Vol. 18 Suppl. 1 2002
Pages S145–S154

Co-clustering of biological networks and gene expression data

Daniel Hanisch[1], Alexander Zien[1], Ralf Zimmer[2] and Thomas Lengauer[3]

[1]Institute for Algorithms and Scientific Computing (SCAI), Fraunhofer Gesellschaft, Schloss Birlinghoven, Sankt Augustin, 53754, Germany, [2]Institut für Informatik, Ludwig-Maximilians-Universität München, Theresienstraße 39, München, 80333, Germany and [3]Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, Saarbrücken, 66123, Germany

Received on January 24, 2002; revised and accepted on March 28, 2002

ABSTRACT

Motivation: Large scale gene expression data are often analysed by clustering genes based on gene expression data alone, though *a priori* knowledge in the form of biological networks is available. The use of this additional information promises to improve exploratory analysis considerably.

Results: We propose constructing a distance function which combines information from expression data and biological networks. Based on this function, we compute a joint clustering of genes and vertices of the network. This general approach is elaborated for metabolic networks. We define a graph distance function on such networks and combine it with a correlation-based distance function for gene expression measurements. A hierarchical clustering and an associated statistical measure is computed to arrive at a reasonable number of clusters. Our method is validated using expression data of the yeast diauxic shift. The resulting clusters are easily interpretable in terms of the biochemical network and the gene expression data and suggest that our method is able to automatically identify processes that are relevant under the measured conditions.

Contact: Daniel.Hanisch@scai.fhg.de

Keywords: gene expression; biological networks; metabolic networks; co-clustering; clustering.

INTRODUCTION

One of the most popular tools for exploratory analysis of gene expression data is clustering of genes and/or experiments. Clustering methods try to group the objects under consideration into (usually disjoint) groups sharing a common characteristic reflected by a similarity measure defined on the objects. Since the influential paper of Eisen *et al.* (1998), clustering methods are routinely used to cluster genes based on a distance measure quantifying the degree of co-regulation.

Furthermore, clustering is also frequently used as the basis for further computational analysis. For example, the function of a gene can be predicted according to known functions of other genes from the same cluster (Eisen *et al.*, 1998). If the arrays are clustered instead of the genes, classes and subclasses of samples can be detected and predicted. By searching the promoter sequences of all genes of a cluster for overrepresented patterns, new transcription factor binding sites can be identified. Gene expression clusters can also be mapped onto metabolic networks in order to find pathways of interest (van Helden *et al.*, 2000).

It appears that this sequential evaluation of the data—by first clustering the gene expression data alone, and incorporating additional information only after the clusters are determined—is suboptimal. One major problem is that the boundaries of the resulting clusters are arbitrary to some degree. Information from other sources could often help in resolving ambiguities or in avoiding erroneous linking based on spurious similarities.

In this paper, we propose a novel method that utilizes information in the form of biological networks in an integrated manner to improve the result of the clustering. Biological networks relate genes, gene products or groups of those (e.g., protein complexes or protein families) to each other in the form of a graph (Bhalla and Iyengar, 1999). In such a graph, nodes represent molecules and edges indicate our knowledge of existing or absent relationships. In general, the edges of biological networks may represent proven facts, but also uncertain information or hypotheses.

There have been previous suggestions for the integrated non-sequential analysis of gene expression data and biological networks. For example, the correct topology of small regulatory networks can be inferred by Bayesian reasoning (Hartemink *et al.*, 2001). Our pathway scoring method (Zien *et al.*, 2000) scores metabolic pathways in terms of gene expression data, but suffers from the combinatorial explosion encountered during pathway generation. Apart from designing improved scoring functions, Kurhekar *et al.* (2002) suggest selecting relevant pathways from a predefined collection, like those given by KEGG (Kanehisa, 1996). While this solves the computational problems, it restricts the analysis to a static view of the biological network as imposed by the pathway collection.

The co-clustering method proposed here, however, is able to extract relevant pathways that cross the boundaries of such categories. Whereas conventional cluster methods rely on gene expression data alone, we propose combining measures derived from gene expression data *and* metrics on biological networks into a single distance function. Thus, if for a set of genes a distance function, δ_{exp}, is defined based on the expression data and another distance function, δ_{net}, is defined based on a biological network, we propose to compute a combined distance function Δ as

$$\Delta = f(\delta_{\text{exp}}, \delta_{\text{net}}).$$

Through the combination of both distance measures, additional structure specified in the biological network can be imposed on the gene expression data. This in turn may lead to increased stability of the clustering solution, if both kinds of information are coherent, i.e., gene expression measurements support relations of the network and *vice versa*. More importantly, co-clustering should lead to biologically more meaningful clusters in many cases. For example, the question concerning which metabolic pathways are activated in the course of an experiment can be examined through gene expression measurements (DeRisi *et al.*, 1997). In this case, reasonably coherent expression of genes participating in an activated pathway can be expected. However, searching for the best correlation in the set of measured genes (as attempted in standard clustering procedures) might not lead to biologically sound results even in the absence of measurement errors.

In the following section we will discuss the application of co-clustering to metabolic networks. This entails representation of the metabolic data, definition of sensible distance functions on expression data and network separately, followed by the definition of the combined distance function. After presentation of the clustering methodology, the performance of our method on a real-world data set is discussed. The paper closes with perspectives for other possible applications of our method and suggestions for further improvement.

APPLICATION TO METABOLIC NETWORKS

The focus of this paper is on the analysis of *co-regulated metabolic pathways* supported by gene expression measurements. Metabolic reactions are an integral part of every organism and comprise fundamental cellular processes such as protein synthesis or energy production. This setting is well suited to assessing the feasibility of our approach as sets of metabolic reactions and associated pathways (e.g., KEGG) and gene expression data known to represent metabolic changes (e.g., DeRisi *et al.* (1997)) are readily available and reasonably well understood.

