Available online at www.sciencedirect.com
ScienceDirect
Journal of Biomedical Informatics 40 (2007) 787–802
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Methodological Review

Towards knowledge-based gene expression data mining

Riccardo Bellazzi [a], Blaž Zupan [b,c,*]

[a] Dipartimento di Informatica e Sistemistica, Università di Pavia, via Ferrata 1, I-27100 Pavia, Italy
[b] Faculty of Computer and Information Science, University of Ljubljana, Trzaska 25, SI-1000, Slovenia
[c] Department of Molecular and Human Genetics, Baylor College of Medicine, 1 Baylor Plaza, Houston, TX 77030, USA

Received 29 September 2006
Available online 21 June 2007

Abstract

The field of gene expression data analysis has grown in the past few years from being purely data-centric to integrative, aiming at complementing microarray analysis with data and knowledge from diverse available sources. In this review, we report on the plethora of gene expression data mining techniques and focus on their evolution toward knowledge-based data analysis approaches. In particular, we discuss recent developments in gene expression-based analysis methods used in association and classification studies, phenotyping and reverse engineering of gene networks.

Keywords: Gene expression data analysis; Data mining; Knowledge-based data mining; Gene association; Classification; Gene networks

1. Introduction

Recent technological advances in high-throughput experimental analysis have had a profound impact on the practices and scope of biomedical research. The volumes of data collected at the genome scale and the requirements for tools to analyze them have largely mobilized the data analysis community. Beyond new applications, the biomedical research of today continuously provides a set of tough challenges for data analysis that go well beyond the sole treatment of large data sets. Biomedicine is a field rich in knowledge, with numerous incentives to formally encode it in an electronic format and share it through usually open and community-maintained data and knowledge bases. Containing information on sequence and sequence structure, gene and protein interactions, function annotation and ontologies, or genetic and metabolic pathways.[1] This information can significantly complement any data analysis and improve its results. The inclusion of additional knowledge sources in the data analysis process can prevent the discovery of the obvious, complement a data-inferred hypothesis with references to already proposed relations, help analysis to avoid overconfident predictions and, finally, allow us to systematically relate the analysis findings to present knowledge.

This review focuses on one of the most active areas of collaboration between biomedical researchers and data analysis developers and practitioners. Since the first publications of gene expression data sets, data analysis methods have played an important, if not a major, role in presenting early DNA microarray results and demonstrating their potential applications. The excellent practice in the field of biomedicine where a number of publications in top-rated journals are accompanied with supplements that include the related experimental data has offered an opportunity for the bioinformatics community to re-analyze the data and compare original analysis techniques with the newly developed ones. In the less than 10 years since the emergence of the field, the result is a plethora of methodological papers published in already established journals now showing renewed interest in data analysis and model-

* Corresponding author. Address: Faculty of Computer and Information Science, University of Ljubljana, Trzaska 25, SI-1000, Slovenia.
E-mail address: blaz.zupan@fri.uni-lj.si (B. Zupan).

[1] See, for instance, the compendium of data bases at the National Center for Biotechnology Information—www.ncbi.nlm.nih.gov/Entrez/ and European Bioinformatics Institute—http://www.ebi.ac.uk/Databases/

1532-0464/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2007.06.005

ing (e.g., *Bioinformatics*, *Journal of Biomedical Informatics*, *Artificial Intelligence in Medicine*, *Journal of Computational Biology*, *Journal of Theoretical Biology* and alike) and in a large variety of new journals (e.g., *PLoS Computational Biology*, *BMC Bioinformatics*, *Applied Bioinformatics*, *Briefings in Bioinformatics*, *Current Bioinformatics, IEEE Transactions on Computational Biology and Bioinformatics*) focusing on bioinformatics and computational approaches to biomedicine. A number of laboratories previously specialized in statistics, informatics, artificial intelligence and data mining are now turning their focus to bioinformatics, computational biomedicine and systems biology.

The field of gene expression data analysis has grown fast: early approaches that mainly involved clustering have quickly been complemented with perhaps more sophisticated but often better-fitting tools for particular analysis tasks. With the growing number of complementary data and knowledge bases, the field has shifted from the application of pure data-oriented methods to methods that aim to include additional knowledge in the data analysis process. These methods are often referred to as intelligent data analysis and form the primary focus of this review. Intelligent data analysis refers to all methods that are devoted to *automatically* transforming data into information exploiting the background knowledge in the domain. Background knowledge is the domain knowledge obtained from the literature, domain experts or from available knowledge repositories. An intelligent data analysis approach usually addresses the problems of the acquisition, encoding and exploitation of background knowledge. Intelligent data analysis, of course, does not exclude user intervention during the information extraction process but aims at reducing interaction through the use of background knowledge, thus reducing the costs of data analysis in terms of both users' time and resources.

Dealing with data coming from DNA microarrays, we should warn the knowledgeable reader that we will not address any of the issues related to microarray image analysis, normalization and data preprocessing. Moreover, of the plethora of data mining approaches currently being applied to problems in bioinformatics (see Allison et al. [1] and Riva et al. [2] for recent critical reviews), we will only review those that can simultaneously deal with various (structured) data sets and sources that explicitly encode the domain knowledge. The paper starts with a review of approaches for finding gene associations, continues with methods for expression-based classification and phenotyping and finishes with techniques for the reverse engineering of gene networks.

2. Gene association studies

It has been widely recognized that the genes involved in the same biological process or with a similar function are likely to be co-expressed [3]. One possible way to perform gene function discovery is thus to group genes with a similar expression profile, consisting of gene expressions measured at either different conditions or in different time points. The functional annotation of a new gene can then be hypothesized on the basis of functional classes of the other, similarly expressed genes. Because of the relative simplicity of available methods and related visualizations that can reveal the underlying data structure, it is not surprising that the area of DNA microarrays data analysis which has so far probably received the greatest attention is the clustering of gene expression profiles [4]. For a critical review of those methods we refer to the paper by Hand and Heard [5]. Over the last couple of years and following the principal idea behind intelligent data analysis, new efforts have been devoted to increase the performance of clustering methods in terms of their robustness and stability by also considering the available knowledge on gene function. For example, such knowledge may be related to the process under study such as the periodicity of the cell cycle [6] or may be codified in knowledge repositories such as Gene Ontology [7], MIPS [8] and KEGG [8].

One seminal work on the combination of heterogeneous data and evidence sources is the development of the system called Magic [9]. Magic uses a Bayesian Network which combines evidence from different data sources to predict if two proteins are functionally related. Following the ideas of probabilistic expert systems in diagnosis, Magic is able to weight different knowledge sources and derive a posterior probability on the hypothesis of functional relationships. The system has been experimentally validated on data sets from budding yeast *Saccharomyces cerevisiae*.

Recently, some attention has been paid to the modification of clustering algorithms for embedding background knowledge. Clustering methods are often divided into three main classes: distance-based, model-based and template-based. Below we provide a survey of approaches that adapt these methods so that they take the additional background knowledge into account. We will use a notation where we suppose that the expression of n genes was measured using DNA microarrays in m different experimental conditions and we will denote the set of expression measurements of the ith gene as $x_i = \{x_{i1}, x_{i2}, \ldots, x_{ij}, x_{im}\}$, where x_{ij} is the jth measurement, with $j = 1, \ldots, m$ and $i = 1, \ldots, n$. We will also call x_i an expression profile of the ith gene.

2.1. Distance-based clustering

The majority of current applications of gene expression clustering are based on an estimation of the distance between expression profiles. The basic idea of these methods is to cluster together those genes which are 'close' to each other according to some distance measure. The most popular method for distance-based clustering is agglomerative hierarchical clustering which derives a hierarchy of clusters ordered in a tree (Fig. 1). The leaves of the tree are the genes, which represent the smallest clusters; at each subsequent node or level of the tree, the two nearest clusters are grouped to form a bigger cluster [3]. The procedure is iterated until a single cluster of all genes in the data is

