Review TRENDS in Genetics Vol.23 No.1

Perturbations to uncover gene networks

Jesper Tegnér[1,2,3] and Johan Björkegren[2,3]

[1] Division of Computational Biology, Department of Physics, Chemistry and Biology, The Institute of Technology, Linköping University, SE-581 83 Linköping, Sweden
[2] Unit of Computational Medicine, Center for Genomics and Bioinformatics, Karolinska Institutet, SE-171 77 Stockholm, Sweden
[3] The Computational Medicine Group, Center for Molecular Medicine, Department of Medicine, Karolinska Institutet, Karolinska University Hospital, Solna, SE-171 76 Stockholm, Sweden

After the major achievements of the DNA sequencing projects, an equally important challenge now is to uncover the functional relationships among genes (i.e. gene networks). It has become increasingly clear that computational algorithms are crucial for extracting meaningful information from the massive amount of data generated by high-throughput genome-wide technologies. Here, we summarise how systems identification algorithms, originating from physics and control theory, have been adapted for use in biology. We also explain how experimental perturbations combined with genome-wide measurements are being used to uncover gene networks. Perturbation techniques could pave the way for identifying gene networks in more complex settings such as multifactorial diseases and for improving the efficacy of drug evaluation.

Introduction

The Human Genome Project identified a surprisingly small number of full-length genes (~30 000) [1,2], a number almost identical to that in mice [3] and not much greater than the number in simpler organisms such as the fruit fly (13 600) [4] and yeast (6000) [5]. Moreover, there are striking similarities in the biochemical building blocks (genes, proteins, metabolites) across these species. The secret of human complexity must in part lie in the non-coding DNA [6] but also in the interactions between genes that are mediated by proteins binding to regulatory regions [7]. Revealing the architecture of gene interactions in networks under normal and pathophysiological conditions is therefore of considerable interest.

Analyzing biological systems: from molecular parts to networks and their dynamic interactions

To understand the exact mechanisms underlying a biological process or a disease, it is first necessary to identify the relevant components (DNA, transcripts, proteins, metabolites) that are involved in the process or disease. Of central importance for obtaining such a 'parts lists' are high-throughput screening technologies, developed in parallel with the genome projects. In humans, relatively reliable whole-genome lists can be extracted using DNA arrays. At the protein level, it is possible to extract the identity of a large number of proteins (although not yet reliable for the entire genome) using two-dimensional (2D) gels and mass spectrometry [8]. Variants in the genome sequence such as single nucleotide polymorphisms (SNPs) could also be important to elucidate because some of these variants might affect gene expression and thus be relevant for the biological question at hand. Nowadays, a large number of genome-wide SNP analyses are available [9].

The second challenge is to determine how these parts interact in networks – the focus of this review. Again, high-throughput technologies have been key, and they have led to a surge of systems biological approaches aimed at elucidating the architecture of functional relationships among genes, proteins and metabolites. Thus far, most of these studies have focused on gene networks (Box 1) in organisms such as *Escherichia coli* [10] and *Saccharomyces cerevisiae* [11,12]. However, scientists are increasingly taking on new challenges, and addressing more complex systems, including mammalian cells [13–15], organs [16] and complex diseases such as cancer [17] and cardiovascular diseases [18].

Finally, the dynamics of gene networks must be worked out. Although not addressed here, this topic will become increasingly important as gene networks are unravelled [14,19]. The dynamic properties of a network include the kinetics of interactions between and among genes and proteins and whether an interaction is activating or repressing. Clearly, computational tools to identify directly these properties from experimental data will be required. In addition, analytical tools will be essential for understanding the dynamic behavior within these networks. Pioneering studies on network dynamics have most notably been performed on the cell cycle by Tyson and colleagues [20].

Gene network identification: hurdles and solutions

The era of modern networks research began in the 1960s following the publication in 1959 of the classic work by Erdos and Renyi [21], who developed the mathematical theory for networks, where nodes are randomly connected by edges. Since then, applied mathematicians and physicists have been busy understanding the interplay between the topology and dynamics of networks [22]. Researchers have studied how to characterize different wiring diagrams in cells, computers and the internet. An important area of

Corresponding author: Tegnér, J. (jespert@ifm.liu.se).
Available online 13 November 2006.

 0168-9525/$ – see front matter doi:10.1016/j.tig.2006.11.003

Box 1. Biological systems as networks

All biological systems can be described in terms of networks that consist of parts (nodes) and their interactions (edges). Because each node receives input from other nodes through the corresponding edges, each node has effectively an input–output function that describes how the inputs are transformed into an activity for the receiving node. The complexity of this function depends on the nature of the network. Within a cell, there are several interconnected networks, such as the protein interaction network and the metabolic network [12,30,66,67]. Here, proteins and metabolites are nodes, and the edges describe either protein–protein interactions or biochemical reactions among metabolites. Most nodes have few connections; however, a few nodes act as hubs (i.e. nodes with a large number of connecting edges) [12]. Other entities, such as genes, can also be considered as nodes in a network where the interactions are mediated by means of the proteins.

