (Larranaga et al. 2006).

Here, we address the question whether these methodologies may have a value in system identification. In particular, we wonder whether they are able to identify components of the circuitry that control phenotypic response. The statistical methodology we have tested is based on a genetic algorithm procedure that selects the subsets of genes that are predictive of a given phenotype. In this case, the phenotype is predicted using a simple non-parametric supervised classification method (K-nearest neighbour (KNN); for further details on the methodology used see the electronic supplementary material).

In order to assess whether this relatively simple methodology has the potential to identify components of a subnetwork closely associated to a phenotypic response, we have developed ordinary differential equation (ODE) simulated datasets. These model networks represent the integration of different modular pathways into an in silico simulated phenotypic response. Figure 2 summarizes the design of these networks in a schematic format. Each network comprises a number of independent modules that can be thought of as representing independent biological pathways collecting signals from the external environment. Some of these are isolated from the other modules; others are connected to a common node (integration node) that, in turn, determines a binary phenotypic state via a threshold effect. Three types of modules are possible. These can be linear pathways (type I), cyclic pathways (type II) and large pathways with a complex topology (type III). There are two different types of integration. These can be either linearly additive or generated by a more complex nonlinear mechanism involving the formation of homo- and heterodimers of the products of the upstream modules.
Once the data are simulated, the expression values of the module nodes and the phenotypic outcome are used as an input to the statistical modelling procedure. The state of the integration node, which has been used to induce phenotypic response, is considered a hidden process that is not measured.

Figure 2. Schematic of the networks used to generate simulated data. (a) The high-level structure of the networks. Each module’s output converged in an integration node whose state determines a binary phenotype. (b) The topologies of the three different types of network modules, (c) the two different kinetics of integration and (d) a histogram representing the distribution of the state concentration of the integration node. The threshold determining the state of the phenotypic node is set on the average of the distribution.

Once the data are simulated, the expression values of the module nodes and the phenotypic outcome are used as an input to the statistical modelling procedure. The state of the integration node, which has been used to induce phenotypic response, is considered a hidden process that is not measured.
Table 1. Results of the classification models applied to networks formed by type I and type II modules.

<table>
<thead>
<tr>
<th></th>
<th>type I, linear</th>
<th>type II, linear</th>
<th>type I, nonlinear</th>
<th>type II, nonlinear</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$N_1$</td>
<td>$N_2$</td>
<td>$N_3$</td>
<td>$N_4$</td>
</tr>
<tr>
<td></td>
<td>0 X X</td>
<td>X X</td>
<td>X X</td>
<td>X X</td>
</tr>
<tr>
<td></td>
<td>X X</td>
<td>X X</td>
<td>X X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>X X X</td>
<td>X X X</td>
<td>X X X</td>
<td>X X X</td>
</tr>
<tr>
<td></td>
<td>X X X X X</td>
<td>X X X</td>
<td>X X X</td>
<td>X X X</td>
</tr>
</tbody>
</table>

Modelling molecular networks and physiology
Using combinations of different modules and integration kinetics, we have generated five different datasets. Four of these are based on the combination between six modules of type I or type II with a linear or nonlinear integration. Type I and type II modules are formed by 10 nodes each. Within a given network, three of them will connect to the integration node. The fifth network type is formed by the combination of 10 type III modules. Four of these will converge into the integration node that combines their outputs with a linear kinetic. Since 428 nodes form each type III module, the whole network has a size that is comparable with an average expression-profiling dataset (further details on the implementation of the ODE model and original datasets are available at http://www.bip.bham.ac.uk/fernando/Royal_Society/Supplementary_Material.pdf). We have used a statistical modelling methodology with a genetic algorithm procedure to select informative variable subsets coupled to a simple non-parametric classifier (KNN; Trevino & Falciani 2006). The results of our analysis are very encouraging. Nodes represented in the models predictive of phenotypic response and developed from the first set of four simulated datasets are part of the modules contributing to the state of the terminal node (table 1). Of great interest is the fact that, when the method has been applied to data simulated from the large network with type III modules, only one out of the 48 selected nodes does belong to a module that is not contributing to the phenotype. Interestingly, nodes identified in the modules contributing to the phenotype are forming interconnected subnetworks in close proximity of the phenotypic node (figure 3a).

