From microarrays to networks: mining expression time series

T. Gregory Dewey

Over the past few years, powerful new methods have been devised that enable researchers to study the expression dynamics of many genes simultaneously (e.g. gene expression profiles using cDNA microarrays). In principle, this potentially vast quantity of data enables the dissection of the complex genetic networks that control the patterns and rhythms of gene expression in the cell. Finding the patterns in those data represents the next major phase in our understanding of the programming and functioning of the living cell. Simple dynamic models can be used to generate gene expression networks. These networks reveal the phenomenological link between the expression of different genes. This review discuss how these networks are generated and outlines several data-mining techniques for extracting relationships and hypotheses in gene expression. These emerging methods can be applied to a range of biological problems.

T. Gregory Dewey
Keck Graduate Institute of Applied Life Sciences
535 Watson Drive
Claremont
CA 91711, USA
tel: +1 909 607 8586
fax: +1 909 607 8598
e-mail: greg_dewey@kgi.edu

▼ High-throughput technologies enable genome-wide interrogation of biological systems. These technologies assist in enumerating the many parameters and variables associated with life processes and reveal the inherent complexities of these processes. The current era is marked by ongoing efforts to assimilate and integrate this avalanche of information into present models of biological functions. The complexity of biological systems is associated not only with the large number of interacting components but also with the complicated dynamics that they exhibit. An emerging problem in bioinformatics is identifying the relationships between the various components of a system and, specifically, how one component impacts on the production of another. The molecular circuitry of gene regulation reveals the dynamics of how one effector or agent influences the entire network. Knowledge of this circuitry implicitly enables the manipulation of gene expression, and has far reaching consequences for drug target identification.

The importance of time-series data

Often, high-throughput methods, such as gene expression arrays, focus on genome-wide profiles of individuals from a population. From a dynamic point of view, this represents a 'snap shot in time' of a potentially heterogeneous population. The complexity of this snapshot arises from two influences. First, no two organisms are alike, even in the same species, and genetic variation will influence the expression profile. Second, each organism has its own history, and this history can lead to a wide range of dynamic states of the system with very different expression profiles. To distinguish between intrinsic genetic variation and the dynamic variation is a challenging task.

There are real advantages to determining expression profiles as a function of time. Just as chemical kinetics yield mechanistic information in a much more straightforward fashion than chemical thermodynamics, expression time-series data are more amenable to network modeling than expression data from a population. In time-series data, an organism is exposed to some perturbation and the response of gene expression is monitored. Time series profiles have been measured in a wide range of systems, including responses to media growth conditions (diauxic shift in yeast [1]), cell-cycle synchronization [2], exposure to vaccines [3], signaling responses to cytokines (M. Bechtel *et al.*, unpublished results), and mechanical stimulation and insect feeding in *Arabidopsis* [4]. While time-series data can be more difficult and expensive to obtain, the distinct advantage is that it is readily amenable to mechanistic interpretation.

The following sections discuss how time-series data obtained from microarray experiments

1359-6446/02/$ – see front matter PII: S1359-6446(02)02440-6

can be modeled in terms of gene networks. These phenomenological networks show the influence of the expression level of one gene on another. They can be directly obtained from the data and can be used as a data-mining and gene classification tool. Also discussed is how networks obtained in this manner can be compared both to each other and to other quite different networks, such as protein–protein interaction networks. Such comparisons provide additional means of mining the underlying expression data.

Networks from time series data

The great advantage of investigating time-series gene expression data is that gene networks can be readily derived from the data using simple dynamic models. These networks describe how the mRNA level of one gene influences the level of another. However, these are not true gene-regulatory networks in the traditional sense because they are not necessarily causal networks. They show a phenomenological link as dictated by the data and the model, rather than a direct causal link. As such, they should be taken as a starting point for data mining and hypothesis generation.