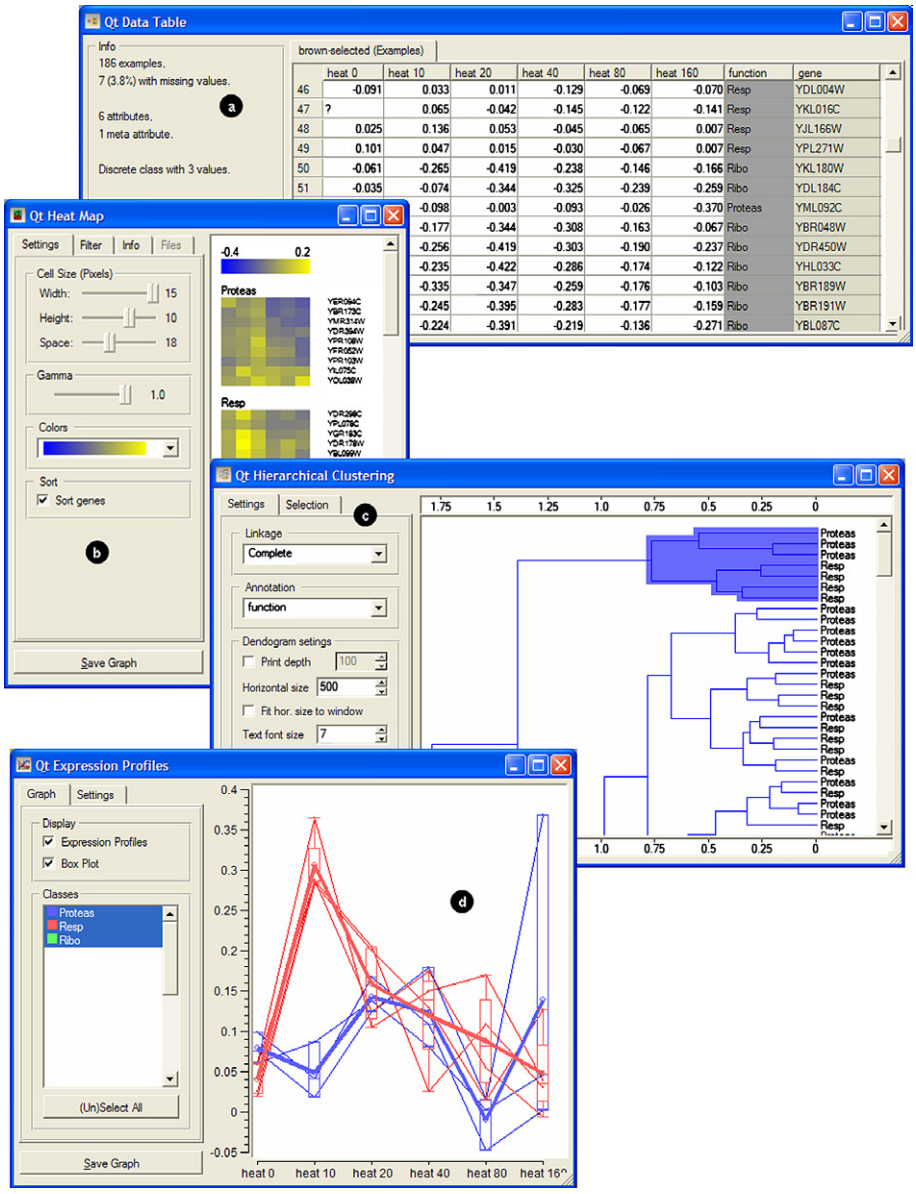

Fig. 1. Snapshots of several standard gene expression analysis and visualization techniques. (a) Time series gene expression data during heat shock from Eisen et al. [3]. Only a subset of genes that belong to either of the three different classes (cytoplasmic ribosomes, proteasome and respiration) as assigned by Brown et al. [35] are included. The spreadsheet displays expression data (first six columns), class label and gene name. (b) The same data displayed using heat map visualization. (c) Dendrogram showing the result of hierarchical clustering with gene classes displayed as dendrogram labels. Topmost dendrogram branch with six genes is selected, their expression profiles are displayed in (d). Snapshots are taken from Orange data mining suite [11].

obtained, which also forms the root node of the tree. Another popular clustering method is *k*-means [10] which partitions the *m*-dimensional space of genes into *k* regions in order to minimize the variance of the data within each region. The algorithm typically starts by selecting *k* different cluster centers, assigning each gene to a cluster with the nearest centre. Then, the mean (centroid) of each cluster is computed and the cluster centers are updated. The genes are reassigned to the clusters and the algorithm is iterated until it converges to a stable solution. Several variants of *k*-means have been proposed in the literature, including *k*-medoids to improve on robustness against the outliers [10].

Gene expression profile-based distance functions are also exploited in a technique called self-organizing maps (SOM) [12]. SOM are maps of a small, usually two-dimensional space in which each point represents a cluster. During the clustering algorithm a mapping function is automatically built in order to assign the genes to one of the points of the map in such a way that the clusters which are close on the map are also similar in the original *m*-dimensional space. A straightforward use of distance function has also been explored in so-called gene co-expression networks [13], which feature a graph of genes as nodes and connections if the gene-to-gene distance is below some preset threshold.

Each of these methods has its distinct advantages: if the number of desired clusters is known in advance, the use of *k*-means may be preferred. Results of hierarchical clustering can be nicely visualized in a dendrogram which may depict the structure of the data and provide a hint as to the number and composition of distinct clusters. While in most reports researchers only present a single dendrogram, this has often been chosen arbitrarily as there are many different ways we can present a tree-hierarchy (consider rotating part of the dendrogram that stems at each of the nodes of the tree). Approaches to optimize the ordering of dendrogram leaves have been proposed, but are either not used or not yet available in major data analysis packages [14]. In an un-optimized dendrogram, the profiles of genes that are distant in the dendrogram with respect to some target gene may in fact be more similar to the profiles of genes that are presented in the dendrogram closer to the target. In this respect, a two-dimensional visualization of SOM may provide more comprehensive information on the similarities of gene expression profiles (for an example of such visualization, see Fig. 2a). Gene co-expression networks can have attractive visualizations with tools like Pajek [15] (Fig. 3) but suffer from the need to set a similarity cut-off that determines which gene pairs will be connected. Finding an appropriate cut-off depends on the data analysis task, as this can balance between coverage (the proportion of genes connected to other genes with a shared function) and accuracy (the proportion of connections of function-related genes that lead to genes with shared functions) [13].

All the methods mentioned above require the definition of the distance between two gene expressions. The most widely used functions in this respect are Euclidean distance and Pearson's correlation. With distance functions playing a central role in distance-based clustering, an interesting research direction is to adapt them by means of incorporating the available background knowledge in computation of the gene distance.

An often used source of available prior knowledge in functional genomics is a repository on gene functions, such as the Gene Ontology (GO) [7]. As defined by the Gene Ontology consortium (www.geneontology.org), the GO is 'a comprehensive structured vocabulary of terms describing different elements of molecular biology that are shared among life forms.' The GO is organized along three main axes, molecular function, biological process and cellular component, where the concepts are represented though a taxonomy, going from the most general to the most specific terms. Such a taxonomy can be easily represented as a graph (see Fig. 2b).

Usually GO annotations are used for enrichment analysis, where we can test if a subset of genes we find based on expression similarity has some interpretable biological significance [16,17]. The result of this analysis may be summarized through a list of GO annotations whose relative frequency is significantly different to that of the reference set (see Fig. 2b for an example). A number of tools for GO terms enrichment analysis are available and were recently reviewed by Khatri and Draghici [18]. In our review, however, we are interested in GO used as an additional background knowledge when the clusters are generated. Obviously, in this case the evaluation of the biological meaning of the clusters must be performed by using an independent knowledge base; in other words GO cannot be exploited for both hypothesis generation and the assessment of clustering results.

To exploit the information available in the GO in the calculation of a distance, some authors have introduced the notion of semantic similarity between genes, computed on the basis of the available taxonomies of concepts. The GOstats package [19] of the R-library Bioconductor (www.bioconductor.org), for instance, allows one to estimate the so-called graph similarity between objects under observation. Given two objects, such as genes or proteins, each object is associated with a sub-graph that is obtained by taking the most specific GO terms annotated with the gene or protein and by finding all their parents up to the root node. Then, the similarity between the two objects is defined through the *union-intersection* method, which computes the number of shared nodes divided by the total number of nodes in the two sub-graphs, or by the *longest shared path* method, which calculates the length of the longest path shared by the two nodes.

Graph similarity does not explicitly take into account the frequency of the terms in the corpus, a deficiency that led to the development of so-called information-based similarity. Denoted with $p(t)$ the relative frequency of a term t or of any child term in the GO (or any other corpus), the similarity between two terms t_1 and t_2 as defined by Lin [20] is:

$$\mathrm{sim}(t_1, t_2) = \frac{2 * \log\left(\min_{t \in S(t_1,t_2)} p(t)\right)}{\log(p(t_1)) + \log(p(t_2))}$$

where $S(t_1, t_2)$ is the set of shared parents of t_1 and t_2. Other variants of this similarity measure have been reported by Resnik [21] and Jiang and Conrath [22]. An evaluation of the different similarity scores can be found in [23]. The values of Lin's similarity score fall in the interval between 0 and 1 so that it is easy to obtain a distance measure defined as $d(t_1,t_2) = 1 - \mathrm{sim}(t_1, t_2)$.

