BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 13 2006, pages 1616–1622
doi:10.1093/bioinformatics/btl127

Gene expression

Effective similarity measures for expression profiles

Golan Yona[1,2,*,†], William Dirks[2], Shafquat Rahman[1] and David M. Lin[3]

[1]Department of Computer Science, [2]Center of Applied Mathematics and [3]Department of Biomedical Sciences, Cornell University, NY, USA

Received on September 27, 2005; revised on March 10, 2006; accepted on March 29, 2006
Advance Access publication April 4, 2006
Associate Editor: Martin Bishop

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 a priori 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/

Contact: golan@cs.technion.ac.il

Supplementary information: Supplementary data are available at *Bioinformatics* online.

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 and Fostel, 2000; McCormick *et al*., 2003; Yoo *et al*., 2003; Diffee *et al*., 2003; Lopez *et 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). In addition, 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 are still considered a substantial source of information on cellular activity and regulation, and the data collected from such studies are 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 owing to the nature of microarray data. Expression data are noisy and in many cases unreliable. Many factors may affect the experiment and 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, e.g. 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 in cells. Our model organism is Yeast, for which extensive experimental information on gene relationships exist.

*To whom correspondence should be addressed.
†Present address: Department of Computer Science, Technion, Haifa 32000, Israel.

2.1 Expression data

We used four different datasets 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 dataset 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 dataset (Time-series 1998) has been normalized by Spellman *et al*. (1998) to correct for experimental variation between the different microarrays.

Our second dataset is the Rosetta Inpharmatics yeast compendium data. This microarray data consist 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 dataset as the Rosetta 2000 dataset.

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 are 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 dataset 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 dataset. This is the subset of genes used in our study. The Stress 2004 dataset contains a smaller number of genes (4699) because of 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 loss 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; e.g. two-thirds of the fly genes have no observable loss of function phenotype under knockout experiments (Miklos and Rubin, 1996).

Table 1. The relation graph

Relation type	#genes	#edges
Interaction	3592 (3454)	5339 (5052)
Sequence	3092 (2852)	19 074 (13 950)
Promoter	213 (213)	2439 (2439)
Pathway	642 (605)	15 789 (13 914)
Total	5079 (4815)	41 902 (34 682)

Number of genes and edges (true relationships) in each dataset. 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.

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 dataset. (Information on the different datasets used in this study is available in the Supplementary Material, Section 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 (*Z*-score-based), local similarity measures that are based on the dynamic programming algorithm (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 2 of the Supplementary Material. We also test the new mass-distance (MD) 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 the more similar are the two profiles.

Given two expression profiles u and 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

$$\mathrm{MASS}(u_i, v_i) = \int_{\min\{u_i, v_i\}}^{\max\{u_i, v_i\}} \mathrm{Prob}_i(x)\,\mathrm{d}x,$$

where $\mathrm{Prob}_i(x) = N_{(\mu_i,\sigma i)}(x)$ 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 versus liver) might exhibit a bi-modal or other distribution. In such cases one could use the empirical distributions when computing the MASS variables

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

where $\mathrm{freq}(x)$ is the empirical frequency of the measurement x.

The MD of u, v is defined as the total volume of samples bounded between the two expression profiles and is estimated by the product over all coordinates

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

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

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

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 datasets 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 datasets (as confirmed by our results over the Rosetta dataset, see next section). The MD 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.