Perhaps the simplest model for analyzing expression time-series is a Markovian linear response model [5,6]. In this model, the expression state at one time point determines the expression state observed at the next time point. The transition between the two states is modeled by:

$$a_i(t) = \sum_{j=1}^{m} \lambda_{ij} a_j(t-1) \qquad [1]$$

where $a_i(t)$ is the expression level of the ith gene at time t after some exposure or treatment, and m different genes are measured. The transition λ_{ij} coefficients are the respective elements of the transition matrix (referred to as the Λ matrix). The matrix elements represent the influence of the expression level of the jth gene on that of the ith gene. The Λ matrix is calculated from a time-series dataset using a generalized matrix inversion technique [5], and is used to construct the underlying gene expression network.

It might, at first, seem overly simplistic to consider a linear model such as that shown in Eqn 1. However, it is important to explore the limitations of linear models before moving on to more complicated non-linear ones. There have been several recent attempts to analyze time-series data for whole-genome expression profiles [5–8]. Interestingly, this does not require the complexity of detailed non-linear models of gene expression, but needs only simple, linear models [5–7]. These previous studies focused on cell-cycle and diauxic shift data in the yeast *Saccharomyces cerevisiae* [1,2]. In both cases, the system is prepared in a given physiological state at the initial time point, and changes in gene expression levels are measured as it moves to a new state. These experiments have some similarity to traditional perturbation–relaxation experiments in physics and chemistry. Given this analogy, it is perhaps not surprising that the time dependence of the expression profiles can be well represented by simple linear response models. Such models can be an effective way of 'reverse engineering' gene networks, especially when additional biological constraints exist [9]. The linear models can also be readily extended to include non-linear terms that incorporate quadratic terms related to time and gene correlations [5].

The matrix generated by these data analyses is a weighted graph showing the interactions between gene expression levels. We simplify the analysis by using a sparse, binary matrix representation of the adjacency matrix [5,6]. This is achieved by applying a threshold to the entries in the transition matrix. The absolute values of the matrix elements are set equal to 1 if they are above a certain threshold, ε, and set equal to 0 below this threshold. For high values of the threshold, the resulting matrix will be a sparse adjacency matrix. This is a di-graph (non-symmetric matrix) showing the connectivity of the biological network. We do not differentiate here between positive and negative values for members in the transition matrix, as we are only interested in the underlying connectivity.

The networks derived by this method show a hierarchical structure that is dominated by a collection of central hubs. These hubs are interconnected; an example of such a structure is shown in Fig. 1 for the cell-cycle data mentioned previously [2]. The figure shows the strongly coupled components (SCC) determined from the adjacency matrix using the 'depth-first' search algorithm from the algorithmic graph theory [10]. The SCC of a graph is a sub-graph with the property that, given any two nodes, a and b, in the sub-graph, there exists a sequence of directed edges from a to b and from b to a. There is also a large number of nodes in the yeast data that have one-way connections (recipients) and are, therefore, not in the SCCs.

A similar global network structure has been observed for all the expression data analyzed to date. The resulting networks have properties common to other large networks, such as the Internet or electrical power grids [6]. They also show some similarities to a class of networks known as 'small world' networks. These expression networks are tightly connected in that the average shortest path from any one node to another is quite low, which enables easy global communication. They have very strong neighborhood structures and are described as 'cliquish'. A third characteristic property is that they show a strong hierarchy of node connectivity. There are few nodes (the hubs) with many connections, and many nodes with few connections. This distribution of edges follows a power law and is sometimes referred to as a 'scale-free' distribution. Metabolic networks showing the connectivity of substrates show high