(b) Identifying molecular networks associated to cancer patho-physiology

The results described in §4a support our initial intuition that multivariate variable selection strategies are able to identify subnetworks that are closely associated to phenotypic responses. However, can we extrapolate this conclusion from simulated data to a real biological system with noisy measurements? In order to address this question, we have used expression-profiling data (Singh et al. 2002) to develop classification models predictive of the ability of prostate cancer cells to penetrate the organ capsule. Capsular penetration is a binary variable and has a great clinical importance as a negative prognostic factor. In this analysis, we want to determine whether genes chosen by the model selection procedure described in §4a would be part of a gene network that has the potential of directly influencing the capsule penetration. On the basis of the previous results, we expect this to be the case. With real experimental data, we do not have access to the structure of the underlying molecular network. However, databases containing information on experimentally supported gene

Figure 3. (Opposite.) Networks identified by classification models. (a) The structure of the subnetworks belonging to (i) module 1, (ii) module 3, (iii) module 4 and (iv) module 5, which have been identified on the large type III networks by the supervised classification procedure described in §4a. The procedure has identified three of the four modules contributing to the phenotype, whereas only one node from one of the six isolated modules ((iv) module 5) has been identified. (b) The structure of the network linking the genes identified by the model we have developed to predict tumour capsular penetration in prostate cancer. The network represents the IL-6 pathway reconstructed using information available in the literature. The legend on the left shows the key to identify node types, type of the interaction and type of supporting information.

Phil. Trans. R. Soc. A (2008)
Figure 3. (Caption opposite.)
interactions can be used to test this hypothesis. In this case, we have used the Ingenuity pathway analysis (IPA) software (Ingenuity Systems, www.ingenuity.com). IPA is an analysis toolbox connected to a database of experimentally supported gene interactions. The IPA software implements a statistical test to assess whether a given list of genes is significantly associated to a molecular subnetwork of known gene interactions. Figure 3b shows the highest scoring network associated to genes represented in a model predictive of capsular penetration. The network represents a pathway centred on IL-6, a cytokine whose activity has been proven to be a prognostic indicator in prostate cancer (Culig et al. 2005). These results support the findings on simulated data that we described in §4a and the view that statistical models can be used on real data to identify components of the molecular networks specifying phenotypic response. Singh et al. (2002) employed a univariate method for the identification of genes predictive of capsular penetration. Although predictive models could be identified, this approach was unable to identify a gene network associated to IL-6. The hypothesis that IL-6 is a central cytokine in regulating gene expression in prostate cancer epithelial cells is amenable to experimental verification. This can be achieved by either increasing or reducing the level of IL-6 (e.g. using purified IL-6 or anti-IL-6 blocking antibodies) in an in vitro or in vivo tumour growth model.

(c) Using biological knowledge for informed model search

In the methodologies we have described, subsets of genes, which are initially randomly selected, are refined to improve prediction accuracy using a search procedure. In this process, only a small portion of the extremely complex solution space is explored. It is therefore reasonable to expect that a large number of potentially interesting solutions are missed by the search process. In this context, we may expect that a variable selection strategy, which is guided by a biologically sensible expectation of where the solutions are, would be more effective than a random search. We have tested this hypothesis in a number of biological systems. In our biologically driven searches, we have assumed that genes belonging to certain high-level functional groupings may have a higher chance of being involved in the process of interest. For example, the process of metastases formation from a solid tumour may more probably be associated to the interaction between extracellular proteins, membrane receptors and transcription factors. A drawback of this approach is that models may be biased towards expected relationships. However, it is reasonable to believe that a biologically driven search identifies highly predictive models that are normally missed by an unbiased search. From a methodological point of view, biologically driven variable selection can be performed using a genetic algorithm with conditional gene inclusion probabilities. However, variable selection procedures based on Bayesian methods (Sha et al. 2004) may offer a better approach to the problem, since biological expectation can be formalized as a Bayesian prior.