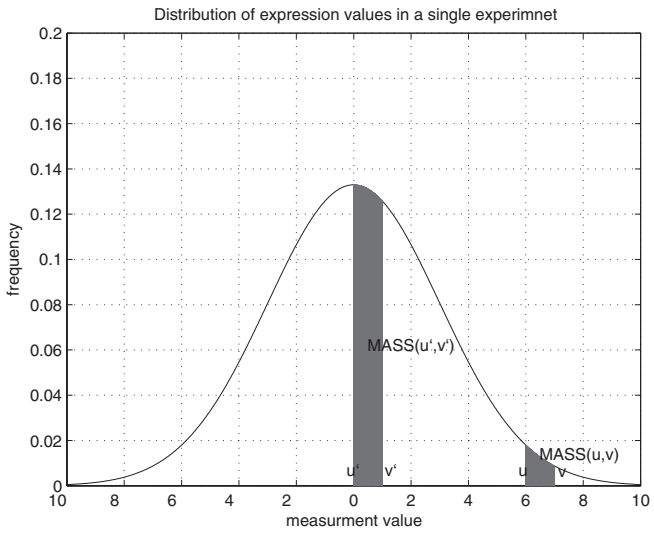

Fig. 1. The mass-distance. Often is the case that the distance between two measurements depends not only on the relative nominal difference between the measurements but also on the background distribution. For example, two measurements u and v are statistically more similar to each other than the two 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.

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 datasets 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 and McNeil, 1982), where the number of true positives is plotted versus 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 dataset 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, Supplementary Material). Introducing shifts in global measures did not improve the performance and even decreased it slightly (by 3%).

All *Z*-score-based measures improved over the original global measures, as is shown in Figure 6b of Supplementary Material. The most drastic improvement was observed with the Euclidean-based *Z*-score measure that outperformed even the Pearson correlation measure and the corresponding *Z*-score 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 with the results with the raw rank correlation values).

The combined measures perform quite well compared with the individual measures (Figure 6c of Supplementary Material shows the results for a subset of the measures attempted). This is not surprising; 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 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 with 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 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 20000 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 a regulatee) for which these measures are most suited, and we intend to revisit this analysis once more data about regulator–regulatee relations are made available.

A summary of the best results from each category (for the Time-series 1998 dataset) is shown in Figure 2. As this graph indicates, all measures contain some information about the true relationships, when compared with what is expected by a random guess (see Section 4 in Supplementary Material). Of all measures tested, the MD 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 Z-score measure (each one is the best in its category, see Figure 6 of Supplementary Material). These are closely followed by the Spearman rank correlation and far behind is the Euclidean metric. Similar results were obtained with two other expression datasets (Fig. 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.

While in general no single measure can be crowned as the best for all possible datasets, the MD measure seems to be the most effective one over two of the datasets tested, with significant improvement in the $ROC_{20\,000}$ values that amounts to 22.3% (Time-series 1998) and 11.4% (Rosetta 2000) over the next best method. The MD measure comes second for the Stress 2004 dataset. The best method over this dataset is the Spearman rank correlation that improves over the MD measure by 9.5%. Both measures are outperformed over the Stress 2000 dataset. Surprisingly, the best metric

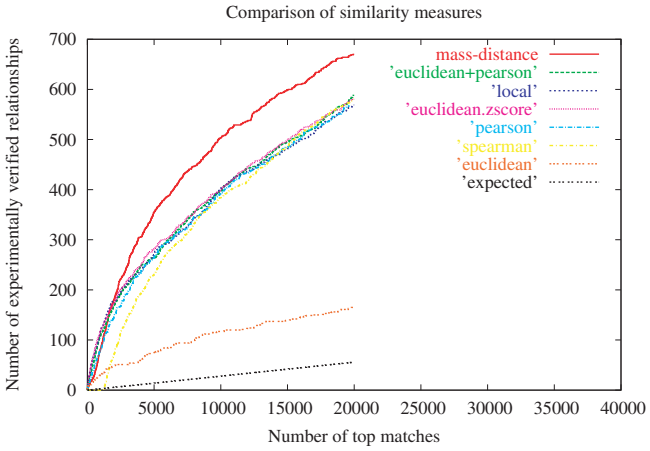

Fig. 2. Performance evaluation on the Time-series 1998 dataset. For clarity, the labels are ordered according to the performance of the corresponding measures. Only a subset of all tested measures is displayed (for more results, see Figure 6 in Supplementary Material). The expected number of relationships is computed under the uniform random setup (see Section 4 in Supplementary Material).

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 MD 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 MD 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 dataset. 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 20 000, compared with 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 e-value < 10 and compute the number of true relationships that are detected (statistical significance of expression similarity is discussed in Section 3 in Supplementary Material). These numbers are compared with 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 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 with the number of such relationships that are expected to happen by chance. Since the experimental data are very partial, this is a lower bound on the

[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 ∼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.

