Analysis of Gene Expression Data with Pathway Scores

Alexander Zien, Robert Küffner, Ralf Zimmer, Thomas Lengauer
Institute for Algorithms and Scientific Computing (SCAI)
GMD - German National Research Center for Information Technology
Schloss Birlinghoven, D-53754 Sankt Augustin, Germany
{Alexander.Zien, Robert.Kueffner, Ralf.Zimmer, Thomas.Lengauer}@gmd.de

Abstract

We present a new approach for the evaluation of gene expression data. The basic idea is to generate biologically possible pathways and to score them with respect to gene expression measurements. We suggest sample scoring functions for different problem specifications. We assess the significance of the scores for the investigated pathways by comparison to a number of scores for random pathways. We show that simple scoring functions can assign statistically significant scores to biologically relevant pathways. This suggests that the combination of appropriate scoring functions with the systematic generation of pathways can be used in order to select the most interesting pathways based on gene expression measurements.

Introduction

Large scale gene expression measurements can now be performed by several established techniques, including EST (expressed sequence tag) sequencing, clustering and counting, e.g. (Okubo *et al.* 1992; Okubo & Matsubara 1997; Ewing & Claverie 2000); SAGE (serial analysis of gene expression), e.g. (Velculescu 1999); DNA-chips, e.g. (Lockhart & others 1996; Chee *et al.* 1996); and micro-arrays, as introduced by Pat Brown's Laboratory at Stanford University, e.g. (DeRisi, Iyer, & Brown 1997). The available techniques are reviewed in (Ramsay 1998; Gerhold, Rushmore, & Caskey 1999). Knowledge of the expression of genes is generally believed to speed up the understanding of living systems on a molecular level. This is especially important in order to find target genes and pathways for drug development, e.g. by the comparison of diseased cells with their healthy counterparts.

Several methods have been proposed in order to interpret large amounts of expression measurement data. The earliest publications focussed on manual interpretation (DeRisi, Iyer, & Brown 1997; Heller *et al.* 1997). The most important basic automatic analysis technique is clustering, e.g. (Eisen *et al.* 1998; Tamayo *et al.* 1999). Several other methods build on this technique: computer visualization (Carr, Somogyi, & Michaels 1997; Michaels *et al.* 1998); semi-automatical coarse-grain function predictions (Chu *et al.* 1998); investigations of promoter sequences for regulatory elements (Zhu & Zhang 2000); coarse-grain genetic network reconstruction (Mjolsness *et al.* 2000); mapping onto metabolic pathways (Fellenberg & Mewes 1999). Methods that do not require clustering include Fourier analysis of measurements of periodic phenomena, e.g. (Spellman *et al.* 1998); principal component analysis, e.g. (Raychaudhuri, Stuart, & Altman 2000); genetic network reconstruction, e.g., for linear (D'haeseleer *et al.* 1999), Boolean (Liang, Fuhrman, & Somogyi 1998) or Bayesian (Friedman *et al.* 2000) models; supervised machine learning techniques, e.g., for disease class prediction (Golub *et al.* 1999) or coarse-grain gene function prediction (Brown *et al.* 2000). All of the automatic methods utilize, at best, a rather broad notion of biological function. To our knowledge, no method, except for human expertise, employs detailed knowledge of (parts of) the biological networks for the evaluation of gene expression data in a systematic and automated way. Pure clustering methods do not exploit prior biological knowledge at all.

Our work is based on the expectation that the use of the available knowledge on biological networks is essential for the development of powerful automatic methods for the evaluation of gene expression data. Surprisingly, to our knowledge, there is only one published attempt to make use of knowledge on biological pathways for the interpretation of gene expression data: the approach described by (Fellenberg & Mewes 1999). They deduce a structure (the clustering) from the expression data and impose it onto the reaction network representing the prior knowledge, resulting in lists of possibly meaningful pathways. However, this method does not provide any quantitative indication for the validity of the generated pathways.

In contrast, we propose to follow the opposite direction. Starting from the known reaction networks, we extract possible pathways and examine how well they are supported by the given expression data. This method can be used to rank candidate solutions. The core idea

is to define scores for putative pathways and scores for genes with respect to a given pathway, both based on gene expression measurements. The scoring function can be designed to indicate any desired property, as long as it is reflected in the available gene expression data. In this paper, we propose scoring functions aiming at three different properties of putative pathways: first, general conspicuousness of the expression patterns of the involved genes; second, synchrony of the expression patterns among the involved genes; and, finally, a combination of both that is intended to indicate whether the pathway is realized in one of the examined cell states. There is a broad range of related questions that can be addressed with this approach.

Methods

Expression data

In this paper, we consider multiple gene expression measurements of cells. Let G be the set of genes common to all investigated cells. For the purpose of this paper, we regard each gene expression measurement as a mapping from each gene $g \in G$ to a positive number. This number represents, as faithfully as current measurement technology permits, the number of mRNA copies of that gene. Let each different measurement be labeled by a time point t, $t \in T$. In this paper, we assume that the measurements form a single time series. A well known example for a time series of length $|T| = 8$ for virtually all yeast genes is the diauxic shift data provided by (DeRisi, Iyer, & Brown 1997). We will use these data for our sample calculations described in the results section.

Let $l_{t,g}$ denote the expression level measured for gene g at time point t. The micro-array technology developed in the Brown Lab allows to perform double measurements. Often, the gene expression levels of a distinguished reference time point t_0 are measured simultaneously to the expression levels for each time point t. Thus, the expression level ratios $l_{g,t}/l_{g,t_0}$ can be determined directly and free from errors resulting from differences between chips and probe concentrations. For each time point t, we take the logarithm of the expression level ratio for each gene g:

$$m_{t,g} = \log\left(\frac{l_{t,g}}{l_{t_0,g}}\right). \qquad (1)$$

This leads to a representation of gene expression data that is symmetric with regard to up- and down-regulation.

Pathway construction

The second essential ingredient of our method is the set of pathways against which the expression data is evaluated. In general, a pathway can be any meaningful substructure of a biological interaction network. In the following, we briefly describe how we obtain such pathways. A more detailed description can be found in (Küffner, Zimmer, & Lengauer 1999).

From the metabolic databases BRENDA (Schomburg, Salzmann, & Stephan 1990 1995), ENZYME (Bairoch 1999), and KEGG/LENZYME (Ogata *et al.* 1999) all reactions are extracted. Since regulatory and signalling relationships are currently not sufficiently covered in databases, we restrict ourselves to metabolic pathways. From the reactions we can construct both universal (organism-independent) or organism-specific networks, represented as PETRI nets that describe all metabolic interactions for which experimental evidence has been found under some condition. We extract possible pathways from the respective network by specifying source- and sink-substrates and topological constraints.

