Challenges and prospects in the analysis of large-scale gene expression data

Jan H. Ihmels and Sven Bergmann

Date received (in revised form): 17th August 2004

Jan H. Ihmels
received his PhD in computational biology from the Weizmann Institute of Science, Israel. He is currently a postdoctoral fellow at the Department of Molecular Genetics of the Weizmann Institute.

Sven Bergmann
received his PhD in theoretical physico from the Weizmann Institute of Science, Israel. He is currently a research associate at the Department of Molecular Genetics of the Weizmann Institute.

Keywords: *systems biology, microarrays, gene expression, clustering*

Abstract

Large heterogeneous expression data comprising a variety of cellular conditions hold the promise of a global view of transcriptional regulation. While standard analysis methods have been successfully applied to smaller data sets, large-scale data pose specific challenges that have prompted the development of new and more sophisticated approaches. This paper focuses on one such approach (the Signature Algorithm) and discusses the central challenges in the analysis of large data sets, and how they might be overcome. Biological questions that have been addressed using the Signature Algorithm are highlighted and a summary of other important methods from the literature is provided.

INTRODUCTION

DNA microarrays have firmly established themselves as a standard tool in biological and biomedical research. Together with the rapid advancement of genome sequencing projects, microarrays and related high-throughput technologies have been key factors in the study of global aspects of biological systems.[1] While genomic sequence provides an inventory of parts, a proper organisation and eventual understanding of these parts and their functions also requires global views of the regulatory relationships between them.[2] Genome-wide expression data offer such a global perspective by providing a simultaneous read-out of the mRNA levels of all (or many) genes of the genome.

To date, most microarray experiments are conducted to address specific biological questions. In the simplest case, such a study may focus on the expression response to the deletion of individual genes or to specific cellular conditions. Already when the experimental design includes several conditions, eg time points along the cell-cycle[3] or several tissue samples, the sheer amount of data necessitates computational tools to extract and organise the relevant biological information. A wide range of approaches has been developed, including numerous clustering algorithms, statistical methods for detecting differential expression, and dimension-reduction techniques (reviewed by Brazma and Vilo[4] and Slonim[5]).

In addition to the specific biological questions probed in individual focused experiments, it is widely recognised that a wealth of additional information can be retrieved from a large and heterogeneous data set describing the transcriptional response to a variety of different experimental conditions.[2] Such comprehensive data have been used to provide functional links for unclassified genes,[3,6–9] to predict novel *cis*-regulatory elements[7,10–12] and to study the structure of the transcriptional program.[12,13]

Large-scale expression data may be produced in systematic efforts to characterise a range of transcription states.[6,8,13] In addition, large data sets can be assembled by collecting published expression profiles and pooling them into one comprehensive database (Figure 1). Until recently, these data appeared in different formats and were scattered among various internet sites.[14] The increasing availability of microarray

Jan H. Ihmels,
Department of Molecular Genetics,
Weizmann Institute,
76100 Rehovot,
Israel

Tel: +972 8934 2201
Fax: +972 8934 4108
e-mail: jan@weizmann.ac.il

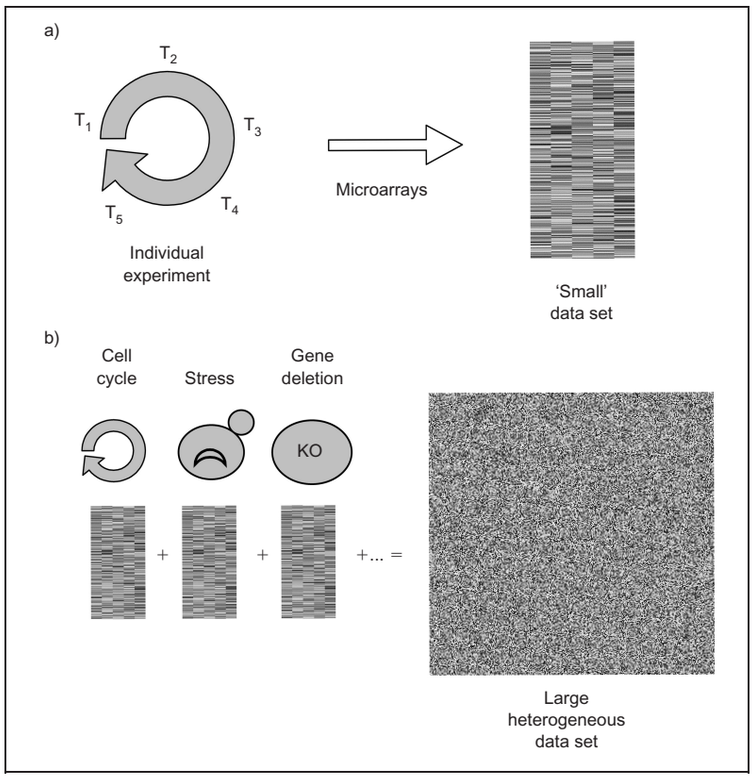

Figure 1: Expression data. (a) Individual microarray experiments addressing specific biological issues give rise to small data sets comprising only a few distinct experimental conditions (eg time points). (b) Large-scale expression data can be generated by pooling profiles from many such individual experiments (or conducting dedicated comprehensive assays). Such data cover not only thousands of genes but also many cellular states by including a heterogeneous collection of experimental conditions

Microarray standards

technology has given rise to an explosion of expression profiles, usually obtained in different laboratories and often using different array technologies. To address this issue, consortiums have been established to define standardised annotations such as the MIAME[15] and MAGE-ML[16] standards, and to create a number of public repositories for array data (Table 1).

Individual genome-wide experiments are global only in the sense that the genes probed with each microarray span all or most of the genome. However, expression patterns change in response to different cellular conditions. Collecting microarray data from a large variety of experimental conditions aims also to span the space of these transcriptional states of the cell. While this is a necessary step towards the elucidation of the transcription program, such data present new and serious challenges. In particular, the context-specific nature of regulatory relationships poses a difficult computational problem (see below). Consequently, a number of different approaches have been proposed in the literature. It is impossible to give a fair and comprehensive exposition of all existing analysis methods. Thus here the main challenges in the analysis of large-scale data are discussed by focusing on one method (the Signature Algorithm) in greater detail. Some biological questions that have been addressed using heterogeneous expression data are highlighted. Other important methods are summarised in Table 2.