Metabolic network acquisition and representation

In our setting metabolic networks consist of a set of chemical reactions, most of which are catalysed by enzymes. Enzymatic reactions which can be assembled into a metabolic network are available in several databases. We choose the KEGG database (Kanehisa, 1996) as it provides an organization of reactions into pathways as well as a large dataset of curated metabolic reactions. The pathways reflect commonly used categories and can be visualized on manually drawn pathway maps. Since our co-clustering procedure does not utilize this information, it is well suited as an independent evaluation guide.

As of December 2001 the KEGG dataset consisted of approximately 4000 metabolic reactions. Each reaction is realized in some organism and is annotated with a reaction identifier, functionally important educts and products, and the classification of the catalysing enzyme. Each reaction may be uni- or bi-directional. We take the combination of identifier and EC classifier to be a unique identifier for a metabolic reaction in the network. We assemble this set of reactions into a network in the form of a PETRI net, similar to the method described by Küffner *et al.* (1999). The constructed PETRI net is essentially a bipartite graph in which one set of nodes (termed places) represents molecules (metabolites and enzymes) and the other set of nodes (called transitions) defines chemical reactions among the molecules. A directed edge between a molecule and a reaction implies that the molecule is product or educt of the reaction depending on the direction of the edge. Note that a catalysing enzyme is educt and product simultaneously, i.e., two edges connect it to an associated transition. Enzyme identifier can occur multiple times as each may be able to catalyse multiple reactions. A unification of these nodes would introduce unwanted shortcuts into the metabolic network which may have no biological meaning. In fact, these nodes may be assigned to different clusters in our co-clustering procedure.

Our co-clustering method should be able to extract biologically plausible subnetworks in the light of the given gene expression data. This should partly correspond, but not be restricted, to defined KEGG pathways. Moreover, we construct a network which is *not* specific to any single organism. Indeed, it would be possible to construct specialized networks for certain organisms based on the KEGG data. Such a network would consist of fewer reactions than the generic one as enzymes might not be present or unknown in the chosen organism. Therefore, it might exhibit a less connected and more cluster-like structure. In this work, however, we report detailed results on only the generic network. Thereby, we emulate the situation in organisms with little prior knowledge. In such organisms known reactions may be connected by generic ones to yield a hypothetical, but plausible, network.

Network distance function

The underlying assumption for our network distance function is that enzymes are related according to their proximity in the network. This is reasonable, as the biological processes under consideration consist of successive metabolic reaction steps which constitute our network.

In order to construct a distance function for the metabolic network δ_{net}, we interpret the derived PETRI net as an undirected graph $G = (V, E)$ with node set V and edge set E. In our case, V is the set of all places and transitions of the PETRI net. Thus, V consists of all molecules, i.e., proteins, metabolites, and reactions of the metabolic network. Furthermore, we define a weight function $w : E \to \mathbb{R}$ which associates a weight with each edge. Let $W \subset V$ be the set of molecules for which gene expression data are available. Then we define $\delta_{\text{net}} : W \times W \to \mathbb{R}$ for two vertices $w_i, w_j \in W$ as the minimum weight of all paths connecting vertices w_i and w_j.

These minimal weights can be computed efficiently using basic graph algorithms. We first eliminate all $|V| - |W|$ superfluous vertices by considering for each such vertex x all pairs of neighbours. Let (v_i, v_j) be such a pair of neighbouring vertices. If the weight $w(p_x)$ of the path $p_x = (v_i, x, v_j)$ is smaller than the weight of the path $p_v = (v_i, v_j)$, we connect v_i and v_j with an edge with weight $w(p_x)$. If all pairs of neighbours have been considered, the node x and all incident edges can be deleted. For the remaining $|W|$ nodes of interest we use the Floyd–Warshall algorithm (Cormen *et al.*, 1992) to compute all shortest paths. For dense graphs, the first step can be time-consuming because $|V|^2$ pairs of vertices have to be considered in the worst case, thus resulting in a worst case running time of $O(|V - W||V|^2)$. In our sparse network, however, this worst case running time is virtually never observed. When nodes are processed in order of increasing degree, it will often suffice to consider only a small fraction of V as most nodes have few incident edges. The subsequent Floyd–Warshall algorithm takes time $O(|W|^3)$.

Another option computing the minimal weights is the Dijkstra shortest-path algorithm (Cormen *et al.*, 1992). This algorithm can compute the shortest path weights of one vertex to all other vertices in asymptotic running time $O(|V|log|V| + |E|)$ for non-negative weight functions. As we need to compute the weights for all vertices in the set of interest W, the overall running time to compute the network distance function is $O(|W||V|log|V| + |W||E|)$. We found that the first strategy is much faster for the networks under consideration. Reasons may be that the worst-case bound of the first strategy is not sharp for our class of graphs and the constant hidden in the O-notation is small.

As the number of interesting vertices is usually smaller than the number of all vertices and because the network under consideration is sparse, computation of this distance function is feasible for even large networks. Furthermore, in our procedure graph distances need to be computed only once for each network.

The simplest plausible choice for the weight function is the uniform weighting $w(e) = 1$ for all $e \in E$. The resulting distance function is termed δ_{net}^{uni}.

Recently it has been shown that metabolic networks exhibit a scale-free structure (Jeong *et al.*, 2000; Fell and Wagner, 2000). Indeed, this is true for the network constructed from KEGG reactions. One characteristic of networks with a scale-free structure is that relatively few high-degree nodes are found. These so-called hubs often constitute unspecific or ubiquitous molecules (e.g., ATP or NH_3). Thus, the connections introduced by such nodes may be biologically unimportant or even misleading. In other words, we prefer clusters in which all genes are reachable from each other in a few steps without hubs. This can be reflected in the distance function using the degree of a vertex v, i.e., the number of edges incident to v, termed $deg(v)$. Our alternative weighting function for an edge e between two vertices v and w is defined as

$$w(e) = \begin{cases} deg(v) & \text{if } v \text{ is a molecule} \\ deg(w) & \text{otherwise.} \end{cases}$$

Note that the graph is bipartite, i.e., every edge connects one molecule with one reaction. We do not use the sum $deg(v) + deg(w)$ as we do not want to penalize reactions with many involved molecules, but only substrates participating in different biological processes. The resulting distance function is termed $\delta_{\text{net}}^{\text{norm}}$. Histograms of the distance distributions for the metabolic network according to both functions are depicted in Figure 1.