Relying on information-based similarity, Kustra and Zagdanski [24] worked on the integration of the semantic and expression-based distance between genes. In order to take multiple annotations (terms) of a gene into account, they computed the weighted average of the similarity scores between all term annotations to obtain the distance $d(g_1, g_2)$ between two genes g_1 and g_2. Then they proposed computing the combined distance using the following convex combination

$$\mathrm{dist}(g_1, g_2) = \lambda \cdot d(x_1, x_2) + (1 - \lambda) \cdot d(g_1, g_2)$$

where $d(x_1, x_2)$ is the distance between the gene expression profiles and λ is a user-defined coefficient in the interval $[0, 1]$ that defines the balance between the expression- and

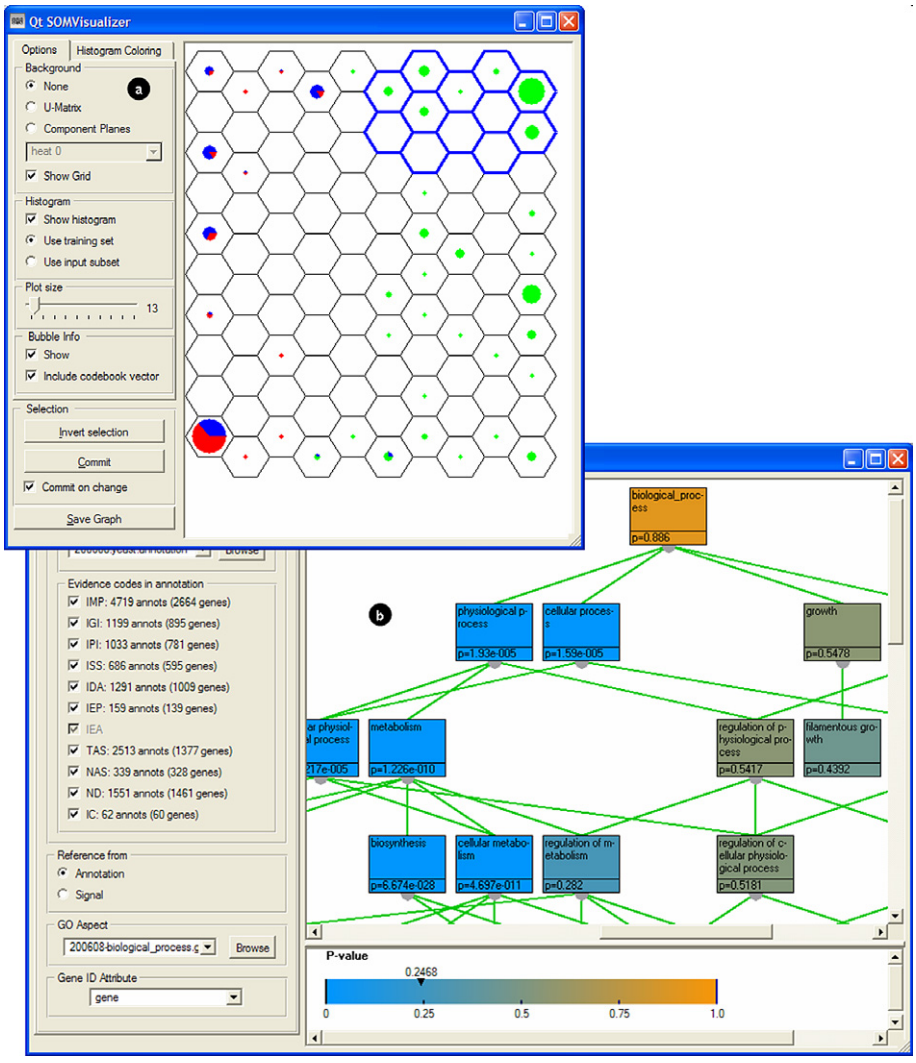

Fig. 2. (a) Visualization of self-organizing map using the data set from Fig. 1. Pie charts in each cell correspond to class distribution. Thirty-eight genes from top-right corner of the map were selected (selected cells are outlined with bold blue line). (b) A snapshot of a GO term browser (shown is biological process aspect, with nodes close to the root), displaying a term enrichment analysis for selected genes. Nodes representing GO terms with relative gene frequencies that deviate most from reference relative frequencies computed from a complete yeast genome are colored in blue. Only the terms with non-zero gene representation are displayed. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

GO-based distance. Kustra and Zagdanski have validated this approach through an analysis of a budding yeast gene expression data set by running a k-medoids algorithm. Their analysis has been compared with the catalogue of protein–protein interaction data set with encouraging results.

Another interesting approach has recently been published by Huang and Pan [25], who modified the distance function used in a k-medoids algorithm to obtain a shrinkage effect when calculating the distance between genes which are known to share a common function. In their case, the distance was computed as $\mathrm{dist}(g_1, g_2) = \lambda \cdot d(x_1, x_2)$ if x_1 and x_2 share a common function, and as $\mathrm{dist}(g_1, g_2) = d(x_1, x_2)$ otherwise. A constant λ is a suitable user-defined shrinkage parameter. Huang and Pan then extract a set of clusters for genes with known functions and assign to those clusters or to new clusters the genes with unknown functions. The results have been favorably evaluated in terms of the accuracy of gene function prediction and studied with simulated and real DNA microarrays data sets collected on yeast.

2.2. Model-based clustering

In model-based clustering each gene expression profile x_i is assumed to be drawn from a probability distribution $f(x_i, \theta)$ which is typically a finite mixture of c components, with each one representing a different cluster:

$$f(x_i, \theta) = \sum_{k=1}^{c} p_k f_k(x_i, \theta_k)$$

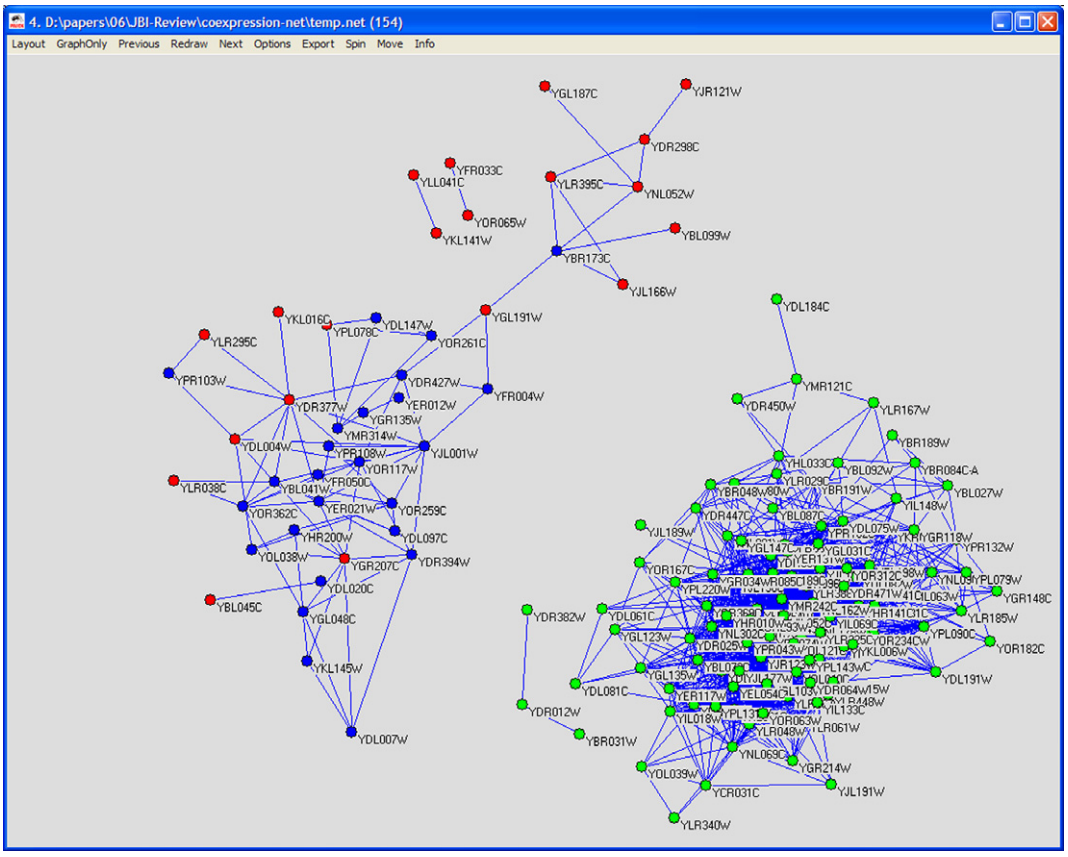

Fig. 3. Pajek [15] visualization of gene co-expression network derived from the same data set as used in Fig. 1. Neighboring gene in the network are connected if the Euclidian distance of their expression profiles is less than 0.2. Nodes representing genes with no connections to other genes (32 genes out of total of 186 genes) are not included in the graph. Nodes in the network are colored with respect to class associated to genes in [35] (green for cytoplasmic ribosomes, blue for proteasome and red for respiration). Genes of the first class (cytoplasmic ribosomes) are nicely clustered, while those of the other two classes are intermixed. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The parameters of this model are the weight coefficients p_k which may be considered as the (prior) probability of each cluster, and the parameters specifying the single cluster distribution $f_k(x_i, \theta_k)$. Although the assumption of finite mixture is the most common, it would also be possible to exploit more complex models such as the Bayesian infinite mixture model or the Dirichlet process mixture model, as reported by Qin [26].

Once a data set of n genes is available, it is possible to compute the likelihood of any configuration of the parameters θ as:

$$L(\theta) = \prod_{i=1}^{n} \sum_{k=1}^{c} p_k f_k(x_i, \theta_k)$$

Within this setup, the goal of finding the clusters is transformed into a problem of likelihood maximization. The probability that the ith gene belongs to the kth cluster can be estimated from the data as $p(x_i \in C_k) \propto p_k f_k(x_i, \theta_k)$. Each gene is then assigned to the cluster that maximizes the probability of gene membership.