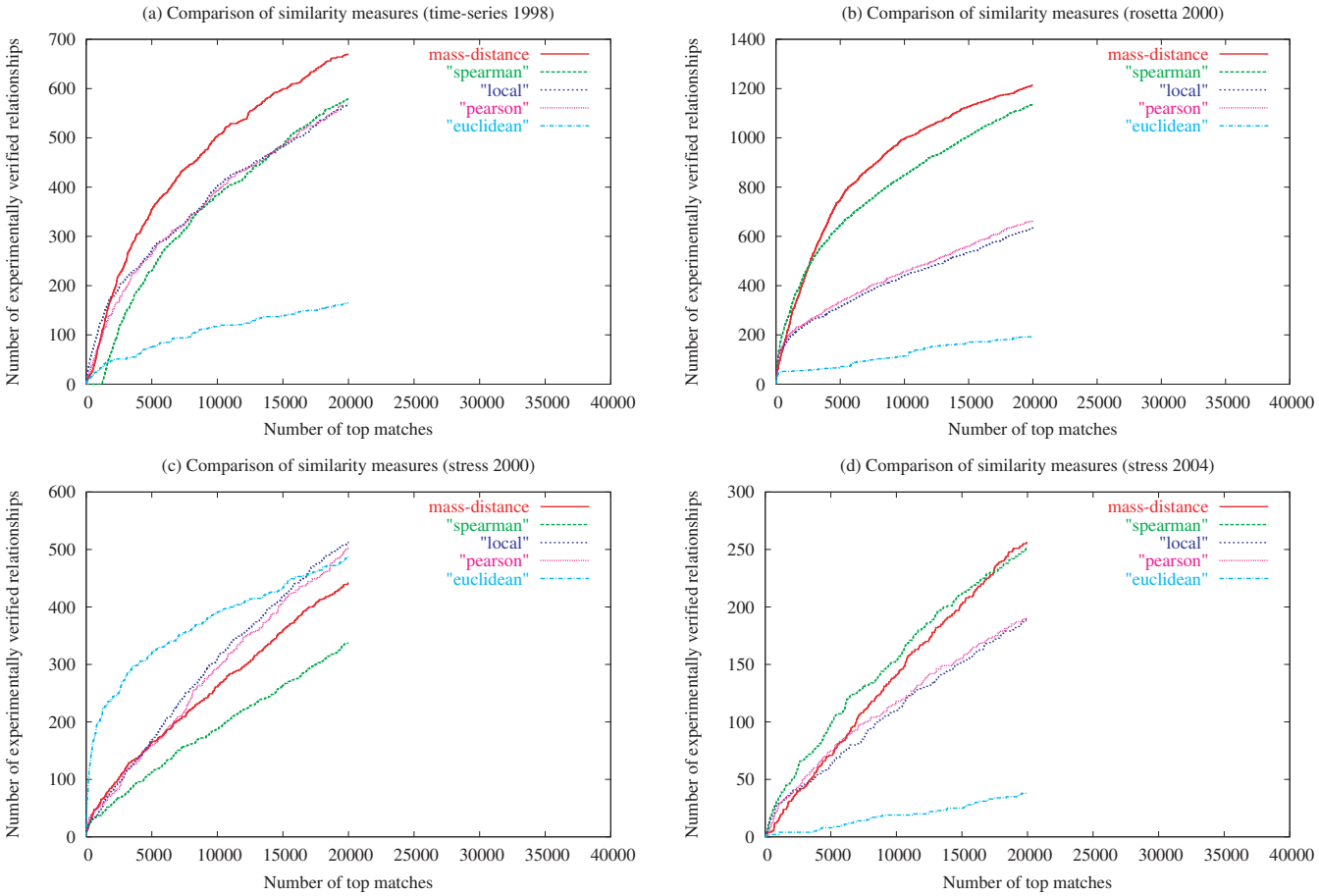

Fig. 3. Performance evaluation of selected metrics on different expression datasets. (*a*) Time-series 1998. (*b*) Rosetta 2000. (*c*) Stress time-series 2000. (*d*) Stress time-series 2004. For each dataset we show the results for the Euclidean metric, the Pearson correlation, the Spearman rank correlation,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

Table 2. Information content of expression similarity

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

Expression similarity was computed using the MD 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 dataset. The second and the third columns are the numbers of edges and true edges above the threshold (e-value < 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.