The generated pathways are closed, i.e. the net production and consumption of all substrates other than source and sink substrates and a definable set of ubiquitous molecules is zero. Together with additional user defined and biologically motivated restrictions, this ensures the generation of biologically meaningful entities and the drastic reduction of the number of generated pathways. The enumeration still results in a large number of pathways, which can be used for model construction or be subject to hypothesis evaluation.

The enzymes performing chemical reactions in pathways are usually labeled by EC numbers as indicators of biological function. Gene expression measurements, however, refer to ORF identifiers. We map the constructed pathways into the space of ORFs according to the MIPS yeast catalogue, subsection EC numbers[1]. Often, there are several proteins for a given EC number. Thus, we may yield a number of different versions of a pathway in the space of ORFs.

Method outline

Our method allows to rate putative pathways according to different properties, examples of which are discussed below. The method can be summarized as follows:

- Given the **Input:**
  - gene expression measurements
  - putative pathways
- Answer the **Questions:**
  - Which pathways show the desired property?
  - How much support does each gene have to belong to a given pathway?
- By producing the **Output:**
  - *Pathway scores*: For each putative pathway as a whole, this is a score that measures to which degree the pathway shows the desired property;
  - *Gene scores* with respect to a given pathway: For each gene (both included and not included in the pathway), this is a score that measures how much the gene shows the desired property with respect to that pathway.

[1] `http://www.mips.biochem.mpg.de/proj/yeast/catalogues/EC/index.html`

In this paper, we first define the gene scoring function, which is based directly on the given gene expression data. Subsequently, a score for the pathway as a whole is computed from the gene scores of those genes that form part of the pathway. Alternatively to building pathway scores on gene scores, it is possible to define pathway scores directly from the gene expression data. Then, for any gene g, a score with respect to a given pathway p can be deduced from those pathway scores. For example, the gene score could be defined as the difference of the scores of the original pathway p and a modified pathway p'. Here, p' is obtained by either removing g from the pathway p or adding g to it, depending on whether g belongs to the pathway p or not, respectively.

In this paper, we derive example pathway scores from gene scores. All scores are defined in such a way that higher values indicate more interesting pathways. We do not claim the presented functions to be optimal in any respect, on the contrary, we believe that much work remains to be done in order to develop suitable scoring systems.

Scoring conspicuousness of expression

First, we construct a scoring function that distinguishes pathways consisting of genes with conspicuous expression patterns, i.e. maximum changes. Therefore, we take advantage of the fact that (DeRisi, Iyer, & Brown 1997) supply a double measurement on the same chip for the reference time point t_0. We calculate, for each gene g, a log ratio value $m_{t_0,g}$ that quantifies pure measurement error. The distribution of these error values is shown in Figure 1. While the distribution function is not necessarily normal, it shares some important characteristics with a normal distribution. The distribution function is sigmoid, and the idealized density function is unimodal and almost symmetric.

Based on this observation, we model the measurement error by a normal distribution. We call this the *null model* for the gene expression values, since it describes which degree of observed change can be expected to arise from measurement errors only and does not indicate biological events. A normal distribution is fitted to the error values according to their mean $\overline{m}_{err}$ and empirical standard deviation s_{err}.

$$\overline{m}_{err} := \frac{1}{|G|}\sum_{g\in G} m_{t_0,g} \quad (2)$$

$$s_{err} := \sqrt{\frac{1}{|G|-1}\sum_{g\in G}(m_{t_0,g}-\overline{m}_{err})^2} \quad (3)$$

As expected, there does not seem to be a significant difference (bias) between two measurements on the same chip. This is indicated by the small value of the mean $\overline{m}_{err}$ (0.089) as compared to s_{err} (0.251). Consequently, we replaced the empirical mean by zero before performing any other calculations.

A gene is considered the more conspicuous, the stronger the observed changes of expression are. For each time point t, we estimate the probability $P_t^0(g)$ of the observed log-relative expression change $m_{t,g}$ of gene g to arise from measurement errors only. We use a two-sided test on the normal distribution Φ that is parameterized with the mean $\overline{m}_{err} = 0$ and the standard deviation s_{err} as calculated for the null model.

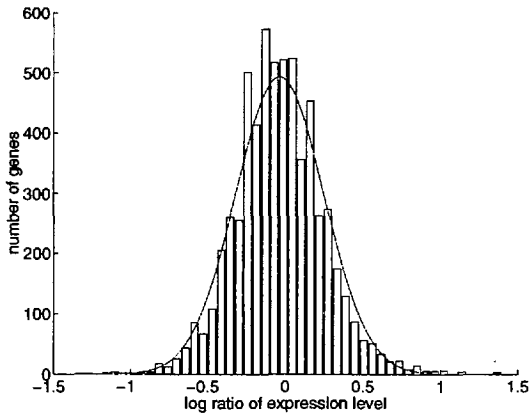

Figure 1: Histogram of the distribution of log-relative expression levels for the double measurement at time point t_0 from (DeRisi, Iyer, & Brown 1997) over all probed yeast genes, superimposed by the normal density function parameterized with the mean and the empirical standard deviation of the expression data.

$$P_t^0(g) := 2\Phi\left(-\left|\frac{m_{t,g}-0}{s_{err}}\right|\right) \quad (4)$$

We can compute a conspicuousness score for each gene g and time point t as follows.

$$\text{score}_t(g) := -\log P_t^0(g) \quad (5)$$

The overall score for the gene g with respect to the complete time series can be computed as the average over the set $T - \{t_0\}$ of time points:

$$\text{score}(g) := \frac{1}{|T|-1}\sum_{t\in T-\{t_0\}} \text{score}_t(g) \quad (6)$$

Since adding the logs is equivalent to multiplying the probabilities, we implicitly assume independence between the different measurements. Other definitions may be more appropriate, e.g. taking the maximum of the values over the time points. Note that, in either definition, the conspicuousness gene score is independent from the actual pathway under investigation.

Now we consider a given pathway that shall be characterized by the set p of involved genes. A score for the complete pathway can be computed, e.g., as the average over the scores of the genes included in the pathway:

$$\text{score}(p) := \frac{1}{|p|}\sum_{g\in p}\text{score}(g) \quad (7)$$

Of course, this scoring function does not provide any information about synchronous regulation of the genes involved in a pathway.

