A comprehensive study of the notion of functional link between genes based on microarray data, promoter signals, protein-protein interactions and pathway analysis

William Dirks and Golan Yona*
Department of Computer Science
and Center of Applied Mathematics
Cornell University

Abstract

It is commonly accepted that genes with similar expression profiles are functionally related. However, so far no clear distinction has been made as for the type of the functional link between genes as suggested by microarray data. Similarly expressed genes can be part of the same complex as interacting partners; they can participate in the same pathway without interacting directly; they can perform similar functions; or they can simply have similar regulatory sequences.

Here we embark on a rigorous study of the notion of functional link as implied from expression data. We analyze different similarity measures of gene expression profiles and assess their usefulness and robustness in detecting biological relationships by comparing the similarity scores with results obtained from databases of interacting proteins, promoter signals, and cellular pathways, as well as through sequence comparisons and pathway modeling. We also introduce new similarity measures we specifically developed for the analysis of expression data. These measures are based on statistical analysis and better discriminate genes which are functionally nearby and faraway.

With the optimized similarity measures we proceed to analyze other aspects of this data. Specifically, we introduce a method of inferring the type of relationship by correlating the expression data with all the other data sets. This method allows us to not only predict when genes are functionally related but also to suggest how they are related. We then cluster the data using clustering algorithms that are specially tailored to deal with noisy data. Finally we propose methods for assessing the significance of clusters and study the correspondence between gene clusters and biochemical pathways.

1 Introduction

The advent of microarray technology has allowed for the large scale analysis of gene expression profiles and is accompanied with a myriad of possible applications. Microarray analysis has been used to monitor the expression of genes as the cell undergoes through a normal physiological process such as the cell cycle [Spellman et al. 1998] in an attempt to determine which genes are involved in this process. It has been used to study differential gene expression patterns under different environmental conditions. These range from chemical treatments [Bammer & Fostel 2000] to cellular stress such as oxidative stress and sheer stress [McCormick et al. 2003, Yoo et al. 2003] and to organismal physiological activities such as dieting and exercise training [Diffee et al. 2003, Lopez at al. 2003]. Others have studied the association between different expression profiles and different cellular conditions. Such associations can help in developing assays that are designed to detect different types of cancers based on the expression patterns of genes [Yeatman 2003, Liu 2003]. Additionally, gene knockout experiments followed by microarray assays have been carried out to determine the role of different genes in cellular processes [Hughes et al. 2000].

The data collected from such studies is often used to suggest possible functional links between genes. Statistical methods to determine differential expression under different conditions give insight into the gene

*corresponding author. email: golan@cs.cornell.edu

functions, and it is purported that genes which are expressed similarly under different conditions or experimental setups are likely to have related biological functions.

While similarly expressed genes may suggest mechanisms of co-regulation, the type of functional link that can be inferred from this data may vary. Proteins might be co-regulated in the cell and expressed similarly because they are part of the same complex as interacting partners; they may participate in the same biological process or pathway without interacting directly; they may simply perform similar functions; or they may have similar regulatory elements without any obvious functional relationship. Additionally there may be other explanations which we are not aware of.

The supposition of a functional link has been explored in specific contexts but has not been thoroughly characterized. For example there is evidence that genes which are similarly expressed are more likely to code for protein products which directly interact with one another [Ge at al. 2001] and that genes with similar expression profiles tend to participate in the same cellular processes [Spellman et al. 1998]. Other studies indicate that genes with similar expression profiles seem to be functionally related based on identical descriptors in the abstracts of papers discussing them [Oliveros et al. 2000, Raychaudhuri et al. 2002]. However, to the best of our knowledge, no study has considered all of these elements simultaneously when analyzing expression profiles.

This type of study is difficult due the nature of microarray data. Expression data is noisy and in many cases unreliable. Many factors may affect the experiment and the measurements thus obscuring signals that might indicate relations between genes. In fact, to enhance the signal most studies employ clustering algorithms, hoping that the coordinated expression profiles of multiple functionally related genes would be more easily detectable than a single pair of similarly expressed genes. These groups of similarly expressed genes are assumed to have related biological functions. However, it is hard to predict what type of functional link exists between similarly expressed proteins, and so far no definitive characterization of the functional link has been obtained. Moreover, in the absence of precise measures for assessing the significance of similarity based on expression profiles, not only it is not clear what type of functional link exists between groups of similarly expressed genes, it is not even clear whether the genes are indeed truly co-regulated.

Our study has three main goals: (1) to develop a reliable and sensitive similarity measure between genes based on expression data, and to develop means for assessing its significance and sensitivity in distinguishing between genes which are functionally nearby and faraway; (2) to develop a method to assess the specific nature of the functional link between proteins as suggested by microarray data, and provide a way to measure the confidence of these proposed links; (3) to design unsupervised learning algorithms that are effective with noisy data (such as microarray data) and to group genes into co-regulated clusters while maximizing the information content of gene clusters.

We explore different measures and combinations of similarity measures between expression profiles, and present two robust clustering algorithms that are suited for extremely noisy data sets. We study their effectiveness in detecting functionally related genes and the correlation with experimentally verified functional relationships extracted from pathway data, protein-protein interaction data, sequence data, structure data, and promoter data.

This paper is organized as follows. We start with a brief overview of the data sets used in our analysis. We then review the different similarity measures used in our analysis and evaluate their performance with respect to the external data sets. Using the optimal similarity measure we then analyze the significantly similar expression profiles. We suggest a method for predicting the type of the functional link, algorithms for grouping noisy expression profiles into gene clusters and methods for assessing their significance and correlation with cell-wide processes and pathways.

2 Data

The main entities of our study are genes, their expression profiles and sets of gene relationships. These relationships define the types of the functional links that cause co-regulation and underlie the complex patterns of expression profiles that we observe in cells.

2.1 Expression data

We obtained the expression data from the publicly available Saccharomyces cerevisiae site [Spellman et al. 1998, Cho et al. 1998]. From this data set we extracted four time series of synchronized S. cerevisiae cells going through the cell cycle. In our analysis each ORF is represented by an extended expression profile derived by concatenating these time series together. We note that this data set has been normalized by Spellman et al. to correct for experimental variation between the different microarrays.

2.2 Sequence data

Our sequence data is the set of protein sequences in the Yeast sequence database with a total of 6298 proteins. Almost all (5902 out of 6298) of the ORFs in the expression data set can be mapped to genes in the Yeast sequence database through the ORF label. This is the subset of genes used in our study.

2.3 Relationships - the functional links

Four possible relationships are studied in this work: protein-protein interactions, pathway membership, promoter co-regulation, and sequence homology. Relationships based on protein-protein interactions and pathway membership explicitly determine the type of function link. Genes that interact or belong to the same pathway are strongly constrained, as co-expression might be essential to sustain the normal function of cells and tissues. On the other hand, genes that are regulated by the same promoters without an other apparent reason, are not necessarily functionally related (the co-regulation could have arose from a duplicated gene event proceeded by the lose or change of function of one of these genes). Or it is possible that the co-regulation may be due to the physical location of the genes, whose adjacency has been maintained throughout evolution without an explicit functional constrain. Finally, even when two genes are not related by either of the relations above, they may be still functionally related if they show significant sequence homology. This might be the underlying explanation behind co-expression if one protein is used as an alternative or as a backup protein (plan B) for the other one. Several examples are observed in known systems and are documented in the literature. For example, two-thirds of fly genes have no observable loss of function phenotype [Miklos & Rubin 1996].

2.3.1 Interaction data

A list of interacting proteins in yeast was compiled from three publicly available databases; Biomolecular Interaction Network Database (BIND) [Bader et al. 2001], Database of Interacting Proteins (DIP) [Xenarios et al. 2001] and BRITE [1]. After eliminating redundancy and merging GenBank entries with identical sequences, we were left with a total of 5636 unique interactions, and 3744 interacting proteins. Of these, 1665 proteins interact with a single protein. The most connected protein is a protein responsible of mRNA catabolism (YJR091C) which interacts with 289 other proteins.

2.3.2 Pathway data

We used the KEGG database [Kanehisa 1996] to assign proteins to pathways. The KEGG database contains information about pathways and other complexes in the cell (e.g. the Ribosome). As of November 2002, there were 127 pathways in the database, consisting mostly or only of enzymes. Enzymes are given by their EC (Enzyme Classification) numbers.

To assign proteins to EC families (and pathways) the yeast sequence database is matched against a composite non-redundant database that contains 933,075 unique sequence entries from 97 different databases (among which are SwissProt, TrEMBL, PIR, PDB, SCOP, DBJ, GenBank, REF, PATAA, PRF and the complete genomes of 77 organisms other organisms). Of the 6298 yeast genes, 832 can be assigned an EC number, either based on their annotation or the annotation of entries with identical sequences from the other databases.

Altogether, Yeast proteins can be assigned to 101 of the 127 pathways. The largest pathway (Purine Metabolism) consists of 88 yeast genes (39 unique enzymes). Of all pathways 63 have at least 3 genes in the yeast database. These are the pathways used in our analysis later.

2.3.3 Promoter data

We use known yeast promoter information from the Saccharomyces cerevisiae Promoter Database [Zhu & Zhang 1999]. This database contains yeast specific promoter information based on published results. As of November 2002, this database contained 237 genes with 113 transcription factors.

2.3.4 Homology clusters

The role of sequence homology in co-expression is believed to be through fail-safe mechanisms that evolved in the cell in the course of evolution. The simplest form of such mechanisms is redundancy, since it provides the cell with improved immunity to gene malfunction. This mechanism can evolve at random through a series of duplication events at the single gene level, or in some cases by duplicating groups of genes or even almost complete genomes. These duplication events may also preserve the promoter region that precedes the gene, thus further simplifying the process of generating a backup system that is concurrent with the main system.

To find all homology relationships we run an all-vs-all comparison of the yeast sequence database using BLAST [Altschul et al. 1990]. The genes are then clustered using the ProtoMap clustering algorithm [Yona et al. 1999]. The sequences cluster into 839 clusters and 3206 singletons.

2.4 Gene relationships as a graph

Each of the four relationships can be thought of in terms of a graph. In this graph each node represents a gene and an edge exists between the nodes if the genes are functionally linked. For protein-protein interactions, an edge exists between genes A and B if the proteins encoded by A and B interact. For the promoter data, an edge exists between A and B if the promoters of A are a subset of the promoters of B or vice versa. For the sequence data, an edge exists between two nodes if they are in the same homology cluster. For the pathway data, an edge exists if the two genes are in the same pathway. These subgraphs are compiled together into a single graph (called the **relation graph**) in which two nodes are connected if a known relationship exists between the two corresponding genes. Table 1 lists the number of genes and edges in each data set.

