BIOINFORMATICS

Vol. 00 no. 00
Pages 1–12

Effective similarity measures for expression profiles

Golan Yona[1]*, William Dirks[1], Shafquat Rahman[1], David M. Lin[2]

[1]Department of Computer Science and Center of Applied Mathematics, Cornell University
[2]Department of Biomedical Sciences, Cornell University

ABSTRACT

It is commonly accepted that genes with similar expression profiles are functionally related. However, there are many ways one can measure the similarity of expression profiles, and it is not clear apriori what is the most effective one. Moreover, 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 conduct a 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. We also introduce variations on similarity measures that are based on statistical analysis and better discriminate genes which are functionally nearby and faraway.

Our tools can be used to assess other similarity measures for expression profiles, and are accessible at `biozon.org/tools/expression/`.

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 a cell undergoes a normal physiological process, such as the cell cycle (Spellman et al. 1998; Shapira et al. 2004), in an attempt to determine the genes involved in this process. It has been used to study differential gene expression patterns under different environmental conditions (Bammer & Fostel 2000; McCormick et al. 2003; Yoo et al. 2003; 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).

While some have criticized the usefulness of microarray technology (Gygi et al. 1999), expression data is still considered a substantial source of information on cellular activity and regulation, and 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 can give insight into the gene functions, and it is purported that genes which are expressed similarly under different conditions or experimental setups are likely to have related biological functions.

However, this type of analysis 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 the absence of precise measures for assessing the significance of similarity based on expression profiles, it is not clear whether genes are indeed truly co-regulated or are functionally linked even when they seem to be similarly expressed. Moreover, the choice of the metric can have a great impact on the analysis, for example, when clustering genes based on microarray data in search for coordinated groups of co-expressed genes. Indeed, it is well known that different representations and distance measures can have significant effect on the quality of the clustering results, as most clustering algorithms rely directly on pairwise distances or similarities between instances. This includes k-means, pairwise clustering (also called hierarchical clustering) and spectral clustering algorithms. Therefore, better pairwise measures are likely to produce better results, i.e. clusters that better correlate with cellular processes.

In this paper we study and compare different similarity measures between genes based on expression data. We assess their accuracy and sensitivity in distinguishing between genes which are functionally nearby and faraway and evaluate their effectiveness in detecting experimentally verified functional relationships extracted from pathway data, protein-protein interaction data, sequence data and promoter data. Our methodology and the tools are also applicable to new similarity measures and datasets.

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

*To whom correspondence should be addressed. Present address: Department of Computer Science, Technion, Haifa 32000, Israel (golan@cs.technion.ac.il) in cells. Our model organism is Yeast, for which extensive experimental information on gene relationships exist.

2.1 Expression data

We used four different data sets in our study. The first is the famous time-series expression data we obtained 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 (*Time-series 1998*) has been normalized by Spellman et al. to correct for experimental variation between the different microarrays.

Our second data set is the Rosetta Inpharmatics yeast compendium data. This microarray data consists of 300 different conditions: 276 deletion mutants, 11 tetracycline regulatable alleles of essential genes, and 13 well-characterized compounds (Hughes et al. 2000). Each ORF is represented by a 300-dimensional vector of the expression values associated with these different conditions. We refer to this data set as the *Rosetta 2000* data set.

Finally, we used two stress time-series datasets. The first (*Stress 2000*) measures the time-wise response to diverse environmental transitions such as temperature shocks, osmotic shocks, amino acid starvation and the presence or depletion of various chemical agents totaling to 129 different experiments (Gasch et al. 2000). The second dataset (*Stress 2004*) measures the expression of genes during the cell-cycle, while under oxidative stress and contains 70 measurements for each gene compiled from four time courses (Shapira et al. 2004).

It should be noted that the Rosetta dataset was generated using oligonucleotide arrays, which differ from the cDNA arrays that were used in (Spellman et al. 1998) and (Shapira et al. 2004). DNA arrays use full gene sequences as opposed to specific oligonucleotides. One drawback of this technology is that sequences of high sequence identity might cross hybridize. We should also note that the Stress 2004 dataset was pre-filtered by its authors, keeping only genes that were successfully amplified. Furthermore, genes that were represented in only one time course were eliminated.

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 genes) of the ORFs in the Time-series expression data set can be mapped to genes in the Yeast sequence database through the ORF label. Of the ORFs reported in the Rosetta and the Stress 2000 datasets, we were able to map 5894 ORFs to genes that exist in both the Yeast sequence database and the Time-series 1998 expression data set. This is the subset of genes used in our study. The Stress 2004 dataset contains a smaller number of genes (4699) due to filtering.

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, might but 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. 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 generating a backup system that is concurrent with the main system.) This process might be the underlying explanation behind co-expression if a protein is used as an alternative or as a backup (“plan B”) protein for another protein. Several examples are observed in known systems and are documented in the literature; For example, two-thirds of the fly genes have no observable loss of function phenotype under knockout experiments (Miklos & Rubin 1996).

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. (Information on the different datasets used in this study is available in the Supplementary Material, section 7.1.)

3 METHODS AND RESULTS

To determine whether two genes have similar expression patterns an appropriate similarity measure must be chosen. We consider several scoring functions and their combinations. These include global measures (such as the Euclidean metric, the Pearson correlation and the Spearman rank correlation), statistical measures (zscore-based), local similarity measures that are based on the dynamic programming algorithm

Relation type	#genes	#edges
Interaction	3592 (3454)	5339 (5052)
Sequence	3092 (2852)	19074 (13950)
Promoter	213 (213)	2439 (2439)
Pathway	642 (605)	15789 (13914)
Total	5079 (4815)	41902 (34682)

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. Numbers in parentheses refer to genes that can be associated with expression profiles.

(Qian et al. 2001) and measures of anti-correlation. Most of these are traditional measures or variations thereof, and are described in detail in section 7.2 of the Supplementary Material. We also test the new mass-distance measure that we introduce in section 3.1. To determine the most effective pairwise similarity measure, all measures are assessed in terms of their ability to detect meaningful pairwise relationships (section 3.2).

3.1 The mass-distance measure

All the measures that are commonly used to assess expression similarity (see Supplementary Material) ignore the specific background distribution of expression values in each experiment. Here we propose the **mass-distance** (MD) measure that adjusts to the background distributions when measuring the similarity of two expression profiles. This measure assesses the distance between two profiles by estimating the probability to observe by chance a vector inside the volume delimited by the profiles. The smaller the volume is, the more similar are the two profiles.