REGULATORY PATTERNS ARE CONTEXT SPECIFIC

The central problem in the analysis of large and diverse collections of expression profiles lies in the context-dependent nature of co-regulation. In general, genes are coordinately regulated only in specific experimental contexts, corresponding to a subset of the conditions in the data set. Most standard analysis methods group genes into clusters based on their similarity across all available conditions. While the underlying assumption of

Table 1: Public repositories for expression data

Repository	URL
EBI ArrayExpress	http://www.ebi.ac.uk/arrayexpress
Stanford Microarray Database (SMD)	http://genome-www.stanford.edu/microarray
NCBI Gene Expression Omnibus (GEO)	http://www.ncbi.nlm.nih.gov/geo
Center for Information Biology gene EXpression database (CIBEX)	http://cibex.nig.ac.jp

Table 2: Overview of methods for the analysis of large-scale expression data, together with the data sets that have been analysed in the original publication

Analysis method; data set analysed (genes by conditions)	Description
Biclustering (Cheng and Church)[17] Yeast (2,884 by 17) Human (4,026 by 96)	The biclustering algorithm aims to identify uniform submatrices corresponding to a set of genes showing consistent up- or down-regulation over a set of conditions. The uniformity of a given submatrix is quantified using a score ('mean squared residue'), and a set of algorithmic procedures is employed to identify large submatrices with high scores. The algorithm is based on the deletion and addition of rows and columns to iteratively improve the score of each bicluster. Discovered biclusters are masked to allow the identification of new clusters in subsequent runs.
Coupled Two-Way Clustering (CTWC)[18–20] Human (1,753 by 72) Human (2,000 by 62)	The Coupled Two-Way Clustering (CTWC) procedure is initialised by separately clustering the genes and condition of the full matrix. Each combination of the resulting gene and condition clusters defines a submatrix of the expression data. Two-way clustering is then applied to all such submatrices in the following iteration. At every step, all pairs of previously identified clusters are used to generate the submatrices for the next iteration. The procedure stops when no new clusters satisfying some criteria, such as stability or size, are identified. Because genes and conditions are clustered over all partitions identified at previous iterations, CTWC is sensitive to context-dependent regulation and suited to the identification of subpartitions. Any standard clustering method can be used in the coupled two-way framework; Getz *et al.* use the Super-Paramagnetic Clustering of Blatt *et al.*[21] algorithm.
SAMBA[22,23] Human (4,206 by 96) Yeast (6,200 by 515) Yeast (6,000 by 1,000)	The SAMBA (Statistical-Algorithmic Method for Bicluster Analysis) method[23] combines graph theory with statistical data modelling. The expression data are modelled as a bipartite graph of conditions and genes, which are connected with edges for significant expression changes. Biclusters are subgraphs of genes whose expression changes significantly over a set of conditions. The method is based on a scoring scheme that assigns weights to the vertex pairs in such a way that finding statistically significant biclusters corresponds to identifying heavy subgraphs in the data. The search heuristic used to identify these subgraphs is guaranteed to find the most significant biclusters. The SAMBA framework has been generalised to model diverse sources of genomic information in addition to expression data.[22] In this extended description, a bipartite graph represents the relationship between genes or proteins and generalised properties, such as expression changes or interactions with another protein.
Gibbs biclustering[24] Human (1,887 by 72)	The biclustering problem is cast into a Bayesian framework and Gibbs sampling is used for parameter estimation. This approach requires discretised expression data. Like the biclustering of Cheng and Church, the method relies on a procedure of masking the genes of discovered biclusters to allow for the identification of multiple biclusters. In contrast to the former method, however, the masking scheme adopted here precludes overlaps in the gene content of the resulting biclusters.
Fuzzy *k*-means[25] Yeast (6,153 by 93)	Fuzzy *k*-means clustering is an alternative approach to capture condition-dependent regulation patterns. Instead of the hard partitioning of standard *k*-means clustering where each gene belongs to exactly one cluster, all genes are associated with all clusters through a continuous membership-degree.
Singular Value Decomposition (SVD)[26,27] Yeast (5,981 by 14) Yeast (4,579 by 22)	SVD yields linear combinations of the rows and columns of the expression data (ie pairs of 'eigengenes' and 'eigenarrays') that describe independent components of the data. Each pair is associated with an 'eigenexpression' level indicating its relative significance. Filtering out insignificant eigengenes (and the corresponding eigenarrays) or those that are inferred to represent experimental artefacts reduces the noise in the expression data. Sorting the genes according to their correlations with the eigengenes means they can be classified into groups corresponding to a similar regulation and function. Similarly, the arrays can be grouped into sets corresponding to similar cellular states or biological phenotypes based on their correlations with the eigenarrays.
Signature Algorithm[7] Yeast (6,200 by 1,000)	The Signature Algorithm is designed to identify groups of co-regulated genes together with the experimental conditions over which the co-regulation is observed ('transcription modules'). The starting point is a set of input genes that partially overlap with a transcription module. Within this set, the algorithm identifies those genes that are co-expressed under a subset of the experimental conditions. Furthermore, it reveals additional genes that display a similar expression pattern under those conditions, but were not included in the original input. Input genes are chosen according to some common feature, such as participation in the same pathway, a common regulatory motif or membership in the same functional category.
Gene Recommender[28] *Caenorhabditis elegans* (11,917 by 533)	The Gene Recommender algorithm aims to find new genes that are co-expressed with a given set of genes. The algorithm follows a similar procedure to the Signature Algorithm, by first selecting a subset of relevant experiments and then using these experiments to rank genes according to their correlation with the query genes.
Plaid Model[29] Yeast (2,467 by 79)	The Plaid Model extends SVD by introducing additional parameters that allow for an improved decomposition of the expression data into potentially overlapping sets of genes and conditions.
Gene Shaving[30] Yeast (4,673 by 48)	In Gene Shaving a cluster is formed by removing iteratively genes whose expression profiles are the least similar to the principal component. The optimal cluster is determined *a posteriori*, by demanding both high-variance clusters and high coherence between the genes in the cluster. Subsequently, new clusters are obtained from the data that are orthogonal to previously identified clusters.

Continued overleaf

Table 2: *Continued*

Analysis method; data set analysed	Description
Probabilistic models[31–33] Yeast (945 by 92) Yeast (528 by 207) **Module networks**[34] Yeast (2,355 by 173)	Probabilistic graphical models, an extension of Bayesian frameworks, provide a general formalism that has been applied to a variety of problems involving diverse sources of genomic data. Applications include a biclustering algorithm that reveals context-dependent relationships that exist over subsets of experimental conditions.[31] Similar approaches have been used to describe dependencies between gene expression and protein interaction data[32] as well as regulatory motifs.[33] In the module network approach,[34] the relationship between regulators and their target genes is modelled explicitly, based on the assumption that regulators themselves are transcriptionally regulated. A regulation program represents the combined up- or down-regulation of a set of regulator genes. The output consists of (refined) modules of co-regulated genes, their predicted regulators and the conditions under which the regulation occurs.

Irrelevant conditions

uniform regulation is reasonable for the analysis of small data sets, it limits the utility of these tools for the analysis of heterogeneous large data sets for two reasons.

First, conditions irrelevant for the analysis of a particular regulatory context contribute noise, hampering the identification of correlated behaviour over small subsets of conditions (Table 3). Second, genes may participate in more than one function, resulting in one regulation pattern in one context and a different pattern in another. Such combinatorial regulation places even greater constraints on the usage of global similarity measures than noisy contributions from irrelevant conditions. It furthermore necessitates the assignment of genes to several overlapping clusters. In contrast, most commonly used clustering techniques yield disjoint partitions, assigning each gene to a single cluster.