Gene expression distance function

Several alternative distance functions for gene expression measurements have been proposed. The most popular choice in the case of time-series data is the Pearson correlation coefficient, as suggested by Eisen *et al.* (1998). This coefficient quantifies the degree of linear dependence between two time-courses of gene expression levels. In this paper, we used log-ratio transformed data, i.e., for each sample of interest, the logarithm of the ratio of the

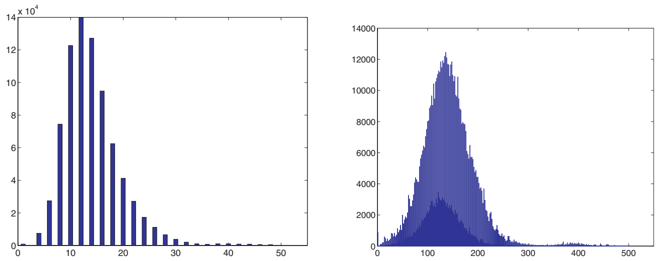

Fig. 1. On the left, histogram of network distances without degree normalization. On the right, histogram of network distances with degree normalization. For both data sets 0.9% of the distances were infinite. These are omitted in the plots.

sample and a control measurement is computed. If G_{ik} represents this value for gene g_i at time point k, then the correlation coefficient ρ is defined as

$$\rho(g_i, g_j) = \frac{1}{N} \sum_k \left(\frac{G_{ik} - \mu_i}{\sigma_i} \right) \left(\frac{G_{jk} - \mu_j}{\sigma_j} \right),$$

where μ_i and σ_i denote mean and standard deviation of the transformed time series data of gene i. The correlation coefficient can be converted easily into a distance measure in the range 0 to 2 by

$$\delta_{\text{exp}}(g_i, g_j) = 1 - \rho(g_i, g_j).$$

This distance function quantifies the degree of dissimilarity for our gene expression data set. Here we consider anti-correlated genes as most distant. The use of correlation as a distance function is reasonable in our setting as we expect a similar expression pattern between genes in successive (or related) reaction steps in metabolic pathways. However, we cannot expect to see perfect correlation of expression in a pathway for two reasons. First, gene expression measurements reflect the amount of mRNA in the sample and, thus, the amount of enzyme to be produced in the near future. Second, expression measurements are noisy with current high-throughput technology. Nevertheless, a coordinated change in the expression patterns of participating genes is to be expected when a metabolic pathway changes its activity level.

CO-CLUSTERING

In the following we define how to combine the distance functions for networks and gene expression data and how to compute the desired clustering. Note that a large part of the following discussion applies not only to metabolic networks but also to more general biological networks.

Combining nodes and genes

The network distance function δ_{net} operates on pairs of enzyme nodes in the graph, whereas the expression distance function δ_{exp} operates on pairwise expression measurements, i.e., genes of an organism. To construct a combined function, a mapping $\mathcal{M}$ that relates genes to enzyme nodes in the graph is required.

For yeast, such a mapping is available from the MIPS database (Mewes *et al.*, 1997). In this database, E.C.-classifications are assigned to all open reading frames (ORFs) with known metabolic function. This mapping is *not* one-to-one. Indeed, one ORF may have several enzymatic functions, and conversely several ORFs may map to one E.C. entry. In addition, each single EC number may correspond to several nodes in the network for reasons given above. To cope with this situation, the combined distance function is defined on the product set of genes and relevant nodes in the network. Members of this product set are termed *objects*. Thus, if G is the set of genes and V is the set of nodes in the network, then the mapping $\mathcal{M} \subset G \times V$ defines the domain of the combined distance function and, thus, the set of objects used in the clustering procedure.

To illustrate this, consider nodes with enzyme classification *Hexokinase* (EC number 2.7.1.1). There are two yeast proteins which are associated with this function, *HXK1* and *HXK2*. However, during diauxic shift, the expression of these two genes is strongly anti-correlated (DeRisi *et al.*, 1997). Conversely, the hexokinase belongs to the group of phosphotranferases. For instance, it can catalyse the conversion of α-D-glucose to α-D-glucose 6-phosphate (KEGG reaction ID R01786) as well as the conversion of D-glucosamine to D-glucosamine 6-phosphate (KEGG reaction ID R01961). While these reactions are similar, it is not clear *a priori* that they are embedded in the same functional context and should therefore share a cluster. Indeed, in terms of the KEGG pathways, these reactions occur in the glycolysis and aminosugars metabolism pathways, respectively. Consequently, our method constructs distinct objects for each gene/reaction pair. For example, the protein HXK1 would map to two distinct objects, i.e., $o_1 = (g_1, v_1) = (\text{HXK1}, \text{EC } 2.7.1.1/\text{R01786})$ and $o_2 = (g_1, v_2) = (\text{HXK1}, \text{EC } 2.7.1.1/\text{R01961})$. The protein HXK2 is treated analogously. The co-clustering results in a clustering of such objects which can be reduced to a clustering of vertices or genes, as necessary.