A standard choice is to exploit a mixture of normal distributions so that the parameters θ are mean vectors and covariance matrices of each cluster. One of the most popular algorithms proposed in the literature to maximize the likelihood is the Expectation Maximization approach, which enables a joint estimate of p_k and θ (see [26] for a complete description of the algorithm). The EM starts with an initial guess at the model parameters $\hat{p}_k, \theta_k$, for every k. Then, in the E-steps the expected value of the posterior probability distribution that x_i belongs to the kth cluster (C_k) is computed as:

$$\hat{p}(x_i \in C_k) = \frac{\hat{p}_k f_k(x_i, \hat{\theta}_k)}{\sum_{k=1}^{c} \hat{p}_k f_k(x_i; \hat{\theta}_k)}$$

In the M-step, the parameter estimates are properly updated. In particular, the prior probability is obtained with the following equation:

$$\hat{p}_k = \frac{\sum_{i=1}^{n} \hat{p}(x_i \in C_k)}{n}$$

By iterating the procedure until convergence it is possible to obtain the Maximum Likelihood Estimate of the cluster parameters and to assign each gene to a cluster by choosing the maximum value of the posterior probability distribution for that gene. The method can be easily extended to cope with the selection of a number of clusters by cross-validation

or by resorting to a Bayesian model selection. An approximation of the latter approach leads to the selection of the model with the highest BIC index [27].

The model-based approach allows one to add background knowledge in several ways. For example, Pan [28] extended the normal mixture model by stratification. In particular, he proposes identifying $g-1$ potential gene groups on the basis of the prior knowledge, for example, looking at the gene biological annotation in the GO. The gth group contains all the genes with an unknown function. The generative mixture model described above is modified by defining a prior probability distribution which is the same for all genes belonging to the same group. Let us note that g is in general different from the number of clusters c adopted in the model. Hence, the revised model for the genes *a priori* classified into the hth group becomes:

$$f_h(x_i, \theta_h) = \sum_{k=1}^{c} p_{hk} f_k(x_i, \theta_k)$$

The EM strategy is suitably modified to take prior information into account. While the posterior probability calculation and the estimate of the parameter set θ remains the same, the update of the hth prior distribution takes into account only the information coming from the genes *a priori* grouped into the hth group as follows:

$$\hat{p}_{hk} = \frac{\sum_{i=1}^{n_h} \hat{p}(x_i^h \in C_k)}{n_h}$$

where n_h is the number of genes belonging to the hth group and x_i^h is the ith gene belonging to this group. The final number of clusters (c) is then selected on the basis of the BIC index.

The approach proposed by Pan has been validated on simulated data and on yeast microarray data. The main problems in the handling of prior knowledge are that: (i) genes may be associated with different functional classes; and (ii) it may be difficult to define the initial number of groups. For example, genes may be involved in different processes and, looking at the gene ontology, it might be very difficult to define the level of classification hierarchy to be chosen in order to derive the groups used when modeling the prior distribution. To solve the first problem the author suggests multiplying the data of each instance of gene labeling with a sole gene function class, thus allowing the same gene to belong to different clusters. The second problem is however handled heuristically: it is suggested to *a priori* choose a number of groups which is not too big or too small and to then rely on the final clustering selected through the BIC index.

2.3. Template-based clustering

When gene expression time series are available, it is often important to recognize and retrieve the data patterns which may correspond to some interesting time-related behavior. If those patterns are known in advance, it is possible to directly apply pattern-matching techniques to solve the problem. However, the available knowledge is often expressed in terms of qualitative patterns or *templates* such as *increase*, *decrease* or *up* and *down*. Genes can then be clustered together according to qualitative similarities in one or more intervals of the overall time series.

The techniques proposed in the literature for performing clustering on the basis of qualitative templates seem promising in terms of providing a suitable means to a domain expert to elicit their knowledge on the problem. These approaches thus usually require a definition of such qualitative templates and an algorithm that matches the templates with the quantitative profiles. Such definitions and mappings are usually knowledge-based since they need to deal with measurement noise and they may be specific to a phenomenon under observation. Knowledge-based temporal abstractions [29,30] may be particularly suitable for performing this kind of qualitative template search.

The main advantage of the qualitative, template-based representation of temporal gene expression profiles is that it enables one to perform clustering by following the same principles used by a human expert, that is, by looking at series with similar (relevant) qualitative behaviors. The principal drawback of the approach is that the various templates to be used in the analysis need to be enlisted in advance, thus in a way forcing the user to exhaustively hypothesize the templates prior to the start of the analysis. In order to overcome these limitations and to perform a completely unsupervised search for the qualitative patterns, two techniques have been proposed in the literature. It must be noted that both of them are currently at the stage of research proposals and that they are not used as frequently as the other approaches presented in this section.

An interesting knowledge-based template clustering approach has been proposed by Hvidsten and colleagues [31]. In their work, whose main goal was to find descriptive rules about the expression behavior of genes from some functional classes, they grouped and summarized the available gene time series by resorting to template-based clustering. They first enumerated all possible subintervals in the time series and labeled all possible subintervals as increasing, decreasing and steady with a temporal abstraction-like procedure. Then, they clustered together the genes matching the same templates over the same subintervals. In this way a single gene may be present in more than one cluster. The overall system has been favorably evaluated on data published by Cho et al. [32] on the cell cycle in human fibroblasts.

Another recent example of intelligent data analysis-inspired template-based clustering is provided by Sacchi and colleagues [30], who modeled time series data as a set of consecutive trend temporal abstractions, identifying the intervals in which one of the basic templates of *increasing, decreasing*, and *steady* matched the data. Clustering is then performed in an efficient way at three different levels of aggregation of the qualitative labels. At the first level, the gene expression time series with the same sequence of

increasing or decreasing patterns are clustered together. At the second level, the time series with the same sequence of increasing, steady or decreasing patterns are grouped, while at the third level the time series sharing the same qualitative labels on the same time intervals are clustered together. The results of this method, known as TA-Clustering, can be visualized as a three-level hierarchical tree and, as such, it is easy to be interpreted (Fig. 4). Sacchi et al. demonstrated the utility of the proposed algorithm on a set of two simulated data sets and on a study of yeast gene expression data.

3. Predictive modeling

Gene association studies relate genes by comparing their expression profiles. The soundness of discovered gene grouping is often judged based on the homogeneity of functional labels shared by function-labeled genes in each group. But, as stated above, since the functions of a subset of genes may be known in advance we could use this information directly to construct the gene expression-to-function mapping. The area of data analysis addressing such problems is called supervised data mining and most often features methods stemming from machine and statistical learning [10,33]. There are several major distinctions between unsupervised and supervised data mining.

Unsupervised approaches require the definition of the distance measure between gene expression profiles, that is, between the expressions obtained in various experimental conditions. Simplest, but also most commonly used distance measure treat these profile elements—features equally, that is, do not use any particular weighting schema

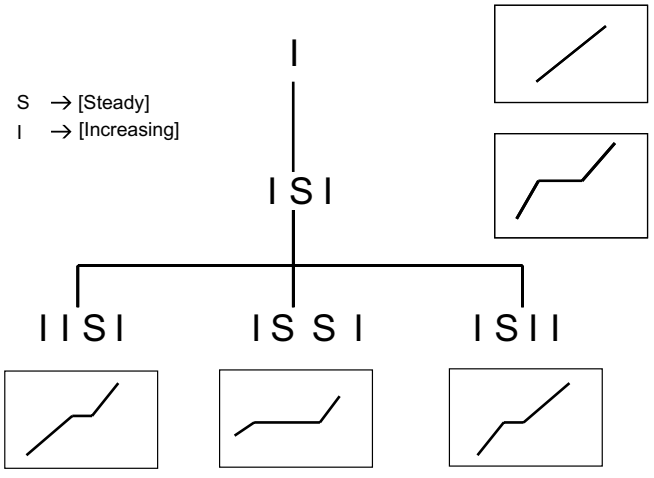

Fig. 4. A schematic representation of the temporal abstraction clustering algorithm applied on three hypothetical time series. At the bottom level, the time series sharing the same qualitative labels on the same time intervals are clustered together. In this case, the three time series are corresponding to three different clusters. At the second level, the time series with the same sequence of increasing and steady patterns are clustered. In this case, the three time series belong to the same cluster. Finally, at the upper (first) level the gene expression time series with the same sequence of increasing/decreasing patterns are clustered together. In the example, the three time series correspond to an increasing pattern.

to favor the result of one experiment over the other when computing the gene-to-gene distance. Supervised machine learning methods, on the other side, aim to find features (experimental conditions) that are most specific with respect to the observed class label (function, group) and favor them in the data-induced models. The quality of supervised data mining is also easier to judge using any of established statistical approaches like cross-validation and bootstrap, and model validation scores like classification accuracy, area-under-ROC or similar [34]. The principal problem the supervised methods need to address is overfitting, whereby the resulting model would well predict the classes from the training data but perform poorly on any new data set. All state-of-the-art supervised data mining include mechanisms to avoid overfitting, while it has become a standard practice in the field to always report performance results on test data sets only, either by indeed acquiring the separate test set on which the induced model is evaluated or by simulating such evaluation through, say, cross validation.

Supervised data analysis methods require class-labeled data. The classes, however, need to be sufficiently well represented; an ideal task for machine learning would, for instance, be a set of a few thousand genes, each labeled with a single function where each function would not be represented with less than, say, a 100 genes. Early demonstrations of the utility of supervised data mining (e.g., [35]) indeed treated such data, but the problems in functional genomics are often more complex. As reported in the previous section, genes may be involved in different pathways, perform different functions and may, for instance, be labeled in each of the aspects of GO using more then one term. GO includes a few thousand terms in each of the three aspects (molecular function, biological process, cellular components) and if the terms were to directly provide for class labels no standard machine-learning approach could sufficiently handle such data. An often explored bypass of the treatment of thousands of terms is to label the genes with only a few parent terms close to the root of ontology. For GO, we could for instance use so-called GO slim terms, which are a collection of high-level GO terms that best represent term-based annotations for a specific organism [7].