The identification of functional modules and molecular components that are predictive of physiological responses is a very important step in modelling complex biological systems from large-scale functional genomics datasets. A further challenge is the identification of the structure of the circuit itself. This can be achieved using a range of network inference techniques.

Phil. Trans. R. Soc. A (2008)
5. Inferring the connectivity of cellular networks

(a) Methodology

In the last few years, a number of methods for network inference have been developed (for a recent review on the subject see Bansal et al. 2007). These include relevance networks (RNs), information-theoretic approaches such as ARACNE (algorithm for the reconstruction of accurate cellular networks), BNs and differential equations-based methods. RNs (Butte & Kohane 2000) are based on pairwise association scores calculated from gene expression values. Several metrics have been used with this methodology, including mutual information and correlation. Networks generated from randomly resampled data can be used to threshold connections generated from the real data and generate a sparse network. ARACNE (Basso et al. 2005) is, in practice, a more advanced version of an RN. It computes the mutual information between all pairs of genes and eliminates potentially indirect connections using the data processing inequality principle. A BN (Hartemink 2006) uses a rather different approach to model gene–gene dependences.

In order to reverse-engineer a gene regulatory network, we need to find the BN that better fits the dataset D. This is achieved by using the BS (a scoring function derived from the Bayes theorem) to evaluate a sample of network topologies. Ultimately, the topology \( T \) that maximizes the score is chosen. The BS is based on the product between the prior probability and the likelihood, which is estimated from the data. The prior probability may represent some prior knowledge of the network topology. This knowledge can be either some experimental evidence of gene–gene interaction or some constraints on the sparsity of the network. Alternatively, the prior may be constant, which is basically an admission of ignorance of the network topology. Searching for an optimal model structure is an NP-hard (non-deterministic polynomial-time hard) problem. For this reason, optimization procedures such as Markov chain Monte Carlo or simulated annealing methods are commonly used to find a local optimal solution. Such searches can lead to many high-scoring networks. Bootstrap, model averaging and other approaches can be used to overcome this problem and select the most likely network edges (Hartemink 2006).

Another approach uses differential equations to reverse-engineer gene networks (D’haeseleer et al. 1999; Tegner et al. 2003; Guthke et al. 2005; di Bernando et al. 2005; Bansal et al. 2006; Bonneau et al. 2006; van Someren et al. 2006). In this formalism, the rate of change in the concentration of a gene product at time \( t \) is described as a function of the concentration of all other gene products in the network. This approach has been traditionally used to model the behaviour of a system when its structure is at least partially known. In using differential equations for reverse engineering, a certain form of relationship (i.e. linear or nonlinear) must be chosen and the parameters should be determined to best fit experimental data. One of the advantages of specifying a given model structure is that it may be designed to better resemble a biologically meaningful mechanism.

One of the first differential equations-based approaches for reverse engineering biological networks is based on a linear function (D’haeseleer et al. 1999). In this approach, the authors consider a simple case where the expression level of one gene is the weighted sum of the expression levels of other genes. Gardner et al. (2003)
used a very similar formalism to reconstruct networks at the steady state. This method (network identification by multiple regression (NIR)) requires knowledge of which genes have been experimentally perturbed. Within the boundaries of this requirement, the algorithm may be more powerful than other competing methodologies. More recently, rather complex nonlinear formalisms have been used to better describe the complexity of biological systems (S-systems and H-systems; Spieth et al. 2006). User-friendly software environments for reconstructing gene networks are available for BNs (Saeed et al. 2003; Hartemink 2006), information-theoretic approaches (Basso et al. 2005), RNs (Saeed et al. 2003) and differential equations-based approaches (Spieth et al. 2006).

(b) Validation of network inference methodologies

Since the first application of reverse-engineering techniques to the analysis of biological systems, many groups have performed validation studies on simulated data or applied them on experimental data.