Scoring synchrony of expression

We construct a second scoring function that quantifies synchrony of expression. To that goal, we can employ any similarity measure for expression time series that may also be used for clustering. Here, we use the correlation coefficient. By considering the absolute magnitude of the correlation coefficient, e.g., synchronous regulation in opposite direction of genes could be taken into account. Depending on the circumstances, different measures of similarity can be optimal.

In order to calculate a score for a gene g with respect to the pathway p, a model of the pathway is constructed, called the *pathway model*. The expression data of the gene is compared to this model, and a score is computed that reflects how well the gene fits the pathway model. Subsequently, a score for the pathway as a whole is computed from the scores of those genes that form part of the pathway. We do not make use of the null model used for the conspicuousness scoring.

Since pathways usually consist of small sets of genes, each of them has a substantial influence on the pathway model. In order to avoid gene scores that mostly reflect self-similarity, each gene from the pathway is scored against a modified model of the pathway, which is obtained by excluding that particular gene. Consequently, we define different pathway models for genes involved in the pathway than for the remaining genes. Let p_g denote the set of genes upon which the pathway model for gene g is built. From the above considerations it follows that $p_g := p - \{g\}$ for genes g included in the pathway and $p_g := p$ otherwise.

This gene scoring function simply quantifies the average similarity, i.e., here, the average correlation coefficient, to the genes on the pathway. Formally, we have:

$$\text{score}_p(g) \quad := \quad \frac{1}{|p_g|} \sum_{h \in p_g} cc(g,h) \; . \tag{8}$$

Here, $cc(g,h)$ denotes the correlation coefficient of the expression time series (excluding t_0 which does not contain biological information) that belong to the genes g and h,

$$cc(g,h) \quad := \quad \frac{\text{cov}_{g,h}}{s_g s_h} \; , \tag{9}$$

where s_g and s_h denote the empirical standard deviations of the sets of values $m_{t,g}$ and $m_{t,h}$, $t \in T - \{t_0\}$, and $\text{cov}_{g,h}$ denotes the empirical covariance of these sets. For this scoring function, the use of p instead of the modified pathway p_g would lead to an increase of the score by $\frac{1}{|p|}$, since $cc(g,g) = 1$. Thus, the correction eases the comparison of differently sized pathways.

Again, the score for the complete pathway can be computed as the average over the scores of the genes included in the pathway:

$$\text{score}(p) \quad := \quad \frac{1}{|p|} \sum_{g \in p} \text{score}_p(g) \tag{10}$$

This scoring function assigns high scores to pathways whenever the involved genes would cluster together nicely. It also assigns a high score to a pathway if the genes are similarly expressed but are, nevertheless, distributed over several clusters. On the other hand, by only considering putative pathways, our approach avoids arbitrarily associating any set of genes with similar expression patterns.

However, this scoring function exhibits one problem that it shares with cluster-based methods: It assigns high scores to pathways consisting of genes with similar, but inconspicuous expression. This is risky, because most genes can be expected to exhibit constant expression during a limited number of measurements, and thus many biologically unrelated genes may be highly correlated.

Combined scoring function

In this scoring function, we combine the ideas of the preceding two functions. In order to do so, we use the error model employed in the conspicuousness score to calculate a modified measure of correlation. In the correlation coefficient (cc, Equation 9), the covariance is scaled with respect to the variances of the two sets of values to be compared. Thus, perfectly synchronuously expressed genes yield the same maximum correlation coefficient of 1, regardless of whether they are significantly regulated at all.

By defining the scoring function to be proportional to the covariance, we achieve scores that reward both synchronuous and strong regulation. We define a modified correlation coefficient cc^* by replacing the denominator by a term which is independent of the genes under consideration:

$$cc^*(g,h) \quad := \quad \frac{\text{cov}_{g,h}}{s_{err} s_{err}} \tag{11}$$

By choosing to scale the covariance to units of the variance arising from measurement errors (s_{err}, Equation 3), the resulting similartity value cc^* has an intuitive meaning: it indicates by which factor the observed covariance exceeds what can be expected by chance.

As before, a score for each gene g is computed by averaging the similarity to the other genes involved in the pathway. Then, a score for the pathway as a whole is computed from the scores of those genes that form part of the pathway.

$$\text{score}_p(g) \quad := \quad \frac{1}{|p_g|} \sum_{h \in p_g} cc^*(g,h) \tag{12}$$

$$\text{score}(p) \quad := \quad \frac{1}{|p|} \sum_{g \in p} \text{score}_p(g) \tag{13}$$

According to the intuition of this scoring function, random pathways should show an average score of 1, which is nicely reproduced by our calculations (Figure 6). This suggests that this kind of score is comparable among different measurement technologies. This has to be investigated in future work.

Further improvements on the performance can be expected with refined scoring functions, e.g. for circumventing the restrictions associated with treating measurements independently as done in the versions described above.

Results

In order to obtain hints on the efficacy of our procedure, we investigate the glycolysis pathway in *S. cerevisiae*. We do not claim that this is a comprehensive performance evaluation. With the currently available data and knowledge on the realization of pathways in specific states it is impossible to generate a comprehensive benchmark for systematic evaluation of our gene expression data analysis method. For such a benchmark a sufficiently large set of realized pathways needs to be available together with many expression measurements related to the state under investigation. In addition, another set of pathways known not to be realized in this state is also required.

For the current evaluation, we relied on published data as a standard of truth. We analyze the behavior of our method for the textbook glycolysis and gluconeogenesis pathway as described in (DeRisi, Iyer, & Brown 1997). The glycolysis pathway consists of ten proteins and is shown in Figure 2. For some nodes, alternative proteins are known that are capable of performing the respective EC function. It is uncertain which of them really belong to the pathway. Figure 3 shows the possible combinations obtained by selecting one protein (or, equivalently, one ORF) for each node, resulting in a total of 36 pathways. We would like to be able to identify (some of) these pathways from the gene expression data, so they should receive high scores.

We make use of the gene expression time series measured by DeRisi et al. (DeRisi, Iyer, & Brown 1997) that is publicly available[2]. For each known yeast gene g, there are pair measurements of both $l_{t,g}$ and $l_{t_0,g}$ for eight different time points t (including t_0). The repeated measurement of the reference time point allows for the compensation of certain types of measurement errors. The time series is optimally suited for the investigation of the glycolysis pathway, as the time points correspond to decreasing concentrations of glucose and a regulation of the glucose processing glycolysis pathway is expected. In fact, the data measured confirm this expectation (DeRisi, Iyer, & Brown 1997), as demonstrated by a manual analysis by the original authors.