Figure Ia illustrates a biological network of interactions between eight hypothetical genes and their respective proteins. An mRNA is first transcribed from DNA and then translated into a protein that can then interact with another protein (protein–protein interaction) or bind to DNA (transcription factor DNA-binding). The effective interaction between the genes is obtained by collapsing Figure Ia into an effective gene network (Figure Ib). A protein interaction with a transcription factor (TF) in the underlying gene–protein network (Figure Ia) renders a regulatory edge to the gene target of the TF in the effective gene network. For example, in the gene–protein network (Figure Ia), protein d interacts with TF g, which in turn regulates the expression of gene E. In the effective gene network (Figure Ib), gene D is therefore a regulator of the expression of gene E. The number of incoming regulatory connections, corresponding to the parameter *k*, representing the in-degree ($k = 4$ for E), is shown in parentheses within the gene symbols in the effective gene network (Figure Ib). In such a 'collapsed' network, the effective input–output function of a given node (gene) is the integration of all the intermediate reactions connecting the two genes in question. When the regulation is dominated by post-translation mechanisms, it might be better to maintain a separate representation of proteins and genes, provided that it is possible to obtain measurement data on the posttranslational modifications of the proteins.

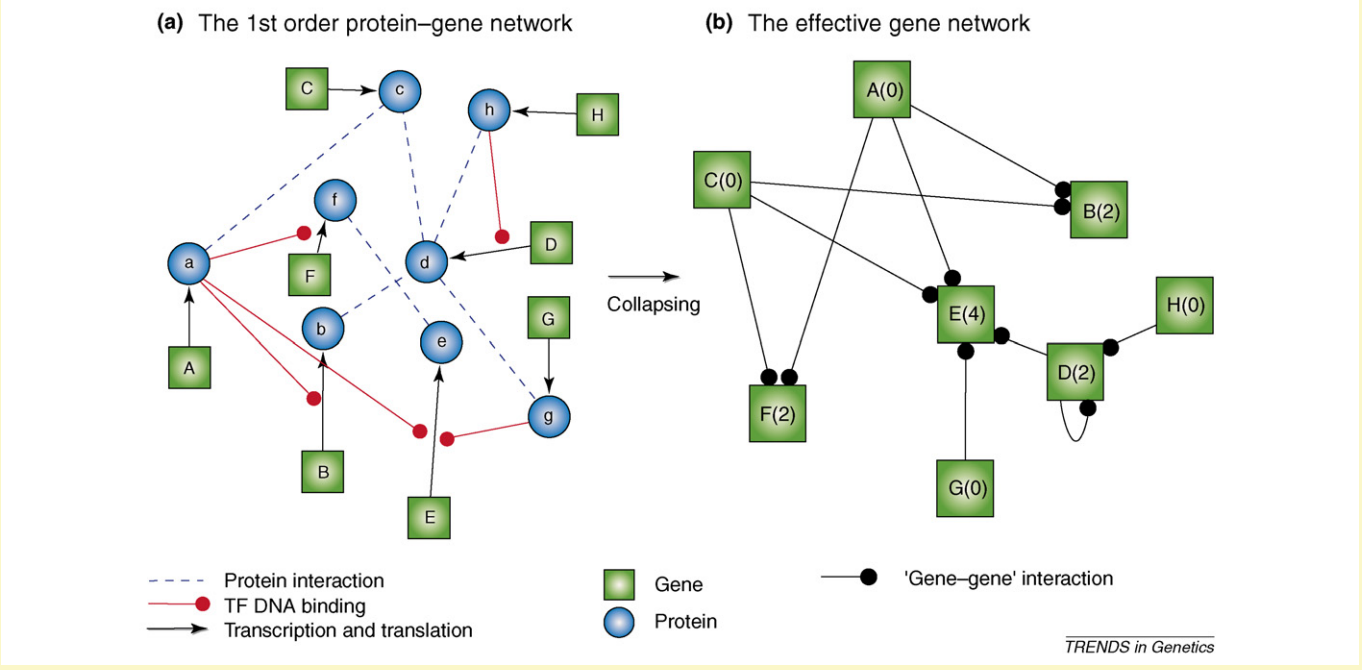

Figure I.

study is how global network dynamics are determined by the input-output properties of individual nodes and the coupling dynamics between nodes. In the late 1990s, the group of Barabasi and co-workers demonstrated that real networks – computer networks, the internet, and metabolic and protein networks – are not randomly organized but are instead characterized by a small number of well-connected nodes and a majority of nodes with only a few interactions [12]. With the rapid development of powerful computers, sequencing of genomes and the parallel development of whole-genome technologies for monitoring the expression of genes and proteins, network identification has become a serious player in biology and medicine.

Identifying gene networks from large-scale dataset measurements is a difficult computational and experimental problem. The main reason for this is the bewilderingly large number of possible wiring diagrams, without even considering the detailed reaction kinetics (i.e. the dynamics of a gene network), even for a small network of only three genes (Box 2). The fundamental challenge for a gene network identification algorithm is the great dimensionality of the number of features (nodes, genes), which gives a large number of different network wiring diagrams, in relation to the small number of experimental samples for use in discriminating between the different networks.

To illustrate this problem, an analogy might be helpful. In algebra, it is not possible to solve an equation system without additional constraints if the number of unknowns $(x_1, x_2, \ldots x_n)$ exceeds the number of equations. Every equation is like a sample that constrains the possible values of the unknowns (nodes) that jointly satisfy the equation. To obtain a unique solution of the equation system it is therefore necessary to have as many equations (samples) as unknowns (nodes). Unfortunately, because there are fewer experimental samples (equations) than genes (unknowns), current cluster analysis techniques [23] can only establish

Box 2. Level of network resolution determines difficulty of network identification

From a combinatorial viewpoint, the problem of identifying a network is difficult. As an example, let us consider a small three-gene network (Figure Ia). The protein products from gene A might or might not influence the expression of gene A. Similarly, gene B or gene C might or might not influence the expression of gene A. In sum, there are eight possibilities for the regulation of gene A and 512 ($8 \times 8 \times 8$) possible directed wiring diagrams in a small three-gene network, where the direction of the arrow indicates causality.