Two recent applications of these techniques appear to be particularly indicative of the ability of reverse-engineering methods to lead to the formulation of a hypothesis on novel biological mechanisms. Basso et al. (2005) applied ARACNE to model B-cell molecular networks from a compendium of 336 expression-profiling experiments representative of a wide selection of normal, transformed and experimentally modified B-cells. This study represents an attempt to perform network inference in very extreme conditions where the interactions of several thousands of components are inferred from only a few hundred samples. ARACNE recovered a topology that resembles high-level properties of biological regulatory networks (i.e. scale free). Using this model, they proposed new candidate Myc target genes. This approach has been shown to be extremely successful since over 90 per cent of the inferred targets were experimentally verified. The high accuracy demonstrated in the case of the targets of Myc is, however, not likely to reflect the overall accuracy of the whole inferred network. The relationship between transcription factors and their targets is in fact direct and it is regulated by a simpler kinetic than the more intricate regulatory mechanisms involving secondary messengers and post-translational modifications.

At the opposite side in the scale of complexity, a pioneering study was performed by Sachs et al. (2005), which used BNs to model 11 components of the Raf signalling pathway. The data analysed by these authors were produced by multiparameter fluorescence activated cell sorting (FACS) analysis. This technique is extremely powerful, since it allows one to analyse hundreds of thousands of individual cells in single experiments. Their model recovers over 85 per cent of experimentally validated directed connections. Unfortunately, this technique can only be applied to measure a very small number of proteins for which good detection reagents are available.

Different inference methodologies have been compared systematically by Werhli et al. (2006) and Bansal et al. (2007). Werhli et al. (2006) have used the Raf pathway as a model system with both simulated and experimental data. They found that BNs produce significantly better results than RNs when edge directions are taken into account. More recently, Bansal et al. (2007) used a number of synthetic networks simulating time course with local (a single gene is perturbed) and global (all genes are perturbed simultaneously in the network)
perturbation experiments to compare the accuracy of ARACNE, BNs, clustering and NIR. Overall, the results of these validation studies suggest that both BNs and RNs have a very strong potential in reverse engineering biological circuits for which a sufficiently large number of experimental samples is available. The authors also show that dynamic BNs are largely ineffective. RNs are generally less accurate than BNs and NIR but have the advantage that they can be applied to very large datasets. On the other hand, BNs take into consideration the interaction of multiple nodes and are more accurate when applied to a smaller number of components. In their analysis, NIR showed a better performance than BN and ARACNE on local perturbation experiments. Although these results strongly support the validity of the differential equations-based methodologies in the analysis of steady-state data, it should be noted that a possible explanation for the extremely high specificity shown (up to 96%) is based on the fact that the data were simulated from systems of differential equations of the same structure as the model used in the inference process.

(c) Observations on reverse engineering using BNs

The results of many of the validation studies performed with BNs and other reverse-engineering techniques are very encouraging and suggest that probabilistic modelling may allow inference of molecular networks from experimental data. There are, however, a number of issues that are important to take into consideration when developing BN models. In this section, we intend to illustrate these using simulated data from networks as described in §4. We have focused our analysis on static BNs as implemented in Banjo, a Java software library developed by Hartemink (2006). Table 2 shows that the results of our analysis with datasets of 100 samples are very much in accordance with what other groups have reported. Overall, the network structure of linear (type I) and cyclic (type II) networks is inferred from the simulated data with good accuracy. Also, the accuracy is much higher than for a subset of randomly selected models. However, as expected, the effectiveness of the inference process is lower for the data derived from networks containing cyclic structures, which theoretically cannot be inferred by static BNs (figure 4b). Also, accuracy drops dramatically if directionality is taken into account. Interestingly, the method appears to be largely ineffective in inferring the connection between modules and the integration node that we use to simulate a phenotypic response (figure 4a,b).

In order to illustrate the effectiveness of the model selection procedure, we have analysed the distribution of the BS of the best models identified by 1000 independent searches performed using simulated annealing. The optimization procedure used in the model selection is reasonably effective, since the distribution of the BS representing the networks selected by the optimization procedure is very different from a distribution of the BS of 1000 randomly selected models (figure 5). We have then compared the distribution of the BS for the selected models with the BS of the real structure (figure 5). Two things are worth noting. The first is that most often the optimization procedure selects structures with a BS that is lower than the real structure, suggesting that the procedure drives the model search towards local maxima. The second is that, interestingly, a small but significant fraction of selected models show a BS higher...
Table 2. Results of the classification models applied to networks formed by type I and type II modules.