Relation type	#genes	#edges
Interaction	3592	5339
Sequence	3092	19074
Promoter	213	2439
Pathway	642	15789
Total	5079	41902

Table 1: **The relation graph**. Number of genes and edges (true relationships) in each data set. The sum of the number of edges in the four categories does not equal the total number of edges because two genes may have multiple types of edges between them, but this relationship is counted only once in the total set.

3 Methods and Results

This section is organized as follow. We first describe the similarity measures that are used in our analysis (section 3.1). To determine the most effective pairwise similarity measure, these measures are assessed in terms of their ability to detect meaningful pairwise relationships, and a reliable threshold is derived, above which similarities are likely to reflect true functional links (section 3.2). Next, we propose a method for predicting the type of the functional link for pairs of genes with significant pairwise expression similarity (section 3.3). The next element is a global study of pairwise similarities. Specifically, based on the significant pairwise similarities genes are grouped into clusters using a clustering algorithm that is especially effective in the presence of noise, as is the case for expression arrays (section 3.4). This way we can expand our predictions of novel gene relationships. Finally, we assess the biological significance of each cluster by validating the grouping with the experimentally verified data set (section 3.5).

3.1 Measures

To determine whether two genes have similar expression patterns an appropriate similarity measure must be chosen. We consider several scoring functions and their combinations. The most common similarity measures used for expression profiles are the Euclidean metric and the Pearson correlation. Other alternative measures have been proposed in the literature, among which are the jack knife correlation [Heyer et al. 1999], the Spearman rank correlation [D'haeseleer et al. 1997], and the dynamic programming scoring scheme described in [Qian et al. 2001]. Here we focus on the two most common measures, as well as on two variants that we introduce. Given two expression vectors **V** and **U** of dimension d, we define the following measures:

- The normalized Euclidean metric

$$Dist_{euc}(\mathbf{V},\mathbf{U}) = \sqrt{\frac{1}{d}\sum_{i=1}^{d}(V_i - U_i)}$$

- Pearson correlation

$$Corr(\mathbf{V},\mathbf{U}) = \sum_{i=1}^{d}\frac{(V_i - <V>)(U_i - <U>)}{\sigma_V \sigma_U}$$

- The **direction distance**. To define this distance we first replace each expression vector **V** with the direction vector $\hat{V}$ such that

$$\hat{V}_i = \begin{cases} +1 & \text{if } V_{i+1} - V_i < 0 \\ -1 & \text{if } V_{i+1} - V_i > 0 \\ 0 & \text{if } V_{i+1} - V_i = 0 \end{cases}$$

  The direction distance is then computed as the normalized l_1 distance between the corresponding direction vectors

$$Dist_{dir}(\mathbf{V},\mathbf{U}) = \frac{1}{2d}\sum_{i=1}^{d}|\hat{V}_i - \hat{U}_i|$$

- The **derivative distance**. This measure is similar to the direction measure, the difference being that $\hat{V}$ is defined as the derivative vector $\hat{V}_i = V_{i+1} - V_i$

Due to missing data the dimension of each vector (expression profile) can vary. Most studies deal with missing data using different variations of extrapolation methods [Alizadeh et al. 2000, Troyanskaya et al. 2001]. Here we decided to use only the available data, to avoid errors one might introduce by extrapolation. When comparing a pair of vectors the dimension d is defined as the number of features both vectors have in common, and the measures are normalized accordingly. In this study we considered only pairs of vectors whose dimension was greater than 11.

3.1.1 Statistical significance

The four measures of similarity defined above do not necessarily entail correlation or co-regulation and a-priori it is not clear when two genes should be considered similarly expressed. Since these measures have no explicit grounding in biology it is unclear whether pairs of genes which score well under these measures are functionally linked. Our goal is to determine, in a statistically meaningful way, which similarities are indeed due to an underlying biological phenomena.

Previously, methods to determine whether two samples are deferentially expressed in microarray experiments were based on the t-test using permutation methods to measure the significance of the t-statistic [Dudoit et al. 2002]. Ours is also a permutation based method. To assess the significance of a similarity score (according to any one of the measures) between two vectors, we permute the entries of one of the vectors and recompute the distance between the two vectors. This permutation step is repeated over and over again (60 times here) to determine the background distribution of distances between random vectors.

This distribution is different for each pair of genes. The background distributions for three selected pairs of genes are shown in Figure 1. These empirical distributions can be very well approximated by the normal distribution, as expected by the central limit theorem given the form of the similarity measures.

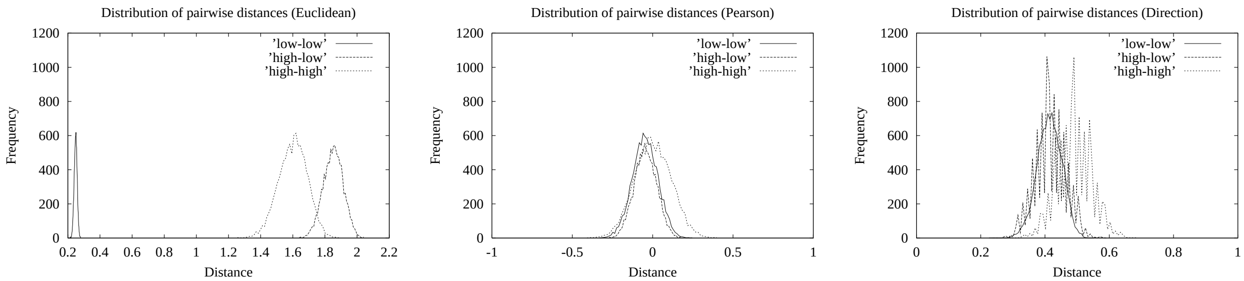

Figure 1: **Background distribution of distances for three pairs of genes** (a) for the Euclidean metric (b) for Pearson correlation (c) for the direction measure. For each metric we show the distributions for three pairs of genes, with different norms; one pair where the two genes have high norm (high-high), another pair where the two genes have low norm (low-low) and a third pair where one gene has high norm and the second has low norm (high-low). The dimensions of the vectors compared are 47 (low-low), 40 (high-low) and 67 (high-high). Despite normalization, all measures are sensitive to the norm (most pronounced with the Euclidean measure).

Given the average distance and the standard deviation of distances between the permuted vectors, the significance of the true distance is assessed in terms of the zscore (the distance from the mean in units of standard deviation). This self-calibrating approach is computationally simple and provides reliable measures as it adjusts to the specific "compositions" of the pair of genes compared. For example, if two genes are inactive throughout the series of experiments their expression similarity will be deemed insignificant when using this self-calibrating method, while otherwise they might be considered similar despite the lack of experimental support. Moreover, the method converts all distances to a uniform scale, independent of the norm and the dimension. This is especially useful for the analysis of noisy and partial expression data. Note that for the Euclidean distance and the direction distance highly negative zscores (small distances) are significant, while for Pearson correlation highly positive zscores are significant. To simplify our analysis we replace the zscores for the Euclidean and the direction distances with their negatives. We refer to the zscores after this transformation as the *normalized zscores*. The overall distributions of the normalized zscores (based on all pairwise comparisons) are shown in Figure 2. It should be noted that the distribution of Euclidean zscores has a slightly longer tail than the other distributions.

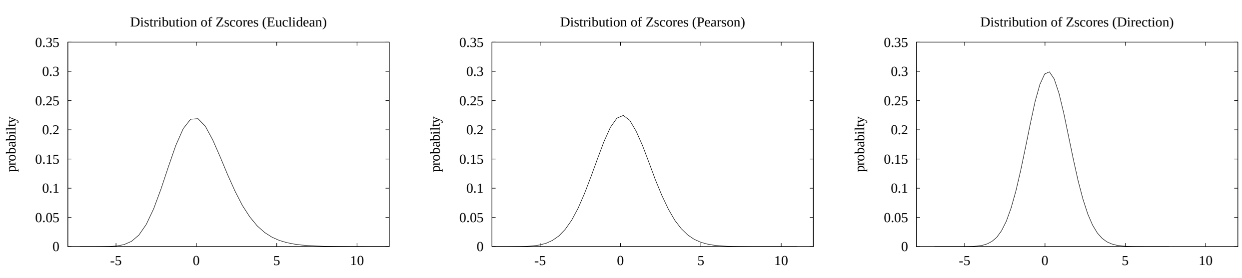

Figure 2: **Distribution of normalized zscores** (a) for the Euclidean metric (b) for Pearson correlation (c) for the direction measure.

While in most cases it is the genes with highly similar expression profiles that we are interested in, one might find the genes that are strongly anti-correlated, those with highly negative zscores, even more interesting. Three such examples are given in Figure 3. In the first case one of the genes is responsible for accelerating protein folding while the other is involved in glycerol metabolism. This may indicate a cellular response (slowing down protein production) to the environment when gycerol is abundant. The second example shows that chorismate aynthase and an integral membrane glucose transporter are inversely related. Chorismate is a molecule used to synthesize some amino acids. Therefore, this relationship might also suggests a cellular response to the environment, when glucose levels are high. In the last pair of genes one of the members is a ribosomal gene (60S ribosomal protein L11) while the other is an RNA helicase belonging to the DEAD-family. This family of RNA helicase has been implicated in processes which involve changes in RNA secondary structure such as translation initiation. Therefore it seems reasonable that there could be some compensating relationship between these two genes.

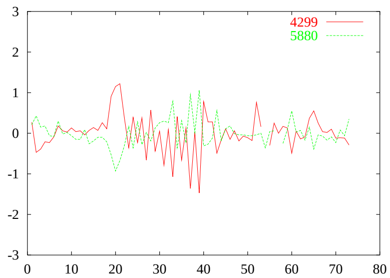

Figure 3: **Anti-correlated genes**. Three pairs of genes that are significantly anti-correlated. Yeast ORFs: YLR449W (4299), YPL206C (5880), YGL148W (2090), YHR096C (2699), YGR085C (2326) and YPL119C (5968)

3.1.2 Combined measures

It is not uncommon that for a given pair of expression profiles one of the measures defined above will indicate similarity while others will result in insignificant values. Actually, there are many thousands of pairs that are reported with significant zscore by only one method. Three such examples are given in Figure 4.