Given two expression profiles $\mathbf{u}$ and $\mathbf{v}$, we consider one coordinate (experiment) i at a time and estimate the total probability mass of samples (genes) whose i-th feature is bounded between the expression values u_i and v_i. The probability mass is computed based on the background distribution for experiment i, as is illustrated in Figure 1. Often, these distributions can be quite reliably modeled using normal distributions. Once the parameters μ, σ of these distributions are estimated, the probability mass between u_i and v_i is computed by the integral

$$MASS(u_i, v_i) = \int_{\min\{u_i,v_i\}}^{\max\{u_i,v_i\}} Prob_i(x)dx$$

where $Prob_i(x) = N(\mu_i, \sigma_i)$ is the normal distribution for experiment i.

The background distribution does not always follow closely a normal distribution. Two-way experiments that measure, for example, the difference in expression between two tissues (e.g brain vs. liver) might exhibit a bi-modal or other distribution. In such cases one could use the empirical distributions when computing the $MASS$ variables

$$MASS(u_i, v_i) = \sum_{\min\{u_i,v_i\} \leq x \leq \max\{u_i,v_i\}} freq(x)$$

where $freq(x)$ is the empirical frequency of the measurement x.

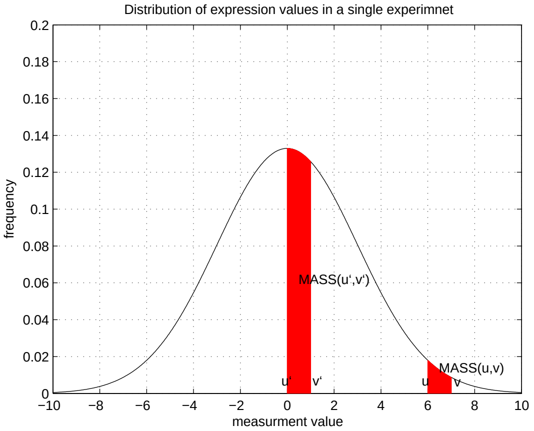


Fig. 1. The mass distance. Often is the case that the distance between two measurements depends not only on the relative nominal difference between the measurements but also on the background distribution. For example, two measurements u and v are statistically more similar to each other than the two measurements u' and v' although $u - v = u' - v'$. That is to say that there are fewer measurements with score between u and v and therefore fewer instances that have similar properties. The probability mass (the shaded area) is given by the integral over the background distribution.

The mass-distance (MD) of $\mathbf{u}, \mathbf{v}$ is defined as the total volume of samples bounded between the two expression profiles and is estimated by the product over all coordinates

$$MD(\mathbf{u}, \mathbf{v}) = \prod_{i=1}^{d} MASS(u_i, v_i). \quad (1)$$

In practice, the MD score is evaluated through its logarithm:

$$-\log MD(\mathbf{u}, \mathbf{v}) = -\sum_{i=1}^{d} \log MASS(u_i, v_i) \quad (2)$$

It should be noted that most of the datasets we used in this paper are time-series datasets where the temporal aspect is important and one usually expects to detect co-expression along a continuous set of experiments. All the metrics we tested are appropriate for this kind of analysis. However, some data sets are generated over a set of experiments that are not expected to form continuous patterns of co-expression (such as the Rosetta dataset). Nevertheless, most of the similarity measures described in this paper can work quite well even for such data sets (as confirmed by our results over the Rosetta dataset, see next section). The mass-distance measure is especially compelling in that respect, as it combines the benefit of a local metric without being restricted to continuous patterns, and can consider arbitrary combinations of experiments.

3.2 Pairwise Analysis

To assess the performance and determine the information content of each of the measures described in the previous sections, 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. This curve is similar to a ROC curve (Hanley & McNeil 1982), where the number of true positives is plotted vs the number of false positives. However, in our case it is very hard or even impossible to determine which pairs of genes are totally *unrelated*. In other words, no similarity can be confidently designated as a false positive.

Detailed results for the Time-series 1998 data set are shown in Figure 6 (Supplementary Material). Of the **global measures**, the Pearson correlation is the most effective one, followed closely by the Spearman rank correlation. Surprisingly, despite its popularity in studies of expression arrays, the Euclidean measure performs poorly (Figure 6a). Introducing shifts in global measures did not improve the performance and even decreased it slightly (by 3%).

All **zscore-based measures** improved over the original global measures, as is shown in Figure 6b. The most drastic improvement was observed with the Euclidean-based zscore measure that outperformed even the Pearson correlation measure and the corresponding zscore measure. For the Spearman rank correlation we computed a significance value based on the tail probability of the coefficient ρ as outlined in (StatLib 1975). The significance value depends on the correlation coefficient as well as the dimension. However, no noticeable improvement was observed when the results were reordered based on this value (compared to the results with the raw rank correlation values).

The **combined measures** perform quite well compared to the individual measures (Figure 6c shows the results for a subset of the measures attempted). This is not surprising; by using a combination of different measures, the hybrid measures capture different types of information about the expression profiles. For example the *EucPear* measure, which seems to work the best, captures two types of information: The Euclidean measure is significant when two expression profiles maintain the same level of expression throughout. It does not capture information about the general correlation of the expression profiles. The Pearson correlation is significant when the vectors change in time in a similar fashion. This is irrespective of the actual expression levels of the profiles. The *EucPear* measure captures both these aspects.

The **local similarity measures** also performed very well compared to all other methods, and the exponential decay that we introduced improved the performance slightly[1]. The dimension-independent extreme-value distribution that is used to assess the significance of the local similarities (see section 7.3 in Supplementary Material) does not affect the performance since it does not change the ranking of the pairwise similarities. Surprisingly, the use of the dimension-specific extreme-value distributions to assess the significance did not improve the performance either.

Another surprising fact is that the **anti-correlation measures** hardly detected any of the edges in our relation graph. For example, of the top 20,000 *anti-correlations* detected with the local similarity measure, only 28 can be associated with one of the relations described in section 2. However, as is also the case with time-delayed (shifted) global similarities and local similarities, this might also indicate that there is another type of relation (such as the one that exists between a regulator and regulatee) for which these measures are most suited, and we intend to revisit this analysis once more data about regulator-regulatee relations will be made available.

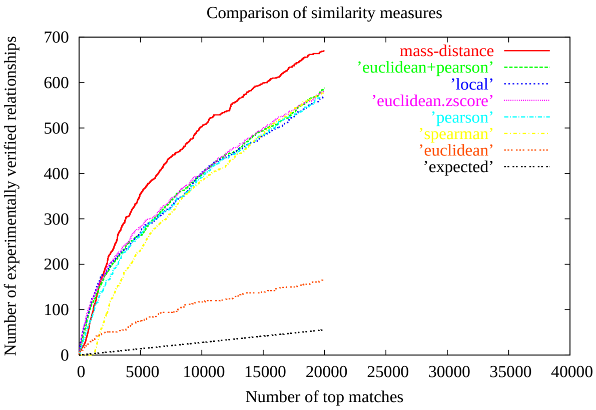


Fig. 2. Performance evaluation on the Time-series 1998 dataset. For clarity, the labels are ordered according to the performance of the corresponding measures. Only a subset of all tested measures is displayed (for more results, see Figure 6 in Supplementary Material). The *expected* number of relationships is computed under the uniform random setup (see section 7.4 in Supplementary Material).