Combinatorial regulation

Combinatorial regulation is considered 'one of the hallmark characteristics of biological systems: the ability to make decisions based on multiple inputs' (Lander[2]). Several examples have been discussed in the literature. Madhani and Fink showed that the transcription factor Ste12, involved in yeast MAPK signalling, can drive separate transcription responses depending on whether it binds as a homodimer or in combination with another transcription factor (Tec1).[35] Yuh *et al*. analysed the combinatorial logic in the control element of a sea urchin gene.[36] In a recent work, the authors discussed the co-regulation of the TCA cycle in *Sacharromyces cerevisiae* and identified two subparts of the cycle that are autonomously co-regulated under different sets of conditions.[7] At the genomic level, a systematic approach has been introduced by Pilpel and coworkers to characterise motif combinations and

Table 3: Correlations observed over subsets of conditions may be masked in large data sets

Co-expression in general occurs only under a subset of the experimental conditions for which the expression levels were recorded. This is particularly relevant for large-scale expression data. To illustrate this point, consider two genes ($g = 1,2$) whose expression levels E_{gc} have been measured under a total of N conditions ($c = 1 \ldots N$). Consider the case where N_s conditions yield a consistent expression pattern while the expression under the remaining N_n conditions is not correlated. For simplicity let us assume that the expression levels are either +1 or −1 and that their average is zero. Then the correlation between the two genes,

$$C = \sum_c E_{1c} E_{2c} / \sqrt{\sum_c E_{1c}^2 \cdot \sum_c E_{2c}^2}$$

has a mean value $<C> = N_s/N$ and standard deviation $\sigma = \sqrt{N_n}/N$. To observe a significant correlation over all conditions, the number of conditions containing the signal has to be larger than the noise: $N_s \gg \sqrt{N_n}$. For example, while the co-expression of two genes in 20 experimental conditions would yield a significantly correlated expression profile in a small sample of 100 arrays, this correlation would be masked by the noise in a large-scale data set of 1,000 experiments.

their synergistic effect on expression patterns.[37,38] It is expected that the degree of combinatorial regulation is elevated in higher eukaryotes,[39] emphasising further the importance of appropriate computational tools as expression profiles are rapidly accumulating also for these organisms.

Classifying conditions over subsets of genes

Similarly, in a large data set, biologically similar conditions may be identified more readily by focusing on specific genes. For example, samples taken from different types of tumour tissue may be distinguished by the differential expression induced in one or several subsets of genes,[18] rather than across the entire genome.

Additional data sources

Co-classification of genes and conditions

To take these considerations into account, expression patterns must be analysed with respect to specific subsets; genes and conditions should be co-classified. The resulting 'transcription modules'[7] or 'biclusters'[17,40] consist of groups of co-regulated genes together with the conditions over which the co-regulation is observed.

Identification of transcription modules

The great challenge lies in the identification of such transcription modules from the expression data. Naively evaluating expression coherence of all possible subsets of genes over all possible subsets of conditions is computationally infeasible, and most analysis methods for large data sets seek to limit the search space in an appropriate way. For example, Getz *et al.* introduced a variant of biclustering based on the idea of performing standard clustering iteratively on genes and conditions.[18] The Coupled Two-Way Clustering (CTWC) procedure begins by separately clustering the genes and conditions of the full matrix. All resulting stable clusters are recorded. In the following iteration, combinations of all previously identified gene and condition clusters are used to define submatrices of the expression data. Two-way clustering is then applied to all such submatrices in the next iteration, and the process is repeated. Thus, instead of considering all possible sets of genes and conditions, clustering is performed only over subsets corresponding to stable clusters. Other biclustering methods[17,24] aim to identify only the most dominant bicluster in the data set, which is then masked in a subsequent run to allow for the identification of new clusters.

INTEGRATING ADDITIONAL DATA SOURCES: SIGNATURE ALGORITHM

For many organisms, another approach to limiting the space of possible solutions is to employ additional biological information. Different types of biological data are rapidly accumulating, including protein–protein interaction data,[41] transcription factor binding information,[42] genomic and promoter sequence, ontologies[43] and protein localisation studies.[44] Such additional data are often noisy and incomplete and cannot be used to infer co-regulation directly. However, they may be used to provide a starting point and a framework for co-expression studies. Co-regulation of genes is generally expected to reflect involvement in related cellular functions or pathways, and to result from shared promoter binding sites for a common set of transcription factors. For example, target-regulator network analyses can be simplified if the set of potential regulators does not include the entire genome but can be restricted to a smaller number of candidates.[34] Similarly, the GRAM (Genetic Regulatory Modules) algorihm[45] by Bar-Joseph *et al.* investigates module–regulator relationships by integrating physical regulator binding data with expression profiles. The algorithm aims to improve the reliability of binding data by allowing lower stringency for those interactions that are supported by co-expression. Likewise, it restricts the analysis of co-expression to genes that are targets to a common set of transcription factors.

The approach highlighted in this

Signature algorithm

review, the Signature Algorithm[7] (Figure 2), offers a more general framework to integrate external data sources with large-scale expression profiles. The method requires an input seed of genes, at least some of which are expected to be co-regulated. Such seeds may be chosen according to some common features, such as participation in the same pathway, a common regulatory motif or membership in the same functional category. The algorithm proceeds in two steps. In the first step, the input seed is used to identify subsets of relevant conditions. To this end, every condition in the data set is scored by the average expression change among the input genes. Conditions that induce a coherent change in at least a subset of the