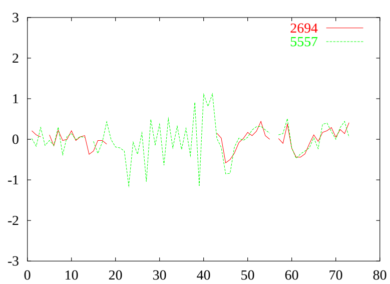

Figure 4: **Significant similarities reported only by one measure** (a) Euclidean metric (b) Pearson correlation (c) direction measure. Yeast ORFs: YHR091C (2694), YOR158W (5557), YJL170C (3103), YJR153W (3426), YKL024C (3639) and YPR036W (6123)

In some cases this might be due to a borderline case or a false "match" while in others it might be a true signal. To improve sensitivity and accuracy we tested several combinations of the three measures. Clearly, if all methods report a pair of genes as significantly similar then our confidence in this similarity increases. Yet, pairs that are reported as highly similar by only one measure will be considered as well. The weakest similarity that is reported by all three methods is shown in Figure 5. Despite the fact that this a borderline similarity according to all three measures, the total probability of observing such a pair by chance is very small.

We test different variations of combined measures: The first variation sums the zscores assigned by the individual measures. We introduce four such measures. The SumAll measure is the sum of the Pearson, direction, and Euclidean permuted measures; the EucPear measure is the sum of the Euclidean and Pearson permuted measures; the EucDir is the sum of the Euclidean and Direction permuted measures; the PearDir is the sum of the Pearson and Direction permuted measures. We also considered a weighted sum of the measures and the maximum similarity over all measures as well as a probability based measure; given the zscores of the individual measures we compute the total probability based on the normal approximations as

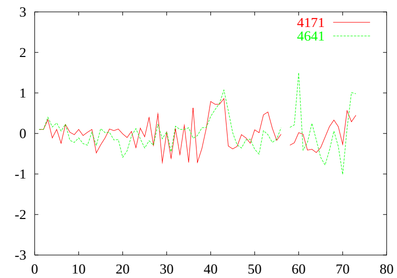

Figure 5: **Combining estimates from three different measures**. This pair (YLR324W and YMR170C) is considered borderline similar by each one of the three methods, however, when considered altogether, the similarity is significant. The function of YLR324W is unknown. YMR170C is an aldehyde dehydrogenase used in aldehyde metabolism and in beta-alanine biosynthesis. It is expressed in response to high osmotic stress. Interestingly, the cell can function even in the absence of this gene. This fact and the similar expression profiles suggest that YLR324W might play a similar functional role as YMR170C and can serve as a substitute.

in Figure 2 with the simplifying independence assumption:

$$p(zscore1, zscore2, zscore3) = p(zscore1/euclidean)p(zscore2/pearson)p(zscore3/direction)$$

Since all measures are based on the normalized zscores, the larger they are the more similar are the expression profiles. Therefore we refer to all these measures as *similarity measures* from here on. We now proceed to evaluate the sensitivity and selectivity of the individual measures (the original and the permuted zscore-based measures) as well as of the combined measures.

3.2 Pairwise Analysis

To assess the performance and determine the information content of all measures the pairwise similarities are examined with respect to the data sets described in section 2. Specifically, for each measure we compute all pairwise expression similarities and sort them in decreasing order of significance. We then plot the number of known relationships detected as a function of the total number of pairwise similarities as we scan the sorted list from the top (Figure 6). This curve is similar to a ROC curve, however, since it is very hard or even impossible to determine which pairs of genes are totally unrelated, there are no experimentally verified false positives. Therefore we plot the number of true positives vs. the total number of pairwise similarities instead of vs. the number of false positives.

As Figure 6 indicates, all measures contain some information about the true relationships, when compared to what is expected by a random guess (see next section). Of the unnormalized measures, the Pearson correlation is the most effective one. Surprisingly, despite its popularity in studies of expression arrays, the original Euclidean measure is the worst one (the performance was even worse without normalization or with other normalizations). However, when using the zscores, it becomes the best individual measure. Of all measures, including the combined measures, the EucPear measure seems to be the most effective one and is chosen as our main similarity measure for subsequent analysis.

3.2.1 Baseline performance

An important element of any learning system is the baseline performance; the performance one would expect to get by a random guess. In our case, the baseline performance would be the number of interactions e that are expected to appear at random. There are two ways to estimate this number. In the **uniform random setup** we assume all the edges in the relation graph are placed uniformly, at random, between genes. Given this setup, one can easily compute e analytically. Specifically, if N is the total number of genes in our set and T is the total number of true relationships (total number of edges), then the number of true edges that are expected to occur in a set of n randomly chosen edges is $e = n\frac{T}{N(N-1)/2}$. The results of the uniform random setup are shown in Figure 6 (labeled "expected").

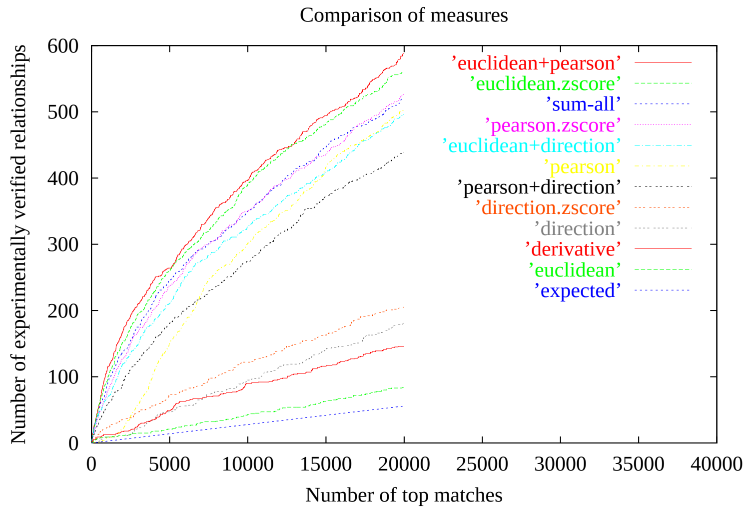

Figure 6: **Performance evaluation.** For clarity, the labels are ordered according to the performance of the corresponding measures.

While computationally very simple, this random setup does not preserve the structure of the gene network. Clearly, this network is far from being random, as some genes are more connected than others, either because they are truly functionally related to multiple genes or because they were better studied than others. Therefore, we consider also the **structure preserving random setup**. Under this setup, the graph structure is frozen (in other words, the edge structure is kept intact and the connectivity of each node remains the same) while the expression profiles associated with the different nodes are shuffled. Given the shuffled graph, we approximate the number of expected relations by computing the number of true edges that exist between the n most similar pairs of expression profiles. This process is repeated hundred times, and the average number of relationships and the standard deviation are computed. Surprisingly, the results are very similar to those obtained with the uniform random setup (results not shown), suggesting that the network connectivity is fairly low, and that the theoretical estimates based on the simplified model are accurate enough.

3.2.2 Setting the threshold of significance

In a typical application a biologist has a particular gene of interest she would like to analyze. By looking at the most similar genes one hopes to get information that would help to characterize this gene of interest. Indeed, the most common tool for protein analysis is a database search, in which the protein sequence is compared with all proteins in the database and its properties are extrapolated from the properties of neighboring sequences. Here we ask if a similar approach can be adopted for expression data.

When measuring the similarity between two expression profiles one would like to know when the similarity can be considered meaningful. The statistical measures we employ unify the different measures into a single scale that is independent of the norm, dimension and vector composition. At high zscore values one expect to find pairs that are truly functionally linked, while near zero zscore values are clearly dominated by unrelated pairs. However, while it is clear that the higher the zscore the more significant the similarity, one would like to have a more precise measure of the probability that the two genes are indeed functionally linked. More specifically, it is useful to have a threshold above which it most likely that the pair of genes are indeed functionally related.

To determine a sensible significance threshold (transition point) for each of the similarity measures we plot the cumulative distribution of pairwise similarities as a function of the zscore (Figure 7). This distribution follows a sigmoidal curve; as the zscore decreases, the number of pairwise similarities increases, however, this increase is not linear, and we observe steep increase below a certain zscore value (this value is different from one method to another), where the third derivative of the sigmoid curve is zero. This increase is explained by

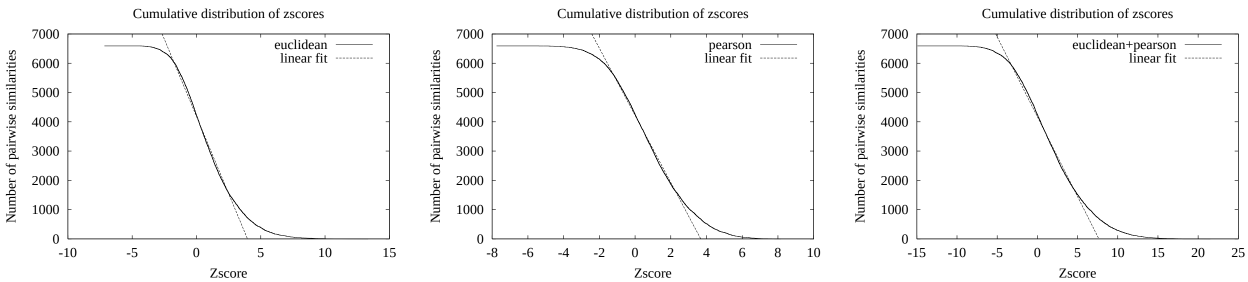

Figure 7: **Setting the significance threshold for pairwise similarity measures**. The cumulative distribution of zscores is plotted for (a) the Euclidean metric (b) Pearson correlation (c) the Euclidean+Pearson measure. Note the sigmoidal shape of the curve. For each measure the significance threshold is defined as the x-interaction point of the linear fit (see text for details).

the abundance of random matches between expression profiles as the zscore threshold loosens. The transition point reflects the transformation from mostly true relationships to a majority of random matches. Therefore, to maintain a favorable signal to noise ratio we set the significance threshold near the transition point. One way to do this is to fit the linear portions of the sigmoidal curve around the transition point with the best fit straight lines and use the intersection point of these lines as the cutoff. A more stringent threshold can be set at the point where the linear fit to the steep part of the curve intercepts the x-axis. This more stringent threshold is the cutoff we chose to use[1]. Two expression profiles with similarity zscore that is greater than this threshold are considered significantly similar.