A summary of the best results from each category (for the Time-series 1998 data set) is shown in Figure 2. As this graph indicates, all measures contain some information about the true relationships, when compared to what is expected by a random guess (see section 7.4 in Supplementary Material). Of all measures tested, the mass-distance measure seems to give the best results. Better results with this measure were obtained when using the empirical estimates for the $MASS$ variables (see section 3.1). Next, we observe four measures that perform almost the same: the EucPear measure, the local similarity measure, Pearson correlation[2] and the Euclidean-based zscore measure (each one is the best in its category, see Figure 6). These are closely followed by the Spearman rank correlation and far behind is the Euclidean metric. Similar results were obtained with two other expression data sets (Figure 3), although the Spearman rank correlation improves significantly over these datasets and comes second or first. Different results were obtained for the fourth dataset (Stress 2000), as discussed below.

[1] In (Qian et al. 2001), both the positive and negative local similarities are considered between each pair of genes and the maximum of the two is defined as their final similarity. We did not detect an improvement with this variation and the performance decreased by about 4%.

[2] The scoring function that is used to compute local similarities is based on Pearson correlation, which explains the similar performance of the Pearson and local similarity measures over all datasets.

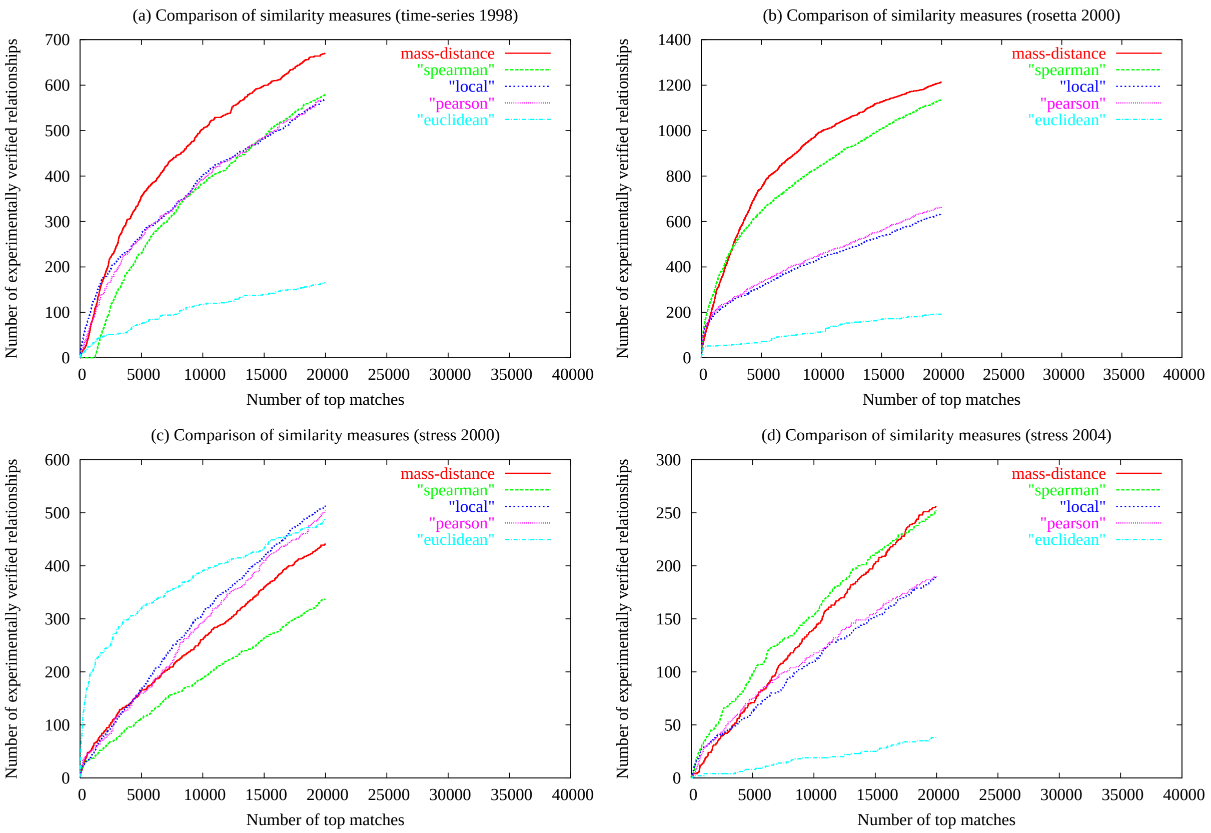


Fig. 3. Performance evaluation of selected metrics on different expression datasets. (a) Time-series 1998. (b) Rosetta 2000. (c) Stress time-series 2000. (d) Stress time-series 2004. For each data set we show the results for the Euclidean metric, the Pearson correlation, the Spearman rank correlation, the local similarity measure and our mass-distance measure.

While in general no single measure can be crowned as the best for all possible datasets, the mass-distance measure seems to be the most effective one over two of the datasets tested, with significant improvement in the ROC_{20000} values that amounts to 22.3% (Time-series 1998) and 11.4% (Rosetta 2000) over the next best method. The mass-distance measure comes second for the Stress 2004 dataset. The best method over this dataset is the Spearman rank correlation that improves over the mass-distance measure by 9.5%. Both measures are outperformed over the Stress 2000 dataset. Surprisingly, the best metric for this dataset is the Euclidean metric (that performs significantly worse than all other metrics on the other datasets). We were baffled by the unusual performance trends that are observed with the Stress 2000 dataset and we suspect that among other factors, these are attributed to effects of sequence cross-hybridization. Estimating the extent of cross-hybridization is not trivial since it is difficult to computationally discern genes that seem co-expressed because they cross-hybridize from truly co-expressed genes. However, our preliminary results indicate that many of the top similarities that are reported with the Euclidean metric over the Stress 2000 dataset are dominated by measurements that deviate only slightly from the mean values for these experiments (in other words, they are insignificant). This is expected for genes that cross-hybridize. Such similarities are down-weighted by the mass-distance measure (that adjusts to the distribution of expression values in each experiment) and therefore it is less sensitive to artifacts or signals that are due purely to cross-hybridization. Because of its solid performance the mass-distance is chosen as our main similarity measure for subsequent analysis.