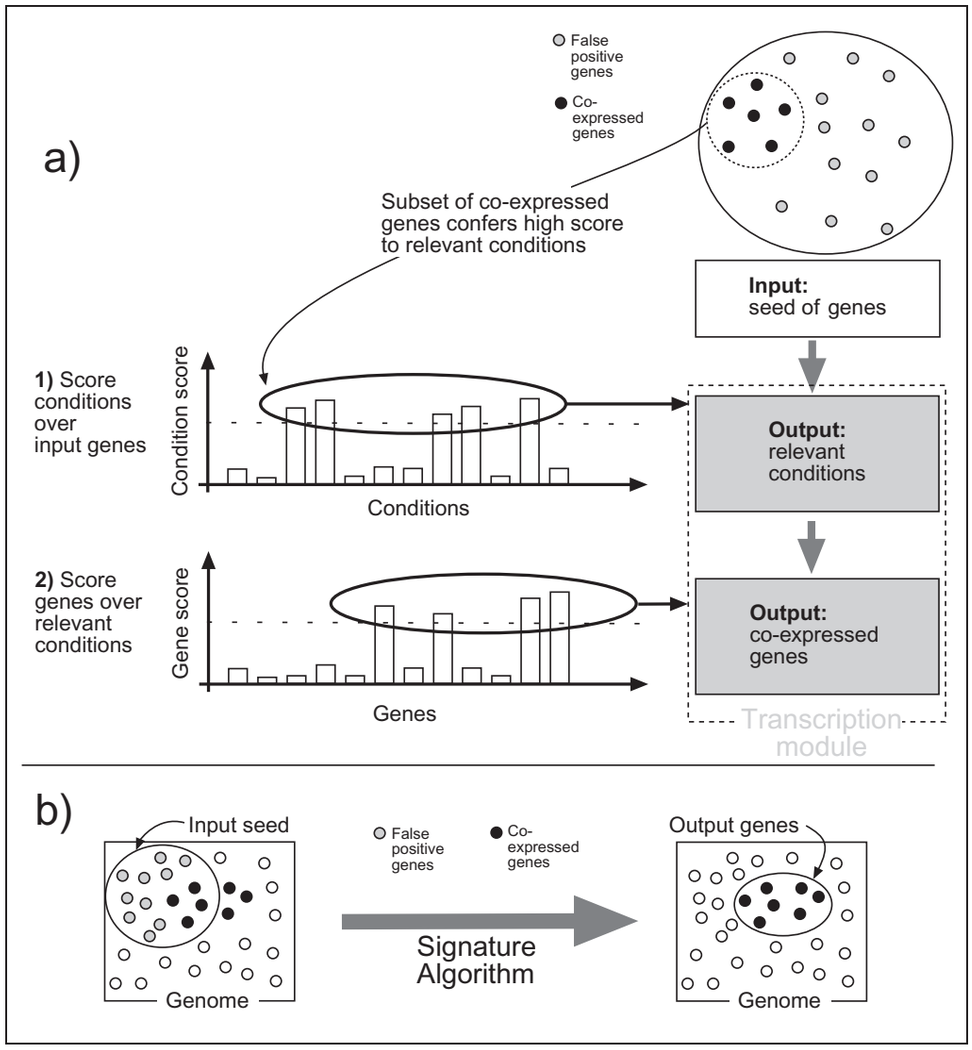

Figure 2: The Signature Algorithm requires as input a set of genes, some of which are expected to be co-regulated based on additional biological information such as a common promoter binding motifs or functional annotation. (a) The algorithm proceeds in two steps: in the first step, this input seed is used to identify the conditions that induce the highest average expression change in the input genes. Only conditions with a score above some threshold are selected. In the second stage of the algorithm, genes that are highly and consistently expressed over these conditions are identified. The result consists of a set of co-regulated genes together with the regulating conditions and is termed a *transcription module*. (b) The output contains only the co-regulated part of the input seed, as well as other genes that were not part of the original input but display a similar expression profile over the relevant conditions

input genes receive higher scores and are selected according to a cut-off parameter. In the second stage of the algorithm, genes that are highly and consistently expressed over the conditions identified in the first step are selected according to a second cut-off parameter (the *gene threshold*). The end result is a *transcription module*, consisting of a set of co-regulated genes together with the regulating conditions. In general, the output set of genes contains the co-regulated part of the input seed, as well as other genes that were not part of the original input but display a similar expression profile over the relevant conditions. Genes in the seed that are not co-expressed (false positives) do not appear in the module. Individual modules are identified independently and thus can naturally overlap both in gene and condition content.

Robustness against 'noise' genes

It is a key property of the algorithm that the identification of relevant conditions (and hence of the output genes) is very robust against the addition of unrelated 'noise' genes to the input. For example, the output obtained from applying the algorithm to a seed 132 of GCN4-controlled genes involved in amino acid biosynthesis is almost identical to the output resulting from a seed containing the same 132 genes mixed together with 2,000 genes randomly picked from the genome. The reason for this noise resistance lies in the different scaling properties of incoherent (noise) and coherent (signal) contributions to the gene and condition scores.[7,50]

Sets of partially co-expressed genes

The sensitivity of the algorithm to co-expressed genes in the input seed and the robustness against 'noisy' false positives make it a useful tool to test and refine hypotheses about sets of genes exhibiting only partial co-regulation. For example,

Metabolic flow central

the mere presence of a motif sequence is not a sufficient criterion to infer regulation by the corresponding transcription factor. Many motif sequences found in the upstream region of a gene have no regulatory function. By applying the Signature Algorithm to all genes containing a particular regulatory motif, those genes that are indeed co-expressed can be distinguished. Importantly, the output provides a connection between the transcription factor binding site, the co-regulated genes and the conditions where the transcription factor is active. By systematically scanning over all hexa-, hepta- and octamers, the method has been used to generate a comprehensive map of transcription factor binding sites in *S. cerevisiae*, together with the associated transcription modules.[7]

Cellular pathways are another example of systems for which often only partial knowledge exists. By applying the Signature Algorithm to a set of genes that are thought to participate in the same function, it is possible to characterise the co-regulated part of the pathway and to retrieve additional genes that are co-expressed and hence likely to be involved in the same function. Identification of the experimental context provides biological insights and makes it possible to identify separate regulation patterns under different cellular conditions.

TRANSCRIPTION CONTROL IN THE METABOLIC NETWORKS

In addition to providing a focus for expression analyses, the integration of the massive body of heterogeneous biological information presents an important challenge in its own right, and is an essential next step towards system-level understanding of cellular processes. In a recent study, large-scale expression data in conjunction with biochemical pathway databases were used to characterise the role of transcription regulation in metabolic pathways in *S. cerevisiae*.[46,47]

The description of metabolism in terms of a set of mostly well-characterised biochemical reactions reveals a highly interconnected network whose connectivity extends far beyond the limits of individual metabolic function. It remains an open question how individual functional units are maintained and isolated from each other in such an

interconnected setting. Although such networks could support flow in many directions simultaneously, metabolic flow patterns follow specific pathways and have been shown to be highly optimised.[48] A number of mechanisms have been implicated in the control of individual pathways, including allosteric interactions, regulation of metabolite concentrations, covalent modifications and transcription regulation of metabolic enzymes.

Transcriptional co-regulation of metabolic genes

The availability of large expression data sets made it possible to go beyond individual biochemical pathways and to gain insights into how and to what extent modulation of enzyme expression shapes metabolic flow patterns on a genomic scale. Thus, the analysis of pathways listed in the KEGG[49] database revealed that in most cases only a subset of the genes associated with each pathway is co-expressed. The Signature Algorithm was used to systematically characterise the co-expressed genes and the experimental conditions associated with each pathway. In many cases, the co-expressed genes were found to be arranged linearly along the central part of the pathway, suggesting that transcription regulation serves to bias metabolic flow towards linear patterns embedded in the non-linear interconnected network. This is further supported by the observation that at divergent junctions in the metabolic network, incoming reactions are predominantly co-regulated with only one of the outgoing reactions. Another recurrent feature that emerged from pathway regulation patterns is that distinct enzymes catalysing the same reaction are often separately co-regulated with alternative reactions at junction points.