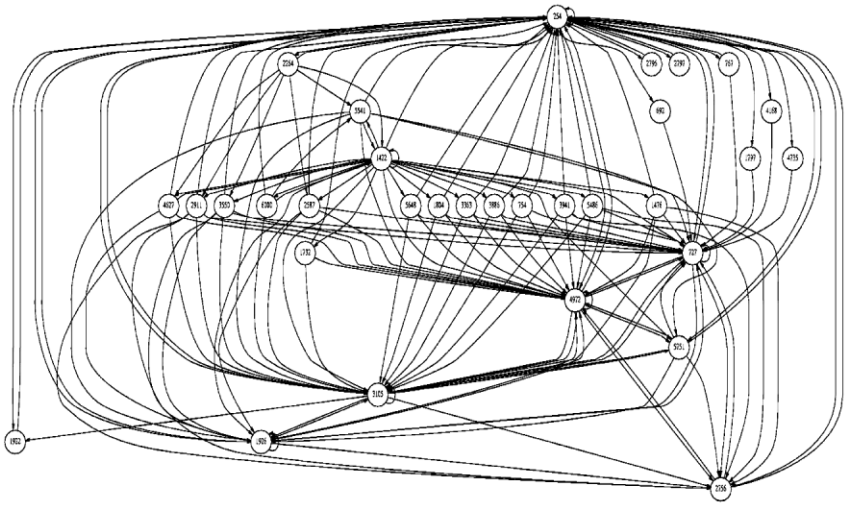

Figure 1. A typical strongly connected component (SCC) constructed from the analysis of yeast time-series data. Nodes represent genes, and directed edges represent the kinetic influence of one gene on another, as determined from the dynamic model. Note the hub-like structure that is commonly seen for yeast data.

'cliquishness' and a scale-free distribution of edges [11]. The yeast protein–protein interaction map also has similar properties. Although similar in some ways, the results for biological networks show consistently different behavior than large man-made networks [6].

Mining networks

A serious challenge in the analysis of gene expression dynamics is the validation of the resulting networks. This can be done in two ways. First, various statistical devices, such as bootstrapping and resampling methods, can be used to manipulate the data and residuals from the model (see [12]). This enables an estimation of the robustness of the parameters and network structure obtained from the data (see [6]). Alternatively, one can identify known genetic regulatory mechanisms to determine whether the resulting networks conform to them. The difficulty with the latter approach is that many of the transcription factors that are involved in such mechanisms are expressed in very low amounts and are outside the dynamic range of current microarray techniques. Nevertheless, literature mining is an essential tool used to validate the correlations seen in the networks.

Gene classification with networks

The goal of the network analysis need not be a strictly quantitative predictive model. Instead, these networks can be used both for data mining, and for providing ways of organizing the data and generating hypotheses from it. These latter applications show considerable promise and provide an alternative to traditional methods such as cluster analysis. Classification is achieved by identifying genes with common network properties. For the yeast cell-cycle data, most of the cell-cycle-regulated genes are in the trees hanging off the hub structure. They are grouped into specific topological regions of the network graph. We have performed a 'topological-sort' [10] on this network and have found the 'depth' of the various cell-cycle-regulated genes in the network. This enables the identification of other cell-cycle-regulated genes by their topological proximity to genes of known regulatory control. This method therefore provides a scheme for classifying genes. In general, this approach yields results similar to cluster analysis, but a comprehensive comparison remains to be done.

Network connectivity and function

A second data-mining technique is to examine the functionality of different genes in the network and observe the connectivity of different functional domains. This is best illustrated by an example from the yeast cell-cycle data in which the cells were synchronized by using the α mating factor. The functional domains are illustrated in Fig. 2. Figure 2a shows a network obtained at a high threshold parameter, ε. The results yield two disconnected graphs. The graph to the left has hubs containing genes involved in cell-cycle regulation. The graph to the right shows several hubs; each one is associated with α pheromone-mediated signaling. When the threshold parameter is lowered in the calculation of the adjacency matrix, one generates a larger network because more genes are now above the threshold. Figure 2b shows a network obtained at a low threshold. Here, the original network (on the left-hand side) remains intact, but additional features appear (on the right-hand side). In particular, a single gene now connects the two disconnected graphs. This generates the hypothesis that this gene is the putative link between the pheromone receptor network and the cell-cycle mechanism. The gene product in question happens to have an unknown function. This data-mining method does not provide a specific role to any given set of linkages, but rather establishes regions of the graph with common functionality, and shows how these functionalities are linked. This general scheme has been useful in interpreting genome-wide data on protein–protein interactions and on single-gene deletion mutations [13,14]. It provides a network view of the connectivity and integration of various biological functional units within a cell.