<table>
<thead>
<tr>
<th></th>
<th>directed sensitivity (%)</th>
<th>directed accuracy (%)</th>
<th>undirected sensitivity (%)</th>
<th>undirected accuracy (%)</th>
<th>random directed sensitivity (%)</th>
<th>random directed accuracy (%)</th>
<th>random undirected sensitivity (%)</th>
<th>random undirected accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>type I, linear (61×100)</td>
<td>38</td>
<td>35</td>
<td>62</td>
<td>75</td>
<td>1.5</td>
<td>4.1</td>
<td>2.9</td>
<td>8.2</td>
</tr>
<tr>
<td>type II, linear (61×100)</td>
<td>51</td>
<td>31</td>
<td>68</td>
<td>59</td>
<td>1.6</td>
<td>4.0</td>
<td>3.1</td>
<td>7.9</td>
</tr>
<tr>
<td>type I, nonlinear (61×100)</td>
<td>37</td>
<td>32</td>
<td>61</td>
<td>72</td>
<td>1.4</td>
<td>4.0</td>
<td>2.8</td>
<td>8.0</td>
</tr>
<tr>
<td>type II, nonlinear (61×100)</td>
<td>50</td>
<td>32</td>
<td>67</td>
<td>58</td>
<td>1.6</td>
<td>4.0</td>
<td>3.2</td>
<td>7.9</td>
</tr>
<tr>
<td>type I, linear (61×1000)</td>
<td>60</td>
<td>60</td>
<td>70</td>
<td>92</td>
<td>1.5</td>
<td>4.1</td>
<td>2.9</td>
<td>8.2</td>
</tr>
<tr>
<td>type III, linear (46×100)</td>
<td>7</td>
<td>23</td>
<td>13</td>
<td>48</td>
<td>2.1</td>
<td>5.4</td>
<td>3.9</td>
<td>10.0</td>
</tr>
</tbody>
</table>
Figure 4. BNs developed from simulated data. The structure of models inferred from data simulated using (a) linear and (b) cyclic networks coupled with linear integration has been shown in a schematic format (additional plots representing other network types are available in figure 4 of the electronic supplementary material). The numbers on the r.h.s. of the edges report the number of models (out of 1000) supporting the connection, whereas the numbers on the l.h.s. report the information for the number of models supporting the connection with its correct direction.
than the real structure. A possible explanation for this observation is that, with the sample size used in this analysis, the BS only partially describes the structure of the underlying network (figure 3 of the electronic supplementary material). In order to achieve high accuracy of both directed and undirected graphs, the size of the dataset needs to be drastically increased (table 2 shows the results for a dataset of 1000 samples). This observation helps in understanding the reasons why BN models of signalling pathways developed using multiparameter FACS analysis are very accurate (§5b).

Unfortunately, in many biological systems, we are unable a priori to identify a small defined signalling pathway and measure so many different samples. Instead, genes are selected using statistical criteria (e.g. using the modularization procedures described in §3c,d or the classification/regression-based approaches described in §4) and are usually measured only in hundreds of biological samples. In this case, the reverse-engineering procedure is challenged by the fact that most of the information that may influence the ‘state’ of the nodes we have identified as predictive cannot be included in the inference process. This scenario is illustrated by applying BN methods to reconstruct the network structure connecting the 48 genes we have identified from the simulated dataset derived from the large network. The accuracy of the inference is acceptable (48%) but not as good as with the other simulated data.

The efficiency of the network inference process, however, can be improved using a priori knowledge on the network structure. BNs offer a good framework for the inclusion of biological knowledge either at the level of prior probability (soft priors) and/or by embedding constraints at the level of model search (hard priors) or by starting the model search from a non-random topology that incorporates connections that are supported by experimental data (Hartemink 2006).