Using the pathway generation method described above (Küffner, Zimmer, & Lengauer 1999), we generate all pathways consuming glucose and producing pyruvate. These pathways are characterized by the types of reactions needed to produce pyruvate from glucose in a number of steps and by the graph structure that these reactions impose on the enzymes and the intermediate substrates. With appropriate constraints, this process results in 541 different pathways on the level of EC numbers. For the computation of scores from expression data, it is necessary to translate the EC numbers into ORFs. Using the assignment of EC numbers to ORFs provided by MIPS, 540 of the 541 pathways contain at least one EC number to which no yeast ORF is assigned, and only one pathway can be mapped into the space of yeast ORFs without gaps. Figure 2 shows this pathway, together with one out of the 900 possible assignments of ORFs to the EC numbers.

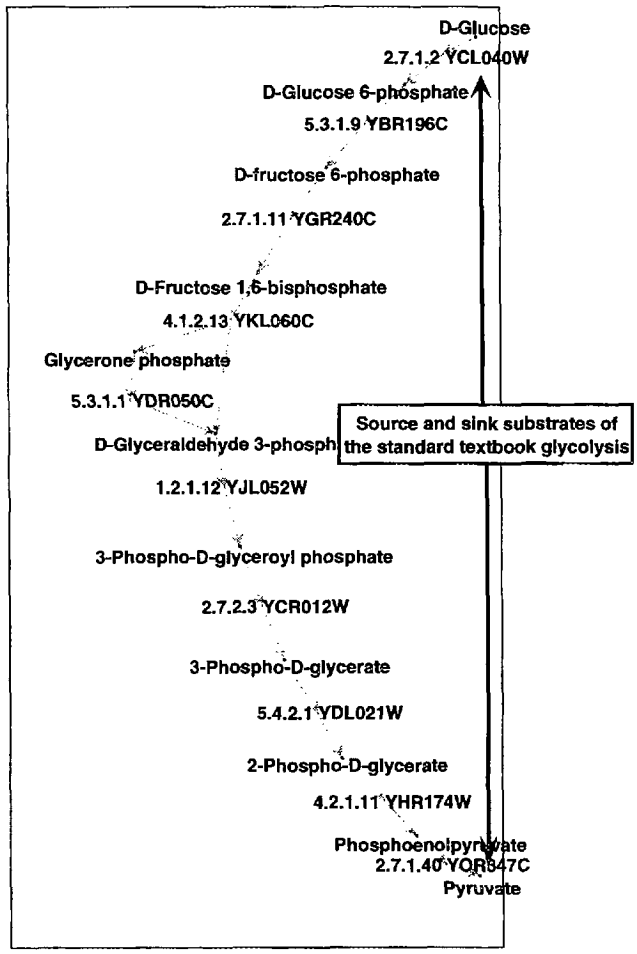

Figure 2: Graphical illustration of the computed glycolysis pathway described in the text. The pathway contains 10 enzymes, each labeled with the associated EC number and the identifier of one yeast ORF that codes for an enzyme which is assigned to that function.

Additionally, we compute scores for 10000 randomly chosen ORF sets of the same size (namely, ten genes). These unstructured sets form a sufficient random model, since the scoring functions used in this paper do not exploit pathway topology.

In the following, we analyze the behavior of the different scoring functions on these three sets of pathways in more detail. First, we plot the distribution of the

[2] `http://cmgm.stanford.edu/pbrown/explore/index.html`

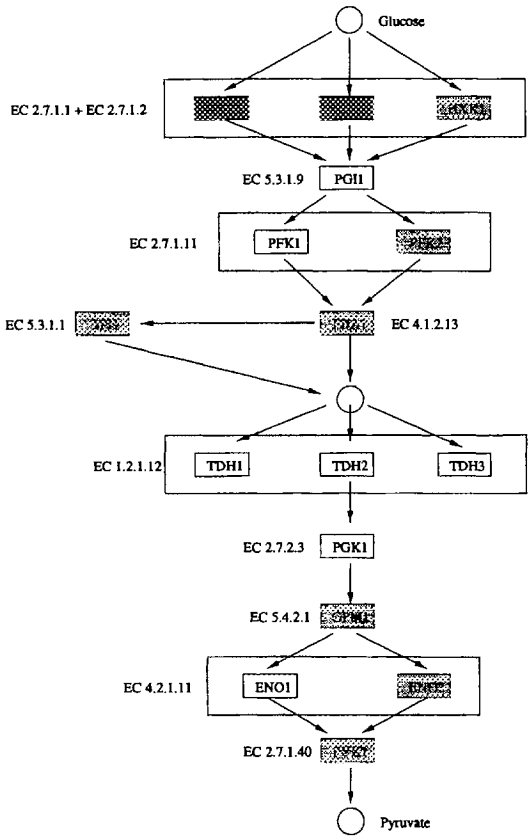

Figure 3: Possible pathways defined by selecting one alternative ORFs for any node/EC-number resulting in 36 pathways altogether. According to the measurements before and after the diauxic shift several of the genes are predominantly expressed in one state but not the other (light gray means only expressed before the shift, dark gray means only after the shift) whereas other proteins do not show significant changes in expression values (white).

scores for the random sets of genes, for the automatically generated 900 glycolysis pathways and for the 36 pathways taken from (DeRisi, Iyer, & Brown 1997). The resulting histograms are shown in Figure 4 for the conspicuousness score, in Figure 5 for the correlation of gene expression on the pathway, and in Figure 6 for the combined scoring function. In all cases, the pathway histograms are quite well distinguished from the distribution of random scores, i.e. most of the glycolysis pathways can indeed be recognized by our method given the current measurements.

As can be seen in Table 1, the correlation function scores those pathways best that are completely activated in the glycolysis, i.e. before the diauxic shift. The most distinguishing position of the pathway is the first enzyme characterized by the expression of HXK2/YGL253W. In contrast, the two other alternatives, HXK1/YCL040W and GLK1/YFR053C were both up-regulated after the diauxic shift and lead to significantly lower scores. For the ranking HXK2 (HXK1/GLK1) is more important than ENO2 (ENO1) and PFK1 (PFK2) while TDH1/2/3 have very little significance. For the other enzymes in the pathway only one ORF is assigned, with no alternatives to be considered. In the described implementation, this scoring scheme seems to prefer pathways which are most active before the diauxic shift as compared to the state after the shift.