Regulation of isozymes

HIERARCHICAL MODULARITY – ITERATIVE SIGNATURE ALGORITHM

Higher order organization of regulatory units

Application of the Signature Algorithm revealed a strong tendency of genes within individual pathways to be coordinately expressed. These functional modules could appear as isolated transcriptional units, or alternatively be embedded in a higher-order structure through coordinated regulation of the pathways themselves. The iterative extension of the Signature Algorithm[50,51] is designed specifically for the analysis of such global hierarchical organisations. In the iterative framework, the output of the Signature Algorithm is repeatedly fed back into the algorithm, until a point of convergence is reached where output and input are identical (Figure 3). The resulting transcription module of genes and conditions is a fixed point of the Signature Algorithm and satisfies a criterion termed self-consistency. The criterion states that the module genes are those genes of the genome that are most coherently co-expressed over the module conditions. The module conditions, in turn, are those conditions in the data set that induce the most

Iterative Signature Algorithm

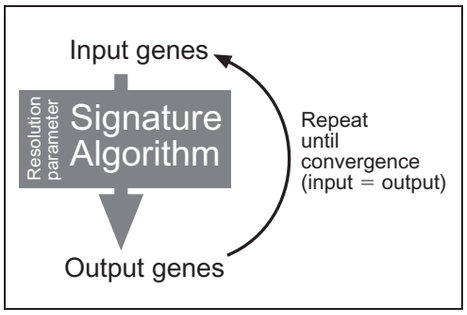

Figure 3: The Iterative Signature Algorithm (ISA) is an extension of the signature algorithm that is designed to reveal hierarchies of co-regulatory units of varying expression coherence. This approach is applicable also in the absence of biologically motivated seeds, in which case the iterative scheme is initialised by many sets of randomly chosen input genes. The output genes determined by the Signature Algorithm are re-used as input until convergence between input and output is reached. Each resulting 'transcription module' is self-consistent: its genes are most coherently co-expressed over the module conditions, which, in turn, induce the most coherent expression of the module genes. Modules at different resolutions can be obtained by changing the co-regulation threshold parameter

coherent expression in the module genes. Such an explicit formulation of the defining property of a module distinguishes the Iterative Signature Algorithm (ISA) method from most clustering algorithms. Modules can be identified in a heuristic search by iterating from a large number of random input seeds. Alternatively, the iterative scheme can also be initiated with biologically motivated sets of genes.[46]

Resolution of modular decomposition

The gene threshold parameter of the Signature Algorithm imposes a minimum co-expression stringency on the output genes. In the iterative framework, this threshold determines the specificity of the fixed points and serves as a resolution parameter. As the resolution is decreased, small modules merge into larger modules with fewer specific functions. Thus, in an application of the method to a yeast data set comprising more than 1,000 expression profiles,[51] iterations from ~20,000 random input seed converged into one of only five fixed points at low resolution, corresponding to the basic functions of the yeast organism. As the resolution parameter is increased, new smaller fixed points with more specific functions arise. Importantly, these new modules can be connected to the more generic modules that they converge to when they are iterated at a lower resolution. Such connections simplify the biological interpretation of modules and reveal hierarchies of co-regulated units of varying expression coherence.

Relevant conditions provide functional context

Higher-order relationships between modules

Computational efficiency

The size of large data sets imposes constraints on the computational implementation of analysis tools. The ISA method is computationally efficient, since computation time scales linearly with the number of genes and conditions.[50] Linear dependence on the number of genes and conditions is an important prerequisite for the capacity to analyse ever larger expression data sets, in particular data from higher organisms, which can include many more genes. Because the Signature Algorithm removes unrelated genes from the input seed, convergence is typically reached within only a few iterations. Importantly, the method does not rely on the calculation of correlation matrices, which requires significant amounts of computation time and memory. The computational bottleneck of the ISA is the adopted heuristic search procedure. The more sophisticated search algorithm PISA (Progressive Iterative Signature Algorithm),[52] based on the sequential elimination of identified modules, can help to further improve its computational efficiency.

Significance of the experimental context

The experimental condition scores provided by the Signature Algorithm quantify the effect of each experimental condition on the expression levels of the module genes and thus provide important insights into the biological function of each module. In addition, experimental conditions can be used to reveal higher-order relationships between modules. For example, in the yeast data, essentially all conditions activating the rRNA processing module genes also repress the stress response module, and vice versa. Thus, the condition sets associated with these two modules are the same, albeit with scores that have opposite signs. In general, modules may be positively or negatively correlated, or their induction can be mutually exclusive. Several examples of relationships of this kind in the yeast modular structure suggest that many distinct modules are not regulated independently but rather change in a coordinated manner.[51] This implies a further reduction of complexity and constrains the outcome of novel microarray experiments. Further exploration of such higher-order dependencies could characterise the scope of possible transcription responses and would represent a first step toward the prediction of expression response following a novel perturbation.

Complexity of the output and visualisation

Module layers

Visualisation methods such as topomaps[6] or hierarchical dendrograms have been useful for the interpretation of the results from standard clustering methods. As the size and complexity of the data increase, the meaningful organisation and visualisation of the extracted features become ever more important. Valuable information beyond the simple enumeration of clusters can be gleaned from the data, including higher-order relationships between clusters,[6,7,50,51] analyses carried out at variable resolutions[18,51,53] and the identification of hierarchical organisation.[46,51,53,54] Visual representations can help to elucidate relationships between genes, conditions or clusters,[6,7,50,51,53] and to integrate additional information about functional annotations or regulators.[7,9,34,37,45]

Module trees

To obtain a snapshot of the relationships between all modules identified at a given resolution, modules can be arranged according to the correlation between their condition scores, such that correlated modules appear close to each other while inversely correlated modules are separated.[51] Such a representation has been utilised in a recent study to obtain a first glance of the differences and similarities in the broad relationships between functional units in different organisms.[53]

In a complementary visualisation, the full modular structure over a range of resolutions is represented in a hierarchical *module tree* (Figure 4). Highly similar modules, identified at adjacent thresholds, are connected by lines and define the branches of the tree. The resulting module trees resemble dendrograms used to represent the results of hierarchical