The most pronounced difference in performance between the different measures is observed over the Rosetta data set. Moreover, the Rosetta dataset seems to contain the maximal information with regard to functional links and more than 1200 true relationships are detected within the top 20000, compared to 670 with the Time-series 1998, 442 with the Stress 2000 and 256 with the Stress 2004 (although the last one is a smaller dataset).

3.3 Information content

To assess the information content of similarities detected with the optimal measure we consider all pairwise similarities that are associated with $evalue < 10$ and compute the number of true relationships that are detected (statistical significance of expression similarity is discussed in section 7.3 in Supplementary Material). These numbers are compared to the number of true relationships one would expect to find by chance. The

Dataset	#edges below threshold	#true-edges below threshold	expected	simulation	ratio (observed/expected)
Time Series 1998	378	33	1.1	0.6 (0.8)	29.2
Rosetta	3689	709	11.0	7.2 (3.8)	64.2
Stress 2000	1532	88	4.6	2.4 (1.9)	19.1
Stress 2004	3860	64	11.8	7.7 (2.8)	5.4

Table 2. Information content of expression similarity. Expression similarity was computed using the mass-distance measure. The total number of pairwise similarities between the 4815 genes is 11,589,705 of which 34,682 (0.3%) are true edges that correspond to experimentally verified relationships. The ratio of true relationships is much higher for significant edges, and especially so for the Rosetta data set. The second and the third columns are the numbers of edges and true edges above the threshold ($evalue < 10$). The next 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). Note that the two random setups give slightly different results, reflecting the non-uniform distribution of edges. The number in parentheses in the fifth column is the standard deviation observed in the simulation. The last column is the ratio between the observed number and the expected number of true relationships.

Dataset	Pathway		Homology		Promoter		Interactions		Total
	observed (expected)	ratio	observed (exp.)	ratio	observed (exp.)	ratio	observed (exp.)	ratio	
Rosetta	197 (4.4)	44.7	519 (4.4)	118.0	31 (0.8)	38.8	7 (1.6)	4.4	709
Stress 2000	36 (1.8)	20.0	58 (1.8)	32.2	1 (0.3)	3	3 (0.7)	4.5	88
Stress 2004	33 (5.5)	6.0	33 (4.0)	8.2	1 (0.9)	1.1	2 (1.7)	1.2	64

Table 3. Correlation of different types of relations with expression similarity. For each dataset we show the breakup of pairs of truly related genes whose expression similarity is assigned $evalue < 10$ (measured with the mass-distance 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 total. For each type we also compute the number of such relationships that are expected to occur by chance (in parentheses) and the ratio observed/expected. Results are not reported for the Time-series 1998 dataset because of the relatively small number of significant edges.

expected number of relationships is computed using the two random setups described in the Supplementary Material. Since not all genes are associated with experimental data, we restrict the analysis to the a meaningful subset of genes for which we have such information. This is the set of 4815 nodes of the relation graph as described in section 2.4.

The results of this analysis are summarized in Table 2. As the table indicates, a substantial number of true relationships is detected, compared to the number of such relationships that are expected to happen by chance. Since the experimental data is very partial, this is a lower bound on the number of true relationships that can be detected based on expression similarity. Note that the number of significant edges vary greatly. This number is correlated with the dimensionality and the quality of the datasets. Of the four datasets, the Rosetta dataset produces the largest number of significant edges, and also has the highest ratio of observed vs. expected true edges. The Time-series 1998 is the second best (ratio-wise).

We also evaluated the correlation of each one of the four relation types (interaction, pathway, promoter, homology) with expression similarity (Table 3). The general trends are similar across the different datasets, and sequence homology is most strongly correlated with expression similarity. The second most substantial source of information is pathway membership. Promoter-based relation is the third strongest signal (ratio-wise) in the Rosetta dataset but is preceded by interaction-based relations in the Stress datasets. Although the number of significant edges in the Stress and Time-series datasets is too small to draw strong conclusions, the general trends seem consistent and similar results were obtained when the analysis was extended to all relationships that were detected within the top 20,000 pairs (see Table 4 in Supplementary Material).

Interestingly, the ratios differ quite markedly over the different data sets. For the Rosetta dataset, sequence homology is more than 2.5 times stronger than pathway membership (appearing almost 118 times more frequently than one would expect by chance). The difference between these two types of relations is much smaller over the other datasets. Also striking is the significant ratio ($observed/expected = 38.8$) for promoter-based relations in the Rosetta dataset.

This is consistent with the nature of the data. The Rosetta dataset tests the effect of various mutations on cell activity. These mutations often affect the regulatory network of the cell. In this view, the strong signal with respect to promoter data is in excellent agreement with regulatory aspects of the cell. The correlation with sequence homology, on the other hand, is not so obvious. However, it is not sequence homology that underlies these functional links, but rather related regulation systems. It is assumed that most sequence homology relationships are due to gene duplication events that for the most part preserved also the promoter sequences and as such are regulated by the same transcription factors (see section 2.3). Unfortunately, promoter information is only sparsely available to verify these relationships. However, the strong correlation of the Rosetta dataset with known promoters lends a strong evidence in support of this hypothesis.

On the other hand, the Rosetta dataset lacks the time aspect that characterizes, for example, the activity of cellular pathways. Time-series data can be more useful in that respect, as is also suggested by our results. Indeed, time-series data has proved to be very effective for pathway prediction in (Popescu & Yona 2005) where it is being used to produce unambiguous assignments of genes to cellular pathways. Note that the signal with respect to interactions is quite weak with both datasets.

4 ASSESSMENT OF NEW MEASURES AND EXPRESSION DATA SETS

It is expected that new measures of similarity based on other principles might be more effective in detecting functional links. Moreover, other expression data sets (be it time-series data or not) might be more informative than the ones we used.

To enable others to use our methodology and test new algorithms or new data sets and compare them to those which were already tested, we constructed a webserver that is available at **biozon.org/tools/expression/**. Users can evaluate a new similarity measure by using the same expression data sets, comparing the expression profiles and uploading the results to our webserver. The results will be evaluated as described in section 3.2 in a graph similar to the one in Figure 3.

We will update our servers to extend the existing data sets and include new relations that indicate other types of functional links. The data will be extracted from the Biozon database (Birkland & Yona 2006). Our model system is yeast, and that is the only constraint right now, in the sense that it limits the application only to yeast expression data. However, in the future we hope to extend this analysis to other model systems.

5 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.