For a measurement on a less well studied set of states or with less characterized genes a natural first question could be which of the possible pathways are the most interesting with respect to the net change of expression within the set of state measurements. This is best addressed with the conspicuousness score. The resulting scores are shown in Table 2. Here, pathways containing the genes HXK1 and GLK1 receive the highest scores as their change in expression level is more significant (though negatively correlated to other genes on the pathway) than the change of HXK2. Again, ENO2 and PFK2 are more important than ENO1 and PFK1.

Another natural question could be which of the possible pathways are both interesting from the level of expression changes and, at the same time, best fitting (e.g. correlated) to the set of genes on the common pathway. Our combined scoring function again scores HXK1 and GLK1 highest (see Table 3), although there is no single state in which these genes participate in the pathway. This hints to the fact that, in the current definition of score, the conspicuousness term dominates over the correlation component. Pathways containing genes known from above to be discriminative for the glycolysis before the shift are ranked in the same order as above. The best pathways contain HXK2, ENO2, and PFK2 and rank TDH1, TDH3, and TDH2 in that order. In general, however, all pathways and genes putatively participating in glycolysis pathways are scored much higher than random sets of genes, indicating that such a combined scoring scheme could be employed for selecting pathways based on both criteria together. A

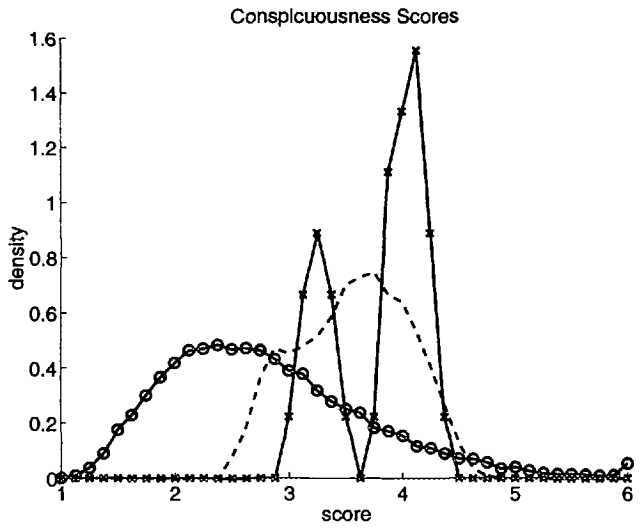

Figure 4: Histograms of pathway scores calculated according to Equation 7 (conspicuousness score). One histogram is shown for each of three sets of pathways in ORF space: the 36 ORF pathways resulting from the glycolysis as described in (DeRisi, Iyer, & Brown 1997) (solid line, crosses), that serve as a substitute for a standard of truth; the 900 possible assignments of yeast ORFs to the reactions of the glycolysis pathway as generated by our methods (dashed line); and 10000 random pathways (solid line, circles). To ensure comparability, all histograms are normalized to resemble probability density functions.

final differentiation of the selected pathways, e.g. in order to assign them as characteristic for a specific state or as discriminating for two or more states should afterwards be based on scoring systems like the correlation score as discussed above.

p-values

Depending on the definition of the scoring function, the score distribution may be biased by the characteristics of the pathways scored, most importantly their size. This hampers the comparison of scores of pathways of different characteristics. In the field of sequence comparison, statistical scores, called *p-values* (probability estimates) or *E-values* (expectation values), that remedy analogous problems, have been an important prerequisite for the success of programs like BLAST and FASTA. In addition to increasing the reliability of decisions, these scores have an intuitive interpretation as probabilities or expectation values of erroneous decisions and can be used to guide the trade-off between sensitivity and specificity. We propose the computation of similar p-values for pathways, for example, by the following brute-force procedure: for each putative pathway under investigation, a large number of random

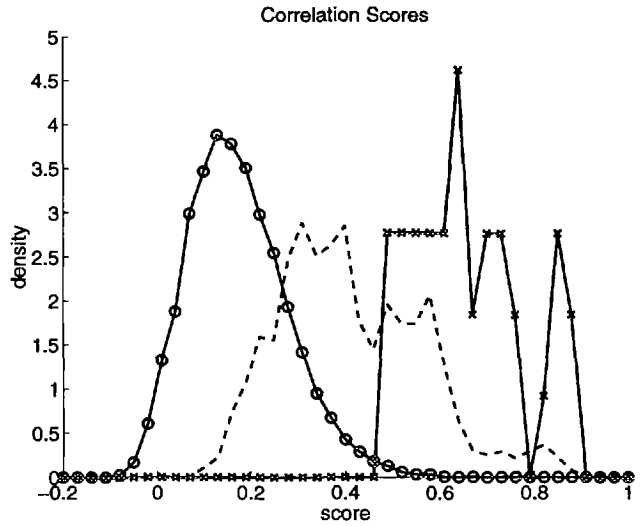

Figure 5: Histograms of pathway scores calculated according to Equation 10 (synchrony score), presented as in Figure 4.

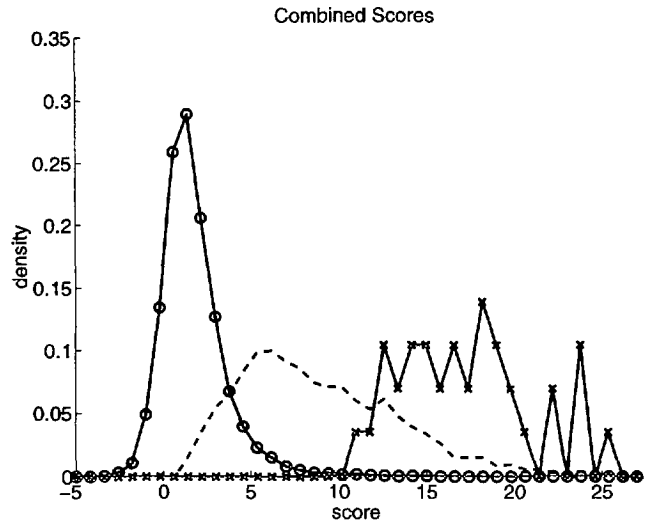

Figure 6: Histograms of pathway scores calculated according to Equation 13 (combined score), presented as in Figure 4.