Instead of trying to fit the data to the present set of supervised data mining methods, an alternative approach is to modify these to appropriately consider the potentially rich and structured knowledge contained in current ontologies. This task is known in machine learning as hierarchical multi-label classification. While it has to some extent been explored in text mining [36,37], Blockeel and co-authors recently proposed an extension to classification trees and tested the method in a task to predict terms from the FunCat [38] protein classification hierarchy.

Predictive modeling has attracted considerable attention in problems related to cancer-gene expression studies, where a typical problem is tumor classification based on a set of tissue-specific gene expression profiles (e.g., [39]). With often small sample sizes (typically a couple of 100 tis-

sues/patients) and very high problem dimensionality (expressions of thousands of genes that characterize the tissue), the chance of reporting overconfident results due to overfitting or the inappropriate use of data mining methods in this domain is very high. This problem was nicely exposed in an excellent review by Simon et al. [40], who also warn about the inappropriate use of clustering in place of supervised data mining methods, and comment on recently frequently reported overly-optimistic results due to preprocessing by feature subset selection prior to cross-validation. Namely, an approach to deal with the curse-of-dimensionality in cancer gene expression studies is to select only the most informative set of genes prior to the modeling. If gene selection is performed prior to cross-validation, the selection of most informative features based on both the training and test set obviously leads to overfitting.

A notable task within gene expression-based classification is gene ranking. In principle, we can reduce the high dimensionality of the task by considering only the most promising genes for classification. In the simplest approach we could rank them according to their differential expression between the classes and consider only top-rated genes when building the classification model. Such a univariate gene-scoring method may fail to identify genes which are on their own not very informative, but due to their interactions—only become informative when considered in a group. Such a failure may be especially damaging for classification techniques like support vector machines that can exploit gene interactions. A number of multivariate gene-scoring schemes have recently been developed for this purpose. We refer interested readers seeking a review and comparison of these approaches to the works of Lai et al. [41] and Jeffery et al. [42].

A subset of genes best-ranked according to how well they can differentiate between different classes may also be checked for their biological significance using their GO annotation. A likely outcome of such analysis is however a list of genes with no unifying biological theme. Subramanian et al. [43] recently proposed a knowledge-based method called Gene Set Enrichment Analysis (GSEA) which uses *a priori* defined sets of gene groups (e.g., genes found in the same metabolic pathways, located in the same cytogenetic band, or sharing the same GO category). Instead of ranking individual genes, GSEA then ranks predefined gene sets. Besides their method, the authors provide an initial database of 1325 biologically defined gene sets, a useful resource yet to be fully exploited in other knowledge-based data analysis approaches.

A plethora of techniques is available for supervised data mining and the selection of a particular method for the task is far from trivial. Some recent reports (e.g., [44]) hinted that a popular data mining technique of support vector machines is the (sole) best choice, but their conclusion is only based on studies of predictive accuracy. Intelligent data analysis favors techniques that can present a discovered model in a readable form prone to interpretation and the study of the discovered relations. Support vector machines comply with this criteria only if linear kernels are used, whereas for more complex kernel functions the interpretation of results is far from trivial and other machine learning methods, like the induction of rules [45] or intelligent visualization techniques [46], may have an advantage in this respect.

4. Genome-wide gene expression profiles as a phenotype

The tissue-specific set of gene expressions as used in cancer-gene expression studies is in a way a replacement of the standard phenotype and provides grounds for tumor classification. Early reports on cancer microarray studies indeed show heat maps of gene expression profiles of different tumor types, provided as evidence that one can make an informed classification through a study of this (visual) fingerprint alone. While such phenotyping is, as we reported in Section 3, frequently used in cancer research, it is also gaining attention in functional genomics. Instead of associating genes through a set of measurements of its expressions under different conditions, one can mutate the gene under consideration and observe an expression of all other genes in a mutated organism. Such a mutant-based transcription profile can be favorable to classical, morphological phenotypes since it can encompass the state of the organism on a much larger scale. The approach has been pioneered in the work of Hughes et al. [47], resulting in a compendium of whole-genome transcription profiles of 300 single deletion mutants in *S. cerevisiae*. The mutant-based, large-scale expression-based phenotypes have provided grounds for the additional characterization of genes when compared to standard gene expression profiles consisting of measurements under different conditions. Hughes et al. present their experimental analysis as a heatmap, that is, a two-dimensional matrix with experiments (mutants) in rows and genes in columns, thus combining the information obtained from the gene expression profile with transcriptional information from its respective mutant. Two-dimensional agglomerative hierarchical clustering [48] was used to order the rows and columns to expose the patterns of over- and under-expression. This method, also referred to as biclustering, was developed in particular to overcome the limitations of standard clustering approaches to the analysis of gene expression data by grouping genes and samples simultaneously (see Prelic et al. [49] for a review and evaluation of different biclustering approaches).

Van Driessche et al. [50] recently showed that whole-genome expression profiles can be used on their own to characterize a mutated gene and to relate it with other genes in order to discover gene-regulation pathways. Their data included transcriptional phenotypes for single and double mutants of *Dictyostelium discoideum*, enabling the so-called universal epistasis analysis [51] by relating two genes according to their single and double-mutant phenotypes (Fig. 5). A similar study was reported by van de Peppel [52].

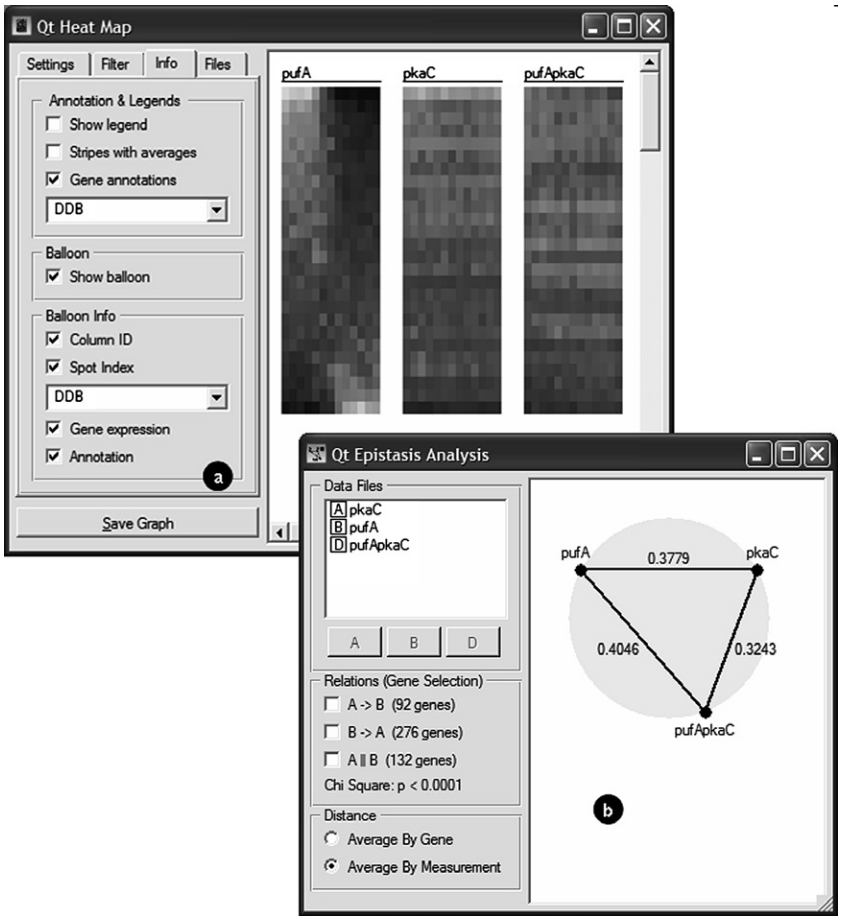

Fig. 5. Gene expression profiles of three *Dictyostelium discoideum* mutants using a data subset from Van Driessche et al. [50] (500 genes, each measured at 13 time points during amoebae development). (a) Heat map of mutant expression profiles showing that *pufA* mutant is different to otherwise similar profile of *pkaC* mutant and double mutant. b) Computational analysis of the differences, indicating for epistasis of gene *pkaC*, hypothesizing a regulation pathway *pufA* → *pkaC*.

The approaches mentioned above were all applied to model organisms where gene silencing through knock-outs is a viable research technique. With new approaches such as those that use RNA interference [53] and exploit a natural mechanism for gene silencing that occurs in organisms ranging from plants to mammals, dedicated computational tools that can deal with large-scale phenotypes that include those based on gene expression have yet to take ground. This also includes their applications for reverse genetic analysis in humans, where RNAi is rapidly being applied to study the function of many genes associated with human disease, in particular those associated with oncogenesis and infectious diseases [54].