However, if we increase the biological resolution by including the sign of the causality, the number of possible wiring diagrams for a three-gene network is even larger. The protein products from gene A can increase or decrease the expression of gene A. Similarly, gene B or gene C could activate or repress gene A (Figure Ib). In sum, there are 27 different activation or repression possibilities for the regulation of gene A and 19 683 ($27 \times 27 \times 27$) networks in a small three-gene system. Similar calculations for a four-by-four or a five-by-five gene network, without considering the signs, result in 65 536 and 33 554 432 possible wiring diagrams, respectively.

Complex as this might seem, such a wiring diagram description is still a simplification of the underlying biology. For each causal wiring diagram, the input–output properties of each node could range from 'simple' linear summation of the inputs to a more complex nonlinear integration of the different input signals. Finally, there is a large number of possible magnitudes for the activation and repression for each possible causal description. In either case, the example of a small-scale gene circuit illustrates that an exhaustive experimental search, such as testing every possible wiring diagram, is not feasible. Hence, any attempt to identify networks from experimental data cannot succeed without a tightly integrated computational and experimental strategy. A computational model of the system can be used to suggest what the next most informative experiment is to learn more about the network structure.

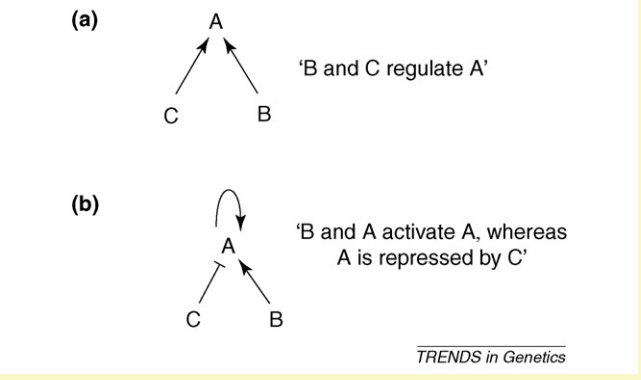

Figure I.

whether a gene A is 'correlated' with gene B (Figure 1a). Clustering of gene expression data provides no direct information on the underlying wiring cellular diagram.

What strategies might we use to solve this key issue? The first would be to increase the number of experiments. During the last couple of years, costs for whole-genome expression studies have steadily decreased, and libraries of hundreds of whole-genome-wide expression profiles for yeast are now available [24,25]. As a rule, the ratio of genes to samples is between 10:1 and 100:1. How much this ratio needs to be lowered to enable accurate network identification is an open question. The answer depends on the computational complexity of the problem and on experimental signal-to-noise levels. In addition, the more complex (nonlinear) the regulatory control mechanisms are, the more samples will be required. Hence, increasing the number of samples will probably not, by itself, solve the network identification problem.

A second strategy is to incorporate prior knowledge. In algebra, the number of possible solutions is severely reduced if only positive integers, rather than any real number, are allowed. Hence, by analogy, prior knowledge can in part compensate for the smaller number of samples in relation to the number of nodes (genes) by reducing the number of possible networks that are consistent with the available experimental observations. For example, there are biological reasons to expect the gene network is sparsely connected [12]. Knowledge of connections that have already been characterized would obviously be useful. These constraints would simplify the problem because not all putative wiring diagrams would be accepted as equally good. Many studies have integrated gene expression data with a prior knowledge network [15,26,27]. Most followed a two-step procedure. First, a prior knowledge network map is constructed by assembling information from a variety of sources, such as transcription factor binding data [28], published literature (text mining) [29], and protein–protein interactions [30] (Figure 1b). This provides a wiring diagram but as a rule there is no information on causality and strengths of the interactions. Next, to improve further the analysis, whole-genome expression data relevant to the biological issue at hand are used to identify a subset of interactions that are considered to be active. However, using prior knowledge in this manner has been severely limited because the gene expression data have only been used to filter out a subset of already characterized interactions. Thus, it has not been possible to infer novel interactions from the prior knowledge network map *per se* or by adding gene expression profiles to the analysis.

A third strategy is to reduce the number of nodes (genes) by collapsing them into functional subgroups based on the correlations in the gene expression activity. This is the strategy behind modular analyses [31]. A large number of nodes (genes) are therefore included within a module. A modular analysis therefore identifies a network of modules and the connections between the modules. Again, this analysis does not provide a causal wiring diagram with the strength of interactions. Although such an analysis is a useful first step for stratifying the genome into functional modules, which can be useful for a later detailed network analysis, these techniques do not by themselves consider the network architecture between the nodes within the modules and thus do not enable a comprehensive gene network identification and will not be considered further.