Genes	HXK1					TDH1				
	GLK1		PFK1			TDH2			ENO1	
Scores	HXK2	PGI1	PFK2	FBA1	TPI1	TDH3	PGK1	GPM1	ENO2	PYK1
0,869	YGL253W	YBR196C	YMR205C	YKL060C	YDR050C	YJR009C	YCR012W	YKL152C	YHR174W	YAL038W
0,867	YGL253W	YBR196C	YMR205C	YKL060C	YDR050C	YGR192C	YCR012W	YKL152C	YHR174W	YAL038W
0,860	YGL253W	YBR196C	YMR205C	YKL060C	YDR050C	YJL052W	YCR012W	YKL152C	YHR174W	YAL038W
0,838	YGL253W	YBR196C	YGR240C	YKL060C	YDR050C	YJR009C	YCR012W	YKL152C	YHR174W	YAL038W
0,837	YGL253W	YBR196C	YGR240C	YKL060C	YDR050C	YGR192C	YCR012W	YKL152C	YHR174W	YAL038W
0,830	YGL253W	YBR196C	YGR240C	YKL060C	YDR050C	YJL052W	YCR012W	YKL152C	YHR174W	YAL038W
0,748	YGL253W	YBR196C	YMR205C	YKL060C	YDR050C	YGR192C	YCR012W	YKL152C	YGR254W	YAL038W
0,748	YGL253W	YBR196C	YMR205C	YKL060C	YDR050C	YJR009C	YCR012W	YKL152C	YGR254W	YAL038W
0,742	YGL253W	YBR196C	YMR205C	YKL060C	YDR050C	YJL052W	YCR012W	YKL152C	YGR254W	YAL038W
0,721	YGL253W	YBR196C	YGR240C	YKL060C	YDR050C	YGR192C	YCR012W	YKL152C	YGR254W	YAL038W
0,721	YGL253W	YBR196C	YGR240C	YKL060C	YDR050C	YJR009C	YCR012W	YKL152C	YGR254W	YAL038W
0,715	YGL253W	YBR196C	YGR240C	YKL060C	YDR050C	YJL052W	YCR012W	YKL152C	YGR254W	YAL038W
0,690		YBR196C	YMR205C	YKL060C	YDR050C	YJR009C	YCR012W	YKL152C	YHR174W	YAL038W
0,687		YBR196C	YMR205C	YKL060C	YDR050C	YGR192C	YCR012W	YKL152C	YHR174W	YAL038W
0,685		YBR196C	YMR205C	YKL060C	YDR050C	YJL052W	YCR012W	YKL152C	YHR174W	YAL038W
0,655		YBR196C	YGR240C	YKL060C	YDR050C	YJR009C	YCR012W	YKL152C	YHR174W	YAL038W
0,653		YBR196C	YGR240C	YKL060C	YDR050C	YGR192C	YCR012W	YKL152C	YHR174W	YAL038W
0,650		YBR196C	YGR240C	YKL060C	YDR050C	YJL052W	YCR012W	YKL152C	YHR174W	YAL038W
0,638	YCL040W	YBR196C	YMR205C	YKL060C	YDR050C	YJR009C	YCR012W	YKL152C	YHR174W	YAL038W
0,637	YCL040W	YBR196C	YMR205C	YKL060C	YDR050C	YGR192C	YCR012W	YKL152C	YHR174W	YAL038W
0,634	YCL040W	YBR196C	YMR205C	YKL060C	YDR050C	YJL052W	YCR012W	YKL152C	YHR174W	YAL038W
0,604	YCL040W	YBR196C	YGR240C	YKL060C	YDR050C	YJR009C	YCR012W	YKL152C	YHR174W	YAL038W
0,604	YCL040W	YBR196C	YGR240C	YKL060C	YDR050C	YGR192C	YCR012W	YKL152C	YHR174W	YAL038W
0,600	YCL040W	YBR196C	YGR240C	YKL060C	YDR050C	YJL052W	YCR012W	YKL152C	YHR174W	YAL038W
0,583		YBR196C	YMR205C	YKL060C	YDR050C	YJR009C	YCR012W	YKL152C	YGR254W	YAL038W
0,582		YBR196C	YMR205C	YKL060C	YDR050C	YGR192C	YCR012W	YKL152C	YGR254W	YAL038W
0,580		YBR196C	YMR205C	YKL060C	YDR050C	YJL052W	YCR012W	YKL152C	YGR254W	YAL038W
0,552		YBR196C	YGR240C	YKL060C	YDR050C	YJR009C	YCR012W	YKL152C	YGR254W	YAL038W
0,551		YBR196C	YGR240C	YKL060C	YDR050C	YGR192C	YCR012W	YKL152C	YGR254W	YAL038W
0,549		YBR196C	YGR240C	YKL060C	YDR050C	YJL052W	YCR012W	YKL152C	YGR254W	YAL038W
0,528	YCL040W	YBR196C	YMR205C	YKL060C	YDR050C	YGR192C	YCR012W	YKL152C	YGR254W	YAL038W
0,527	YCL040W	YBR196C	YMR205C	YKL060C	YDR050C	YJR009C	YCR012W	YKL152C	YGR254W	YAL038W
0,525	YCL040W	YBR196C	YMR205C	YKL060C	YDR050C	YJL052W	YCR012W	YKL152C	YGR254W	YAL038W
0,498	YCL040W	YBR196C	YGR240C	YKL060C	YDR050C	YGR192C	YCR012W	YKL152C	YGR254W	YAL038W
0,496	YCL040W	YBR196C	YGR240C	YKL060C	YDR050C	YJR009C	YCR012W	YKL152C	YGR254W	YAL038W
0,495	YCL040W	YBR196C	YGR240C	YKL060C	YDR050C	YJL052W	YCR012W	YKL152C	YGR254W	YAL038W

Table 1: The 36 pathways (see Figure 3) scored by the correlation function (see Formula 10). Dark shading of ORF IDs indicates up-regulation during diauxic shift, light shading indicates down-regulation, no shaing indicates unchanged expression. The pathways that are realized in the glycolysis before the diauxic shift are scored highest. The most distinguishing position of the pathway is the first enzyme characterized by the expression of ORF YGL253W which is up-regulated (in contrast to YCL040W and YFR053C). The ranking resulted in the following decreasing order of significance values: HXK2 (HXK1/GLK1) >> ENO2/YHR174w (ENO1/YGR254w) > PFK1/YMR205c (PFK2/YOR240c); the genes/ORFs TDH1/YJL052c, TDH2/YJR009w, TDH3/YGR192w show no influence.