5. Gene networks

One of the most intriguing possibilities offered by collections of genome-wide expression data sets is to infer a hypothesis on the gene-regulatory mechanisms. The regulation of gene transcriptional activity is a complex process which involves protein and protein complexes, intracellular signaling activity and, as recently discovered, specific control molecules such as micro-mRNA. The majority of the proposed computational models to describe regulation at the genome scale therefore provide an abstract view of the overall behavior, typically neglecting the details of the biochemical regulations. Gene-regulation relationships are most often represented in terms of a network. Following Schlitt and Brazma [55], we can distinguish between four different levels of detail for those networks, which may be integrated in a single, final model: (1) parts list, which are collections, descriptions and systematizations of network elements in a certain organism or for particular biological processes (like the list of transcription factors); (2) topology models, which are networks in which nodes represent genes and arcs represent the relationships between genes; (3) control models, which express the effect of one gene on another gene in terms of control action, such as activation and repression; and (4) dynamic models, which aim at defining a model able to predict the gene expression activity on the basis of the knowledge of the expression of other genes and, if available, other information such as an external stimulus. Once a suitable set of genes is selected, that is, a part list has been defined, several methods can be applied to infer the topology, control and gene expression dynamics directly from DNA microarray

data. In particular, the availability of time series gene expression data has stimulated researchers applying data mining approaches to infer control and dynamic models. An interesting review of a number of those approaches is provided by de Jong [56], while problems related to the analysis of time-oriented data are reviewed by Bar-Joseph [57].

Three different types of approaches and related network representation formalisms to infer gene-regulation networks from gene expression data have been largely exploited: Boolean networks, differential equations and Bayesian Networks. With Boolean networks, the binary discretized expression of a gene at time $t+1$ is modeled as a logical function of all modeled genes at time t. One of the first and perhaps most often cited approaches in this area is REVEAL (REVerse Engineering ALgorithm) [58], which infers Boolean networks from data under some limitations in terms of a possible number of antecedents of inferred logical rules. As another approach, using differential and difference equations assumes that the gene relationships can be described with a parametric dynamic model where the time derivative of each gene, that is, the rate of gene expression synthesis or degradation, is modeled as a function of the expression of other genes and of the gene itself [59]. Recently, the approach which has received more interest is to use the inference of causal probabilistic networks, in particular Bayesian Networks (BNs) and Dynamic Bayesian Networks (DBNs). Such networks are particularly appropriate for modeling gene expression data since they are probabilistic in nature so they can appropriately consider both measurement and modeling errors. Together with a model of the dynamics of the system, BNs and DBNs provide a topological model which describes the relationships between genes [60].

A current key question is whether it is really possible to infer such gene networks from gene expression data only. In fact, two crucial limits hamper the use of pure data-driven approaches. First, the amount of available data for each gene, that is, the number of time points where gene expression was recorded, is still insufficient to appropriately explore the space of all possible regulatory networks. In the case of learning a BN with 25 genes where the expression was discretized to a binary presentation (up- and down-regulation), it is possible to show that by limiting the number of genes that regulate the expression of a particular gene to four, a sufficient number of measurement samples exceeds 8000. Although such a problem can be partially circumvented by more parsimonious models (such as continuous models) [61], the limitations in the gene expression data sampling forces the algorithm to be biased towards simpler regulation structures to avoid overfitting. Second, the control or dynamic models chosen to describe the system always represent a great simplification of the molecular processes under analysis. This makes the model search feasible from a computational viewpoint but may lead to unstable solutions or, even worse, to obtain a set of equivalent models that describe the data equally well. This later problem is known in the modeling literature as an unidentifiability problem and cannot be solved even if an infinite number of samples is available. The unidentifiability problem is related to the lack of information which may be useful to disambiguate models, such as data on processes occurring to the cell during measurement (activated proteins, intra-cellular signals and external stimula).

Due to the problems mentioned above, the modeling community is currently particularly interested in the integration of different data sources in the learning process and exploiting the available background knowledge to guide the model search. Over the last few years several papers have proposed solutions that can take accumulating experience into account in at least two ways. First, information available in biological databases, such as protein–protein interactions, promoter sequences and transcription factors binding sites, has been used to derive a prior hypothesis on the gene network structure, which is then revised and updated by exploiting the gene expression data [62]. An interesting example is given by the work of Li and colleagues [63], who combined text mining to search associations in the PubMed literature with linear regression modeling to verify the proposed associations with gene expression data. A validation of the network extracted on the angiogenesis process, performed by comparing the results with the pathway information contained in the KEGG database, showed that literature mining can be greatly improved by gene expression data and that, at the same time, the extraction of a gene expression network may be successfully guided by an automated literature search. A second interesting approach is to integrate different gene expression data sets by summarizing the relations within a network that is consistent with the considered data sets. An example of this approach is an implementation by Wang and colleagues [64], who integrated different gene expression data sets, relying on differential equation models to describe the process dynamics and on linear programming to find the network structure. Since the majority of published work on data and knowledge integration for deriving regulatory networks exploits probabilistic modeling, below we review several issues of learning BNs and DBNs by combining background knowledge and data.

The algorithms to infer BNs and DBNs from data must typically infer their structure, that is, infer the relations between genes in the network. From a Bayesian viewpoint, this means identifying the structure or graph G with the highest posterior probability distribution given a data set X. The resulting scoring metric is:

$$p(G|X) = \frac{p(X|G)p(G)}{P(X)} \propto p(X|G)p(G)$$

The metric depends on two terms: $p(X|G)$ is the marginal likelihood, which expresses how likely the model is with respect to the available data; and $p(G)$ is the prior probability of the model. The marginal likelihood is computed by

averaging the likelihood over the possible values that the parameter set θ of the conditional probability distributions of a structure G may assume, so that:

$$p(X|G) = \int_\theta p(X, \theta|G)\,\mathrm{d}\theta = \int_\theta p(X|G, \theta)P(\theta|G)\,\mathrm{d}\theta$$

such an integral can be solved in close form when the variables are discrete [65] and when the model is conditionally Gaussian [66]. Since the search in the directed acyclic graph space—assumed for the structure of the models—is super-exponential, several heuristic algorithms have been proposed in the literature. The best known approach from this field is the original K2 search proposed by Cooper and Herskovits [65]. Other approaches include genetic algorithms [67] and Monte Carlo Markov Chain techniques [68].

The BN framework allows one to introduce background knowledge in several different ways. Imoto et al. [69] modulated the prior probability of each model $p(G)$ as a function of the available background knowledge. The prior knowledge is modeled with a Gibbs distribution

$$p(G) = Z^{-1}\exp\{-\lambda E(G)\}$$

where $E(G)$ is the energy of the network, Z is a normalizing constant and λ is a suitable hyperparameter. Thanks to the locality property of the BNs, $E(G)$ is decomposed in order to take into account the force of prior evidence for an arc, expressed in terms of the energy of the arc.

$$E(G) = \sum_{i=1}^{n} E_i = \sum_{i=1}^{n} \sum_{j \in pa(x_i)} E_{ij}$$

where E_{ij} is the energy of the arc going to the ith gene from its jth parent. Through a set of elegant algebraic transformations, the problem of specifying prior knowledge is then transformed into the specifications of a number of hyperparameters. The search process is then guided by the prior knowledge using the usual Bayesian scoring metric. The method was tested on simulated data and yeast data. A similar approach has been applied by [70] to infer gene networks by combining gene expression data and the structure of a gene promoter region.

Later on, the same group [71] proposed an approach for the joint learning of gene-regulatory networks and protein–protein interaction networks in terms of Bayesian and Markov (undirected) networks (Fig. 6). Given a set of data coming from DNA microarrays (X) and a set of data of interaction networks (Y), they computed the posterior probability of the gene-regulatory network (G_r) and of the protein interaction network (G_p) as:

$$p(G_r, G_p|X, Y) \propto P(X|G_r)P(X|G_p)P(G_r|G_p)P(G_p)$$

They tested their approach on a mutant expression data set [47] and protein–protein interaction data [72] and additionally relied on the background knowledge contained in the MIPS functional category database. The approach has been evaluated against an external reference knowledge source (KEGG). The results showed an increase in accuracy in recovering protein–protein interactions of about 10% with respect to learning without gene expression information, and an improvement in recovering correct regulatory interactions with respect to learning without protein–protein interaction information.

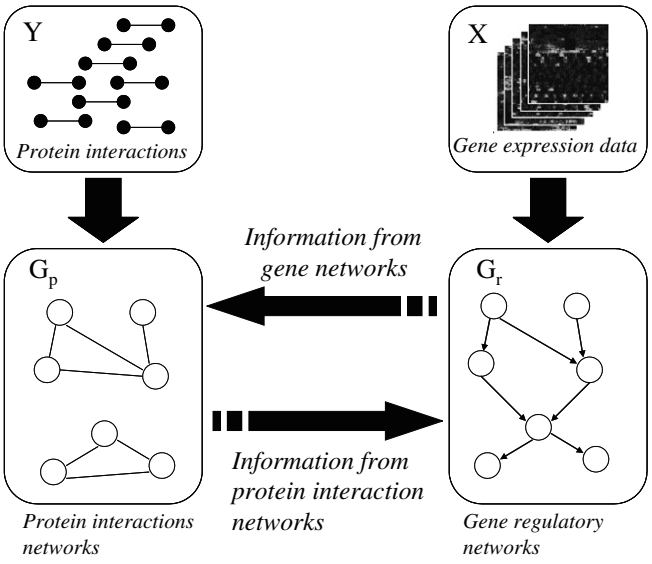

Fig. 6. The conceptual view of the approach proposed by Nariai et al. [71]. Gene regulatory networks and protein interaction networks are jointly built by borrowing strengths from both kinds of available data. The figure is adapted from [71].