We note that in this analysis we restrict our attention to a meaningful subset of the genes. Namely, the subset of genes for which we have some information about their protein-protein interactions, pathway membership, and sequence homology. This set consists of 490 genes with extensive information about their gene relationships (we did not consider promoter information as there is very little data about promoters and its inclusion reduced the set to a mere 63 genes). However, it should be noted that even for this subset the experimental information is partial as not all possible pairwise relationships are documented and of those that are, some might be false (see discussion). When the procedure was repeated with the full data set a similar cutoff was obtained.

3.2.3 Information content

To assess the information content of the optimal measure we consider all pairwise similarities that are above the significance threshold (a different threshold is selected for each method, based on its unique distribution of pairwise similarities) and compute the total and relative number of true relationships that are detected. These numbers are compared to the number of true relationships one would expect to find by chance. The expected number of relationships is computed using the two random setups described in 3.2.1. The results are summarized in Table 2. Note that the two random setups give almost identical results (suggesting again that the theoretical estimates are accurate enough). As the table indicates, we detect a substantial number of true relationships, compared to the number of such relationships that are expected to happen by chance. Since the experimental data is partial, this is a lower bound on the number of true relationships that can be detected with this method.

Interestingly, of the different types of true relationships, the one type that is most strongly correlated with expression data is pathway membership (Figure 8). The second most substantial source of information is, surprisingly, sequence homology. Ratio-wise, sequence homology is the strongest signal as it appears almost 4 times more frequently than one would expect by chance. Promoter-based relation is the second strongest signal ratio-wise, appearing almost 3 times more often than expected by chance.

[1]One can also use Bayesian decision theory to set a threshold at the point where $P(true/zscore) = P(false/zscore)$. However, analysis of the two distributions resulted in a more relaxed threshold, as the distributions highly overlap, and was rejected. Note that many of the possible relationships between genes haven't been verified experimentally, therefore it is hard to characterize the true distributions of true and false relationships.

Measure	Threshold	#edges above threshold	#true-edges above threshold	expected	simulation
Euclidean+Pearson	7.7	5907	677 (11.5%)	343	346 (33)

Table 2: **Information content of the EucPear similarity measure**. The total number of edges in the subset of 490 genes is 113492 of which 6591 are true, experimentally verified relationships. The last two columns report the number of true edges one would expect to detect by chance, estimated using the uniform random setup (expected) and the structure preserving random setup (simulation). Number in parentheses in last column are the standard deviations observed in the simulation.

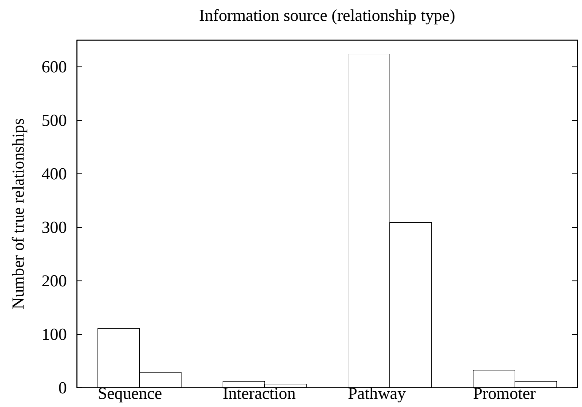

Figure 8: **Information source**. Breakup of the 677 pairs of truly related genes that have significant expression similarity (measured with the Euclidean+Pearson measure) according to the type of relationship. Note that some pairs of genes are related by more than one type of a functional link, therefore, the sum of all true edges exceeds 677. For each type we also compute the number of such relationships that are expected to occur by chance.

3.3 Predicting the type of the functional link

A significant similarity between two expression profiles suggests a functional link between the two corresponding genes. Given such a pair we would like to determine the type of the underlying relationship. We start by comparing the sequences of the two genes to the sequences of all known pairs of related genes (the relation graph described in section 2.4). Our assumption is that if the pair is significantly similar to another pair of genes that are known to be related then they might be related as well with the same type of relationship. Our assumption is based on the hypothesis that biological systems must be redundant to some extent. Since biological systems can be non-deterministic, irregular and failure/error-prone one might expect some kind of an error-correcting mechanism to exist to ensure successful operation and to sustain life. Redundancy is perhaps the simplest mechanism, that is essentially "implemented" through gene duplication events. While unplanned, these random duplications could have increased the robustness of certain genomes that were better fit to survive under different conditions. Indeed, our hypothesis is supported by an analysis of the genomes of even the smallest organisms that were fully sequenced, Mycoplasma Pneumoniae (MP) with 677 genes and Mycoplasma Genitalium (MG) with 470 genes. Using the ProtoMap clustering algorithm [Yona et al. 1999] the 677 genes in the MP genome cluster into 411 sequence clusters (approximated redundancy level of almost 40%) and the 470 genes in the MG genome cluster into 338 clusters, a redundancy level of 28%. For yeast, the redundancy level is about 35%. This is a lower bound on the number of gene duplication events considering possible ancient gene duplications.

The sequence similarity of a given candidate gene pair with a pair of functionally linked genes is estimated as the maximal geometric average of the pairwise similarities of the constituent genes. For example, given a pair 1 of possibly linked genes A and B, and a pair 2 of genes C and D that are known to interact then the **support** of pair 1 with respect to pair 2 is determined as the geometric average of the homology scores of A to C and B to D (or the reverse order) conditioned that both are significant.

In our analysis we consider all four possible types of relationships between genes. For each type of relationship we report the maximum support that we can mine in the data sets for the specific pair of genes in question. To determine if the support scores are significant we again run a simulation study. As in the structure preserving random setup described in section 3.2.1, we shuffle the expression profiles between the nodes (genes) while leaving the graph structure and the sequence associated with the nodes unchanged. We then determine the maximum support for each pair of genes whose expression similarity is above the threshold and compute the average and the standard deviation. This simulation breaks the relationship between a gene and its expression pattern and will randomize which genes are considered similar by expression thus generating a background distribution for the maximum support scores.

Pair type	Gene1	Gene2	Definition1	Definition2	Predicted Link
Predicted	YPR137C-A	YHR214C-C	Ty1 Element Transposition DNA recombination, proteolysis and peptidolysis	Ty Element Transposition,	Interaction
Support	YOL103W-B	YPL257W-A	Ty Element Transposition DNA recombination, proteolysis and peptidolysis	Ty Element Transposition,	(Genes interact)
Predicted	YAR009C	YOL103W-B	RNA-directed DNA polymerase activity, peptidase activity, ribonuclease activity	Ty element transposition, DNA recombination, proteolysis and peptidolysis	Interaction
Support	YBR012W-B	YPL257W-A	RNA-directed DNA polymerase activity, peptidase activity, ribonuclease activity	Ty element transposition	(Genes interact)
Predicted	YLR355C	YOR108W	leu+val+ile biosynthesis	leu+val+ile biosynthesis	Pathway
Support	YLR355C	YNL104C	leu+val+ile biosynthesis,	leu+val+ile biosynthesis,	(Same pathway)
Predicted	YLR028C	YLR058C	purine biosynthesis protein ADE17, de novo biosynthesis of purine, his+purine+pyrimidine biosynthesis	Serine hydroxymethyltransferase, formylTHF biosynthesis, pathways of chorismate, folic acid biosynthesis serine and glycine biosynthesis	Pathway
Support	YMR120C	YLR058C	purine biosynthesis protein ADE17, de novo biosynthesis of purine, his+purine+pyrimidine biosynthesis	Serine hydroxymethyltransferase, formylTHF biosynthesis, pathways of chorismate, folic acid biosynthesis serine and glycine biosynthesis	(Same pathway)
Predicted	YER065C	YIR031C	Isocitrate lyase, glyoxylate cycle serine-isocitrate lyase pathway	Malate Synthase 2 allantoin catabolism	Promoter
Support	YER065C	YNL117W	Isocitrate lyase,glyoxylate cycle serine-isocitrate lyase pathway	Malate synthase 1, glyoxylate cycle malate synthase activity	(Same promoter)
Predicted	YBR018C	YDR009W	UDP uridylyltransferase (EC 2.7.7.12) lactose degradation galactose metabolism colanic acid biosynthesis	GAL3 galactokinase (EC 2.7.1.6) transcription regulation, galactose metabolism	Promoter
Support	YBR018C	YBR020W	UDP uridylyltransferase (EC 2.7.7.12) lactose degradation galactose metabolism colanic acid biosynthesis	GAL1 galactokinase (EC 2.7.1.6) lactose degradation galactose metabolism colanic acid biosynthesis	(Same promoter)

Table 3: **Predicting the type of the functional link**. For each pair of genes with significantly similar expression profiles we search for the maximal support in the set of true edges. For example, genes YER065C and YIR031C are predicted to have the same promoter as their maximal support comes from the gene pair YER065C and YNL117W that are known to have the same promoter. Since the support genes are defined based on sequence similarity the definitions are often very similar to those of the query genes, however, in many cases the relationship that exists between the support genes has not been established for the query genes (for validation results see Table 4).

Table 3 lists gene pairs which have significant expression similarity for which we can predict the type of functional link with high confidence score. Although very simplistic, this method does seem to suggest interesting types of functional links between genes. When examining Table 3 a few interesting observations can be noted. For example, while YLR355C and YOR108W are not in our pathway database they are predicted to be in the same pathway. Indeed the annotations suggest that they are in the same pathway, lending credence to our method. With all the predictions the support genes provide a model of the predicted relationship. For example, the genes YPR137C-A and YHR214C-C are likely to interact in the same way as the known interacting genes YOL103W-B and YPL257W-A.

To validate our method, we also tested our predictions on the dataset of true relationships. For each pair, the corresponding genes were excluded from the list and the same procedure described above was used to predict the type of the functional link. Prediction was made only if significant sequence homology was detected with one or more of the remaining functionally related pairs. The results are summarized in Table 4. Surprisingly, almost 80% of the pathway relationships can be predicted reliably (99.4% accuracy) even when the pair is eliminated from the edge set. Also surprising is the level of sequence redundancy or duplication. Note that even promoter-based relationships are detected effectively, when a prediction can be made (81% accuracy, compared to 4.2% expected by chance, where the other 19% are possibly screened by other types of underlying functional links that exist between the same pairs of genes, e.g. pathway membership).