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 versus 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 datasets. For the Rosetta dataset, sequence homology is >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.

Table 3. Correlation of different types of relations with expression similarity

Dataset	Pathway observed (expected)	ratio	Homology observed (expected)	ratio	Promoter observed (expected)	ratio	Interactions observed (expected)	ratio	Total
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

For each dataset we show the breakup of pairs of truly related genes whose expression similarity is assigned *e*-value < 10 (measured with the MD 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.

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 have proved to be very effective for pathway prediction in Popescu and 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 DATASETS

It is expected that new measures of similarity based on other principles might be more effective in detecting functional links. Moreover, other expression datasets (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 datasets and compare them with those which were already tested, we constructed a web server that is available at biozon.org/tools/expression/. Users can evaluate a new similarity measure using the same expression datasets, comparing the expression profiles and uploading the results with our web server. 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 datasets and include new relations that indicate other types of functional links. The data will be extracted from the Biozon database (Birkland and 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 are 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 datasets (Popescu and 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 are 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, e.g. 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 *Z*-score-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 MD measure that significantly outperforms the other metrics on some datasets. 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 MD 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. 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 MD 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 become 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 these 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 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 become available. Indeed, when tested against a new set of protein–protein interactions, many of the putative functional links were confirmed as interacting proteins (e.g. of the top 20 000 pairwise relationships over the Time-series 1998 dataset, 159 are due to new interactions compared with 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 versus time-series). The results are expected to be even better for higher quality datasets and as more data become available. Finally, our tools can be applied also to other types of expression data and other similarity measures through our web server at biozon.org/tools/expression/

[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.

ACKNOWLEDGEMENTS

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.

Conflict of Interest: none declared.

REFERENCES

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

Birkland,A. and Yona,G. (2006) The BIOZON Database: a hub of heterogeneous biological data. *Nucleic Acids Res.*, **34**, D235–D242.

Cho,R.J. *et al.* (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.

Diffee,G.M. *et al.*, Seversen,E.A., Stein,T.D. and Johnson,J.A. (2003) Microarray expression analysis of effects of exercise training: increase in atrial MLC-1 in rat ventricles. *Am. J. Physiol. Heart Circ. Physiol.*, **284**, H830–H837.

Gasch,A.P. *et al.* (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.

Gygi,S.P. *et al.* (1999) Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.*, **19**, 1720–1730.

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

Hughes,T. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Liu,E.T. (2003) Classification of cancers by expression profiling. *Curr. Opin. Genet. Dev.*, **13**, 97–103.

Lopez,I.P. *et al.* (2003) DNA microarray analysis of genes differentially expressed in diet-induced (cafeteria) obese rats. *Obes. Res.*, **11**, 188–194.

McCormick,S.M. *et al.* (2003) Microarray analysis of shear stressed endothelial cells. *Biorheology*, **40**, 5–11.

Miklos,G. and Rubin,G. (1996) The role of the genome project in determining gene function: insights from model organisms. *Cell*, **86**, 521–529.

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

Qian,J. *et al.* (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,M. *et al.* (2004) Disruption of yeast forkhead-associated cell cycle transcription by oxidative stress. *Mol. Biol. Cell*, **15**, 5659–5669.

Spellman,P.T. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

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

Yeatman,T.J. (2003) The future of clinical cancer management: one tumor, one chip. *Am. Surg.*, **69**, 41–44.

Yoo,M.S. *et al.* (2003) Oxidative stress regulated genes in nigral dopaminergic neurnol cell: correlation with the known pathology in Parkinson's disease. *Brain Res. Mol. Brain Res.*, **110**, 76–84.