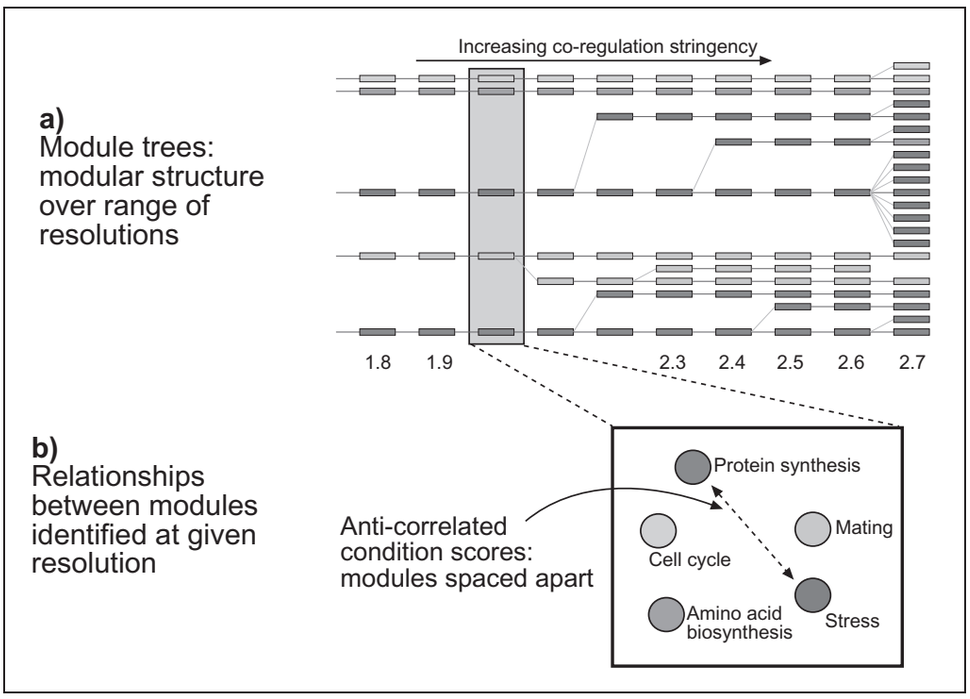

Figure 4: Visualisation of large-scale expression data. (a) Module trees summarise the transcription modules identified by the ISA at different resolutions. Branches represent modules (rectangles) that remain fixed points over a range of thresholds. Modules that emerge at a higher threshold converge into an existing module when iterated at a lower threshold (thin transversal lines). (b) All modules identified at the same resolution are represented on a plane such that their distances reflect the regulatory relations. Modules induced under similar conditions are closer to each other, while large distances indicate inverse activation

clustering. An important distinction is that modules associated with distinct branches may have common genes.

Related methods

Related eigenvalue-based methods

The search for gene sets that remain invariant under application of the Signature Algorithm bears formal similarity[50] with eigenvalue-based methods such as Principal Component Analysis (PCA) or Singular Value Decomposition (SVD). The use of SVD for analysing genome-wide expression data was proposed by Holter *et al*.[26] and Alter *et al*.[27] SVD yields linear combinations of the rows and columns of the expression data (ie pairs of 'eigengenes' and 'eigenarrays') that describe orthogonal components of the data. Each pair is associated with an 'eigenexpression' level indicating its relative significance. Filtering out insignificant eigengenes (and the corresponding eigenarrays) or those that are inferred to represent experimental artefacts reduces the noise in the expression data. By sorting the genes according to their correlations with the eigengenes, one can classify them into groups corresponding to a similar regulation and function. Similarly, the arrays can be grouped into sets corresponding to similar cellular states or biological phenotype based on their correlations with the eigenarrays. Other related algorithms include Spectral Biclustering,[55] Correspondence Analysis,[56] Plaid Model[29] and Gene Shaving (Table 2).[30]

By definition, eigenvectors in SVD must be orthogonal, a limitation that is absent in the ISA. Another important difference between the ISA and SVD is the threshold that is applied in each iteration of the ISA in order to include only the most pertinent genes and conditions. This threshold provides an efficient means to extract robust modules from noisy expression data.[50] However, its introduction breaks the linearity of the problem and prevents the use of optimised diagonalisation algorithms. The heuristic search originally proposed for the ISA method[50,51] cannot guarantee an exhaustive identification of all transcription modules encoded in the expression data, a problem which may be overcome by the PISA method[52] mentioned above.

INTEGRATING EXPRESSION DATA FROM DIFFERENT SPECIES

Large-scale expression data are now becoming available for many organisms.[57] Integrating such data with genomic sequence information promises exciting new insights into biological and evolutionary principles. One application is the use of conserved co-expression to enhance functional annotations.[53,58] For example, if two genes have similar expression profiles in different organisms, then this significantly strengthens a functional link between the products of these genes.[53,58]

Conserved co-expression improves functional assignments

A comparative analysis employing expression data from *S. cerevisiae*, *Caenorhabditis elegans*, *Escherichia coli*, *Arabidopsis thaliana*, *Drosophila melanogaster* and *Homo sapiens* revealed that functionally related genes are indeed frequently co-expressed in these organisms.[53] The Signature Algorithm was applied to seeds containing the orthologues of co-expressed yeast genes with known cellular functions. In most cases a co-expressed subset of these orthologues could be identified. These genes are likely to participate in a function similar to that of the original yeast genes. Moreover, this approach also provides functional predictions for genes that have similar expression patterns but no sequence similarity with the original genes.

Comparison of transcription programs

The modular structures of the expression data were characterised by first identifying the transcription modules in each data set (using the ISA) and subsequently their organisation in each transcription program. The relative importance of conserved modules to the transcription program varies significantly

between organisms. Moreover, a significant number of modules are composed primarily of genes that are organism-specific.

When the co-expression of modules (rather than of individual genes) was compared, similarities and differences were revealed in higher-order structures of the various transcription programs. Specifically, it was asked whether pairs of modules that are (anti-)correlated in one organism, exhibit the same regulatory relationship in another organism. Studying a set of eight representative modules related to core cellular processes among the six diverse species, the available expression data indicated that relatively few of these relationships have been conserved among the six diverse species.

Comparison of expression networks

Global properties of expression data can also be studied by constructing 'expression networks' where genes with similar expression profiles are connected.[59] Despite the small proportion of conserved relationships, it was found that basic topological properties of such networks are conserved in all six organisms. This includes power-law connectivity distributions (with similar exponents), increased likelihood of connecting genes of similar connectivity, and a high degree of clustering.[53] In addition, highly connected genes were significantly more likely to be essential and conserved.

Synthetic expression data

DISCUSSION

Performance of various methods

The specific challenges that arise in the analysis of large heterogeneous data sets have been addressed with a number of different methods (Table 2). Central issues include the capacity to identify condition-specific regulation, to assign genes and conditions to multiple overlapping clusters and computational efficiency. In addition to the Signature Algorithm and ISA, Coupled Two-Way Clustering,[18] biclustering by Cheng and Church,[17] probabilistic graphical models[31] and the SAMBA algorithm[22,23] all have the ability to identify potentially overlapping clusters in a condition-specific manner. While the Gibbs sampling variant of biclustering,[24] introduced by Sheng *et al*., groups genes with respect to subsets of conditions, the proposed masking procedure precludes the assignment of genes to multiple clusters. Conversely, fuzzy *k*-means clustering[25] allows for the assignment of genes to multiple clusters, but evaluates similarity between genes across the entire data set.

Apart from a consideration of such basic features, a comparative assessment of different clustering methods is difficult. Tanay *et al*. compare the performance of their biclustering method[23] with that of Cheng and Church[17] by projecting the respective solutions against the known correct partitions in the data. However, for large and heterogeneous data sets, underlying 'true' solutions are not available. The authors recently attempted to quantify biological coherence based on the conservation of putative *cis*-regulatory binding sites between four related yeast species, and compared the outputs of several clustering methods applied to the same data set.[51] An alternative approach is the analysis of synthetic gene expression data. Several of the methods discussed above used *in-silico* data to test algorithm performance in a controlled setting. A systematic assessment of different algorithms applied to the same large synthetic data set containing overlapping and context-specific modules would be valuable for comparison, as well as for improvement and fine-tuning of each individual method.