The computational perturbation approach

The computational perturbation approach originates from engineering sciences, such as control theory and physics. Instead of using only observational data, as is generally the case in biological and medical sciences, engineering sciences combine a computational representation of the system of interest with controlled perturbations. By perturbing the system and analyzing the response, parameters in the computational model can be identified. The chosen level of detail at which the computational

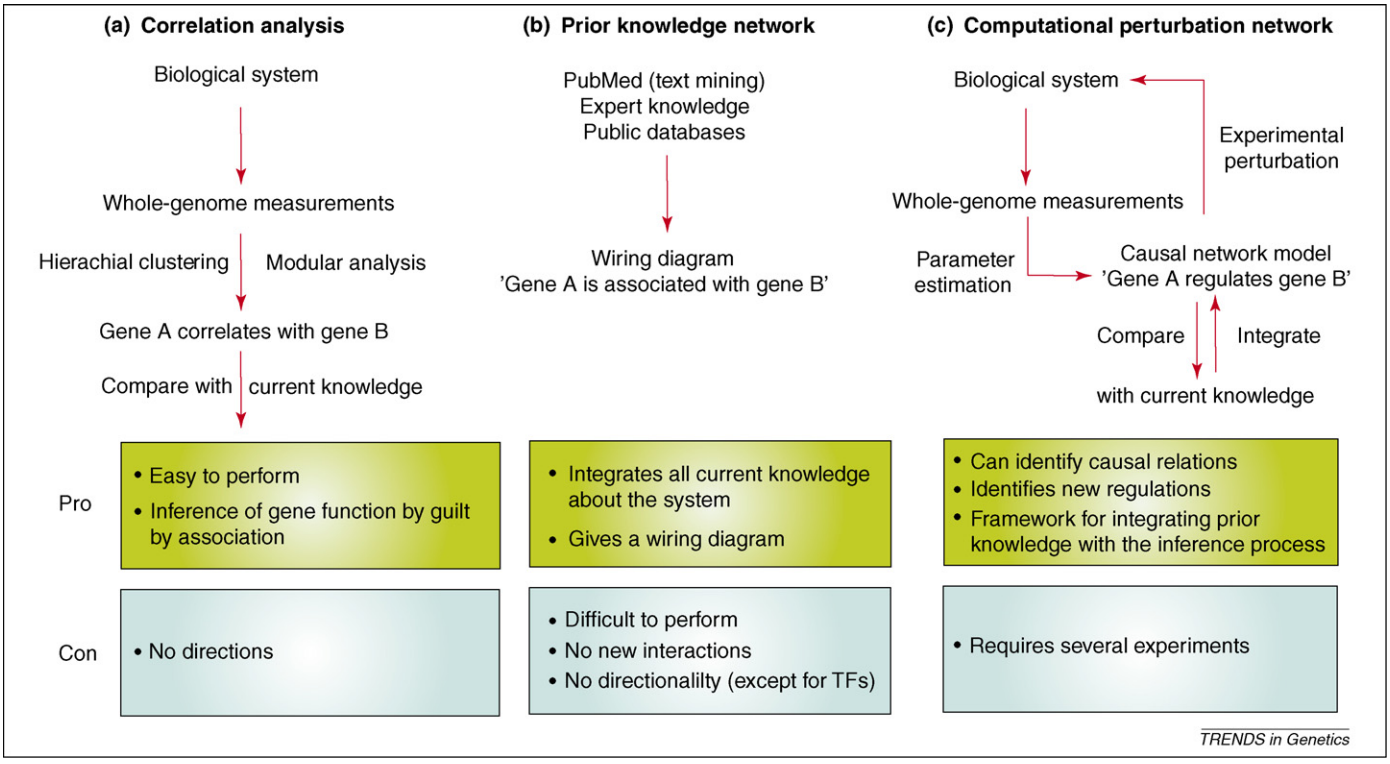

Figure 1. Flowcharts illustrating the difference between correlation analysis **(a)**, prior knowledge networks **(b)** and a computational perturbation network identification scheme **(c)**. Pros and cons for each scheme are given; example references include [24,31] (a); [15,27,29] (b) and [41,43,50] (c).

model reflects the true system depends on the particular engineering discipline and the application.

In control theory, for example, tools for system identification are based on a computational representation of the true system, referred to as a transfer function, which is determined by comparing the input (perturbations) with the system response (measurements). The perturbation experiments determine the parameters of the transfer function, which then can be used to predict the system response to different input data.

The above procedure used in control theory is also the principle for using perturbation techniques in biology – that the system (represented by a computational model) responsible for transforming an input signal to an output signal can be characterized by sending different input signals (perturbations) and monitoring the system response. The perturbation approach can therefore be viewed as a way to increase the number of samples nonrandomly by choosing those that will generate the most useful information. Importantly, engineering sciences have (as a rule) focused on describing a system well enough to predict the response. In applying the perturbation approach to biological problems, however, the intrinsic structure of the system is the main interest.

Several studies [32–34] have demonstrated that the perturbation approach can be applied to identify biological networks. The principle for this application (Figure 1c) is the use of a computational model that is more detailed than a transfer function, to represent the underlying system (i.e. a biological network). Identifying a network from data is then equivalent to determining the parameters of the computational model from experimental data. Here, a wiring diagram is obtained where each edge has a number that represents the strength of the activation or repression between the nodes (genes). Such a map provides molecular hypotheses that can be experimentally investigated. Importantly, this analysis is not limited to edges that have previously been described. The idea of perturbing a system and measuring the response is well established in experimental biology. However, measuring the system response using high-throughput technologies provides novel challenges because of the large amounts of data produced. These challenges can be addressed by using an underlying computational model and reducing the identification problem to a parameter estimation problem in an engineering style. Thus, the computational model facilitates the choice of selecting genes to perturb and sets the stage for integrating prior knowledge directly into the network inference procedure.

Several molecular methods (Figure 2a) can be used as perturbation tools, including knockouts [24], small interfering RNA (siRNA), and overexpression of selected genes [25]. To monitor the response of the system (i.e. the network), whole-genome measurements, most commonly gene expression profiling, are being used. An important issue in gene network identification from whole-genome expression measurements is to estimate the noise levels. The noisier the data, the more samples are required to identify the network [32].