Combining distances

The combined distance function should assign a small distance to pairs of objects which are close in the network *and* show similar expression patterns. Objects which are far apart in the network and thus are presumably used in different biological context should be far apart according to the combined function. The same holds true for objects which are not co-regulated or even oppositely regulated, as we want to extract co-regulated pathways. The largest

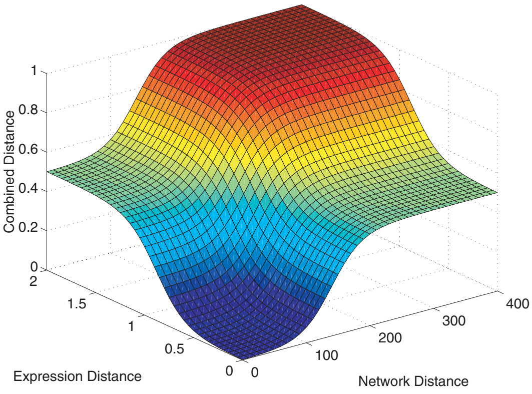

Fig. 2. Graph of combined distance function Δ^{norm}.

distance should be assigned to objects which are far apart according to *both* distance measures. In addition, the combined function should be robust against noise in the data. For example, perfect correlation should not result in extraordinarily small distance in comparison with a good or moderate correlation, as such differences might often be due to measurement noise. This in turn could lead to artifacts in the clustering procedure. Analogously, erroneously missing links between enzymes of interest in the network should not lead to prohibitively high distances. This robustness can be achieved by saturation at the extremes of the parameter space.

One function capable of combining the individual distance functions into a joint one with the above properties is the sum of two logistic curves. This sigmoidal function attains its maximum for minimal x- and y-values and gradually declines with increasing x- and y-values. As our distance function needs to assign minimal values for minimal parameter values, the functional form of the combined function Δ for two objects $o_i, o_j \in \mathcal{M}$ with corresponding genes g_i, g_j and corresponding enzyme nodes in the graph v_i, v_j is

$$\Delta(o_i, o_j) = 1 - 0.5 \times (\lambda_{\exp}(g_i, g_j) + \lambda_{\text{net}}(v_i, v_j)),$$

$$\text{where } \lambda_\Psi(x_i, x_j) = \frac{1}{1 + e^{-s_\Psi(\delta_\Psi(x_i, x_j) - \nu_\Psi)}}$$

for $\Psi \in \{\exp, \text{net}\}$. The parameters $\nu_\Psi, s_\Psi \in \mathbb{R}$ control the shape of the logistic curve. Essentially, a one-dimensional logistic curve is a smooth threshold function with value $1/2$ at point ν_Ψ. The parameter s_Ψ controls the slope of the curve. We set the parameter ν_Ψ to the mean of the distance distributions of the network and expression distances, respectively. The parameter s_Ψ is chosen heuristically to yield a moderate slope ($s_\Psi = 6/\nu_\Psi$). The resulting combined distance function is shown in Figure 2. A validation of this combination can be found in the Results section.

As pointed out above, genes and vertices may occur multiply in distinct objects. It needs to be considered whether pairs of these objects are assigned a sensible distance. The first case is that one vertex is assigned to several genes, i.e., more than one gene may fulfill the desired function or even all genes are needed to fulfill the enzymatic activity (e.g., as subunits in a complex). Then we leave the network distance of zero as implied by the distance function, δ_{net} unchanged. This means that these objects share a cluster if their expression profiles do not disagree strongly. In the second case, when one gene is mapped to several vertices in the network, we set the expression distance to the mean of all expression distances, i.e., to a normalized value of $1/2$. The rationale for this is that distinct vertices in the network catalyse different reactions, and thus a single gene may play a role in different biological contexts. A perfect correlation of the expression measurements, however, might prevent the assignment to different clusters.

Clustering methodology

From the large number of available cluster methods, we chose a version of *hierarchical average linkage clustering*. Since the influential work of Eisen *et al.* (1998), this is one of the most popular clustering techniques for the analysis of gene expression measurements. Starting from the set of objects as singleton clusters, the method successively joins clusters with smallest average pairwise distance. The result of a hierarchical clustering procedure is a binary tree (or a dendrogram) in which each inner node represents a joining step of the procedure. To produce a set of clusters, the tree is cut by removing all nodes after a chosen joining step.

One major problem arising in cluster analysis is determining the appropriate number of clusters and thus the cutting point for the dendrogram. Various statistical measures exist for this purpose. We selected the *silhouette-coefficient* (Rousseeuw, 1987). This coefficient measures the quality of a clustering by a comparison of the tightness and separation of clusters. Let i be any object in the clustering and A its corresponding cluster. Then

$$a(i) = \frac{1}{|A| - 1} \sum_{j \in A, j \neq i} \Delta(i, j)$$

measures the average distance of i to all other objects in the cluster A. Then we compute for each cluster $C \neq A$

$$d(i, C) = \frac{1}{|C|} \sum_{j \in C} \Delta(i, j)$$

to quantify the distance to other clusters. The minimum value,

$$b(i) = \min_{C \neq A} d(i, C),$$

gives the distance of i to the second-best cluster. The silhouette value $s(i)$ of i is then defined as

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}.$$

This value in the range $[-1; 1]$ quantifies how well object i fits into cluster A. If the $s(i)$ is smaller than zero, the nearest cluster would be a better choice. The *average silhouette value* is the average of the $s(i)$ over all objects in the clustering and is a measure for the quality of the clustering. It has to be noted that other measures may be important for the assessment of cluster quality as well. For example, we are interested in clusters of a certain size or compactness, which is not directly measured by the average silhouette coefficient. Nonetheless, it provides a good aid in choosing reasonable cutting points. We will plot this average silhouette value for different cut points of our dendrogram to find sensible clusterings.

RESULTS

To evaluate the performance of our method, we use the gene expression time series measurements conducted by DeRisi *et al.* (1997) for the organism *S. cerevisiae* (yeast). In this data set, measurements for seven different time points are taken. In this experiment, yeast is inoculated into a sugar-rich medium. When the sugar is progressively depleted, the metabolism of yeast switches from anaerobic growth to aerobic respiration. This so-called diauxic shift involves changes in several metabolic processes which should be detectable by our method. DeRisi and co-workers manually analysed several pathways related to the diauxic shift which can serve as a standard of truth for the validation of our method.