These results are surprising as they suggest a very high level of active redundancy. It is known that biological systems can employ homologous genes to perform the same function, under different cell conditions and environments. For example, aceB and glcB both are Malate synthases but are deferentially used depending on the presence of glyoxylate [Rison et al. 2002], and it is suggested that at any given time, a single gene is being used. However, here the expression arrays suggest that there are some groups of homologous

	Total	Predictions made	Predicted correctly (%)	Expected (%)
Promoter	2439	650	527 (81.1%)	37 (5.7%)
Interaction	5339	748	643 (86.0%)	94 (12.5%)
Pathway	15789	12627	12548 (99.4%)	4575 (37.0%)
Sequence	19074	18636	18621 (99.9%)	8336 (44.7%)

Table 4: **Predicting the type of the functional link - validation**. This test is independent of expression data. Each true edge is eliminated in turn from the edge set and the procedure is run with all remaining edges in search for maximal support (see text for details). A prediction is made only if significant sequence similarity is detected with a true edge.

genes that are co-expressed, therefore all or some of them are available at the cell's disposal to perform the designated function at any given time. At this point it is not clear if indeed the cell maintains all these genes for possible use (thus supporting our hypothesis), or if there a subsequent mechanism in the cell that prunes this set of genes to a single one, the one that is relevant to the cell conditions. Indeed, there are many factors that can affect a gene product from the moment it is being transcribed to the moment it is being translated and folded into a stable protein product, and the final protein level might be highly uncorrelated with the expression level. However, the overwhelming evidence for redundancy makes it unlikely that all of these gene duplicates that are initially co-expressed are indeed suppressed at a later stage. In other words, the results suggest that some kind of a fail-safe mechanism is being implemented in practice.

3.4 Cluster Analysis

While the pairwise relationships account for the individual similarly expressed gene pairs, they are not sufficient to analyze the complexity of the whole gene set. It is the coordinated nature of the cell regulation mechanism that differentiates expression analysis from sequence or structure analysis. The degree or extent of the similarity under different conditions will determine the co-dependency of gene pairs. However, on another level, the coordinated pattern of expression hints at more global processes, the nature of which cannot be determined based on individual gene pairs.

Cluster analysis is an unsupervised technique commonly used to detect groups of similarly expressed genes. There is a large number of papers in the literature describing different clustering techniques and their application to expression data. However, expression data is very noisy and in some cases unreliable, which might greatly affect clustering algorithms that are based purely on the pairwise distances between expression profiles. Here we present two clustering algorithms that are designed to work effectively with noisy data and maximize the signal to noise ratio.

Our goal is to cluster the data into meaningful groups under the constraint that the information content of clusters is maximized. We propose two different approaches to tackle this problem. The first (strict clustering) uses the original vector representation and the EucPear similarity measure, and attempts to maximize the information content by maintaining a certain level of confidence about all pairwise relationships in a cluster. The second (iterative clustering) diverges from the classical representation and gains its power by employing a dynamic representation that samples the complete space of expression profiles to determine the whereabouts of each individual expression profile. We also provide the means to evaluate the clusters and assess the effectiveness of each method. Our benchmark can be used to evaluate any other clustering algorithm in an objective manner.

3.4.1 Strict clustering

In strict clustering, clusters are constrained to be strongly connected. A threshold is chosen first (as in section 3.2.2) and for every cluster we require that the average similarity of a node to every other node in that cluster is more significant than the threshold chosen. This procedure generates clusters of very high quality in the sense that we have high confidence in the cluster membership of each individual gene. The algorithm can be viewed as a heuristic approximation to the max-clique problem, where the resulting clusters are in essence pseudo-cliques (pseudo, as the average similarity of each member with all other members is above the threshold, as opposed to all the individual pairwise similarities).

Strict clustering is implemented by first considering the node with the maximal number of neighbors and grouping it and all its neighbors into one candidate cluster. The node with the worst average within-cluster similarity is then considered for elimination. If its average similarity is below the significance threshold then the node is thrown out. The process is continued until all nodes satisfy the above condition. Other nodes are considered for inclusion if their average similarity score and the average cluster similarity score exceed the threshold, adding the maximally similar node at a time. This procedure is repeated with the remaining nodes until all nodes are exhausted. The clustering procedure ends when every node is in a cluster (possibly a singleton).

3.4.2 Iterative clustering

Our iterative clustering method is an hierarchical algorithm based on an iterative modification of data representation until convergence at each level of the hierarchy [Dubnov et al. 2000]. The algorithm begins with the proximity matrix of the pairwise similarities[2]. Pairwise similarities that are undefined due to missing data are approximated. If the distance between genes A and B is unknown then this number is approximated by $min\{dist(C,B), dist(D,A)\}$, whichever one is more significant, where C is the closest gene to A whose distance to B is known and D is the closest gene to B whose distance to A is known.

After the preprocessing of the proximity matrix the clustering procedure precedes in two steps. In the first step each row is normalized so its Euclidean norm is one[3]. Next, a new estimate of the distances between genes is calculated from the proximity matrix. Specifically, the distance between genes i and j is estimated by the distance between rows i and j of our proximity matrix[4]. This procedure is repeated until the matrix converges to a binary matrix containing only zeros and some positive constants. In most cases this final matrix can be permuted to a block diagonal matrix with two blocks. If this is the case the two blocks are considered as a split of the data into two clusters. If the final matrix converges to more than two blocks one of the blocks is chosen arbitrarily to be one of the clusters and the remaining genes are classified to the other cluster. This process is repeated with each one of the two clusters generated. A cross-validation protocol is employed [Dubnov et al. 2000] to assess the stability and the prominence of the split and splitting is stopped when the validation score (the percentage of points on which the two validation sets agree) falls below 0.6 (although all splits are stored for further evaluation, as described in the next section) or when the size of the cluster falls below 4.

This clustering algorithm has a major advantage that makes it most appealing in our case; it is highly robust to noisy data (see [Dubnov et al. 2000]). By repeatedly estimating distance profiles, the geometry of the gene space is being re-evaluated, and fluctuations are being averaged out. One way of explaining the success of this method is the fact that $\binom{n}{2}$ constraints (the set of distances with respect to the complete set of genes) are being considered when evaluating the distances between pairs of genes. Thus, even when the direct distance between genes i and j is "corrupted", the distance profiles of these genes with respect to all other genes constrain the relative position of genes i and j in the gene space, so that their true distance can be measured indirectly. These distance profiles themselves can be noisy as well (assuming random noise), and yet induce reliable distances, even for small references sets as in demonstrated in [Dubnov et al. 2000].

3.5 Assessment of clusters

While clustering algorithms can be very effective in discovering patterns in unlabeled data, when applied simple-mindedly, they can lead to false conclusions. It is important to note that the application of any clustering algorithm will result in some partitioning of the data into groups, and the choice of the clustering algorithm may greatly affect the outcome. Moreover, clustering algorithms are likely to become trapped in a local minimum, and their output may vary a great deal, depending on the starting point. Most importantly, determining the "correct" number of clusters (classes) within the data set is considered one of the most difficult problems in unsupervised learning, and one usually preset this number arbitrarily, or select it based on hierarchical cluster analysis. However, in the absence of a natural criterion by which to stop the process of splitting or merging data points, again, usually an external, user-specified parameter is used (e.g., when the distance between clusters exceeds a certain threshold). Consequently, the recovered structure does not

[2] To compare the effectiveness of this method to the previous one only genes which are members of non-singleton clusters when using strict clustering were considered.

[3] Other normalizations can be applied but without any apparent significant impact [Dubnov et al. 2000]

[4] The choice of the metric used to compare the rows can vary. Here we employed the Euclidean metric.

necessarily fit the "real" structure of the data, and its validity is questionable. Finally, is it not clear how to compare clusterings which are obtained by different approaches, in the absence of an apparent general quality measure. Known quality measures are usually application specific, but no obvious measure exists for gene expression data. Therefore, given a model, it is hard to evaluate its validity and its ability to generalize to new samples.

Here we suggest two criteria to assess the quality of different clustering models. The first is based on internal index of validity, specifically the complexity of the model and its ability to explain new instances. The second is based on external index of validity, specifically, the correlation with the experimental data. We corroborate the evidence as provided by the two measures to select the best clustering.

Next we describe these two indices in detail. Our procedures were applied to test and evaluate the clusterings that are produced with the iterative clustering algorithms at different cross-validations levels (starting from 0.98 and ending at 0.6) and the clustering that is produced with the strict clustering algorithm.

3.5.1 Internal index of validity - the Bayesian approach

We consider each possible clustering (partition) of the data as a unique model, denoted by h. Posed in probabilistic terms, the problem of selecting the best model (best clustering) can be phrased as a problem of maximizing the posterior probability of the model given the data, i.e.

$$P(h/D) = \frac{P(h)P(D/h)}{P(D)}$$

The optimal hypothesis h^* is the one that maximizes $P(h/D)$ (minimizes $-P(h/D)$) or equivalently

$$h^* = \arg\min_h[-P(h)P(D/h)] = \arg\min_h[-\log_2 P(h) - \log_2 P(D/h)]$$

Note that this Bayesian approach resembles the minimum description length principle. The description length of a given **model (hypothesis) and data** is defined as the description of **the model** plus the description of the **data given the model**. This heuristic calls for selecting models that minimize the overall description length, thus compromising between complex and overly specific models and simple models that are too general and fail to provide reliable and accurate description of the data. By the Shannon theorem $-\log_2 P(x)$ is related to the shortest description of a string x and therefore the first term can be considered as the model description length, while the second term $\log_2 P(D/h)$ is the likelihood of the data given the model and can be considered as the description of the data given the model (this is a measure of the uncertainty that is left in the data, given the model).

The second term is simply the minus log-likelihood of the data given the model

$$\log_2 P(D/h) = \log_2 \prod_{i=1}^{n} p(x_i/h) = \sum_{i=1}^{n} \log_2 p(x_i/h)$$

In our hard clustering model, each sample is classified to a single cluster w_i, whose prototype is estimated as the average of all samples that are classified to that cluster. The prototype is assumed to be corrupted by some normally distributed noise, with mean zero and standard deviation that can be estimated from the data. Thus, each cluster is parametrized by a d-dimensional vector of means and standard deviations at the different time steps. The probability of an individual sample to belong to cluster w_i is given by the product of the d normal distributions characterized by the the individual means and standard deviations at each time step

$$p(\vec{x}/w_i) = p((x_1, x_2, .., x_d)/w_i) = p(x_1/\mu_{i1}, \sigma_{i1})p(x_2/\mu_{i2}, \sigma_{22})...p(x_d/\mu_{id}, \sigma_{id})$$

where each $p(x_j/\mu_{ij}, \sigma_{ij})$ is a one-dimensional normal distribution (j being the time step). This equation can be easily generalized to the soft clustering case, where each gene has a nonzero probability to belong to each one of the clusters.