While many have pointed out the problems with interpreting microarray data, microarray analysis is still very instrumental to gene function prediction, and is useful for prediction of interactions and pathways, especially when combined with other data sets (Popescu & Yona 2005). However, despite the many publications that are concerned with microarray analysis, some basic questions remained unanswered.

Most studies that utilize microarrays use 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. Therefore, the choice of the metric can greatly affect the analysis results, for example, when searching for clusters of co-expressed genes.

The goal of this study is to evaluate the quality of different similarity measures for expression profiles and determine which measure(s) is the most effective for detecting functional links. We do so by comparing the expression similarity results with sequence similarities and information on promoters, pathways and interacting proteins. Our results clearly indicate that there are substantial differences in performance between the different measures, and that the popular measures (Euclidean, Pearson correlation) are sometimes significantly inferior to the other measures we tested. We conclude that combined similarity measures (and especially the *EucPear* measure), the zscore-based Euclidean metric, the Pearson correlation measure and the local similarity measure perform better than the commonly used Euclidean metric. The Spearman rank correlation, that has not received much attention so far in studies of mRNA expression data, performs even better than these metrics and is a strong contender for the most effective one. A solid performer is the mass-distance measure that significantly outperforms the other metrics on some data sets. Since it adjusts to the distribution of expression values in each experiment it is also less susceptible to chance similarities that are due to average or typical expression values. Another advantage of the mass-distance measure is its flexibility, as it can produce good results even when the genes are co-expressed in an arbitrary subset of the experiments.

To associate significance values with expression similarity scores we model the background distributions (section 7.3). These significance values (evalues) provide a natural and useful measure of importance and relevance for a pair of supposedly similarly expressed genes. All significant pairwise similarities (with the mass-distance measure) that were computed over the Time-series 1998, Rosetta 2000 and Stress 2004 datasets are available at `biozon.org`. These similarities can help characterize genes of unknown function[3].

Our analysis of *anti-correlated* genes suggests that while they are very effective for the study of causal networks in cells, they are not as effective for the study of direct functional links between genes. This might change as data on other types of functional links becomes available, and we intend to update our analysis accordingly.

While in this study we focused on Yeast, our results are likely to extend to other organisms. Our choice of Yeast as a model organism was motivated by the myriad of experimental data available for the Yeast genome. The availability of pathway data as well as lists of protein-protein interactions and promoter information allowed us to pose basic questions about the utility of this data and the similarity measures that are used to evaluate expression similarity. However, it should be noted that although the effective similarity measures detected a significant number of true relationships, their percentages (out of all similarities that are reported as significant) are still quite low. This is to be expected given the relatively little experimental information that is available. For example, when we consider genes for which there is some information about homology, interactions *and* pathways we are left with a subset of only 239 genes. When this subset is restricted to those genes which also have

[3] For example, TrEMBL Q07992 (`biozon.org/Biozon/Profile/272323`) is an uncharacterized Yeast ORF protein (documented as an unnamed protein in GenPept and probable membrane protein in PIR). However, examination of proteins with similar expression profiles (`biozon.org/Biozon/Similar/Expression/001250000629`) suggests that this protein possesses some ribosomal activity as it is strongly linked to other ribosomal proteins.

some promoter information only 32 genes remain. Even for this subset the experimental information that is available today is partial. However, we believe that many of our predictions will be supported by experiments, as more data becomes available. Indeed, when tested against a new set of protein-protein interactions, many of the putative functional links were confirmed as interacting proteins (for example, of the top 20,000 pairwise relationships over the Time-series 1998 dataset, 159 are due to new interactions compared to 11.4 that are expected to occur by chance).

Despite the lack of *complete* experimental data we were able to detect significant differences between the measures. Moreover, we were able to characterize more precisely the type of the functional link that is most strongly predicted with different types of datasets (mutation-wise vs. time-series). The results are expected to be even better for higher quality datasets and as more data becomes available. Finally, our tools can be applied also to other types of expression data and other similarity measures through our webserver at `biozon.org/tools/expression/`.

6 ACKNOWLEDGMENTS

The authors thank Sarah Teichmann for valuable discussions. This work is supported by the National Science Foundation under Grant No. 0218521, as part of the NSF/NIH Collaborative Research in Computational Neuroscience Program.

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. *Nucl. 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.

[Birkland & Yona 2006]Birkland, A. & Yona, G. (2006). The BIOZON Database: a Hub of Heterogeneous Biological Data. *Nucl. Acids Res.* **34** D235-D242.

[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.

[Dembo & Karlin 1991]Dembo, A. & Karlin, S. (1991). Strong limit theorems of empirical functionals for large exceedances of partial sums of i.i.d variables. *Ann. Prob.* **19**, 1737-1755.

[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.

[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.

[Gasch et al. 2000]Gasch, A.P., Spellman, P.T., Kao, C.M, Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D. & Brown, P.O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11**, 4241-4257.

[Gumbel 1958]Gumbel, E. J. (1958). "Statistics of extremes". Columbia University Press, New York.

[Gygi et al. 1999]Gygi, S.P., Rochon, Y., Franza, B.R. & Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. *Mol. Cell Biol.* **19**, 1720-1730.

[Hanley & McNeil 1982]Hanley, J.A. & McNeil, B.J. (1982). The meaning and use of the area under the Receiver Operating Characteristic (ROC) curve. *Radiology* **143**, 29-36.

[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.

[Popescu & Yona 2005]Popescu, L. & Yona, G. (2005). Automation of gene assignments to metabolic pathways using high-throughput expression data. *BMC Bioinformatics* **6** 217-.

[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.

[Shapira et al. 2004]Shapira, M., Segal, E. & Botstein, D. (2004). Disruption of Yeast Forkhead-associated Cell Cycle Transcription by Oxidative Stress. *Mol. Biol. Cell* **15**, 5659-5669.

[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.

[StatLib 1975]Algorithm AS 89. (1975). Tail probabilities for Spearman's rho. *Applied Statistics algorithms* **24**, 377.

[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.

7 SUPPLEMENTARY MATERIAL

7.1 Datasets

7.1.1 Interaction data A list of interacting proteins in yeast was retrieved from the Biozon database at `biozon.org`. This list 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 (BRITE). After eliminating redundancy and merging 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.

7.1.2 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. In this study we used only the 127 metabolic pathways, 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 the Biozon composite non-redundant database that contains over 2 million unique sequence entries from multiple databases (among which are SwissProt, TrEMBL, PIR, PDB, SCOP, DBJ, GenBank, REF and others). Of the 6298 yeast genes, 903 can be assigned an EC number, either based on their annotation or the annotation of entries with identical sequences from 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.

7.1.3 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.