Le and colleagues [68] described the effect of using background knowledge on the learning of a BN from a set of simulated data describing glucose homeostasis. The value of the inclusion of prior knowledge was evaluated by clamping a number of arcs in the gene-regulation network to the correct ones as obtained from background knowledge, and by looking at the sensitivity, *i.e.* the proportion of true edges found over the total number of true edges. The results showed that even a relatively small proportion of clamped edges may improve the sensitivity of the algorithm and, at the same time, reduce the number of expression profiles required for learning.

A different approach was presented by Bernard and Hartemink [73], who derived a model for defining the prior probability $p(G)$ of a DBN from a transcription factor database. In particular, they derived the probability of an arc connecting two genes in the network by analyzing the data on transcription binding sites. By assuming that the evidence that a transcription factor regulates a gene is expressed through a p-value, they computed the prior probability as the composition of local models, with each one expressing the probability of an edge being present given its p-value. Since the scoring metric can be decomposed into local models, the search procedure is easily modified taking into account different priors for each different gene model. The effectiveness of the algorithm has been shown on simulated data on the gene-regulatory network of the yeast cell cycle, producing similar findings as Le Phillip et al. [68].

6. Issues of model accuracy and evaluation

The outcome of data analysis, being descriptive or predictive in nature, is a model of the elements and their interactions being considered in the analysis. The model therefore represents a hypothesis (rather, a set of hypotheses) that were inferred from the data and background knowledge. These hypotheses were inferred from experimental data that may often include a substantial component of noise. How accurate are they? Are all components of the model equally well founded in the data or can some be more trusted than others? Did the method avoid overfitting the data, or are the results a by-product of randomness?

The components of inferred models can be tested computationally, verified through a literature search, and/or eventually tested with *in vitro* and *in vivo* experiments.

Verification through tracing the findings in published literature may increase our confidence in the resulting model and, if executed on a global model scale, identify a set of relations that may be either new or false. The true benefit of knowledge-based data analysis stems from automation and the incorporation of this phase within the analysis procedure. Compared to an often infeasible and unsystematic subsequent manual search, a computer-based comparison with knowledge bases may reveal which data-based findings are consistent with present knowledge, inconsistent with present knowledge pointing to either faults in the data or a need for a revision of current theories, or findings where our inferred hypothesis is new and requires further experimental confirmation. Experimental tests in laboratory (*in vivo* or *in vitro*) represent the crucial benchmark for hypothesis verification in molecular biology since they allow to study the biological system under controlled conditions. Such tests are usually performed on model organisms. The availability of new experimental procedures, such as RNA-interference, allows researchers to experimentally verify interactions and regulation mechanisms which are hypothesized with computational procedures.

Experimental testing in the laboratory is, of course, the most expensive validation and should only be used for the most promising or most interesting hypotheses. When models include larger sets of diverse hypotheses, procedures to rank these in terms of their computationally-estimated accuracy and the cost-benefit of their laboratory validation may be of benefit.

This leads us back to computational assessments of hypotheses. Standard statistical modeling techniques, often assuming that data were randomly drawn from some known distribution, can provide the means to estimate confidence intervals of predictions and components of the models (e.g., model coefficients). However, other techniques, mostly stemming from engineering and artificial intelligence, do not provide such tools. There, the estimation of the success of the modeling technique is based on data sampling. For instance, a standard technique in machine learning and data mining called cross-validation [74] splits the data into, say, ten subsets of equal size, and then averages the model performance measure across ten different experiments, each time using a distinct subset for testing and remaining subsets for training the models. Other popular alternatives to cross validation include bootstrap and leave-one-out [74]. Note that such procedures only assess the performance of the modeling method, giving us an indirect indication of the reliability of the target model that is developed from the complete data set. The computational evaluation of the target model and its components requires a separate validation set that is often expensive to obtain, but which is required for a systematic performance analysis. Computational validation has become a standard practice in data mining. Statistics, machine learning and data mining communities are also converging on testing procedures and measures to assess the predictive performance [34] and have become standard equipment of all major data analysis software packages.

Ideally, all three validation procedures described above should be used. The modeling methods should be tested through cross-validation on the training data, choosing the best-performing approach for development of the final model to be then validated on an external data set and whose findings should be compared with present knowledge. The scientifically most promising hypothesis would then need to undergo laboratory testing. The current practice in bioinformatics most often includes a range of evaluation procedures, with reports published in computer-science or statistical journals often only resorting to an assessment through cross-validation, and reports published in biomedical journals most often relying on experimental laboratory testing or testing through an independent, external data set. Scientifically, the value of modeling techniques is only measured through the hypothesis tested in the laboratory, while for the methodological development and comparison of methods cross-validation can suffice.

Models in systems biology often encompass a large set and variety of relations between the components (genes, sequence motif, proteins etc.) of observation. There is a risk of considering only the most promising part of the model, assessing the quality of the approach and resulting model through subsequent tests of this biased sample. Despite showing the utility of the approach to explorative data analysis, generalizing such findings to the entire model is wrong and misleading. The performance of system-based, data mining approaches can only be assessed through systematic testing and validation. The community has well recognized a major importance of the field, and has recently initiated a number of projects and activities to specifically address these issues. Examples of these are a DREAM (Dialogue on Reverse Engineering Assessment Methods, http://www.nyas.org/ebriefreps/main.asp?intEBriefID=534) initiative with the goal to provide reference material, gold standards and metrics for the evaluation of reverse engineering methods, competitions like CoEPrA (Comparative Evaluation of Prediction Algorithms,

http://www.coepra.org/), and dedicated conferences like CAMDA (Conference on Critical Assessment of Microarray Data Analysis).

7. Conclusion

The obvious trend in gene expression data analysis, as also emphasized in this review, is a departure from the utility of standard, off-the-shelf data analysis toolboxes to specialized approaches that can, besides the target data set, include additional information from other available knowledge and data sources. This transition is accelerated with community efforts to standardize data and knowledge storage formats and access protocols, and efforts to make the corresponding bases publicly available, most often through web-based access. The level of standardization and homogenization of data and knowledge bases is far from the point where we can speak about uniform bioinformatics and computational biology platforms, and with the present rate of the development of new approaches and paradigms in biotechnology this is unlikely to take place soon. Yet, the environment is ripe for research-based implementations of integrative tools that support knowledge-based data mining which, as reviewed in this paper, is a fast growing field. As recently proposed in the Science 2020 report [75], it is this integration that can provide the cornerstone of research in the coming years. Developments in this area will require even closer interdisciplinary collaboration and will lead not only to the integration of data and knowledge, but also to computer-supported experimental and knowledge generation platforms, thereby closing the loop of data gathering, hypothesis generation and experiment-based testing [76,77].

Acknowledgments

The authors would like to acknowledge the help given by the International Medical Informatics Association and its Working Group on Intelligent Data Analysis and Data Mining, which they are chairing. Authors would like to thank Uros Petrovic, Tomaz Curk, Fulvia Ferrazzi and Lucia Sacchi for their help in preparation of graphical material. The work was supported by a Slovenian-Italian Bilateral Collaboration Project. RB is also supported by the Italian Ministry of University and Scientific Research through the PRIN Project 'Dynamic modeling of gene and protein expression profiles: clustering techniques and regulatory networks', and BZ by the Slovenian Research Agency's Program and Project Grants.

References

[1] Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet 2006;7:55–65.
[2] Riva A, Carpentier AS, Torresani B, Henaut A. Comments on selected fundamental aspects of microarray analysis. Comput Biol Chem 2005;29:319–36.
[3] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998;95:14863–8.
[4] Datta S. Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics 2003;19:459–66.
[5] Hand DJ, Heard NA. Finding groups in gene expression data. J Biomed Biotechnol 2005;2005:215–25.
[6] Andersson CR, Isaksson A, Gustafsson MG. Bayesian detection of periodic mRNA time profiles without use of training examples. BMC Bioinformatics 2006;7:63.
[7] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 2000;25:25–9.
[8] Mewes HW, Amid C, Arnold R, Frishman D, Guldener U, Mannhaupt G, et al. MIPS: analysis and annotation of proteins from whole genomes. Nucleic Acids Res 2004;32:D41–4.
[9] Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). Proc Natl Acad Sci USA 2003;100:8348–53.
[10] Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2001.
[11] Curk T, Demsar J, Xu Q, Leban G, Petrovic U, Bratko I, et al. Microarray data mining with visual programming. Bioinformatics 2005;21:396–8.
[12] Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA 1999;96:2907–12.
[13] Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. Science 2003;302:249–55.
[14] Bar-Joseph Z, Gifford DK, Jaakkola TS. Fast optimal leaf ordering for hierarchical clustering. Bioinformatics 2001;17(Suppl 1):S22–9.
[15] Batagelj V, Mrvar A. Pajek—analysis and visualization of large networks. In: Jünger M, Mutzel P, editors. Graph drawing software. Berlin: Springer; 2003. p. 77–103.
[16] Bolshakova N, Azuaje F, Cunningham P. A knowledge-driven approach to cluster validity assessment. Bioinformatics 2005;21:2546–7.
[17] Kong SW, Pu WT, Park PJ. A multivariate approach for integrating genome-wide expression data and biological knowledge. Bioinformatics 2006;22(19):2373–80.
[18] Khatri P, Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics 2005;21:3587–95.
[19] Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. Bioinformatics 2006.
[20] Lin D. An information-theoretic definition of similarity. In: Proc. 15th international conference on machine learning. Madison, WI, USA: Morgan Kaufmann; 1998. p. 296–304.
[21] Resnik P. Using information content to evaluate semantic similarity in a taxonomy. In: Proc. 14th international conference on machine learning. Morgan Kaufmann; 1995. p. 444–53.
[22] Jiang JJ, Conrath DW. Semantic similarity based on corpus statistics and lexical taxonomy. In: International conference on research in computational linguistics. Tapei, Taiwan: Academia Sinica; 1997.
[23] Guo X, Liu R, Shriver CD, Hu H, Liebman MN. Assessing semantic similarity measures for the characterization of human regulatory pathways. Bioinformatics 2006;22:967–73.
[24] Kustra R, Zagdanski A. Incorporating Gene Ontology in Clustering Gene Expression Data. In: 19th IEEE symposium on computer-based medical systems. IEEE Computer Society; 2006. p.555–63.
[25] Huang D, Pan W. Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data. Bioinformatics 2006;22:1259–68.