The first term $-\log_2 P(h)$ (the approximated model description length) is estimated by measuring the uncertainty we have about the model. Since the standard deviations within clusters are not explicitly used in the clustering algorithm, we can assume that each cluster in our model is characterized by only d parameters (the prototype means), and we have to account for the uncertainty in estimating the value of each one. Since each parameter is estimated as a sum of i.i.d. random variables drawn from an arbitrary probability distribution with mean μ and variance σ^2 (estimated from the data) as in $\theta = \frac{1}{n}\sum_{i=1}^{n} X_i$ then by the central limit theorem we know that as $n \to \infty$ then $\theta \to \mu$ with standard deviation $\sigma_\theta = \frac{\sigma}{\sqrt{n}}$. This standard deviation (in bits) is a measure for the uncertainty in the value of the estimator. The overall uncertainty is a sum over the uncertainties of all parameters for all clusters.

In Figure 9 we plot the description length of the different clustering models we produce; with strict clustering and with the iterative clustering (using thresholds in the range of 0.6 and 0.98)[5]. As the graph indicates, the description length decreases as the threshold of the iterative clustering decreases (note the major decrease around 0.83) until it reaches a plateau around 0.75. Without any further information one might suggest 0.75 as the optimal threshold, were the plateau starts. However, there is no clear reason to prefer 0.75 over 0.6 for example, based on the MDL principle alone, and each one of these clusterings is a viable model. In the next subsection we address the problem of reducing this set of models to a single one.

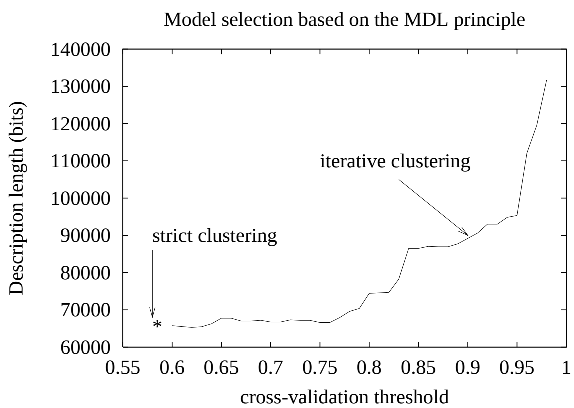

Figure 9: **Model selection based on the posterior probability/MDL principle**. We tested all clustering models that we produced with the strict clustering algorithm and the iterative algorithm (using a variable threshold). For each model we computed the overall description length (the sum of the description length of the model **and** the data given the model). This is proportional to the posterior probability of the model (see text). As the graph indicates, all models with thresholds ranging from 0.6 to 0.75 can be considered good models, as well as the model generated with the strict clustering algorithm (denoted by "*").

3.5.2 External index of validity - the relation graph

While the internal validity index described above can help in discerning better models it does not necessarily produce clusterings that best describe the experimental data. Moreover, as this principle is a heuristic, it can only guide the designer by pointing to more effective models, however usually the designer is left with several models, all seem to describe the data equally well. Indeed, in our case, all models with thresholds ranging from 0.75 to 0.6 are good models according to this principle. However, the clustering profile changes quite drastically, ranging from 158 clusters (at threshold 0.75) to 319 clusters (at threshold 0.6).

To reduce this set to a single model, we re-evaluate the clustering results and select the most significant one given the experimental data using the relation graph described in section 2.4. Specifically, we count the number of true edges within each cluster and estimate the probability to obtain this partition of edges amongst clusters by chance. In other words, the significance of the clustering is determined by how much this partition deviates from the partition one would expect to get by assigning edges to clusters at random.

Formally, we are given a cluster set $\mathbf{C} = C_1, C_2, ..., C_k$ with $n_1, n_2, .., n_k$ genes and $e_1, e_2, ..., e_k$ edges, that partitions the total set of true edges such that t_i edges fall in cluster i ($1 \leq i \leq k$). In addition we have

[5]The optimal clustering is likely to be composed of clusters at varying thresholds, however, this means an exponentially large set of possible clusterings to test. To make this problem tractable we reduced the set to those with a uniform threshold.

e_{cross} edges across clusters, of which t_{cross} are true. The probability to obtain this partition by chance is

$$P(t_1, t_2, ..., t_k, t_{cross}/\mathbf{C}) = \binom{T}{t_1 t_2 ... t_k t_{cross}} p_1^{t_1} p_2^{t_2} ... p_k^{t_k} p_{cross}^{t_{cross}}$$

where $E = \sum_i e_i + e_{cross}$ is the total number of edges, $T = \sum_i t_i + t_{cross}$ is the total number of true edges, and $p_i = e_i/E$ (we refer to set $\{p_i\}$ as the **random type**, and to set $\{t_i\}$ as the **sample type**).

However, this probability in itself is not enough to ascertain the significance of the overall partition of genes into clusters. To assess the significance one needs to know how probable it is to get partitions that are less likely by chance than the one in question (in other words, one needs to know the weight of the tail of the distribution). Alternatively, one can use a concentration of measure result such as standard deviation, when known, to determine significance. Given a significance measure, one can use it as an external validation index and pick the clustering with the minimal total random probability (maximum significance).

A common significance measure for the multinomial distribution is the following statistic

$$Q = \sum_i \frac{(t_i - Tp_i)^2}{Tp_i} + \frac{(t_{cross} - Tp_{cross})^2}{Tp_{cross}}$$

This statistic has an approximate chi-square distribution with k degrees of freedom. However, as many clusters are fairly small, the underlying statistical assumptions are violated and the chi-square approximation does not hold. Instead we use an algorithm that was recently introduced in [Bejerano 2003] for efficient exact computations of pvalues for the multinomial distribution using the likelihood ratio statistics[6]. We implemented this algorithm to compute the probability that a type sampled according to the random type will obtain a higher likelihood ratio than that of the observed type. We repeat this calculation for each one of the clustering models we produce and compare them based on their significance (their pvalue). The results are given in Figure 10. Note that the most significant clusterings are those produced with thresholds 0.83 and 0.69. The first corresponds to a major decrease in the MDL (see Figure 9); however, it is the second one that obtains the better score when considering both the MDL principle and the pvalue, and therefore is selected for further analysis.

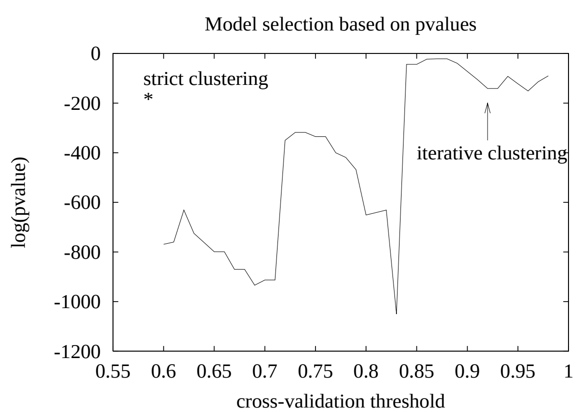

Figure 10: **Model selection based on pvalue**. For each model we compute the significance (pvalue) of the sample type (the observed distribution of true edges) given the clustering model (see text). Both clustering techniques perform well and the resulting clusters are strongly correlated with the experimental data sets (strict clustering is denoted by a "*"). Of all models, those that correspond to thresholds of 0.83 and 0.69 are the most significant ones. However, corroborating the evidence from the MDL computations (Figure 9) we conclude that the best model is obtained at threshold of 0.69.

[6]Given two probability distributions P_1, P_2 and a type $T = \{t_i\}$, the likelihood ratio of the type is defined as $-2 * \log \frac{P(T/P_1)}{P(T/P_2)}$. In our case P_2 is defined based on the empirical sample type, and P_1 is the random type. High positive ratio indicates that T is more similar to the observed type than to the random type, and when considering the two hypotheses it is less likely to emerge from the random type.

3.5.3 Interesting clusters

Our best clustering model is the one produced at a cross-validation threshold of 0.69 (see section 3.4.2). This clustering consists of 213 clusters with the number of genes per cluster ranging from 5 to 141. We were interested in those clusters that are most strongly correlated with the experimental data. As before, one can compute the pvalue of individual clusters and pick the most significant ones. Given a cluster i with n_i genes and e_i edges, of which t_i edges are experimentally verified edges, and $e_{i,cross}$ cross edges of which $t_{i,cross}$ are experimentally verified, we define the probability of the type

$$P(i) = \binom{T}{t_i t_{i,cross}} p_i^{t_i} p_{i,cross}^{t_{i,cross}} (1 - p_i - p_{i,cross})^{T - t_i - t_{i,cross}}$$

and its likelihood ratio. We then compute the probability that a type sampled according to the random type will have higher likelihood ratio.

In Figure 11 we plot the significance of individual clusters. We examined the annotations for genes in the clusters with the top pvalues (Table 5). While one can not assign a singular function to each cluster we observe the themes of these clusters. For example, the second most significant cluster contains mostly dehydrogenases, while the third most significant cluster contains genes which are involved in DNA repair, mitosis and meiosis. The fourth contains many ribosomal genes as well as translation initiation factors and tRNA Synthetases and the fifth contains mitochondrial genes. The collective expression profiles for the 5 most significant clusters are shown in Figure 12

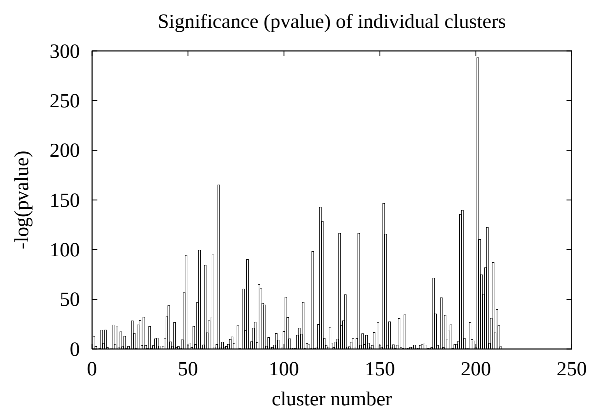

Figure 11: **Significance of individual clusters**. Significance is measured in terms of the pvalue (see text). Note that cluster number does not indicate cluster size. Using the Bonferroni correction we set the significance threshold at $< 1e-3$. Of the 213 clusters 143 have significance value $< 1e-3$.