7.1.4 Homology clusters To find all homology relationships we run an all-vs-all comparison of the yeast sequence database using BLAST (Altschul et al. 1990), discarding similarities with evalue > 0.1. The genes are then clustered using the ProtoMap hierarchical clustering algorithm (Yona et al. 1999). The sequences cluster into 839 clusters and 3206 singletons. The largest cluster contains 113 proteins, almost all of which are protein kinases.

7.2 Measures of expression similarity

7.2.1 Global measures 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 Spearman rank correlation (D'haeseleer et al. 1997) and the jack knife correlation (Heyer et al. 1999). We tested the Euclidean metric, Pearson correlation, Spearman correlation and two other variants[4]. The measures are simple and straight-forward and their definitions are provided just for the sake of completeness.

Given two expression vectors $\mathbf{v}$ and $\mathbf{u}$ of dimension d, we define the following basic measures:

- The normalized Euclidean metric

$$Dist_{euc}(\mathbf{v},\mathbf{u}) = \sqrt{\frac{1}{d}\sum_{i=1}^{d}(v_i - u_i)^2}$$

- Pearson correlation

$$Corr(\mathbf{v},\mathbf{u}) = \frac{1}{d}\sum_{i=1}^{d}\frac{(v_i - <\mathbf{v}>)(u_i - <\mathbf{u}>)}{\sigma_V \sigma_U}$$

- Spearman rank correlation. To define this measure first rank the experiments in each vector based on the expression values. Then

$$Spearman(\mathbf{v},\mathbf{u}) = \rho = 1 - \frac{6\sum_{i=1}^{d} x_i^2}{d^3 - d}$$

  where x_i is the difference in rank for the i-th experiment
- The **direction distance**. To define this distance we first replace each expression vector $\mathbf{v}$ 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[5]. In this study we considered only pairs of vectors whose dimension was greater than 11.

Note that all measures but the Spearman rank correlation are normalized by the dimension. To eliminate the dependency on the dimensionality when using Spearman correlation we also tested a variant that replaced the correlation coefficient with its statistical significance as described in (StatLib 1975). The statistical significance of a specific correlation value is estimated by computing the probability p of getting a value greater than or equal to $\sum_i x_i^2$ in d experiments. For identical profiles, the Spearman rank correlation is 1 and so is p.

Shifts: In time-series data similar expression patterns might be shifted since one gene might activate another (directly or indirectly). To allow for time-delayed responses we consider shifts. Since the typical cellular reaction times are almost instantaneous, it is unlikely to observe a delay of minutes or hours (the typical time interval between consecutive measurements in time-series expression data) and therefore we consider only shifts of one time point to the left (-1) or to the right (+1) (these shifts can be adjusted based on the data set). Given two expression profiles and a similarity or distance measure, their time-shift-similarity is defined as the *maximum* over their similarity with shifts -1,0 and 1 (where 0 stands for the similarity of the non-shifted profiles).

7.2.2 Zscore based measures Measures such as the Pearson correlation or the Euclidean distance do not necessarily entail correlation or co-regulation and pairs of genes which score well under these measures are not necessarily functionally linked. To determine which similarities are more likely to be due to an underlying biological phenomena one needs to assess these similarities in a statistically meaningful way.

Previously, methods to determine whether two samples are differentially 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). Here we apply similar permutation methods. 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 (100 times here) to determine the background distribution of distances between random vectors[6]. This distribution is different for each pair of genes (see Figure 4).

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. 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

[4] The jack-knife measure was not tested, as it seems more effective for detecting outliers than for detecting similarity, as suggest in (Heyer et al. 1999).

[5] The performance was worse without normalization.

[6] The empirical distributions can be very well approximated by the unimodal normal distribution, as expected by the central limit theorem, given the functional form of the similarity measures.

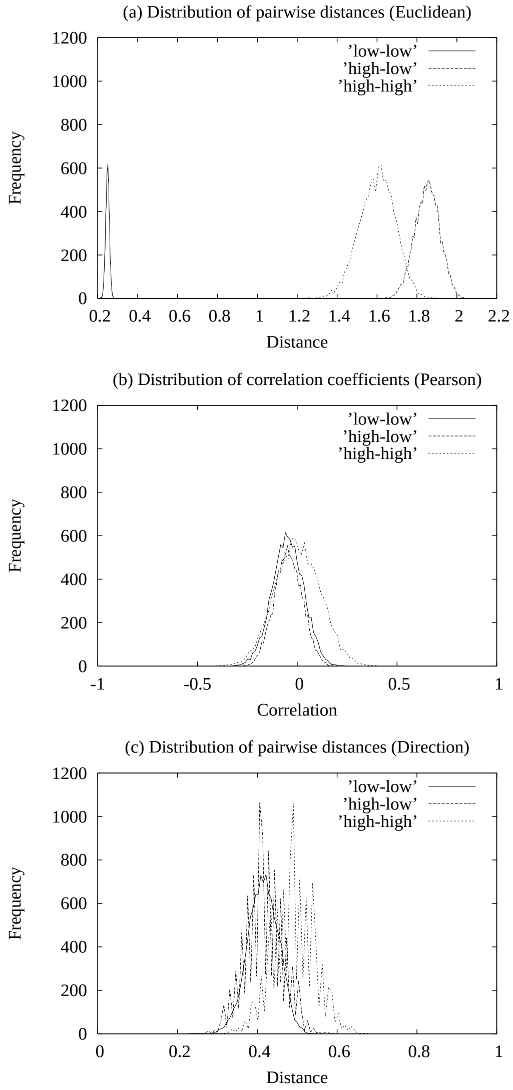


Fig. 4. Background distribution of distances for three pairs of genes (a) with the Euclidean metric (b) with Pearson correlation (c) with the direction measure. For each metric we show the distributions for three pairs of genes of different norms (the Euclidean norm of the common features in these gene pairs was calculated to determine whether the gene had high or low norm). For example, the low-high curve corresponds to a pair of expression profiles, one with low norm and one with high norm, and is based on 100 permutations of the expression vectors. Despite dimension-normalization, all measures are sensitive to the norm (most pronounced with the Euclidean measure), and can differ markedly from one pair of genes to another (especially with the popular Euclidean metric), emphasizing the inadequacy of the raw similarity/distance measures.

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*.