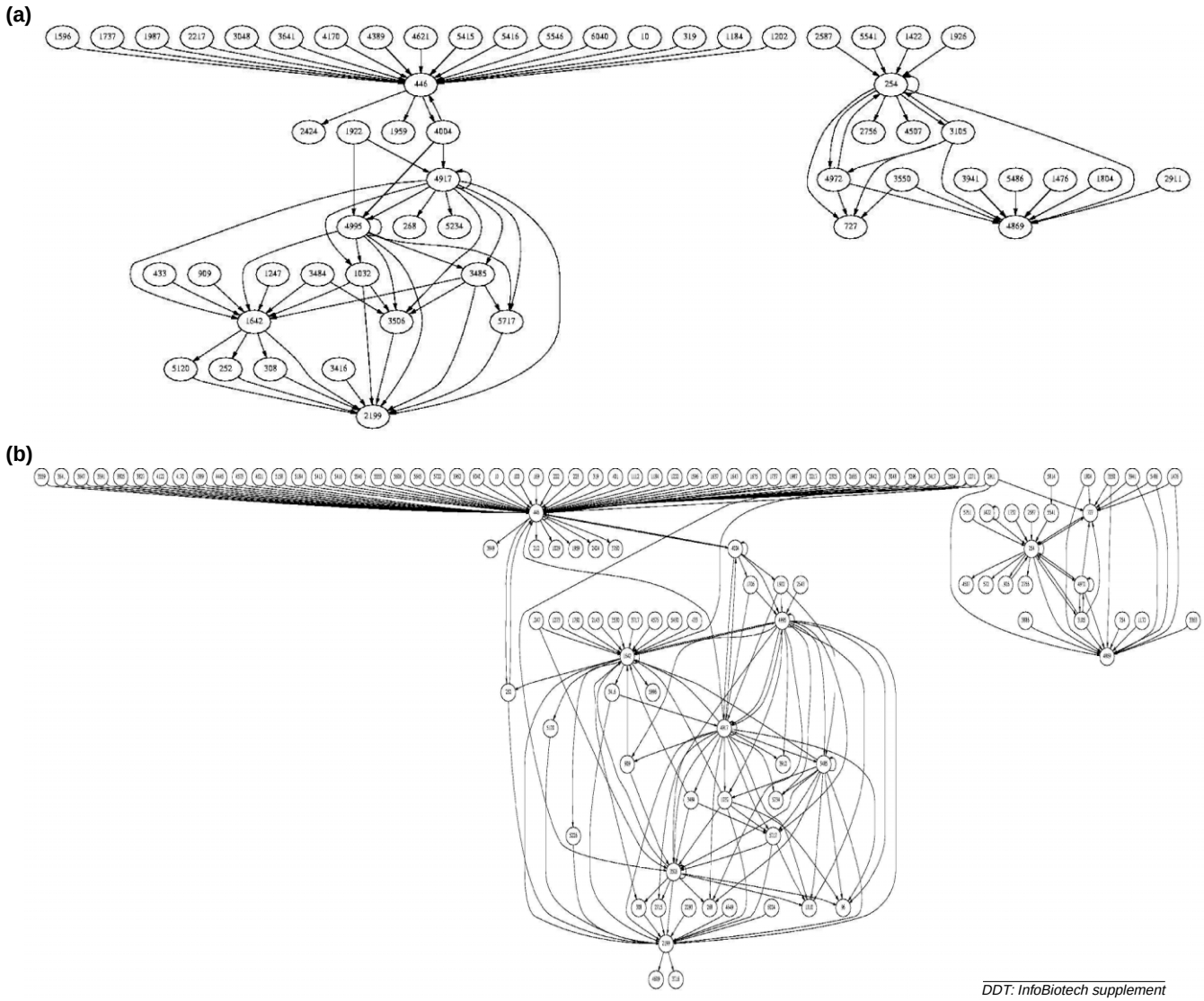

Figure 2. Connectivity of networks at two different thresholds. Networks obtained from yeast cell-cycle data are shown. (a) Disconnected graphs obtained at high threshold settings. Cell-cycle genes are separate from α pheromone response genes. (b) The corresponding network at low threshold. The two sets of genes are now linked via a single connection.