Perturbation themes: insights and challenges

Box 2 illustrates one of the simplest computational models to represent genes and their interactions using a three-gene example. Recently, several studies using different levels of biological resolution have addressed how networks of several hundred genes can be identified (see ref. [35] for a recent mathematical review). To increase

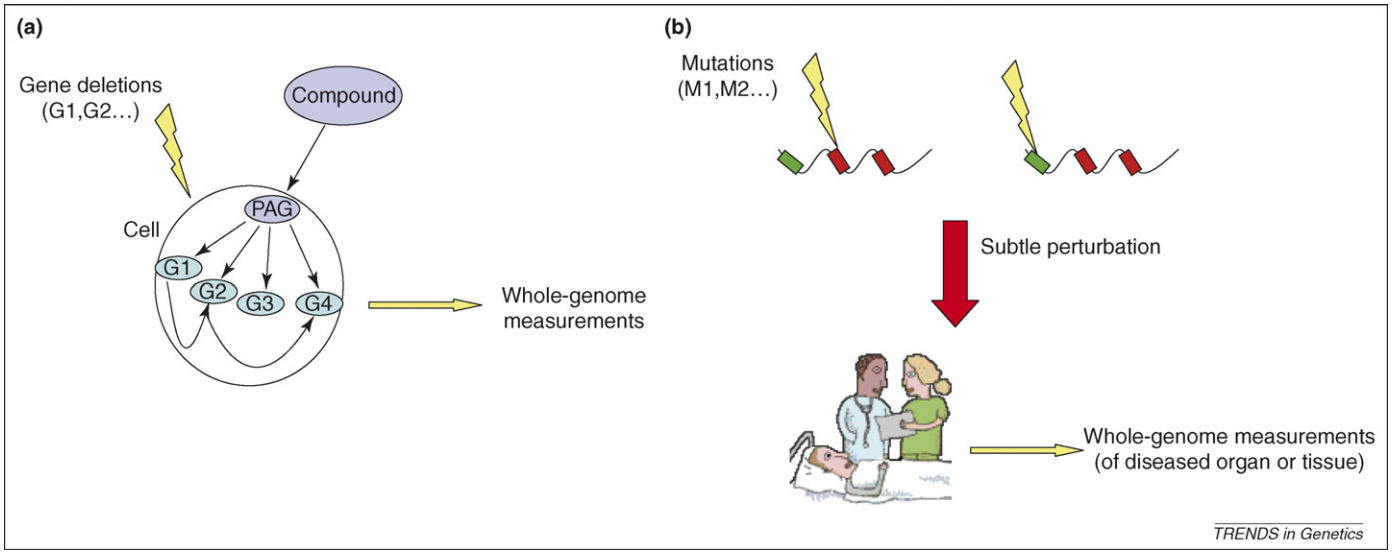

Figure 2. Future potential for perturbation-based gene network identification strategies. **(a)** Compound target identification. Network identification based on a perturbation approach has been proposed to be useful for identifying the primary affected gene (PAG) of compounds [61]. PAGs will be the most frequently differentially expressed network gene(s) over several whole-genome measurements of targeted cells within a given compound. If the network architecture is known, the position of the PAGs in the network can be calculated [68]. If the network architecture is unknown, repeated whole-genome measurements in cell lines incubated with the compound in combination with different mutant strains will still enable the identification of PAGs using the CutTree algorithm [62]. The basic logic for this proposal is that the PAGs will be affected by the compound regardless of the nature of the deletions. Currently, perturbations are executed mainly in the form of gene deletions in cell lines of yeast or prokaryotic organisms. In humans and animal model systems, the genome-wide deletion strategy is prohibited. **(b)** Etiology of complex disease. We and others [65] suggest that, instead of deletions, naturally occurring disease-related mutations can be considered as subtle disease perturbations (as compared with deletions from which disease networks can be extracted) and, hence, useful for uncovering gene networks underlying the etiology of complex diseases.

further the biological resolution beyond a graphical model, differential equations are useful because they incorporate gradations of gene activity and different strengths of genetic interactions. Tegnér and colleagues introduced the concept of integrating differential equations with experimental perturbations to identify gene networks [32] whereas Akutsu and others developed a perturbation approach using a discrete model [34,36]. Friedman and colleagues pioneered a probabilistic perturbation formulation using Bayesian models [33]. The discrete representation models the interactions as either on or off whereas the interactions in Bayesian models encode probabilities between random variables for explaining the data. Despite these differences in the underlying model representation all algorithms basically fit the expression profiles to the computational model and thereby identify the parameters describing the network [35,37]. Indeed, several common insights and challenges have been revealed in these studies.

Perturbations to reduce sample number

Despite the large number of possible wiring diagrams for even a small network (Box 2), the perturbation approach enables us to identify the correct wiring diagram with surprisingly few samples in relation to the size of the network. As a rule, for a network of N genes, the number of samples required to identify the correct wiring diagram is proportional to a constant multiplied by $\log(N)$ [32,34,36,38]. The constant includes both a term for the signal-to-noise ratio and k, a parameter that indicates the number of incoming regulatory connections to the genes (Box 1). Using a reduced model for data analysis and assuming reliable measurements of the activity, a typical network of say 10 000 genes can be identified with fewer than 100 perturbations. The analysis [32,34,36,38] demonstrated that by choosing a set of perturbations that collectively activate a large part of the network, the large number of genes in the network is not, by itself, a limiting factor for solving the network identification problem.

