Beyond Synexpression Relationships: Local Clustering of Time-shifted and Inverted Gene Expression Profiles Identifies New, Biologically Relevant Interactions

Jiang Qian, Marisa Dolled-Filhart, Jimmy Lin and Mark Gerstein*

Department of Molecular Biophysics and Biochemistry,
Yale University, 266 Whitney Avenue, PO Box 208114,
New Haven CT 06520-8114, USA.

*To whom correspondence should be addressed

The complexity of biological systems provides for a great diversity of relationships between genes. The current analysis of whole-genome expression data focuses on relationships based on global correlation over a whole time course, identifying clusters of genes whose expression levels simultaneously rise and fall. There are, of course, other potential relationships between genes, which are missed by global clustering. These include activation, where one expects a time-delay between related expression profiles, and inhibition, where one expects an inverted relationship. Here we propose a new method, which we call local clustering, for identifying these time-delayed and inverted relationships. It is related to conventional gene expression clustering in a fashion formally similar to the way local sequence alignment (Smith-Waterman algorithm) is derived from global alignment (Needleman-Wunsch). We applied our method to the yeast cell-cycle expression dataset and were able to detect a considerable number of additional biological relationships between genes, beyond those resulting from conventional correlation. We related these new relationships between genes to their similarity in function (determined from the MIPS scheme) or their having known protein-protein interactions (determined from the large-scale two-hybrid experiment), finding that genes strongly related by local clustering were considerably more likely than random to have a known interaction or a similar cellular role. This suggests that local clustering may be useful in functional annotation of uncharacterized genes. We examined many of the new relationships in detail. Some of them were already well-documented examples of inhibition or activation, which provide corroboration for our results. For instance, we found an inverted expression profile relationship between genes *YME1* and *YNT20*, where the latter has been experimentally documented as a bypass suppressor of the former. Other relationships were new, often involving uncharacterized yeast genes and thus suggesting functions for many of them. In particular, we found a time-delayed expression relationship between *J0544* (which has not yet been functionally characterized) and four genes associated with the mitochondria. This suggests that *J0544* may be involved in the control or activation of mitochondrial genes. Our clustering program and a detailed website of clustering results is available at:http://bioinfo.mbb.yale.edu/expression/cluster.

Key words: expression profile; local clustering; time-shifted; inverted; bioinformatics

Introduction

The massive datasets generated by microarray experiments presents a challenge to those interested in studying the regulatory relationship between genes (Ermolaeva *et al.*, 1998; Gaasterland & Bekiranov, 2000; Hegde *et al.*, 2000; Kim *et al.*, 2000; Shalon *et al.*, 1996). Up to now, one of the main challenges has been to devise methods for grouping together genes that have similar expression profiles; this is done to determine clusters of genes that are transcribed together as cellular conditions vary. The most obvious use of such clusters is an improved understanding of transcription regulatory networks within genomes. Genes with similar expression profiles are likely to be subject to identical, or related, transcriptional control. This fact has been used to search for binding site motifs common to corregulated genes (Bussemaker *et al.*, 2001; Hughes *et al.*, 2000a).

There are further applications of gene clusters, especially in combination with other data sources such as cellular localization, metabolic functions, and intermolecular interactions of the protein products (Drawid *et al.*, 2000; Jansen & Gerstein, 2000). Microarray technology allows for studying the entire genome, while the other types of data are available only for about a third of the genes. Therefore, researchers have attempted to predict protein functions and interactions by expression clustering. This is based on 'guilt by association' (Altman & Raychaudhuri, 2001), the premise that proteins with similar expression profiles (i.e., synexpression relationship) probably hold similar functions or interact with each other (Eisen *et al.*, 1998; Gerstein & Jansen, 2000; Marcotte *et al.*, 1999; Niehrs & Pollet, 1999).

Given the central importance of gene clusters in the studies just described, computational methods have been devised to (i) assess the similarity between pairs of expression profiles from different genes, and then (ii) group together those genes with similar profiles. Effectively, the two aims are analogous to approaches in protein sequence analysis, where there are methods for assessing sequence similarity between pairs of proteins (e.g. BLAST (Altschul *et al.*, 1990)) and then grouping them into homologous families (e.g. Pfam (Bateman *et al.*, 2000) or Protomap (Yona *et al.*, 2000)).

The most common algorithms for grouping genes with related profiles are hierarchical clustering (Eisen *et al.*, 1998; Wen *et al.*, 1998), self-organizing maps (Tamayo *et al.*, 1999; Toronen *et al.*, 1999), and K-means clustering (Tavazoie *et al.*, 1999). Hierarchical methods were originally derived from algorithms used to construct phylogenetic trees, and group genes in a "bottom-up" fashion; genes with the most similar expression profiles are clustered first, and those with more diverse profiles are included iteratively. In contrast, the self-organizing maps and K-means methods employ a "top-down" approach in which the user predefines the number of clusters for the dataset. The clusters are initially assigned randomly, and the genes are regrouped iteratively until they are optimally clustered. Bayesian and neural networks provide additional approaches toward clustering (Friedman *et al.*, 2000).