Strategies to deal with hidden information have also been developed. They use methodologies that assume the existence of hidden factors in the inference process (Ong et al. 2002; Rangel et al. 2004; Beal et al. 2005; Sabatti & James 2005). An extensive evaluation on the benefit of including hidden factors in the modelling, however, is still needed.
6. Conclusions: towards a strategy for system identification of complex biological systems

In this paper, we have described a number of strategies that can be used to identify the structure of complex molecular networks. We have shown that this can be done at different levels. At a higher level, in the organization of molecular networks, we have functional modules defined as groups of genes whose overall activity is associated to a specific function. We have described examples of module identification procedures that have been demonstrated to be effective in predicting physiological or pathological response. Using simulated and experimental data, we have also shown that, although modularization algorithms are not explicitly used, classification and regression models are very effective in identifying connected subnetworks that are closely associated to the expression of a phenotype. These methodologies provide the means to discover the ‘important components’ of a system but do not reveal their underlying network structure. At a lower level in the organization of molecular networks, reverse-engineering methodologies are very effective in inferring the structure of an underlying network but have limitations in terms of the number of components that can be analysed, the type of topology that can be discovered and their ability to capture the integration of different signals converging from independent pathways.

We have also seen that both variable selection and network inference methods benefit from the inclusion of biological knowledge. This can be either in the form of a general expectation of the type of components that are important to the process (i.e. the interaction between extracellular matrix and membrane receptors) or as a priori information on the possible structure of the interaction network.

From our overview, it appears that system identification of biological systems cannot be performed by a single technique. Instead, it should be based on the combination of the different methodologies we have described and involve the integration of different sources of biological knowledge and experimental data. An example of a possible strategy is summarized in figure 6.

Models developed using this strategy can be used to simulate the behaviour of a complex system (e.g. including different levels of biological organization) to generate in silico predictions of the effects of gene knockouts or overexpression experiments. Ultimately, models representative of the complexity of molecular and physiological processes may involve the integration between multiple models based on different formalisms.

Modelling languages, which allow the standardization of the model representation, such as SBML (Finney & Hucka 2003) and CElLML (Cuellar et al. 2003), have been developed and are (at least in the case of SBML) widely used in the community.

Modelling languages are useful to represent and manipulate biological models. However, model integration involves the development of a framework, which allows the use of heterogeneous models, performs model and data management and aims to use these models in a synergetic fashion (Hetherington et al. 2007).

Recently, a number of approaches that fit this definition have been proposed in the systems biology community (as reviewed in Hetherington et al. 2007). Some examples are as follows: E-Cell project (Takahashi et al. 2003); the Virtual Cell (Loew & Schaff 2001); the XS system (Antoniotti et al. 2003); and the Beacon project (Hetherington et al. 2007).
None of these frameworks have yet been applied to the integration of probabilistic and deterministic models. However, Hetherington et al. (2007) have proposed the use of model interaction connectors to integrate BNs and

Figure 6. Overall analysis strategies for system identification. Schematic of a possible strategy for combining the different statistical methods described in this paper. Experimental data are first used for the identification of gene subsets or modules that may be associated to phenotypic responses. In this process, biological knowledge can be used for both modularization (A) and variable selection in the development of the classification/regression models (B). Classification/regression models are used in combination with phenotypic information (C and D). Once gene subsets or modules predictive of physiological responses have been identified, their network structure can be inferred using network inference techniques, which use experimental data and may incorporate biological knowledge (E).
differential equations-based models. A model interaction connector allows the simultaneous execution of different models. Both the E-cell and Beacon projects use this methodology.

An important challenge in the development of a future generation of system identification methodologies is the ability to incorporate measurements from different levels of cellular and molecular organization (e.g. by explicitly integrating the relationship between gene expression, translation and post-translational events). Ultimately, this will allow the proper integration of different ‘omics’ measurements into the network inference process.

The work described in this paper has been partially supported by the European Commission (BIOBRIDGE project). We acknowledge the support of the NERC Postgenomics and Proteomics programme award NE/C507702/1 and the NERC International Opportunity award NE/D000793/1. V.T. is a recipient of a fellowship from the Darwin Trust of Edinburgh. R.C. is funded by a BBSRC PhD studentship. Computer simulations were carried out using the Birmingham Biosciences High Performance Compute Cluster funded by the BBSRC REI grant BB/D524624/1.

References


Phil. Trans. R. Soc. A (2008)