Of the 6101 yeast ORFs measured in this experiments, 642 have known metabolic functions, according to the MIPS database (Mewes *et al.*, 1997). 884 nodes in the metabolic network derived from the KEGG database correspond to these enzymatic functions. Due to the multiplicity of the mapping, we arrive at 1571 objects to be clustered.

In a first step, we compare the relative quality of the five defined distance functions, i.e., the gene expression distance function δ_{exp}, the normalized and non-normalized network distance functions $\delta_{\text{net}}^{\text{uni}}$ and $\delta_{\text{net}}^{\text{norm}}$, and the two combined distance functions Δ^{uni} and Δ^{norm}. DeRisi *et al.* (1997) found that the glycolysis pathway is influenced significantly by the diauxic shift. From this pathway we selected ORFs (and corresponding objects) which score highest in our pathway scoring method (Zien *et al.*, 2000), i.e., show high co-regulation and constitute a complete reaction chain (YGL253W, YBR196C, YMR205C, YKL060C, YJR009C, YDR050C, YCR012W, YHR174W, YAL038W). As a figure of merit for the quality of a distance function, we plot the fraction of distances among *all* objects against the fraction of pairwise distances of glycolysis objects found among these. In other words, we investigate which fraction of relationships within the glycolysis pathway is already considered when we inspect a certain fraction of all distances. The resulting graph is shown in Figure 3.

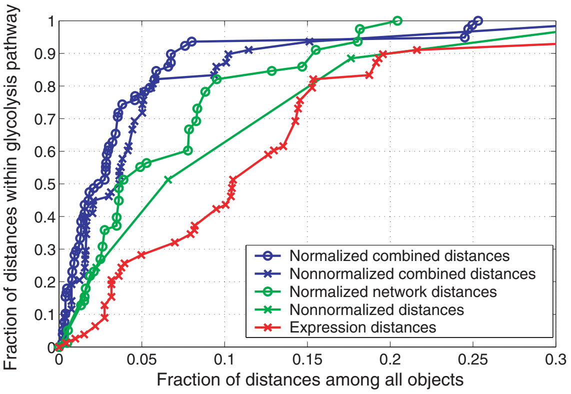

Fig. 3. Relative quality of defined distance functions. The higher the curve, the better the respective distance function discriminates related from unrelated objects.

Although each distance function is far better than random, the combined distance functions clearly perform best. For instance, to consider 90% of distances among glycolysis objects, we need to inspect only approximately 6% of the overall data, in contrast to 23% for the expression distance function. Note that the expression distance function δ_{exp} begins with a moderate slope indicating that many co-regulated objects are present. By utilizing network information, pairs of objects that are co-regulated but far apart in the network, are not considered at an early stage. The steep slope of the combined functions shows that this is successfully accomplished. In addition, the normalized functions perform better than their non-normalized counterparts indicating that the scale-free structure of metabolic networks can provide additional distance information compatible with the definition of the glycolysis pathway and probably with the notion of metabolic pathways in general.

In the following discussion, we focus on the normalized combined distance function. As already hinted at by the above evaluation, the normalized distance function leads to biologically more plausible clusters. In particular, linking of clusters via ubiquitous nodes, such as NH_3, CO_2 or ATP, does not occur. After construction of the dendrogram based on Δ^{norm}, we computed the average silhouette value for a range of possible cutting points (cf. Figure 4) to find a clustering reflecting the desired output, i.e., clusters corresponding to pathways with coordinated change in gene expression.

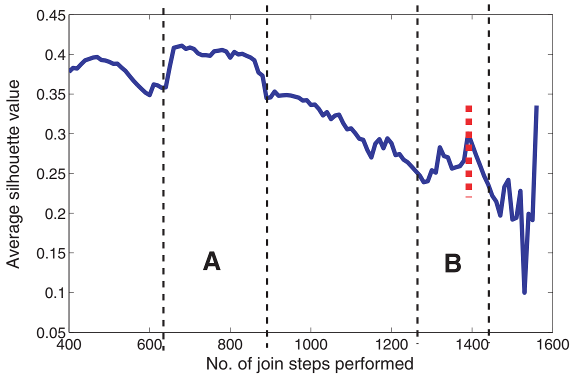

Fig. 4. Average silhouette values for clusterings resulting from different cut-off values. The normalized network distance function is used. Cutting points are marked with dashed lines.

Part *A* of Figure 4 has highest average silhouette values. Cutting points in this interval result in clusters of objects with multiple representations, i.e., objects sharing the same E.C. number and similar expression patterns or one gene mapped to several closely related nodes in the network. For instance, most subunits of V-type and F-type ATPases are co-regulated and cluster together during this phase. Interestingly, these types of ATPases exhibit different regulation patterns. Whereas F-type ATPase, also known as ATP synthase, is upregulated during diauxic shift, V-type ATPase is downregulated. ATPases are proton transporters using ATP in the process. In contrast to V-type ATPase, F-type ATPase (located in the inner membrane of mitochondria) is usually driven in reverse by chemiosmosis to produce ATP. This becomes increasingly necessary as the ATP supply through the glycolysis chain ceases. Other examples include the formation of cytochrome c oxidase which is upregulated to enhance the capability of yeast to produce ATP in the respiratory chain, and the upregulated succinate dehydrogenase which plays a role in the TCA cycle. The average size of clusters during this period, however, is small (2 objects per cluster). Since the focus of this study is the identification of co-regulated pathways, i.e. clusters exceeding a certain size, we inspect further local maxima of the silhouette coefficient.

For very late cutting points we find an artificial increase of the average silhouette value. At this point, huge clusters have formed which do not exhibit either tight correlation of expression nor small network distance. However, unconnected small clusters are present which are still far apart from the main component, thus resulting in a high average silhouette value.