Cluster Number	P-value	Number of Genes	Cluster Summary
201	6.51e-294	44	Large Diverse Cluster, cytochrome c oxidase subunits ATP synthase subunits, mitochondrial proteins
66	6.72e-166	14	Dehydrogenases
152	2.42e-147	88	DNA replication, DNA repair, Mitosis, Meiosis
119	1.50e-143	141	Translation Related Proteins, Ribosomal Proteins tRNA Synthetases, Translation Initiation Factors
193	2.20e-140	113	Mitochondrial proteins
192	3.43e-136	116	Large Diverse Cluster, Transferases, Hydrolases
120	3.13e-129	15	Diverse Cluster, Lysases, Transferases, Ligases
206	4.04e-123	19	Hydrolases
139	3.17e-117	8	Mitosis
129	3.36e-117	9	tRNA Synthetases
153	1.99e-116	50	Ribosomal Proteins, Helicases
202	6.01e-111	5	Diverse Cluster with Oxidoreductases
56	2.40e-100	31	Diverse Cluster, Hydrolases, Oxidoreductases Mitochondiral Proteins
115	6.94e-99	21	RNA processing, Growth Regulation
63	1.68e-95	8	Oxidoreductases, Fe containing proteins

Table 5: **Most significant clusters**. A summary of the significance, size, and type of genes in each cluster.

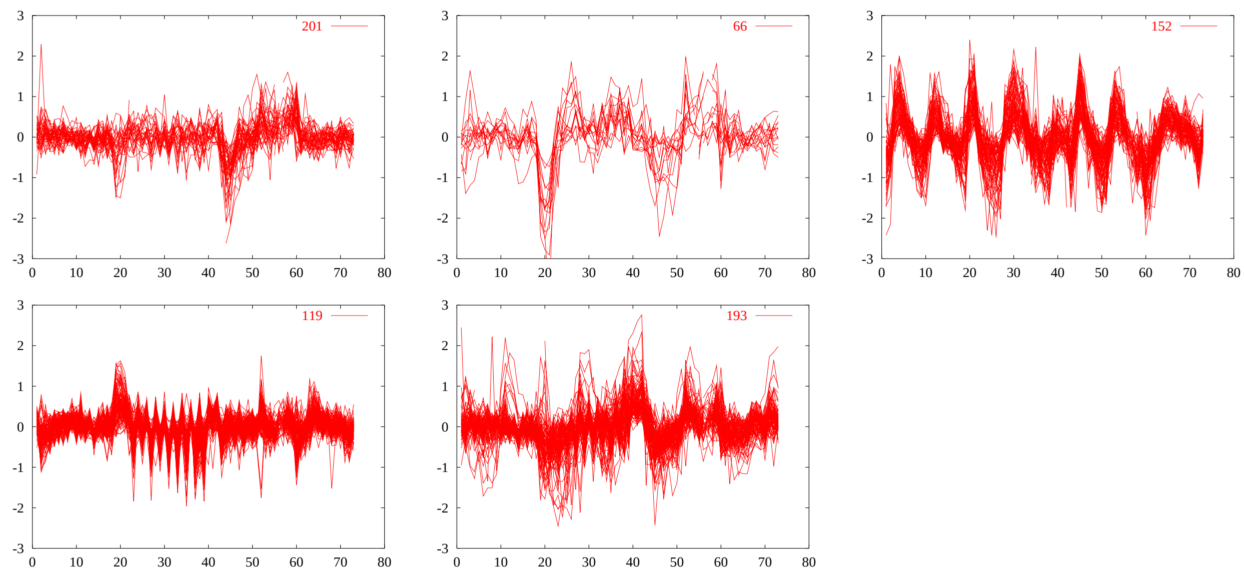

Figure 12: **Clusters of expression profiles** The collective expression profiles for the 5 most significant clusters are shown. Note the very distinctive expression profile of each cluster.

3.6 Pathway association

Our results in the previous section indicate that almost all clusters are indeed strongly correlated with the experimental data and are the consequence of true functional links between genes. The strongest signal seems to be the correlation of clusters with pathways.

It is interesting to analyze the mapping between pathways and clusters. Specifically, we were interested in finding if specific pathways can be mapped to a single or a few clusters. If a non-random mapping exists, then the corresponding clusters can help in refining the pathway structure as they might suggest activation and inhibition between the component proteins (for example, if the consensus profiles of clusters are correlated when a time shift is introduced).

To assess the significance of a mapping between a pathway and clusters we analyze the distribution of the pathway proteins among the clusters and compare the distribution to the distribution one would obtain if the genes were to be placed into the clusters at random. This random process has a multinomial distribution where the probability p_i to fall into a cluster C_i is estimated by the number of all pathway proteins that are classified to C_i divided by the total number of pathway proteins. Given a pathway with n_i proteins in cluster i $(1 \le i \le k)$ the probability of the mapping is

$$P(n_1, n_2, ..., n_k) = \binom{N}{n_1 n_2 ... n_k} p_1^{n_1} p_2^{n_2} ... p_k^{n_k}$$

where $N = \sum_i n_i$ is the total number of proteins in the pathway.

Pathway	pvalue
Purine metabolism	6.2e-75
Oxidative phosphorylation	1.8e-64
Pyrimidine metabolism	3.3e-58
Glycolysis / Gluconeogenesis	1.2e-45
Starch and sucrose metabolism	1.2e-45

Table 6: **Pathways that are strongly correlated with clusters.** The pvalue indicates the significance of the mapping of pathway proteins into clusters.

To determine the significance of a pathway distribution among the clusters we can again compute the pvalue as in section 3.5.2. The 5 most significant mappings are listed in Table 6. A similar analysis of gene distribution can also be done for the homology clusters.

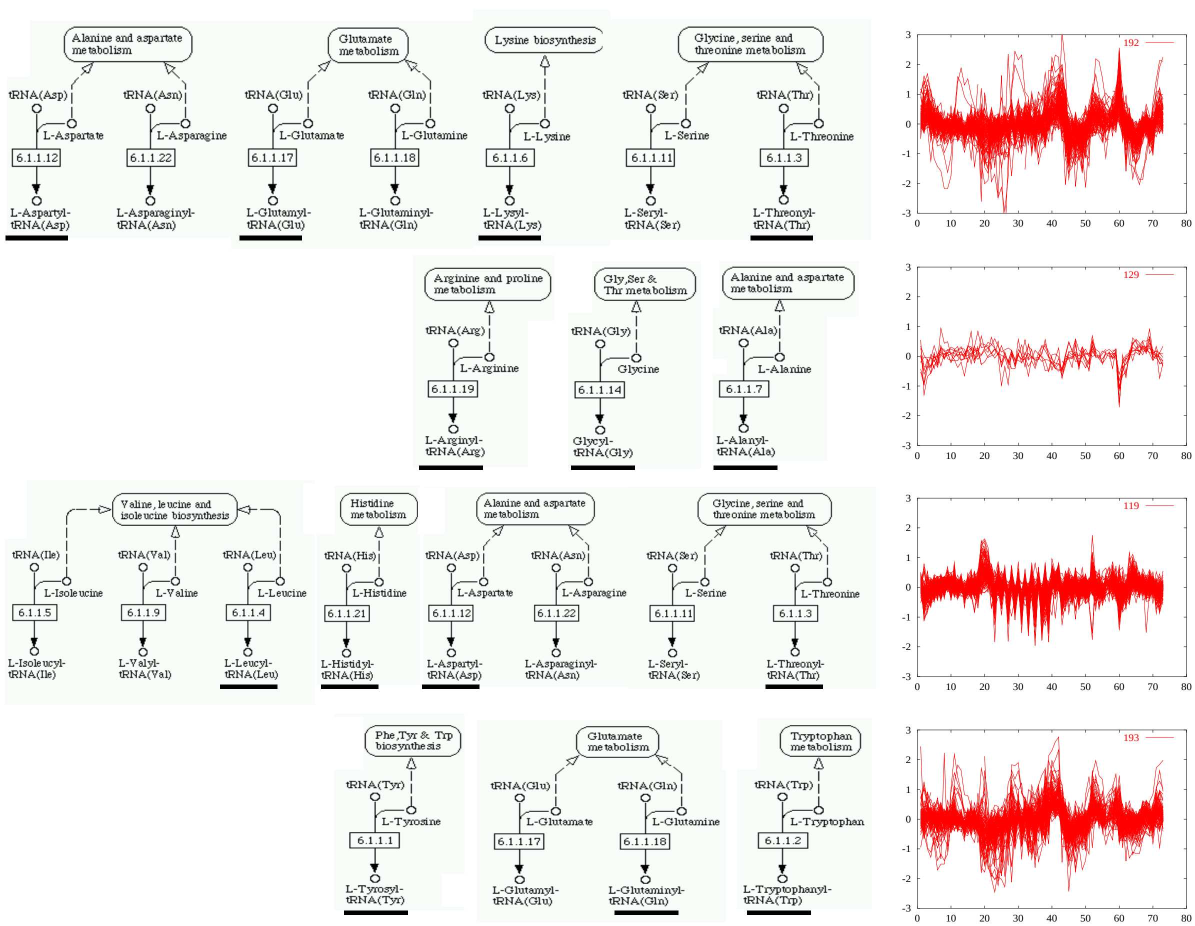

Figure 13: **The aminoacyl-tRNA biosynthesis pathway**. The amino acids synthetases cluster into 4 major clusters: 192, 129, 119 and 193. The specific synthetases that are clustered to each cluster are underlined. (Diagrams are courtesy of the KEGG database [Kanehisa 1996].)

To make our discussion more concrete, we examined two significant pathways and the cluster profiles which correspond to the largest groupings of genes from these pathways. First we considered the aminoacyl-tRNA biosynthesis pathway (pvalue of $1.06e-42$). Interestingly, the amino acid synthetases are clustered into 4 major clusters, each with a unique expression profile (Figure 13). In other words, different amino acids are synthesized at different frequencies and in different parts (times) of the cell cycle. It is unclear why amino acid production is implemented this way, and this phenomena requires further study.

We also examine the oxidative phosphorylation pathway. This pathway is composed of 5 major complexes. While some genes from the clusters intermingle between the complexes there is some separation of the different complexes into different clusters, suggesting again that their production is being coordinated. This analysis of the distribution of pathway genes among the different clusters could be performed with more rigor for more complicated pathways. This may aid in the understanding of known biochemical pathways and the discovery of new ones.