Comparative network analysis

A third method for data mining is through network–network comparisons. Once a network is determined from gene expression analysis, it is often useful to compare it with other networks. These networks could be derived from other microarray experiments or could be of a completely different origin, such as protein–protein networks or gene-deletion networks. The adjacency matrix formalism provides a particularly simple means of overlaying two different networks. This overlay can be done by taking the inner product of the adjacency matrix for the two networks. For two networks represented by adjacency matrices, Λ_1 and Λ_2, the nodes with common connection are given by:

$$\Lambda_{\text{common}} = \Lambda_1 * \Lambda_2 \qquad [2]$$

where * represents the inner product, rather than matrix multiplication. The inner product is simply the multiplication of matrix elements with identical indices. As a connection is represented by '1' and no connection by '0', only instances where both networks have connections survive this operation.

As an illustration of how this device might be used, the network determined from cell-cycle data (α pheromone dataset) in yeast was compared with the protein–protein interaction map determined by a yeast two-hybrid approach [15]. In

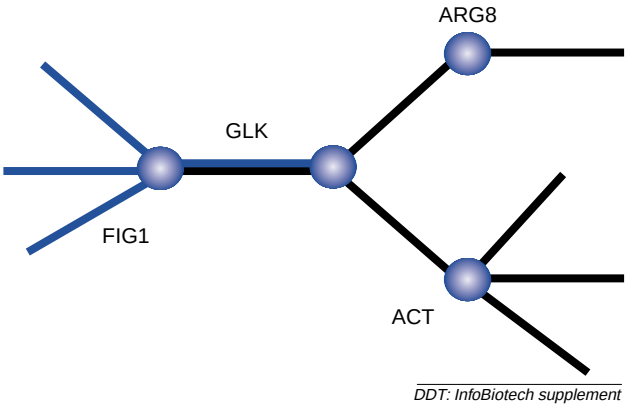

Figure 3. Overlay of a gene expression network (blue lines) with a protein–protein interaction map (black lines). FIG1 is a pheromone response gene that is a hub in the gene expression network. ARG8 is a zinc-finger transcription factor.

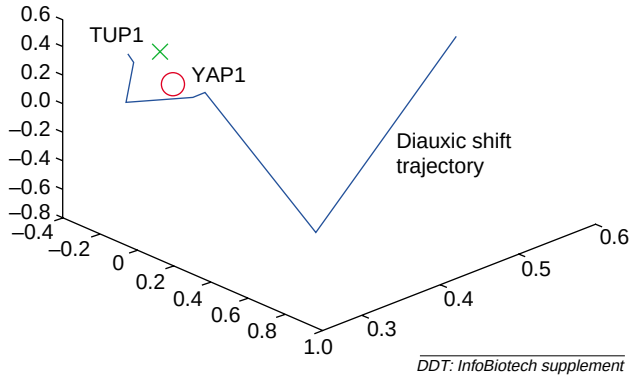

Figure 4. The trajectory of yeast diauxic shift data in an abstract 3D space. The dimensions of the space are derived from a principal component analysis of the data. Each point in the trajectory is a point in the time series, starting from the left (anaerobic state) to the right (aerobic state). The cross represents the expression of the TUP1 deletion mutant and most closely resembles the anaerobic state. The circle represents the expression of the YAP1 overexpressing mutant and is closest to an intermediate metabolic state that occurs during the diauxic shift.

general, there is not a significant overlap between the two graphs. This small overlap is not necessarily a failure of the expression methodology; there is actually little overlap between yeast two-hybrid maps obtained from different groups, or with different methodologies. Nevertheless, in five separate instances there are hubs that show overlap similar to that shown in Fig. 3. In each of these cases, a hub (FIG1 in this example) connects with a second species (GLK) that is also connected to a transcription factor (ARG8). The gene expression network did not contain ARG8 but the protein–protein network does. The microarray experiment might not have the dynamic range to pick up changes in the transcription factor and so these are likely to be silent in the network. This inability to pick up minor messages is a limitation of the technology and not the analysis. This situation will hopefully be remedied as the technology improves. However, the overlap with the protein–protein map shows a potential correlation between expression network hubs and specific transcription factors. This provides the molecular biologist with a very specific hypothesis generated from the two networks. There would be great use in discovering molecular markers that mirror the expression of transcription factors at a higher level.