More interestingly, Figure 4 shows an alternative cutting point in section *B*, which leads to a clustering with pathway-like clusters with an average size of 9 objects per cluster. Table 1 depicts the largest resulting clusters which exhibit relatively high co-regulation.

These clusters are able to paint a picture similar to the one extracted manually by DeRisi and co-workers. Their main observations are covered by clusters *E* (down-regulated glycolysis pathway), *D* (up-regulated TCA cycle) and *J* (up-regulated carbohydrate storage pathways). During growth in a sugar-rich medium, the yeast cell employs the glycolysis pathway for energy production. This pathway constitutes the main part of cluster *E*, which is illustrated in more detail in Figure 5. The expression pattern of cluster *E* shows down-regulation of genes during diauxic shift. This is due to the fact that the yeast cells turn to ethanol as an alternative energy source, when the sugar in the medium is exhausted. This pathway is marked in Figure 5 with bold edges.

Cluster *E* does not contain only edges from the glycolysis pathway. Parts of the pentose phosphate pathway, which constitutes an alternative for conversion of glucose 6-phosphate into pyruvate for energy production, are included in the cluster. Moreover, reactions are included which convert other types of sugar (e.g., sucrose or UDPglucose) to α-D-Glucose. At the end of the glycolysis pathway, we find reactions that channel phosphoenolpyruvate to the phenylalanin, tyrosin and trypsin metabolism. All of these additional pathways also exhibit down-regulation during diauxic shift.

In contrast, we find up-regulation during diauxic shift for all genes in cluster *D*, which is composed of mainly the TCA cycle. This cycle is essential in aerobic growth, as it provides energy using acetyl-CoA as its source. In this cluster, we find additional reactions from the glutamate metabolism. Here, 2-oxoglutarate is transaminated to 4-aminobutyrate which in turn can be transformed to succinate by succinsemialdehyde. This reaction chain avoids oxidative decarboxylation in favour of nitrogen-containing products.

In cluster *A*, we find DNA- and RNA-polymerases and the V-type ATPase, already discussed before, together with some supporting reactions from the purin and pyrimidine metabolism. This cluster shows consistent down-regulation. DNA- and RNA-polymerase activity as well as ATPase is reduced due to scant energy resources.

Cluster *H*, in contrast, contains the F-type ATPase, parts of the purin metabolism and parts of the riboflavin metabolism. F-type ATPase is used for ATP production, and riboflavin metabolism may be activated to produce riboflavin and, in turn, FAD, which is used for energy production in the TCA cycle. This cluster contains genes enabling alternative ways of energy production in response to the declining supply of glucose.

The final cluster *J* incorporates one key player responsible for the switch from glycolysis to gluconeogenesis (FBP1), together with pathways which support the channelling of glucose away to the carbohydrate storage pathways (e.g., starch metabolism). Again, this cluster

Table 1. This table shows the 10 largest clusters with gene expression distance smaller than 0.3 sorted by combined average distance. Main constituent KEGG pathways are listed for each cluster; single nodes may belong to pathways not listed here. Column *regulation* describes characteristic expression pattern during diauxic shift. The three following columns give the number of objects, distinct ORFs and distinct EC classifiers, respectively. The last two columns give the average value for expression and normalized network distance within the respective cluster

Id	Pathways	regulation	# objects	# ORFs	# EC	$\delta_{\text{net}}^{\text{norm}}$	δ_{exp}
A	Purin and Pyrimidine metabolism with complexes DNA / RNA polymerases, V-type ATPase; part of Aminosugar metabolism	down	174	71	27	49.84	0.24
B	Sterol biosynthesis and Glycoprotein biosynthesis; fragment of Fructose metabolism	down	51	38	22	51.83	0.29
C	Purin and Histidine metabolism; parts of Folate biosynthesis and Pyrimidine metabolism	down	75	57	51	71.50	0.26
D	TCA cycle and Glutamate metabolism	up	76	45	30	72.29	0.23
E	Glycolysis, Pentose phosphate pathway; Starch metabolism; start of Phenylalanin, Tyrosin and Trypsin metabolism	down	83	50	38	70.82	0.29
F	Phenylalanin, Tyrosin and Trypsin metabolism; part of Folate biosynthesis	down	41	25	22	74.43	0.25
G	Amino acid metabolisms: Valine,Leucine, Isoleucin metabolism; Glycine, Serine, Thrionine, Methionine metabolism; Selenoamino acid metabolism	down	121	68	59	79.28	0.25
H	Purin metabolism with F-type ATP synthase	up	52	33	19	70.86	0.28
I	Pyruvate metabolism; Selenoamino acid metabolism; Valine, Leucine, Isoleucine degradation	down	39	18	17	83.83	0.28
J	Starch and Sucrose metabolism; Glycerolipid metabolism; part of Glycolysis, Fructose, Mammose and Galactose pathway	up	94	60	51	79.85	0.29

corresponds well to a set of genes manually identified by DeRisi and co-workers to be involved in the described processes.

It has to be noted that outliers, i.e., reactions that are not connected to the main component of a cluster, are usually present. This situation is to be expected as clustering is a heuristic procedure and the hierarchical clustering algorithm employed here is susceptible to noise-induced instability. The big picture of the clusters presented here was stable against changes of the parameters. This leads to the conclusion that resampling or bootstrapping methods (e.g., Kerr and Churchill (2001)) should be applied to detect the reliable cores of the computed clusters. Nonetheless, the generated clustering, as shown above, helps to quickly obtain a picture of metabolic changes indicated by the gene expression data.

In contrast to the results shown above, we evaluated clusters based on only expression or network distances. Figure 6 shows a comparison of clustering based on three distance functions: δ_{exp} (expression only), $\delta_{\text{net}}^{\text{norm}}$ (normalized network only) and Δ^{norm} (combined).