4 Discussion

Microarray technology has become one of the industry standards for high throughput analysis of large pools of gene data. Data collected using microarray assays is used in many forms of analysis, the most typical of which is search for similarly expressed genes. Assuming that the similar patterns are the consequence of an underlying biological process, and that the co-expression of the genes is essential for that specific process, one can infer functional kinship between the genes, be it protein-protein interactions, or biochemical pathways, for example.

However, despite the substantial success in linking expression data to functional aspects of genes, the type of the functional link has not been precisely defined and characterized. Moreover, most studies used one measure or another to quantify the similarity of expression profiles without objectively assessing their merit, and without an underlying statistical justification. This is especially important as microarray data is noisy, and it is hard to discern real signals from random fluctuations and coincidental regularities.

The goal of our study was to explore extensively the notion of functional link based on microarray expression data, sequence similarity and information on promoters, pathways and interacting proteins, while addressing the aforementioned problems. Our results can be divided and summarized as follow

- **pairwise analysis**: (1) we developed a reliable and sensitive similarity measure between genes based on expression data (EucPear) that captures the functional relationship encoded in these gene expression profiles (2) we developed means for assessing the significance of similarity between expression profiles; using these measures we detected significantly correlated genes as well as significantly anti-correlated genes (3) we predict the specific nature of the functional link between proteins as suggested by microarray data.

- **cluster analysis**: (1) we clustered genes reliably into clusters of co-expressed genes, addressing specifically the noisy nature of the data. (2) we developed means for assessing the quality of the results of different clustering algorithms (3) we analyze the mapping between pathways and clusters.

All our tests indicated that the EucPear measure is the most effective one. However, it should be noted that although our chosen measure detected a significant number of true relationships the percentages are still quite low. This is to be expected given the relatively little experimental information that is available. When we consider genes for which there is some information about interactions, pathways, and sequence we are left with a subset of 490 genes. When this subset is restricted to those genes which also have some promoter information only 63 genes remain. Even for this subset the experimental information that is available today is partial. However, since our measures and analysis is statistically verified, we believe that many of our predictions will be supported by experiments, as more data becomes available.

Our clustering algorithms further enhance the signal and detect groups of strongly co-regulated genes. This analysis can present new genes which may relate to each other but are missed in the pairwise analysis. The algorithms were especially optimized to deal with noisy data and therefore we believe that the quality of the grouping is self-assured. The internal validation based on the MDL principle and the external validation using the experimentally verified datasets further verify the quality of the clustering results. The external measure directly assesses the correlation of the clusters with the underlying structure of functional links, while the internal measure addresses issues such as model complexity and generalization. It is the combination of the two measures that leads to a selection of a better model overall. This objective assessment of the results can be used not only to assess our clustering algorithms but also to compare different clustering algorithms and to select the best clustering profile of many that are generated by a specific algorithm.

While our results may not present this data in a whole new perspective it is important to note that this was not our goal. We intentionally picked a data set that has been studied extensively. Our interest in this dataset originated from the relatively low-quality of the expression data (noise and missing values) that posed a challenge for any data mining and learning algorithm. Most dominantly, our choice was motivated by the myriad of experimental data available for the Yeast genome. The availability of pathway data as well as extensive lists of protein-protein interactions and promoter information allowed as to pose basic questions about the utility of this data and the similarity measures that are used to evaluate expression similarity. We propose a rigorous methodology for the analysis of expression data, from the first step of comparing these profiles, to the prediction of the functional link, to cluster analysis and the assessment of the clustering results. Our tools that were optimized to deal with noisy and unreliable data are expected to work even better with a higher quality data sets.

5 Acknowledgments

The authors thank David Lin and Sarah Teichmann for valuable discussions. This work is supported by the National Science Foundation under Grant No. 0218521 to Golan Yona.

References

[Alizadeh et al. 2000] Alizadeh A, Eisen M, Davis RE, Ma CA, Lossos I, Rosenwald A, Boldrick J, Sabet H, Tran T, Yu X, Powell J, Yang L, Marti G, Moore T, Hudson J, Chan WC, Greiner TC, Weissenberger DD, Armitage JO, Levy R, Grever MR, Byrd JC, Botstein D, Brown PO & Staudt LM. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503-511.

[Altschul et al. 1990] Altschul, S. F., Carrol, R. J. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.

[Bader et al. 2001] Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F., Pawson, T. & Hogue, C. W. (2001). BIND - The Biomolecular Interaction Network Database. *Nucleic Acids Res.* **29**, 242-245.

[Bammer & Fostel 2000] Bammer, G. & Fostel, J. (2000). Genome-wide expression patterns in Saccharomyces cerevisiae: comparison of drug treatments and genetic alterations affecting biosynthesis of ergosterol. *Antimicrobial Agents and Chemotherapy*, **44**, 1255-1265.

[Bejerano 2003] Bejerano, G. (2003). Efficient exact p-value computation and applications to biosequence analysis. *In the proceedings of RECOMB 2003*, 38-47, ACM press.

[1] [BRITE] BRITE website: http://www.genome.ad.jp/brite/

[Cho et al. 1998] Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T., Gabrielian, A.E., Landsman, D., Lockhart, D.J. & Davis, R. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2**, 65-73.

[D'haeseleer et al. 1997] D'haeseleer P., Wen, X., Fuhrman, S. & Somogyi, R. (1997). "Information Processing in Cells and Tissues", Holcombe, M. & Paton, R. eds., pp 203-135, Universal Academy Press, Tokyo, Japan.

[Diffee et al. 2003] Diffee, G.M., Seversen E.A., Stein, T.D.& Johnson, J.A. (2003). Microarray expression analysis of effects of exercise training: increase in atrial MLC-1 in rat ventricles. *American Journal of Physiology Heart and Circulatory Physiology* **284**, 830-837.

[Dubnov et al. 2000] Dubnov, S., El-Yaniv, R., Gdalyahu, Y., Schneidman, E., Tishby, N. & Yona, G. (2001). A new nonparametric pairwise clustering algorithm based on iterative estimation of distance profiles. *Machine Learning*, **47**, 35-61

[Dudoit et al. 2002] Dudoit, S., Yang, Y.H., Callow, M.J. & Speed, T.P. (2002). Statistical Methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, **12**, 111-139.

[Ge at al. 2001] Ge, H., Liu, Z., Church, G. & Vidal, M. (2001). Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nature Gen.* **29**, 482-486.

[Goto et al. 1997] Goto S., Bono, H., Ogata, H., Fujibuchi, W., Nishioka, T., Sato, K. & Kanehisa, M. (1997). Organizing and computing metabolic pathway data in terms of binary relations. *Pac. Symp. Biocomput.* 175-186.

[Heyer et al. 1999] Heyer, L.J., Kruglyak, S. & Yooseph, S. (1999). Exploring Expression Data: Identification and Analysis of Coexpressed Genes. *Genome Research* **9**, 1106-1115.

[Hughes et al. 2000] Hughes, T., Marton, M., Jones, A., Roberts, C., Stoughton, R., Armour, C., Bennett, H., Coffey, E., Dai, H., He, Y., Kidd, M., King, A., Meyer, M., Slade, D., Lum, P., Stepaniants, S., Shoemaker, D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M. & Friend, S. (2000). Functional discovery via a compendium of expression profiles. *Cell* **102**, 109-126.

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

[Liu 2003] Liu, E.T. (2003). Classification of cancers by expression profiling. *Current Opinion in Genetics and Development* **13**, 97-103.

[Lopez at al. 2003] Lopez, I.P., Marti, A., Milagro, F.I., Zulet, Md Mde L., Moreno-Aliaga, M.J., Martinez, J.A. & De Miguel, C. (2003). DNA microarray analysis of genes differentially expressed in diet-induced (cafeteria) obese rats. *Obesity Research* **11**, 188-194.

[McCormick et al. 2003] McCormick, S.M., Frye S.R., Eskin, S.G., Teng, C.L., Lu, C.M., Russell, C.G., Chittur, K.K. & McIntire L.V. (2003). Microarray analysis of shear stressed endothelial cells. *Biorheology*, **40**, 5-11.

[Miklos & Rubin 1996] Miklos, G. & Rubin, G. (1996). The role of the Genome Project in determining gene function: insights from model organisms. *Cell* **86**, 521-529.

[Oliveros et al. 2000] Oliveros, J.C., Blaschke, C., Herrero, J., Dopazo, J. & Valencia, A. (2000). Expression Profiles and Biological Function. Genome *Informatics* **11**, 106-117.

[Qian et al. 2001] Qian, J., Dolled-Filhart, M., Lin, J., Yu, H. & Gerstein, M. (2001). Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *J. Mol. Biol.* **312**, 1053-1066.

[Raychaudhuri et al. 2002] Raychaudhuri, S., Schutze, H. & Altman, R.B. (2002). Using text analysis to identify functionally coherent gene groups. *Genome Res.* **12**, 1582-1590.

[Rison et al. 2002] Rison, S.C., Teichmann, S.A. & Thornton, J.M. (2002). Homology, pathway distance and chromosomal localization of the small molecule metabolism enzymes in Escherichia coli. *J. Mol. Biol.* **318**, 911-932.

[Spellman et al. 1998] Spellman, P.T., Sherlock, G., Zhang, M., Iyer, V., Eisen, M., Brown, P., Botstein, D. & Futcher, B. (1998). Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. *Mol. Bio. Cell* **9**, 3273-3297.

[Troyanskaya et al. 2001] Troyanskaya O., Cantor M., Sherlock G., Brown P., Hastie T., Tibshirani R., Botstein D. & Altman R. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520-525.

[Xenarios et al. 2001] Xenarios, I., Fernandez, E., Salwinski, L., Duan, X.J., Thompson, M.J., Marcotte, E.M. & Eisenberg, D. (2001) DIP: The Database of Interacting Proteins: 2001 update. *Nucl. Acids Res.* **29**, 239-241.

[Yeatman 2003] Yeatman, T.J. (2003). The future of clinical cancer management: one tumor, one chip. *The American Surgeon*, **69**, 41-44.

[Yona et al. 1999] Yona, G., Linial, N. & Linial, M. (1999). ProtoMap: Automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. *Proteins*, **37**, 360-378.

[Yoo et al. 2003] Yoo, M.S., Chun, H.S., Son, J.J., DeGiorgio, L.A., Kim, D.J., Peng, C. & Son J.H. (2003). Brain Research. *Molecular Brain Research*, **110** 76-84.

[Zhu & Zhang 1999] Zhu, J. & Zhang, M. (1999). SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics* **15**, 607-611.