Network sparseness

Computational analysis [32,34,36,38] showed that k (the number of incoming regulatory connections to a gene) strongly influences the number of required whole-genome measurements to identify the gene network. The rationale behind this result is that the number of possible regulatory combinations for a given gene increases dramatically when k increases and there is a large number of genes (N) in the network (see also Box 2). For example, with 100 genes and three inputs per gene, there are $\sim 10^6$ different input combinations for each gene in the network, resulting in $\sim 10^{600}$ different network wiring diagrams. The number of samples required to identify a gene network from expression profiles is linear with k [32,36,38]. Current network algorithms cannot identify gene networks if k is large. Thus, current studies have as a rule constrained k to be less than five inputs despite some evidence that k could be large [39]. Fortunately, there is ample evidence that gene networks are sparse and that $k << N$ is a reasonable approximation [12,28,40]. An important future challenge will be to develop efficient network identification algorithms when k is large.

Signal-to-noise ratio

The number of samples required to identify gene networks is also linearly proportional to the inverse of the signal-to-noise ratio [32]. Hence, the measurement resolution must be sufficient to distinguish whether two distinct genes, x_i and x_j, have different gene expression values after the perturbation. The computational analysis therefore clearly underlines the importance of increasing

the signal-to-noise ratio by using replicates and adopting sound algorithms for normalization and probe-level analysis of microarray data.

Linear computational models and beyond

An underlying computational model with greater biological resolution generally has more parameters. Its identification therefore requires more data points and thus more experiments to identify the gene network. For this reason, it is advantageous to use simplified models that require less data but are flexible enough to account for the underlying gene network. The choice of model complexity depends on the accuracy of the measurement technology used and the amount of prior knowledge of the system under study. One can therefore question how realistic these simplifications are. Experimental tests of the approach using linear differential equations data [32,36,38] demonstrated that despite these simplifications of the biological regulatory complexity, linear differential equations are sufficient to recover a nine-gene SOS, DNA repair gene network in *E. coli* from gene expression data [41]. The free parameters in the underlying linear differential equation, representing connections between genes, were identified by fitting the gene expression data to the model by regression. In the case of the SOS *E. coli* network, the inferred connections were compared with previously described connections. This important experimental validation together with another large-scale experimental validation in yeast [42] demonstrates the usefulness of a linear representation of the underlying gene regulation. Both the SOS network and the yeast network were good enough to predict the mode of action from the expression profile of the respective chemical compounds. Recent network identification studies have shown that the inferred networks are in accordance with current knowledge [13,43–47].

However, the next step is clearly to develop network identification algorithms that can recover the nonlinear dynamics of cellular networks. However, this requires the incorporation of nonlinear aspects of the gene or protein regulation. *In silico* studies are necessary to determine initially how to design experiments optimally using multiple simultaneous perturbations to reveal the nonlinear regulation of the nodes (genes, proteins or metabolites). This direction of research requires not only more experimental data but also clever incorporation of conditions such as prior knowledge on both interactions and functional forms of the node regulation to reduce the possible solutions space. This is a major future challenge, because these properties will most probably be important for performing a detailed dynamic analysis of smaller regulatory networks. A recent study, using a Bayesian model formulation, demonstrated the usefulness of focusing on a gene of interest, a seed gene, and then reconstructing the local network around the seed gene [48]. This strategy, which can be combined with a modular analysis, reduces the number of genes and could provide a sound basis for developing locally correct dynamical network models.

General applicability of the perturbation approach

Recent studies have demonstrated the general applicability of a perturbation approach [11,13,49–59]. One particularly illustrative example is the analysis of protein networks by Sachs *et al.* [50]. These investigators used a perturbation approach using simultaneous measurements of multiple phosphorylated proteins and phospholipid components in human immune cells. The cells were experimentally perturbed, and an underlying Bayesian model enabled the group to recover several known connections and to predict novel interactions. Importantly, these predicted interactions were verified experimentally. This study demonstrates the usefulness of integrating an underlying computational model with experimental perturbations and measurements. Finally, as this case study also illustrates, there is nothing inherent in the perturbation approach that limits its applicability to gene networks.

Prior knowledge

To reduce the number of possible wiring diagrams for a given system, it is important to incorporate prior biological knowledge. Current approaches [15,26,27] use a prior knowledge network (Figure 1b) together with gene expression data to define which passive edges from the prior knowledge network are to be considered active. Here, prior knowledge and the gene expression data are treated separately. A perturbation approach can in principle readily incorporate prior biological knowledge directly into the inference algorithm (Figure 2b). However, how to incorporate a prior knowledge network into an inference algorithm remains a challenge. In principle, integrating such data requires an underlying mathematical model for analysis because different types of data (transcription factor binding, protein binding data) can be used to estimate the parameters in the model (i.e. network structure). Such an algorithm can exclude a large number of solutions (network structures) that are not compatible with current knowledge; because prior knowledge originating from different data types is directly used, algorithms thereby increase inference power. An important challenge ahead is therefore to design a new generation of network identification algorithms that can infer novel interactions while harnessing the power of a large prior knowledge database integrated with the inference algorithm.