For every feasible cutting point of the resulting dendrograms, we evaluate the average distance of each object to all objects within the same cluster. The average of this value over all objects provides a measure for the compactness of a clustering according to a specific distance function. We compute this value separately for all three distance functions. Figure 6 indicates that clustering based on network distance or expression distance alone is not sufficient to arrive at co-regulated pathway-like clusters. When gene expression distance is clustered, the corresponding network distance is high and *vice versa*. This means that we would either end up with sets of well co-regulated genes which are scattered over the network or with a compact part of the network which is not co-regulated. The combined distance function, however, is able to yield clusters with low average distance according to network and expression distance function simultaneously. This shows that our method successively incorporated joint information on regulation and network proximity into the clustering process.

DISCUSSION AND FUTURE WORK

We demonstrate a general method for the coupled analysis of gene expression data and biological networks. While its aim is similar to that of pathway scoring (Zien *et al.*, 2000), our method for co-clustering of networks and expression data avoids the necessity of defining pathways beforehand. Thus, we neither have to deal with the combinatorial explosion encountered when enumerating all possible pathways, nor have to rely on prior knowledge of possible areas of interest, nor do we have to adopt a static view of the network. Therefore, our method is a novel approach that allows for an entirely exploratory joint

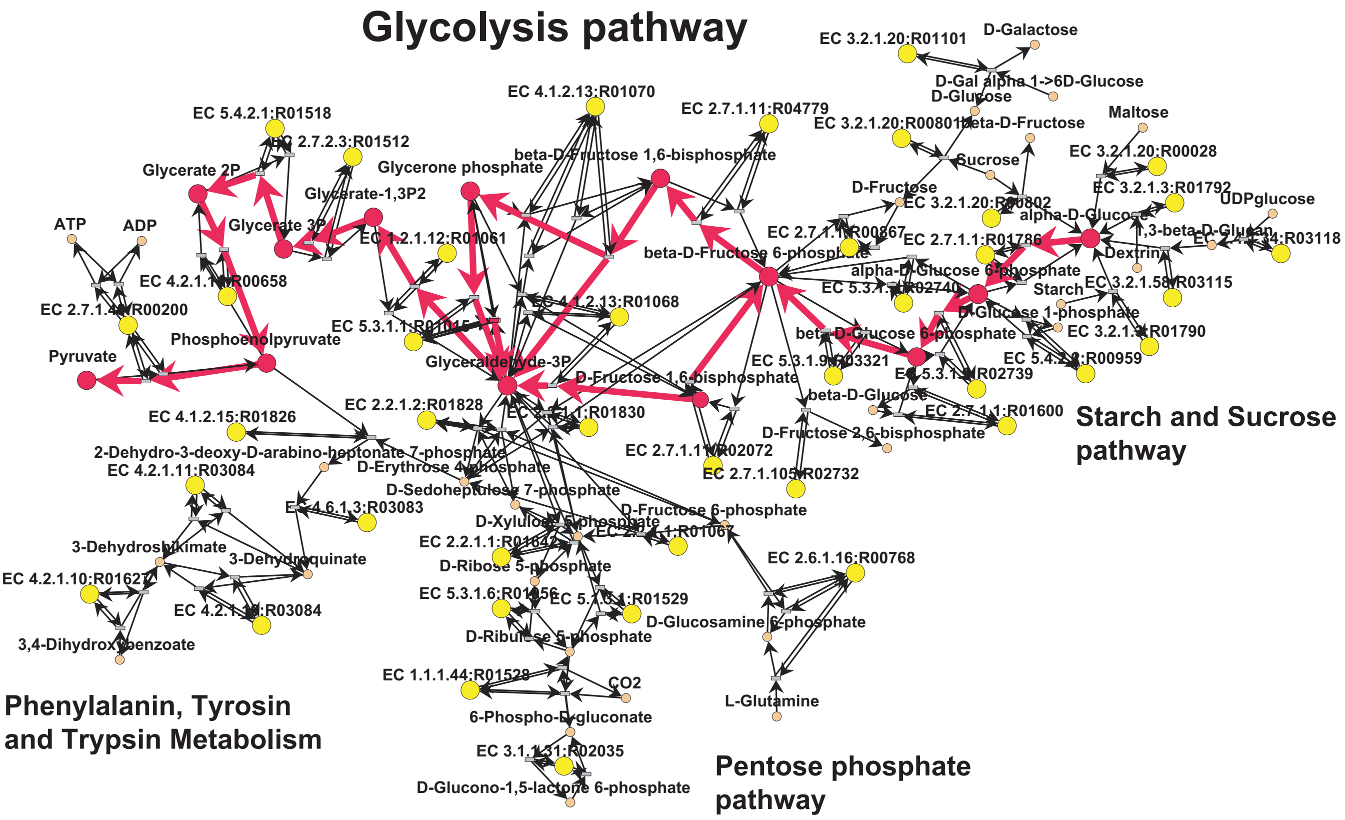

Fig. 5. This diagram shows the main part of Cluster *E* of Table 1. The cluster contains the down-regulated glycolysis pathway (marked with thick edges) and fragments of several directly connected reactions with co-regulated catalysing enzymes. Some unconnected nodes have been omitted. Either these are nodes with duplicate EC identifiers catalysing different reactions or nodes connected via few missing enzymes to the main component.

analysis of gene expression data and biological networks.

Sequential evaluation methods (e.g., van Helden *et al.* (2000)) to construct metabolic pathways from clusters of gene expression face the problem of determining the correct number of clusters based solely on gene expression data. To the best of our knowledge, this problem has not been addressed in this context. It may be useful to infer a small number of clusters and map these (therefore large) clusters onto metabolic networks. However, this direct approach leads to a static view of the network. To arrive at a dynamic view, post-processing of the mapping is mandatory, which leads to the same problems our approach solves in an integrated manner.