Genes	HXK1									
	GLK1		PFK1						ENO1	
Scores	HXK2	PGI1	PFK2	FBA1	TPI1	TDH1	PGK1	GPM1	ENO2	PYK1
4,358		YBR196C		YKL060C	YDR050C	YJL052W	YCR012W	YKL152C		YAL038W
4,274		YBR196C	YGR240C	YKL060C	YDR050C	YJL052W	YCR012W	YKL152C		YAL038W
4,248		YBR196C		YKL060C	YDR050C	YJL052W	YCR012W	YKL152C		YAL038W
4,164		YBR196C	YGR240C	YKL060C	YDR050C	YJL052W	YCR012W	YKL152C		YAL038W
4,153		YBR196C		YKL060C	YDR050C	YJL052W	YCR012W	YKL152C	YGR254W	YAL038W
4,068		YBR196C	YGR240C	YKL060C	YDR050C	YJL052W	YCR012W	YKL152C	YGR254W	YAL038W
4,043		YBR196C		YKL060C	YDR050C	YJL052W	YCR012W	YKL152C	YGR254W	YAL038W
3,958		YBR196C	YGR240C	YKL060C	YDR050C	YJL052W	YCR012W	YKL152C	YGR254W	YAL038W
3,490		YBR196C		YKL060C	YDR050C	YJL052W	YCR012W	YKL152C		YAL038W
3,405		YBR196C	YGR240C	YKL060C	YDR050C	YJL052W	YCR012W	YKL152C		YAL038W
3,284		YBR196C		YKL060C	YDR050C	YJL052W	YCR012W	YKL152C	YGR254W	YAL038W
3,200		YBR196C	YGR240C	YKL060C	YDR050C	YJL052W	YCR012W	YKL152C	YGR254W	YAL038W

Table 2: The 36 pathways of Figure 1 reduced to 12 pathways via equivalencing TDH1, TDH2, and TDH3 (see Figure 3) ranked according to conspicuousness expression score (see Formula 7). Pathways containing the genes HXK1/YFR053c and GLK1/YCL040c receive the most significant scores as their change in expression level is more significant than the change of HXK2/YGL253w.

Genes	HXK1									
	GLK1		PFK1						ENO1	
Scores	HXK2	PGI1	PFK2	FBA1	TPI1	TDH1	PGK1	GPM1	ENO2	PYK1
25,143		YBR196C		YKL060C	YDR050C	YJL052W	YCR012W	YKL152C		YAL038W
23,555		YBR196C	YGR240C	YKL060C	YDR050C	YJL052W	YCR012W	YKL152C		YAL038W
20,737		YBR196C		YKL060C	YDR050C	YJL052W	YCR012W	YKL152C	YGR254W	YAL038W
19,740		YBR196C		YKL060C	YDR050C	YJL052W	YCR012W	YKL152C		YAL038W
19,326		YBR196C	YGR240C	YKL060C	YDR050C	YJL052W	YCR012W	YKL152C	YGR254W	YAL038W
18,078		YBR196C	YGR240C	YKL060C	YDR050C	YJL052W	YCR012W	YKL152C		YAL038W
17,663		YBR196C		YKL060C	YDR050C	YJL052W	YCR012W	YKL152C		YAL038W
16,065		YBR196C	YGR240C	YKL060C	YDR050C	YJL052W	YCR012W	YKL152C		YAL038W
15,852		YBR196C		YKL060C	YDR050C	YJL052W	YCR012W	YKL152C	YGR254W	YAL038W
14,367		YBR196C	YGR240C	YKL060C	YDR050C	YJL052W	YCR012W	YKL152C	YGR254W	YAL038W
13,859		YBR196C		YKL060C	YDR050C	YJL052W	YCR012W	YKL152C	YGR254W	YAL038W
12,438		YBR196C	YGR240C	YKL060C	YDR050C	YJL052W	YCR012W	YKL152C	YGR254W	YAL038W

Table 3: The 12 pathways of Figure 2 (see Figure 3) scored by the combined scoring function (see Formula 13). Similar to Table 2 pathways containing HXK1 and GLK1 receive the highest scores.

pathways with the same characteristics are generated and scored. Then, the p-value of the pathway under investigation is taken as the fraction of random pathways that achieve the same or a greater score. This p-value is an estimate of the fraction of false positives to be expected when assuming that the pathways under investigation are realized in the specific cell states represented by the current measurement. However, no notion of false negatives is represented in this figure.

According to this procedure, the best scoring pathways from Tables 1-3 yield the following p-values: 0.0606 for the conspicuousness score, and less than 0.0001 for both the correlation score and the combined score. The best scoring automatically generated pathways achieve values of 0.0365 for the conspicuousness score, and again less than 0.0001 for the other scoring schemes.

Discussion

One of the most popular techniques for the analysis of gene expression data is clustering. Clustering deduces a structure (the set or hierarchy of clusters) from the data without employing prior knowledge. This structure is, to a certain degree, always arbitrary, due to the high noise level of expression measurements and the lack of clear cluster boundaries, as shown in (Raychaudhuri, Stuart, & Altman 2000).

We propose a method that performs a detailed analysis of expression data with respect to biologically meaningful units, namely possible biochemical pathways. It is a general, automatic procedure to rate those pathways according to evidence from expression measurements, thereby allowing to test hypotheses that are relevant for drug target discovery and for guidance for further experimentation. Thus, the possible applications of our method go significantly beyond other known methods, e.g. clustering or function prediction. To our knowledge, the only other method that is capable of testing hypotheses on biological networks is that presented in (Friedman *et al.* 2000), which, however, does not yet make use of prior knowledge.

Interesting related work can be found in (Marcotte *et al.* 1999). Here, protein-protein interactions are predicted on a broad data basis, including gene expression measurements. Note that these interactions may overlap with, but are not identical to the edges in the graphs representing metabolic pathways. Interactions that are predicted by the methods of (Marcotte *et al.* 1999) can be fed into our method as hypotheses, and be re-evaluated in the context of the metabolic network. The same holds for experimentally detected interactions, which recently have been determined for yeast with a comprehensive Yeast2hybrid screen (Uetz *et al.* 2000).

Certain improvements are required to make our approach more useful. More work is required on the development of refined scoring functions. Most obviously, the graph structure of the pathways should be exploited. Another important point is to be able to take into account more complicated structures of measurements than linear time series. Also, we believe that our method will profit from advances in the definition of (regulatory) networks. We envision that, with further improvements and extensions implemented, the basic idea behind our approach will be useful for applications like the search for drug targets.

Acknowledgements

We thank our collegue Joannis Apostolakis for helpful comments. Part of this work has been funded by the BMBF under contract no. TargId 0311615.

References

Bairoch, A. 1999. The ENZYME data bank in 1999. *Nucleic Acids Research* 27(1):310–311.