Using perturbations to improve drug evaluation and understand complex diseases

Interestingly, causal gene networks have proven to be useful in two practical applications beyond the basic science quest of characterizing and understanding molecular networks. First, integration of computational techniques with experimental perturbations is increasingly being used to meet challenges in drug evaluation and development [60]. For instance, it has become clear that compounds with many targets interacting in several pathways can be investigated using gene network inference based on experimental perturbations [61] (Figure 2a). Knowledge of the gene network architecture could also be useful to assess possible side effects of existing drugs and compounds. In essence, combining gene expression profiles associated with administration of a particular drug or compound with the network architecture enables a 'backwards' calculation of the primary mechanism of action of the drug or compound [41,42]. This is a consequence of a

linear modelling approach, because here it is sufficient to have only two of the following three components: the input perturbation (action of the drug or compound), the network model and the systemic whole-genome response (output). However, a recent study generalized this approach by demonstrating that an additional input perturbation delivered in parallel to the (unknown) drug or compound perturbation enables the mechanism of action to be inferred from the system response even without a network model [62] (Figure 2a).

Second, uncovering gene networks could also be essential for understanding complex diseases such as cancer and cardiovascular disease. Thus far, the 'candidate gene approach' has governed the search for novel targets and diagnostics. Also, fuelled by the successes of identifying genetic variation underlying single-gene disorders (mendelian traits), positional cloning approaches have been used to search for common variants underlying complex traits, thus far without success [63]. Recently, however, Schadt and coworkers elegantly showed that integrating genetic variants underlying complex quantitative trait loci (QTLs) with gene expression traits (eQTLs) greatly improves the chance of identifying genes of importance for complex traits [64]. The authors suggested that DNA variants should be viewed as subtle perturbations and therefore might be useful for identifying gene networks underlying complex traits [65].

The use of genetic variants as subtle perturbations is mainly limited to identifying genes whose expression is governed by genetic variants. It is likely that a substantial number of genes that govern a complex trait are not influenced (at least directly) by genetic factors but instead by environmental factors and by the disease itself (reactive gene expression). Currently, we are testing the slightly controversial notion that using the underlying phenotypes of complex disorders as perturbations can reveal gene networks underlying complex traits.

Concluding remarks

The principles underlying the use of perturbations are fundamental. For instance, if deep in a forest you find a living creature that you have never seen before, you will probably pick up a stick and poke it carefully to see its reaction. From the reaction you will most probably learn something about this creature that was not apparent from merely looking at it. Using large-scale perturbations of biological systems in combination with detailed molecular monitoring of the system response and a computational model for data analysis is likely to open up new and exciting perspectives of the molecular networks governing life and disease, systems about which we so far know little.

Acknowledgements

We thank the reviewers for useful comments. J. Lundström is acknowledged for help with Figure I in Box 1. This work is funded by the Swedish Foundation for Strategic Research, the Swedish Society for Medicine and the Swedish Research Council.

References

1 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
2 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
3 Waterston, R.H. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–562
4 Adams, M.D. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science* 287, 2185–2195
5 Goffeau, A. *et al.* (1996) Life with 6000 genes. *Science* 274, 546, 563–547
6 Carninci, P. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science* 309, 1559–1563
7 Kitano, H. (2002) Computational systems biology. *Nature* 420, 206–210
8 Wittmann-Liebold, B. *et al.* (2006) Two-dimensional gel electrophoresis as a tool for proteomics studies in combination with protein identification by mass spectrometry. *Proteomics* 6, 4688–4703
9 International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* 437, 1299–1320
10 Dobrin, R. *et al.* (2004) Aggregation of topological motifs in the *Escherichia coli* transcriptional regulatory network. *BMC Bioinformatics* 5, 10
11 Ideker, T. *et al.* (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292, 929–934
12 Barabasi, A.L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113
13 Basso, K. *et al.* (2005) Reverse engineering of regulatory networks in human B cells. *Nat. Genet.* 37, 382–390
14 Ma'ayan, A. *et al.* (2005) Toward predictive models of mammalian cells. *Annu. Rev. Biophys. Biomol. Struct.* 34, 319–349
15 Calvano, S.E. *et al.* (2005) A network-based analysis of systemic inflammation in humans. *Nature* 437, 1032–1037
16 Crampin, E.J. *et al.* (2004) Computational physiology and the Physiome Project. *Exp. Physiol.* 89, 1–26
17 Rhodes, D.R. and Chinnaiyan, A.M. (2005) Integrative analysis of the cancer transcriptome. *Nat. Genet.* 37 (Suppl.), S31–S37
18 Ghazalpour, A. *et al.* (2004) Thematic review series: the pathogenesis of atherosclerosis. Toward a biological network for atherosclerosis. *J. Lipid Res.* 45, 1793–1805
19 Ehrenberg, M. *et al.* (2003) Systems biology is taking off. *Genome Res.* 13, 2377–2380
20 Tyson, J.J. *et al.* (2001) Network dynamics and cell physiology. *Nat. Rev. Mol. Cell Biol.* 2, 908–916
21 Erdos, P. and Renyi, A. (1959) On random graphs. *Publicationes Mathematicae* 6, 290–297
22 Strogatz, S.H. (2001) Exploring complex networks. *Nature* 410, 268–276
23 Quackenbush, J. (2001) Computational analysis of microarray data. *Nat. Rev. Genet.* 2, 418–427
24 Hughes, T.R. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell* 102, 109–126
25 Mnaimneh, S. *et al.* (2004) Exploration of essential gene functions via titratable promoter alleles. *Cell* 118, 31–44
26 Lee, I. *et al.* (2004) A probabilistic functional network of yeast genes. *Science* 306, 1555–1558
27 Luscombe, N.M. *et al.* (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* 431, 308–312
28 Lee, T.I. *et al.* (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298, 799–804
29 Jenssen, T.K. *et al.* (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.* 28, 21–28
30 Rual, J.F. *et al.* (2005) Towards a proteome-scale map of the human protein–protein interaction network. *Nature* 437, 1173–1178
31 Segal, E. *et al.* (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* 34, 166–176
32 Tegnér, J. *et al.* (2003) Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proc. Natl. Acad. Sci. U. S. A.* 100, 5944–5949
33 Friedman, N. (2004) Inferring cellular networks using probabilistic graphical models. *Science* 303, 799–805
34 Ideker, T.E. *et al.* (2000) Discovery of regulatory interactions through perturbation: inference and experimental design. *Pac. Symp. Biocomput.* 5, 302–313
35 Gardner, T.J. and Faith, J.J. (2005) Reverse-engineering transcriptional control networks. *Physics of Life Reviews* 2, 65–88
36 Akutsu, T. *et al.* (1999) Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Pac. Symp. Biocomput.* 4, 17–28