7.2.3 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. To improve sensitivity and accuracy we tested several combinations of the zscore-based measures: **(1)** A sum of the zscores assigned by the individual measures. The *SumAll* measure is the sum of the Pearson, direction, and Euclidean permuted measures. Other variations considered partial sums of these measures. For example, the *EucPear* is the sum of the Euclidean and Pearson permuted measures

$$EucPear(\mathbf{v},\mathbf{u}) =$$
$$Zscore[Corr(\mathbf{v},\mathbf{u})] - Zscore[Dist_{euc}(\mathbf{v},\mathbf{u})]$$

(2) A weighted sum of the measures (where the weights are defined based on the performance of the individual measures, as described in the next section); **(3)** The maximum similarity over all measures; **(4)** A probability-based measure: given the zscores of the three individual measures we compute the probability to observe these zscores by chance, based on the normal approximations of the background distributions (see section 7.3 in Supplementary Material), with the simplifying independence assumption $pvalue(z_1, z_2, z_3) = pvalue_{euc}(z_1)pvalue_{pear}(z_2)pvalue_{dir}(z_3)$.

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.

7.2.4 Measures of local similarity The measures described above are all global measures that take all measurements into account. However, it is quite likely that genes will be expressed similarly only along a subset of the experiments. To compute the local similarity of two expression profiles we implemented a method similar to the one described in (Qian et al. 2001) that is based on a dynamic programming algorithm[7]. To detect *local* similarities one needs a scoring function that can assign positive and negative scores to individual pairs of measurements v_i and u_j. This is achieved by taking the product $v_i \cdot u_j$, measuring correlation between the two values. To avoid bias due to differences in magnitudes, the vectors are first normalized such that each has zero mean and standard deviation of one. Given this scoring function, a dynamic programming algorithm is then invoked to find the most similar subsets of measurements. Unlike sequence comparison, gaps are not allowed. With non time-series data, a gapped match might align arbitrarily different conditions. If the data is a time-series dataset, the meaning of a gap is that a time delay is introduced. However, if two genes are co-expressed (possibly with a time-shift) it is unlikely that one will re-synchronize itself to the other, once they departed (a gap was introduced). Shifts are allowed in time-series data, but to avoid arbitrary time shifts we introduce a modulation that decreases the similarity score exponentially as a function of the time delay such that

$$score(i,j) = v_i u_j e^{|i-j|}$$

[7] Our program is available at `biozon.org/ftp/software/exim/`

7.2.5 Correlation vs. anti-correlation Not all functional links manifest themselves as similarities. 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 even more interesting. That might happen for example when there is a regulatory relation between two genes (e.g. when one gene suppresses another). Since it is either the correlation or the anti-correlation that might indicate a functional link we studied also significant anti-correlations. Of the **global** similarity/distance measures only the Pearson correlation measure seems to be effective for detection of anti-correlation. To detect **locally** anti-correlated genes we use the same algorithm described in section 7.2.4 with a small modification to the scoring function suggested in (Qian et al. 2001): specifically, the score of matching measurement i with measurement j is $-v_i \cdot u_j$. As before we introduce the time-shift modulation such that

$$score(i,j) = -v_i u_j e^{|i-j|}$$

If the expression values are given in the form of log-ratio then all measures can be adjusted to measure anti-correlation by mapping one of the expression profiles to its negative.

7.3 The statistical significance of expression similarity

Each of the similarity measures described in the paper can be associated with a significance measure. The significance value is computed based on the distribution of pairwise similarities for randomly selected pairs.

To estimate the significance of *local similarities* we derived the background distributions of similarity scores for a large population of expression profiles. This empirical distribution can be modeled with the extreme value distribution (Gumbel 1958; Dembo & Karlin 1991). Unlike the estimates in (Qian et al. 2001), we adjust for the size of the search space (number of common features) and derive different distributions for different dimensions. While the dimension-independent significant measure maintains the monotonicity of scores, the dimension-normalized measure might change the order in which hits are reported as higher score no longer entails more significant match.

All other measures follow the normal distribution as is expected by their functional form[8]. The distributions of distances for the mass-distance measure over the Time-series 1998 and the Rosetta 2000 datasets are shown in Figure 5a,c. From these distributions we derive the pvalue of any expression similarity score S (Figure 5b,d). The evalue of a similarity score S is obtained by multiplying the pvalue by the number of tests (pairwise comparisons), that is $evalue(S) = \frac{N\cdot(N-1)}{2} pvalue(S)$ where $N = 5894$ is the number of genes with expression profiles. The evalue is a measure of the expected number of occurrences of chance similarities with a score S or higher. An evalue of 0.1 for example means that on average one needs to perform $10 \cdot \frac{N\cdot(N-1)}{2}$ pairwise comparisons (i.e. compare all pairs of genes in 10 genomes as big as the yeast genome) before encountering one such chance similarity.

7.4 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 relations 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. That is, the number of true edges that are expected to occur in a set of n randomly chosen edges is $e = n\frac{T}{N(N-1)/2}$ where N is the total number of genes in our set and T is the total number of true relationships (total number of edges).

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 a 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.