Integration of additional data sources

Large-scale gene-expression data sets hold the promise of a global view of the transcription program. Substantial progress has been made toward this goal, including the identification of co-expressed units of genes, the modular and hierarchical structure of the transcription network and higher-order relationships between distinct transcription modules. An essential next step toward a better understanding of the system-level properties of cellular networks is the analysis of expression patterns in the

context of additional sources of genomic information. This review has summarised some of the results obtained by integrating expression data with promoter sequence, pathway databases or genomic sequence, using the Signature Algorithm. Another promising and more general framework is the SAMBA algorithm,[22,23] which combines graph theory with statistical data modelling to represent relationships between genes or proteins and generalised properties. The SAMBA formalism identifies modules of genes with correlated behaviour across various genome-wide data sets. These data sets may consist of expression data (biclustering),[23] but can also integrate more diverse data sources such as protein interactions, phenotypic measurements or transcription factor binding data.[22] Probabilistic graphical models represent another very general approach that has been successfully applied to a variety of problems involving diverse sources of genomic data.[31–34] Probabilistic models have been used to identify condition-specific relationships between genes,[31] to identify groups of genes that are both co-expressed and code for interacting proteins ('pathways'),[32] to study relations between expression and regulatory motifs[33] and to explicitly model the regulator–target gene relationships of a transcription network.[34] Other works combining gene expression with diverse types of genomic data include the GRAM algorithm described above[45] and studies by Kemmeren *et al.*[60] and Schlitt *et al.*[61] Integrative approaches such as these are likely to play a central role in future research efforts. Going beyond the pure decomposition of expression matrices into transcription units, such analyses promise new insights into global features of the transcription program and of the interplay between different levels of cellular organisation.

Acknowledgments

We thank Naama Barkai and Judith Berman for helpful comments. This work was supported by an NIH grant and the Israeli Ministry of Science. SB is a Koshland fellow.

References

1. Kitano, H. (2002), 'Systems biology: A brief overview', *Science*, Vol. 295, pp. 1662–1664.
2. Lander, E. S. (1999), 'Array of hope', *Nature Genet.*, Vol. 21, pp. 3–4.
3. Tavazoie, S., Hughes, J. D., Campbell, M. J. *et al.* (1999), 'Systematic determination of genetic network architecture', *Nature Genet.*, Vol. 22, pp. 281–285.
4. Brazma, A. and Vilo, J. (2000), 'Gene expression data analysis', *FEBS Lett.*, Vol. 480, pp. 17–24.
5. Slonim, D. K. (2002), 'From patterns to pathways: Gene expression data analysis comes of age', *Nature Genet.*, Vol. 32(Suppl.), pp. 502–508.
6. Kim, S. K., Lund, J., Kiraly, M. *et al.* (2001), 'A gene expression map for *Caenorhabditis elegans*', *Science*, Vol. 293, pp. 2087–2092.
7. Ihmels, J., Friedlanders, G., Bergman, S. *et al.* (2002), 'Revealing modular organization in the yeast transcriptional network', *Nature Genet.*, Vol. 31, pp. 370–377.
8. Hughes, T. R., Marton, M. J., Jones, A. R. *et al.* (2000), 'Functional discovery via a compendium of expression profiles', *Cell*, Vol. 102, pp. 109–126.
9. Wu, L. F., Hughes, T. R., Davier Wala, A. P. *et al.* (2002), 'Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters', *Nature Genet.*, Vol. 31, pp. 255–265.
10. Bussemaker, H. J., Li, H. and Siggia, E. D. (2001), 'Regulatory element detection using correlation with expression', *Nature Genet.*, Vol. 27, pp. 167–171.
11. Hughes, J. D., Estep, P. W., Tavazoie, S. and Church, G. M. (2000), 'Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*', *J. Mol. Biol.*, Vol. 296, pp. 1205–1214.
12. Wang, W., Cherry, J. M., Botstein, D. and Li, H. (2002), 'A systematic approach to reconstructing transcription networks in *Saccharomyces cerevisiae*', *Proc. Natl Acad. Sci. USA*, Vol. 99, pp. 16893–16898.
13. Gasch, A. P., Spellman, P. T., Kao, C. M. *et al.* (2000), 'Genomic expression programs in the response of yeast cells to environmental changes', *Mol. Biol. Cell*, Vol. 11, pp. 4241–4257.
14. Brazma, A., Robinson, A., Cameron, G. and Ashburner, M. (2000), 'One-stop shop for microarray data', *Nature*, Vol. 403, pp. 699–700.

15. Brazma, A., Hingamp, P., Quackenbush, J. *et al.* (2001), 'Minimum information about a microarray experiment (MIAME) – toward standards for microarray data', *Nature Genet.*, Vol. 29, pp. 365–371.

16. Spellman, P. T. *et al.* (2002), 'Design and implementation of microarray gene expression markup language (MAGE-ML)', *Genome Biol.*, Vol. 3, pp. research0046.1–0046.9.

17. Cheng, Y. and Church, G. M. (2000), 'Biclustering of expression data', *Proc. Int. Conf. Intell. Systems Mol. Biol.*, Vol. 8, pp. 93–103.

18. Getz, G., Levine, E. and Domany, E. (2000), 'Coupled two-way clustering analysis of gene microarray data', *Proc. Natl Acad. Sci. USA*, Vol. 97, pp. 12079–12084.

19. Getz, G. and Domany, E. (2003), 'Coupled two-way clustering server', *Bioinformatics*, Vol. 19, pp. 1153–1154.

20. Getz, G., Gal, H., Kela, I. *et al.* (2003), 'Coupled two-way clustering analysis of breast cancer and colon cancer gene expression data', *Bioinformatics*, Vol. 19, pp. 1079–1089.

21. Blatt, M., Wiseman, S. and Domany, E. (1996), 'Superparamagnetic clustering of data', *Phys. Rev. Lett.*, Vol. 76, pp. 3251–3254.

22. Tanay, A., Sharan, R., Kupiec, M. and Shamir, R. (2004), 'Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data', *Proc. Natl Acad. Sci. USA*, Vol. 101, pp. 2981–2986.

23. Tanay, A., Sharan, R. and Shamir, R. (2002), 'Discovering statistically significant biclusters in gene expression data', *Bioinformatics*, Vol. 18(Suppl. 1), pp. S136–144.

24. Sheng, Q., Moreau, Y. and De Moor, B. (2003), 'Biclustering microarray data by Gibbs sampling', *Bioinformatics*, Vol. 19(Suppl. 2), pp. II196–II205.

25. Gasch, A. P. and Eisen, M. B. (2002), 'Exploring the conditional coregulation of yeast gene expression through fuzzy *k*-means clustering', *Genome Biol.*, Vol. 3, pp. research0059.1–0059.22.

26. Holter, N. S., Mitra, M., Maritan, A. *et al.* (2000), 'Fundamental patterns underlying gene expression profiles: Simplicity from complexity', *Proc. Natl Acad. Sci. USA*, Vol. 97, pp. 8409–8414.