37 de Jong, H. (2002) Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.* 9, 67–103
38 D'Haeseleer, P. *et al.* (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16, 707–726
39 Davidson, E.H. *et al.* (2002) A genomic regulatory network for development. *Science* 295, 1669–1678
40 Guelzim, N. *et al.* (2002) Topological and causal structure of the yeast transcriptional regulatory network. *Nat. Genet.* 31, 60–63
41 Gardner, T.S. *et al.* (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* 301, 102–105
42 di Bernardo, D. *et al.* (2005) Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat. Biotechnol.* 23, 377–383
43 Gustafsson, M. *et al.* (2005) Constructing and analyzing a large-scale gene-to-gene regulatory network – lasso-constrained inference and biological validation. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 2, 254–261
44 Thorsson, V.H. *et al.* (2005) Reverse engineering galactose regulation in yeast through model selection. *Stat. Applicat. Genet. Mol. Biol.* 4 Article 28
45 Papin, J.A. *et al.* (2005) Reconstruction of cellular signalling networks and analysis of their properties. *Nat. Rev. Mol. Cell Biol.* 6, 99–111
46 Rung, J. *et al.* (2002) Building and analysing genome-wide gene disruption networks. *Bioinformatics* 18 (Suppl. 2), S202–S210
47 Nilsson, R. *et al.* (2006) Transcriptional network dynamics in macrophage activation. *Genomics* 88, 133–142
48 Pena, J.M. *et al.* (2005) Growing Bayesian network models of gene networks from seed genes. *Bioinformatics* 21 (Suppl. 2), ii224–ii229
49 Kremling, A. *et al.* (2004) A benchmark for methods in reverse engineering and model discrimination: problem formulation and solutions. *Genome Res.* 14, 1773–1785
50 Sachs, K. *et al.* (2005) Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 308, 523–529
51 Shmulevich, I. *et al.* (2002) Gene perturbation and intervention in probabilistic Boolean networks. *Bioinformatics* 18, 1319–1331
52 Wagner, A. (2001) How to reconstruct a large genetic network from n gene perturbations in fewer than n(2) easy steps. *Bioinformatics* 17, 1183–1197
53 Wagner, A. (2004) Reconstructing pathways in large genetic networks from genetic perturbations. *J. Comput. Biol.* 11, 53–60
54 Hardy, K. *et al.* (2005) Transcriptional networks and cellular senescence in human mammary fibroblasts. *Mol. Biol. Cell* 16, 943–953
55 Perkins, T.J. *et al.* (2004) Inferring models of gene expression dynamics. *J. Theor. Biol.* 230, 289–299
56 Markowetz, F. *et al.* (2005) Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. *Bioinformatics* 21, 4026–4032
57 Yeang, C.H. *et al.* (2004) Physical network models. *J. Comput. Biol.* 11, 243–262
58 Ideker, T. (2004) A systems approach to discovering signaling and regulatory pathways – or, how to digest large interaction networks into relevant pieces. *Adv. Exp. Med. Biol.* 547, 21–30
59 Gadkar, K.G. *et al.* (2005) Iterative approach to model identification of biological networks. *BMC Bioinformatics* 6, 155
60 Burley, S.K. and Park, F. (2005) Meeting the challenges of drug discovery: a multidisciplinary re-evaluation of current practices. Keystone Symposium 'Meeting the Challenges of Drug Discovery', Vancouver, Canada, 15–19 January 2005. *Genome Biol.* 6, 330
61 Csermely, P. *et al.* (2005) The efficiency of multi-target drugs: the network approach might help drug design. *Trends Pharmacol. Sci.* 26, 178–182
62 Hallen, K. *et al.* (2006) Detection of compound mode of action by computational integration of whole-genome measurements and genetic perturbations. *BMC Bioinformatics* 7, 51
63 Xiong, M. *et al.* (2005) A systems biology approach to genetic studies of complex diseases. *FEBS Lett.* 579, 5325–5332
64 Schadt, E.E. *et al.* (2005) An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* 37, 710–717
65 Schadt, E.E. *et al.* (2005) Embracing complexity, inching closer to reality. *Sci. STKE* 2005, pe40
66 Ramani, A.K. *et al.* (2005) Consolidating the set of known human protein–protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol.* 6, R40
67 Bork, P. *et al.* (2004) Protein interaction networks from yeast to human. *Curr. Opin. Struct. Biol.* 14, 292–299
68 Ambesi-Impiombato, A. and di Bernardo, D. (2006) Computational biology and drug discovery: from single-target to network drugs. *Curr. Bioinf.* 1, 3–13