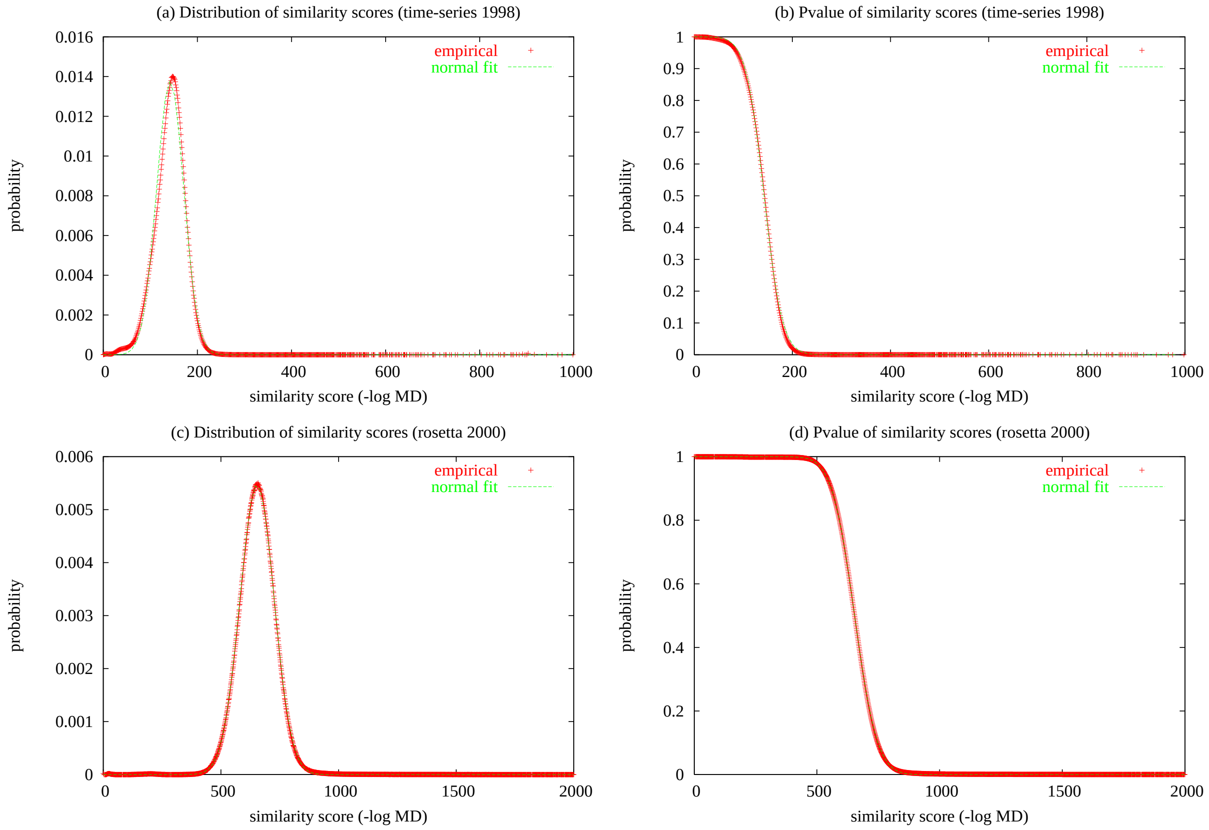
7.5 Comparison of similarity measures

In Figure 6 we report the results of comparing different groups of similarity measures that belong to different categories.

[8] For example, the mass-distance measure in it logarithmic form is a sum of all the $MASS$ values over all experiments. Since each experiment is typically distributed normally, so are the $MASS$ random variables, and so is the total sum. Therefore, the resulting mass-distance is also normally distributed.

Dataset	Pathway		Homology		Promoter		Interactions		Total
	observed (expected)	ratio	observed (exp.)	ratio	observed (exp.)	ratio	observed (exp.)	ratio	
Rosetta	448 (23.9)	18.8	765 (24)	31.9	61 (4.3)	14.2	19 (8.7)	2.2	1214
Stress 2000	179 (23.9)	7.5	274 (24)	11.4	5 (4.3)	1.2	19 (8.7)	2.2	442
Stress 2004	139 (28.5)	4.9	119 (21)	5.7	8 (4.5)	1.8	11 (8.8)	1.2	256
Time Series 1998	124 (23.9)	5.2	533 (24)	22.2	32 (4.3)	7.4	20 (8.7)	2.3	670

Table 4. Correlation of different types of relations with expression similarity. For each dataset we show the breakup of pairs of truly related genes whose expression similarity is among the top 20,000 (measured with the mass-distance measure), according to the type of relationship. The results are weaker than those reported in Table 3, where only relations with significant expression similarity are analyzed. However, they confirm the trends observed in Table 3. 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 total. For each type we also compute the number of such relationships that are expected to occur by chance (in parentheses) and the ratio observed/expected.



(a) Distribution of similarity scores (time-series 1998)

(b) Pvalue of similarity scores (time-series 1998)

(c) Distribution of similarity scores (rosetta 2000)

(d) Pvalue of similarity scores (rosetta 2000)

Fig. 5. Significance of expression similarity. The distribution of mass-distances and their pvalue for the Time-series 1998 dataset (a,b) and the Rosetta dataset (c,d).

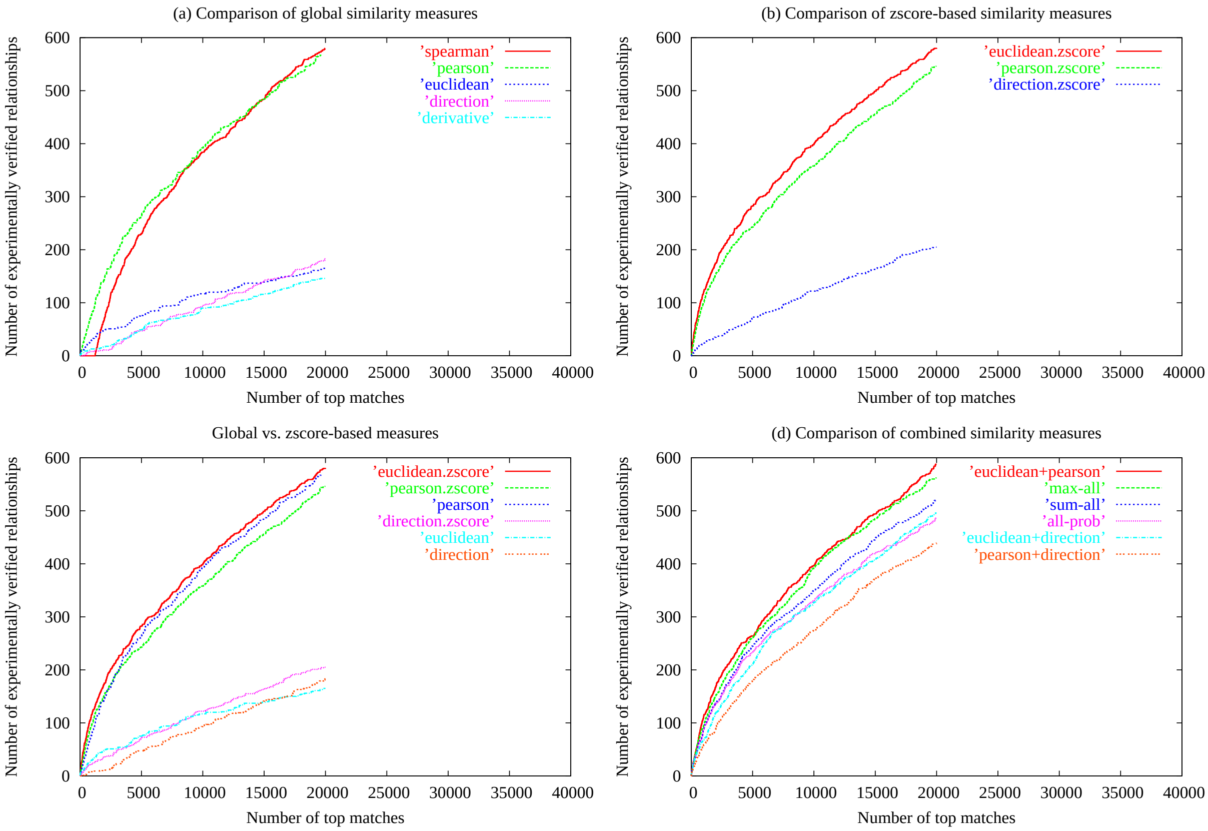


Fig. 6. Performance evaluation of different similarity measures on the Time-series 1998 dataset. (a) Global measures. (b) Zscore-based measures. (c) Global vs. zscore-based measures. (d) Combined measures. See section 3.2 for details.