Challenges and applications for drug discovery

Gene expression time-series can be used to generate networks associated with drug response and provide potential data-mining tools for target discovery. Several strategies can be used for generating such networks. First, one can measure the expression profiles at different times in response to a drug. The time-series analysis will then directly yield a phenomenological, correlative network. Alternatively, one can provide a stimulus to the system in the presence and absence of an antagonist. The parallel time series can be used to generate two different networks that can then be compared using the methods described above. This is the strategy currently being used in our laboratory to explore the response of signalling pathways to IL-1β in a mammalian cell culture system (M. Bechtel *et al.*, unpublished data). These methods can be used to identify the 'major players' in the drug response and generate hypotheses about specific targets.

The ultimate challenge of this approach to inferring gene networks from expression time-series is practical and not conceptual. We will see increasingly sophisticated models and mining techniques whose behavior and validation will become better understood (see [9]). The limitation on the methodology for the foreseeable future will be caused by the expense and effort that it takes to generate time-series data. For quantitative models of gene expression, one needs approximately an order-of-magnitude more time points than are typically measured in most microarray studies, which, currently, is not a realistic option. It is also frequently important to see the variation of response to a drug or stimulus from different members in a population. Again, it is not realistic to do extensive time-series measurements on many members of a population.

Concluding remarks

A very real challenge in this field will be to relate the networks derived from time-series measurements to individual measurements from a population. We are currently exploring methods to map the time-series measurements into an abstract phase space that represents the trajectory of the stimulus response. Individual profiles can also be represented in this phase space

and this gives a measure of how close the observed individual is to the response trajectory. This indicates which 'state' of the dynamic process the individual most closely resembles. This method is illustrated in Fig. 4, in which the trajectory for the diauxic shift data is shown and compared to two mutants grown to stationary phase in culture. As can be seen, specific mutants most closely resemble specific dynamic states along the trajectory. This enables the mutant profile to be related to the network dynamics. Establishing a database of time series and their trajectories could extend this methodology. These could provide a reference set of expression states to which population data can be compared. Linking time-series and population data will be a crucial step in sorting through the complexity of drug response, and will provide a means to distinguish between variations owing to history and variations owing to genetics.

Acknowledgement

I gratefully acknowledge funding from NIH grant 1R01 GM63912-01.

References

1 DeRisi, J. *et al.* (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680–686

2 Spellman, P.T. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, 9, 3273–3297

3 Boldrick, J.C. *et al.* (2002) Stereotyped and specific gene expression programs in human innate immune responses to bacteria. *Proc. Natl. Acad. Sci. U. S. A.* 99, 972–977

4 Raymond, P. *et al.* (2000) Differential gene expression in response to mechanical wounding and insect feeding in *Arabidopsis*. *Plant Cell* 12, 707–719

5 Dewey, T.G. and Galas, D. (2001) Dynamic models of gene expression and classification. *Func. Integr. Genomics* 1, 269–278

6 Bhan, A. *et al.* A duplication growth model of gene expression networks. *Bioinformatics* (in press)

7 Holter, N.S. *et al.* (2001) Dynamic modeling of gene expression data. *Proc. Natl. Acad. Sci. U. S. A.* 98, 1693–1698

8 Heyer, L.J. *et al.* (1999) Exploring expression data: identification and analysis of coexpressed genes. *Genome Res.* 9, 1106–1115

9 Yeung, M.K.S. *et al.* (2002) Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci. U. S. A.* 99, 6163–6168

10 Cormen, T.H. *et al.* (1997) *Introduction to Algorithms*, MIT Press

11 Wagner, A. and Fell, D. The small world inside large metabolic networks. *Proc. Roy. Soc. London Ser. B.* (in press).

12 Manly, B.F.J. (1997) *Randomization, Bootstrap and Monte Carlo Methods in Biology*, Chapman & Hall

13 Tucker, C.L. *et al.* (2001) Towards an understanding of complex protein networks. *Trends Cell Biol.* 11, 102–106

14 Tong, A.H.Y. *et al.* (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* 294, 2364–2368

15 Uetz, P. *et al.* (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* 403, 623–627