Brown, M. P. S.; Grundy, W. N.; Lin, D.; Cristianini, N.; Sugnet, C. W.; Furey, T. S.; Ares, M.; and Haussler, D. 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences of the USA* 97(1):262–267.

Carr, D. B.; Somogyi, R.; and Michaels, G. 1997. Templates for Looking at Gene Expression Clustering. *Statistical Computing & Statistical Graphics Newsletter* 8(1):20–29.

Chee, M.; Yang, R.; Hubbell, E.; Berno, A.; an d David Stern, X. H.; Winkler, J.; Lockhart, D.; Morris, M.; and Fodor, S. A. 1996. Accessing genetic information with high-density DNA arrays. *Science* 274:610–614.

Chu, S.; DeRisi, J.; Eisen, M.; Mulholland, J.; Botstein, D.; Brown, P. O.; and Herskowitz, I. 1998. The Transcriptional Program of Sporulation in Budding Yeast. *Science* 282:699–705.

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.

D'haeseleer, P.; Wen, X.; Fuhrman, S.; and Somogyi, R. 1999. Linear Modeling of mRNA Expression Levels During CNS Development and Injury. In *Proceedings of the Pacific Symposium on Biocomputing '99*, volume 4, 41–52.

Eisen, M. B.; Spellman, P. T.; Brown, P. O.; and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the USA* 95(25):14863–14868. Genetics.

Ewing, R. M., and Claverie, J.-M. 2000. EST Databases as Multi-Conditional Gene Expression Datasets. In *Proceedings of the Pacific Symposium on Biocomputing '00*, volume 5, 427–439.

Fellenberg, M., and Mewes, H. W. 1999. Interpreting Clusters of Gene Expression Profiles in Terms of Metabolic Pathways. In *Proceedings of the German Conference on Bioinformatics '99*. Poster.

Friedman, N.; Linial, M.; Nachman, I.; and Pe'er, D. 2000. Using Bayesian Network to Analyze Expression Data. In *Proceedings of the Forth Annual Conference on Research in Computational Molecular Biology (RECOMB'00)*, 127–135.

Gerhold, D.; Rushmore, T.; and Caskey, C. T. 1999. DNA chips: promising toys have become powerful tools. *Trends in Biochemical Sciences* 24(281):168–173.

Golub, T. R.; Slonim, D. K.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, J. P.; Coller, H.; Loh, M. L.; Downing, J. R.; Caligiuri, M. A.; Bloomfield, C. D.; and Lander, E. S. 1999. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science* 286:531–537.

Heller, R. A.; Schena, M.; Chai, A.; Shalon, D.; Bedilion, T.; Gilmore, J.; Woolley, D. E.; and Davis, R. W. 1997. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proceedings of the National Academy of Sciences of the USA* 94(6):2150–2155.

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*, 141–147.

Liang, S.; Fuhrman, S.; and Somogyi, R. 1998. REVEAL, A General Reverse Engineering Algorithm for Inference of GeneticNetwork Architectures. In *Proceedings of the Pacific Symposium on Biocomputing '98*, volume 3, 18–29.

Lockhart, et al. 1996. Expression Monitoring by Hybridization to High-Density Oligonucleotide Arrays. *Nature Biotechnology* 14:1675–1680.

Marcotte, E. M.; Pellegrini, M.; Thompson, M. J.; Yeates, T. O.; and Eisenberg, D. 1999. A combined algorithm for genome-wide prediction of protein function. *Nature* 402(6757):83–86.

Michaels, G.; Carr, D.; Askenazi, M.; Fuhrman, S.; Wen, X.; and Somogyi, R. 1998. Cluster Analysis and Data Visualization of Large-Scale Gene Expression Data. In *Proceedings of the Pacific Symposium on Biocomputing '98*, volume 3, 42–53.

Mjolsness, E.; Mann, T.; Castao, R.; and Wold, B. 2000. From Coexpression to Coregulation: An Approach to Inferring Transcriptional Regulation among Gene Classes from Large-Scale Expression Data. In *Advances in Neural Information Processing Systems '99*, volume 12. To appear.

Ogata, H.; Goto, S.; Sato, K.; Fujibuchi, W.; Bono, H.; and Kanehisa, M. 1999. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* 27:29–34.

Okubo, K., and Matsubara, K. 1997. Complementary DNA sequence (EST) collections and the expression information of the human genome. *FEBS Letters* 403(3):225–229.

Okubo, K.; Hori, N.; Matoba, R.; Niiyama, T.; et al. 1992. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nature Genetics* 2:173–179.

Ramsay, G. 1998. DNA chips: State-of-the art. *Nature Biotechnology* 16:40–44.

Raychaudhuri, S.; Stuart, J.; and Altman, R. 2000. Principal Components Analysis to Summarize Microarray Experiments: Application to Sporulation Time Series. In *Proceedings of the Pacific Symposium on Biocomputing '00*, volume 5, 452–463.

Schomburg, D.; Salzmann, D.; and Stephan, D. 1990-1995. *Enzyme Handbook, Classes 1-6.* Springer.

Spellman, P.; Sherlock, G.; Zhang, M.; Iyer, V.; Anders, K.; Eisen, M.; Brown, P.; Botstein, D.; and Futcher, B. 1998. Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. *Molecular Biology of the Cell* 9(12):3273–3297.

Tamayo, P.; Slonim, D.; Mesirov, J.; Zhu, Q.; Kitareewan, S.; Dmitrovsky, E.; Lander, E. S.; and Golub, T. R. 1999. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the USA* 96:2907–2912.

Uetz, P.; Giot, L.; Cagney, G.; Mansfield, T. A.; judson, R. S.; Knight, J. R.; Lockshon, D.; Narayan, V.; ; Srinivasan, M.; Pochart, P.; Qureshi-Emili, A.; Li, Y.; Godwin, B.; Conover, D.; Kalbfleisch, T.; Vijayadamodar, G.; Yang, M.; Johnston, M.; Fields, S.; and Rothberg, J. M. 2000. A comprehensive analysis of protein-protein interactions in *saccharomyces cerevisiae. Nature* 403(6770):623–631.

Velculescu, V. E. 1999. Tantalizing Transcriptomes – SAGE and Its Use in Global Gene Expression Analysis. *Science* 286:1491–1492.

Zhu, J., and Zhang, M. Q. 2000. Cluster, Function and Promoter: Analysis of Yeast Expression Array. In *Proceedings of the Pacific Symposium on Biocomputing '00*, volume 5, 476–487.