[26] Qin ZS. Clustering microarray gene expression data using weighted Chinese restaurant process. Bioinformatics 2006;22:1988–97.
[27] Raftery AE. Bayesian model selection in social research [with discussion]. In: Marsden PV, editor. Sociological methodology. Cambridge, MA: Blackwell; 1995. p. 111–95.
[28] Pan W. Incorporating gene functions as priors in model-based clustering of microarray gene expression data. Bioinformatics 2006;22:795–801.
[29] Shahar Y. A framework for knowledge-based temporal abstraction. Artif Intell 1997;90:79–133.
[30] Sacchi L, Bellazzi R, Larizza C, Magni P, Curk T, Petrovic U, et al. TA-clustering: cluster analysis of gene expression profiles through temporal abstractions. Int J Med Inform 2005;74:505–17.
[31] Hvidsten TR, Laegreid A, Komorowski J. Learning rule-based models of biological process from gene expression time profiles using Gene Ontology. Bioinformatics 2003;19:1116–23.
[32] Cho RJ, Huang M, Campbell MJ, Dong H, Steinmetz L, Sapinoso L, et al. Transcriptional regulation and function during the human cell cycle. Nat Genet 2001;27:48–54.
[33] Mitchell TM. Machine learning. New York: McGraw-Hill; 1997.
[34] Hand DJ, Mannila H, Smyth P. Principles of data mining. Cambridge, Mass.: MIT Press; 2001.
[35] Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci USA 2000;97:262–7.
[36] Moskovitch R, Cohen-Kashi S, Dror U, Levy I, Maimon A, Shahar Y. Multiple hierarchical classification of free-text clinical guidelines. Artif Intell Med 2006;37:177–90.
[37] Rousu J, Saunders C, Szedmak S, Shawe-Taylor J. Learning hierarchical multi-category text classification models. In: De Raedt L, Wrobel S, editors. 22nd international conference on machine learning. ACM Press; 2005. p. 744–51.
[38] Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, et al. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res 2004;32:5539–45.
[39] Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 2001;7:673–9.
[40] Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 2003;95:14–8.
[41] Lai C, Reinders MJ, van't Veer LJ, Wessels LF. A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. BMC Bioinformatics 2006;7:235.
[42] Jeffery IB, Higgins DG, Culhane AC. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. BMC Bioinformatics 2006;7:359.
[43] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 2005;102:15545–50.
[44] Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 2005;21:631–43.
[45] Gamberger D, Lavrac N, Zelezny F, Tolar J. Induction of comprehensible models for gene expression datasets by subgroup discovery methodology. J Biomed Inform 2004;37:269–84.
[46] Mramor M, Leban G, Demsar J, Zupan B. Conquering the curse of dimensionality in gene expression cancer diagnosis: tough problem, simple models. In Proc. of artificial intelligence in medicine (AIM-2005). Aberdeen, UK; 2005. p. 514–23.
[47] Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, et al. Functional discovery via a compendium of expression profiles. Cell 2000;102:109–26.
[48] Cheng Y, Church GM. Biclustering of expression data. Proc Int Conf Intell Syst Mol Biol 2000;8:93–103.
[49] Prelic A, Bleuler S, Zimmermann P, Wille A, Buhlmann P, Gruissem W, et al. A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics 2006;22:1122–9.
[50] Van Driessche N, Demsar J, Booth EO, Hill P, Juvan P, Zupan B, et al. Epistasis analysis with global transcriptional phenotypes. Nat Genet 2005;37:471–7.
[51] Hughes TR. Universal epistasis analysis. Nat Genet 2005;37:457–8.
[52] van de Peppel J, Kettelarij N, van Bakel H, Kockelkorn TT, van Leenen D, Holstege FC. Mediator expression profiling epistasis reveals a signal transduction pathway with antagonistic submodules and highly specific downstream targets. Mol Cell 2005;19:511–22.
[53] Clayton J. RNA interference: the silent treatment. Nature 2004;431:599–605.
[54] Cheng JC, Moore TB, Sakamoto KM. RNA interference and human disease. Mol Genet Metab 2003;80:121–8.
[55] Schlitt T, Brazma A. Modelling gene networks at different organisational levels. FEBS Lett 2005;579:1859–66.
[56] de Jong H. Modeling and simulation of genetic regulatory systems: a literature review. J Comput Biol 2002;9:67–103.
[57] Bar-Joseph Z. Analyzing time series gene expression data. Bioinformatics 2004;20:2493–503.
[58] Liang S, Fuhrman S, Somogyi R. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. Pac Symp Biocomput 1998:18–29.
[59] D'Haeseleer P, Liang S, Somogyi R. Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics 2000;16:707–26.
[60] Friedman N. Inferring cellular networks using probabilistic graphical models. Science 2004;303:799–805.
[61] Sebastiani P, Abad M, Ramoni MF. Bayesian networks for genomic analysis. In: Dougherty ER, Shmulevich I, Chen J, Wang ZJ, editors. EURASIP book series on signal processing and communications: genomic signal processing and statistics. New York, NY: Hindawi; 2005. p. 281–320.
[62] Xing B, van der Laan MJ. A causal inference approach for constructing transcriptional regulatory networks. Bioinformatics 2005;21:4007–13.
[63] Li S, Wu L, Zhang Z. Constructing biological networks through combined literature mining and microarray analysis: a LMMA approach. Bioinformatics 2006;22:2143–50.
[64] Wang Y, Joshi T, Zhang XS, Xu D, Chen L. Inferring gene regulatory networks from multiple microarray datasets. Bioinformatics 2006.
[65] Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Mach Learn 1992;9:309–47.
[66] Geiger D, Hackerman D. Learning Gaussian networks. In: de Mantaras RL, Poole D, editors. Tenth conference on uncertainty in artificial intelligence. San Francisco, CA: Morgan Kaufmann; 1994. p. 235–43.
[67] Larrañaga P, Sierra B, Gallego MY, Michelena MJ, Picaza JM. Learning Bayesian networks by genetic algorithms: a case study in the prediction of survival in malignant skin melanoma. In Keravnou E, Garbay C, Baud R, Wyatt CJ, editor, Artificial intelligence in medicine Europe. Grenoble, France; 1997. p. 261–72.
[68] Le Phillip P, Bahl A, Ungar LH. Using prior knowledge to improve genetic network reconstruction from microarray data. In Silico Biol 2004;4:335–53.
[69] Imoto S, Higuchi T, Goto T, Tashiro K, Kuhara S, Miyano S. Combining microarrays and biological knowledge for estimating gene networks via bayesian networks. J Bioinform Comput Biol 2004;2:77–98.
[70] Tamada Y, Kim S, Bannai H, Imoto S, Tashiro K, Kuhara S, et al. Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. Bioinformatics 2003;19(Suppl. 2):II227–=0?>II236.

[71] Nariai N, Tamada Y, Imoto S, Miyano S. Estimating gene regulatory networks and protein–protein interactions of Saccharomyces cerevisiae from multiple genome-wide data. Bioinformatics 2005;21(Suppl. 2):ii206–12.
[72] Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 2002;415:141–7.
[73] Bernard A, Hartemink AJ. Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data. Pac Symp Biocomput 2005:459–70.
[74] Witten IH, Frank E. Data mining: practical machine learning tools and techniques with Java implementations. 2nd ed. San Francisco, CA: Morgan Kaufmann; 2005.
[75] Towards 2020 Science. Available at http://research.microsoft.com/towards2020science.
[76] King RD, Whelan KE, Jones FM, Reiser PG, Bryant CH, Muggleton SH, et al. Functional genomic hypothesis generation and experimentation by a robot scientist. Nature 2004;427:247–52.
[77] Zupan B, Holmes JH, Bellazzi R. Knowledge-based data analysis and interpretation. Artif Intell Med 2006;37:163–5.