There are several meaningful options for the distance functions and corresponding parameter settings. Though our choice was motivated heuristically, we were able to successfully validate it on the gene expression experiment conducted by DeRisi *et al.* (1997). Networks built from metabolic reactions are well suited as validation scenario because large, curated networks are available (Kanehisa, 1996). Co-regulation can be expected to be the most important relationship on the gene expression data and much is already known about metabolic pathways which enables for a kind of evaluation of the results. However, for a full quantitative analysis, a gold standard is required, e.g., a database of activated pathways for specific expression measurements. Unfortunately, this gold standard is not available yet.

Our co-clustering method is by no means restricted to metabolic networks. We expect to be able to base analysis on regulatory networks, networks of interacting proteins, or hypothetical networks mined from the scientific literature. As such networks may be huge, the ability of co-clustering to locate areas of interest will be most useful for their analysis. After this exploratory step, the relevant subnets can be corrected or complemented by human experts and then analysed in more detail.

Another advantage of co-clustering is its algorithmic generality. By designing the distance function to reflect additional biological knowledge, one is still free to select from a wide variety of proposed clustering algorithms. Additionally, the apparatus designed for the assessment

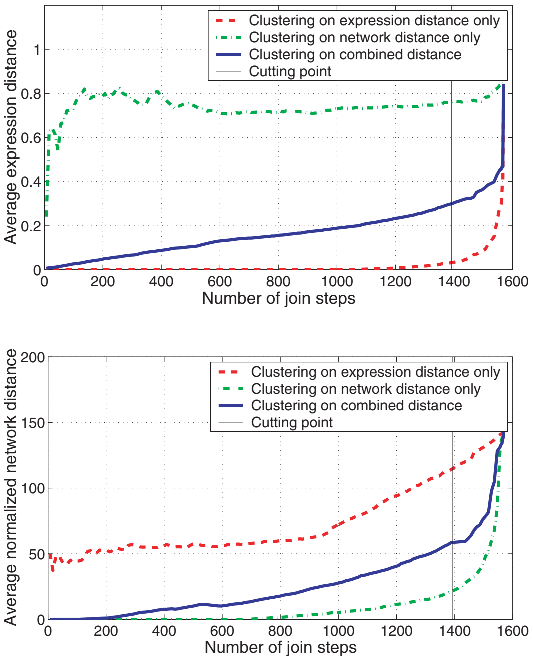

Fig. 6. Comparison of clustering based on three distance functions: expression distances only, normalized network distances only and normalized combined distances. Upper figure shows average intra-cluster *expression distance*, lower figure shows average intra-cluster *normalized network distance*.

of good cluster solutions is also applicable. Consequently, additional knowledge of biological networks can readily be used in applications for gene clustering.

The distance function is at the core of our approach. While the correlation metric is most frequently used with gene expression data and it appears to be appropriate in the context of metabolic networks, other metrics may be necessary for the meaningful analysis of regulatory, binding or co-occurrence networks. Many of the signaling cascades that convey information on regulation of other genes can be expected to be realized by post-translational processes. Here, the metric should focus on the presence or absence of expression of the signaling proteins and on the co-expression of the downstream regulated genes. The network distance function may also have to be adapted to regulatory relations. A crucial point is the proper combination of the two individual distance functions. If sufficiently many pathways are known to be relevant in advance, this knowledge may be utilized to automatically fit the logistic curves or even to learn an appropriate functional form by employing machine learning methods. By further developing the method along these lines, we may be able to extend the capabilities of our method from metabolic networks to the entire cellular networks: to direct the scientist quickly to the biological meaning behind the expression data.

ACKNOWLEDGEMENTS

The authors wish to thank Juliane Fluck for help with the evaluation and Lars Kunert for his suggestions on graph algorithms. Part of this work was supported by Aventis Pharma, Frankfurt.

REFERENCES

Bhalla,U.S. and Iyengar,R. (1999) Emergent properties of networks of biological signaling pathways. *Science*, **283**, 381–387.

Cormen,T.H., Leiserson,C.E. and Rivest,R.L. (1992) *Introduction to Algorithms*, Vol. 13, The MIT Press, Cambridge, MA.

DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–685.

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*, **95**, 14863–14868. Genetics.

Fell,D.A. and Wagner,A. (2000) The small world of metabolism. *Nat. Biotechnol.*, **18**, 1121–1122.

Hartemink,A.J., Gifford,D.K., Jaakkola,T.S. and Young,R.A. (2001) Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. In *Proceedings of the Pacific Symposium on Biocomputing '01*. pp. 422–433.

Jeong,H., Tombor,B., Albert,R., Oltvai,Z.N. and Barabisi,A.-L. (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.

Kanehisa,M. (1996) Toward pathway engineering: a new database of genetic and molecular pathways. *Science and Technology Japan*, **59**, 34–38.

Kerr,K.M. and Churchill,G.A. (2001) Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *Proc. Natl Acad. Sci. USA*, **98**, 8961–8965.

Küffner,R., Zimmer,R. and Lengauer,T. (1999) Pathway analysis in metabolic databases via differential metabolic display (DMD). In *Proceedings of the German Conference on Bioinformatics '99*. pp. 141–147.

Kurhekar,M.P., Adak,S., Jhunjhunwala,S. and Raghupathy,K. (2002) Genome-wide pathway analysis and visualization using gene expression data. In *Proceedings of the Pacific Symposium on Biocomputing '02*. pp. 462–473.

Mewes,H.W., Albermann,K., Heumann,K., Liebl,S. and Pfeiffer,F. (1997) MIPS: a database for protein sequences, homology data and yeast genome information. *Nucleic Acids Res.*, **25**, 28–30.

Rousseeuw,P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.

van Helden,J., Gilbert,D., Wernisch,L., Schroeder,K. and Wodak,S. (2000) Applications of regulatory sequence analysis and metabolic network analysis to the interpretation of gene expression data. *LNCS*, **2066**, 155–172.

Zien,A., Küffner,R., Zimmer,R. and Lengauer,T. (2000) Analysis of gene expression data with pathway scores. In Altman,R. *et al.*, (eds), *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*. La Jolla, CA, pp. 407–417. AAAI.