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

28. Owen, A. B., Stuart, J., Mach, K. *et al.* (2003), 'A gene recommender algorithm to identify coexpressed genes in *C. elegans*', *Genome Res.*, Vol. 13, pp. 1828–1837.

29. Lazzeroni, L. and Owen, A. (1999), 'Plaid models for gene expression data', Technical report, Stanford University, Statistics.

30. Hastie, T., Tibshirani, E., Eisen, M. B. *et al.* (2000), '"Gene shaving" as a method for identifying distinct sets of genes with similar expression patterns', *Genome Biol.*, Vol. 1, p. RESEARCH0003.

31. Segal, E., Taskar, B., Gasch, A. *et al.* (2001), 'Rich probabilistic models for gene expression', *Bioinformatics*, Vol. 17(Suppl. 1), pp. S243–252.

32. Segal, E., Wang, H. and Koller, D. (2003), 'Discovering molecular pathways from protein interaction and gene expression data', *Bioinformatics*, Vol. 19(Suppl. 1), pp. i264–271.

33. Segal, E., Yelensky, R. and Koller, D. (2003), 'Genome-wide discovery of transcriptional modules from DNA sequence and gene expression', *Bioinformatics*, Vol. 19(Suppl. 1), pp. i273–282.

34. Segal, E., Shapira, M., Regev, A. *et al.* (2003), 'Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data', *Nature Genet.*, Vol. 34, pp. 166–176.

35. Madhani, H. D. and Fink, G. R. (1997), 'Combinatorial control required for the specificity of yeast MAPK signaling', *Science*, Vol. 275, pp. 1314–1317.

36. Yuh, C. H., Bolouri, H. and Davidson, E. H. (1998), 'Genomic *cis*-regulatory logic: Experimental and computational analysis of a sea urchin gene', *Science*, Vol. 279, pp. 1896–1902.

37. Pilpel, Y., Sudarsanam, P. and Church, G. M. (2001), 'Identifying regulatory networks by combinatorial analysis of promoter elements', *Nature Genet.*, Vol. 29, pp. 153–159.

38. Sudarsanam, P., Pilpel, Y. and Church, G. M. (2002), 'Genome-wide co-occurrence of promoter elements reveals a *cis*-regulatory cassette of rRNA transcription motifs in *Saccharomyces cerevisiae*', *Genome Res.*, Vol. 12, pp. 1723–1731.

39. Werner, T. (2001), 'The promoter connection', *Nature Genet.*, Vol. 29, pp. 105–106.

40. Hartigan, J. (1975), 'Clustering Algorithms', Wiley, New York.

41. Gavin, A. C., Bosche, M., Krause, R. *et al.* (2002), 'Functional organization of the yeast proteome by systematic analysis of protein complexes', *Nature*, Vol. 415, pp. 141–147.

42. Lee, T. I., Rinaldi, N. J., Robert, F. *et al.* (2002), 'Transcriptional regulatory networks in *Saccharomyces cerevisiae*', *Science*, Vol. 298, pp. 799–804.

43. Mewes, H. W., Amid, C., Arnold, R. *et al.* (2004), 'MIPS: Analysis and annotation of

proteins from whole genomes', *Nucleic Acids Res*., Vol. 32 Database issue, pp. D41–44.

44. Huh, W. K., Falvo, J. V., Gerke, L. C. *et al.* (2003), 'Global analysis of protein localization in budding yeast', *Nature*, Vol. 425, pp. 686–691.
45. Bar-Joseph, Z., Gerber, G. K., Lee, T. I. *et al.* (2003), 'Computational discovery of gene modules and regulatory networks', *Nature Biotechnol*., Vol. 21, pp. 1337–1342.
46. Ihmels, J., Levy, R. and Barkai, N. (2004), 'Principles of transcriptional control in the metabolic network of *Saccharomyces cerevisiae*', *Nature Biotechnol*., Vol. 22, pp. 86–92.
47. Segrè, D. (2004), 'The regulatory software of cellular metabolism', *Trends Biotechnol*., Vol. 22, pp. 261–265.
48. Edwards, J. S., Ibarra, R. U. and Palsson, B. O. (2001), '*In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data', *Nature Biotechnol*., Vol. 19, pp. 125–130.
49. Kanehisa, M., Goto, S., Kawashima, S. and Nakaya, A. (2002), 'The KEGG databases at GenomeNet', *Nucleic Acids Res*., Vol. 30, pp. 42–46.
50. Bergmann, S., Ihmels, J. and Barkai, N. (2003), 'Iterative signature algorithm for the analysis of large-scale gene expression data', *Phys. Rev. E Stat. Nonlinear Soft Matter Phys*., Vol. 67, p. 031902.
51. Ihmels, J., Bergmann, S. and Barkai, N. (2004), 'Defining transcription modules using large-scale gene expression data', *Bioinformatics*, Vol. 20(13), pp. 1993–2003.
52. Kloster, M., Tang, C. and Wingreen, N. (2004), 'Finding regulatory modules through large scale gene expression data analysis', Also published online: Bioinformatics, Advanced Access, 28 October, 2004.
53. Bergmann, S., Ihmels, J. and Barkai, N. (2004), 'Similarities and differences in genome-wide expression data of six organisms', *PLoS Biol*., Vol. 2, p. E9.
54. Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998), 'Cluster analysis and display of genome-wide expression patterns', *Proc. Natl Acad. Sci. USA*, Vol. 95, pp. 14863–14868.
55. Kluger, Y., Basri, R., Chang, J. T. and Gerstein, M. (2003), 'Spectral biclustering of microarray data: Coclustering genes and conditions', *Genome Res*., Vol. 13, pp. 703–716.
56. Fellenberg, K., Hauser, N. C., Brors, B. *et al.* (2001), 'Correspondence analysis applied to microarray data', *Proc. Natl Acad. Sci. USA*, Vol. 98, pp. 10781–10786.
57. Sherlock, G., Hernandez-Boussard, T., Kasaiskis, A. *et al.* (2001), 'The Stanford Microarray Database', *Nucleic Acids Res*., Vol. 29, pp. 152–155.
58. Stuart, J. M., Segal, E., Koller, D. and Kim, S. K. (2003), 'A gene-coexpression network for global discovery of conserved genetic modules', *Science*, Vol. 302, pp. 249–255.
59. Farkas, I., Jeong, H., Vicsek, T. *et al.* (2003), 'The topology of the transcription regulatory network in the yeast, *Saccharomyces cerevisiae*', *Phys. A: Stat. Mech. Appl*., Vol. 318, pp. 601–612.
60. Kemmeren, P., van Berkum, N. L., Vilo, J. *et al.* (2002), 'Protein interaction verification and functional annotation by integrated analysis of genome-scale data', *Mol. Cell*, Vol. 9, pp. 1133–1143.
61. Schlitt, T., Palin, K., Rung, J. *et al.* (2003), 'From gene networks to gene function', *Genome Res*, Vol. 13, pp. 2568–2576.