Prior to clustering, users must define all the pair-wise similarities between the individual expression profiles. Up to now, the most popular measure that has been employed is the Pearson correlation coefficient; given a pair of genes, this method compares the expression levels at each time-point and measures the variation across the whole profile. The score, the coefficient *r*, ranges from –1 to 1, where –1 signifies perfect negative correlation, 0 indicates no correlation and 1 a perfect positive correlation. Gene pairs with scores approaching 1 are considered to have similar expression profiles, as shown in Figure 1A. Other measures include squared Pearson correlation coefficient (D'haeseleer *et al.*, 1997), Spearman rank correlation (D'haeseleer *et al.*, 1997), jackknife correlation coefficient (Heyer *et al.*, 1999), and Euclidean distance (Wen *et al.*, 1998).

The major drawback in these measures is that they ignore many additional relationships implicit in expression time courses. For instance, a gene may control or activate another gene downstream in a pathway; in this case, the expression profiles may be staggered, indicating a time-delayed response in the transcription of the second gene. Other genes may have an inhibitory relationship -- i.e. as one rises the other falls in response -- and we can expect their expression profiles to be inverted with respect to each other (or inverted with a time-delay). The current methods using correlation coefficients fail to detect these important relationships. First, they only assess global similarities between expression profiles, thereby missing staggered relationships. Second, negative correlations have not previously been considered, thus inhibition has been ignored. Here, we propose a new algorithm; it is based on the dynamical programming method for local sequence alignment (Smith & Waterman, 1981) and hence we call it local clustering. Its development from traditional gene expression clustering method (e.g. (Eisen *et al.*, 1998)) is strongly suggested by the way local sequence alignment (Smith & Waterman, 1981) followed on the original global approach (Needleman & Wunsch, 1970).

Using local clustering, we can identify expression profiles that have one of the following relationships:

1) Simultaneous correlation (Figure 1A) – The expression profiles of the two genes are synchronous and coincident. Genes with such profiles are expected to be subject to identical transcriptional regulation, which are sometimes called synexpression (Niehrs & Pollet, 1999). This is the only type of relationship currently detected using the traditional correlation coefficient.

2) Time-delayed correlation (Figure 1B) – The profiles of the two genes are similar, but one is time shifted, or out of phase, with respect to the other. The expression of some genes may be delayed compared to others due to a time lag in their transcription control.

3) Inverted correlation (Figure 1C) – The profiles of the two genes are inverted, with one of the profiles flipped on the time axis relative to the other. These profiles may exist where the expression of one gene inhibits or suppresses the expression of the other.

4) Inverted and time-delayed correlation – This combines time-shifted and inverted correlations, so in addition to being inverted, the profile of one gene is staggered with respect to the other.

As a test of the effectiveness and accuracy of our algorithm, we applied it to a yeast cell cycle dataset published by Cho *et al.* (1998). Affirmatively, our algorithm detected simultaneous correlations, as well as time-shifted, inverted and inverted-time-shifted relationships. Many of our predicted interactions were confirmed with published gene pair relationships. Furthermore, the algorithm proposes highly correlated gene pairs representing novel pairs of gene relationships.

Algorithms and Datasets

Local Alignment between Pairs of Expression Profiles

We use a degenerate dynamical programming algorithm to find the time shift and inverse correlation between expression profiles. This algorithm versus the traditional methods based on the correlation coefficient, is analogous to local alignment (Smith & Waterman, 1981) versus global alignment (Needleman & Wunsch, 1970) for protein sequences. The algorithm does not allow gaps between consecutive time points in the current version. However, there are some obvious extensions, which we explore later in the discussion section.

Suppose there are *n (1,2,…n)* time-point measurements in the profile. First, the expression ratio is normalized in "Z-score" fashion, so that for each gene the average expression ratio is zero and standard deviation is 1. The normalized expression level at time point *i* for gene *x* is denoted as x_i. Consider a matrix of all possible similarities between the expression ratio for gene *x* and gene *y*. This matrix can also be called a "score matrix". In our algorithm, it is defined as $S(x_i, y_j) = x_i y_j$. For simplification, it will be referred as $S_{i,j}$ for comparison of any two genes.

Then, two sum matrices $E_{i,j}$ and $D_{i,j}$ are calculated as $E_{i,j} = \max(E_{i-1,j-1} + S_{i,j}, 0)$ and $D_{i,j} = \max(D_{i-1,j-1} - S_{i,j}, 0)$. The initial conditions are $E_{0,j} = 0$ and $E_{i,0} = 0$, and the same initial conditions are also applied to the matrix of *D*. The central idea is to find a local segment that has the maximal aggregated score, i.e., the sum of $x_i \cdot y_j$ in this segment. Finally, a maximal value $S=V_{m,n}$ in matrix $E_{i,j}$ and $D_{i,j}$ is found as the match score for these two expression profiles. If the maximum is off diagonal in the matrix, i.e., $m \neq n$, these two expression profiles have a time-shift relationship. A maximal value from $D_{i,j}$ indicates these two profiles have an inverted relationship. At the end of this procedure, one obtains a match score and a relationship, i.e., simultaneous, time-delayed or inverted. Obviously, for the gene pairs with a very low match score, even though they are also assigned a relationship, we can classify them as "unmatched".

Figue 1E is the corresponding matrix $E_{i,j}$ for the expression profiles shown in Fig. 1B. The matrix $D_{i,j}$ for these expression profiles is not shown here because the maximal value is not in this matrix. The match score for these expression profiles, a score of 19, is highlighted in the dark gray cell. There is a time delay (time shift) in their relationship because the match score of 19 is not on the main diagonal of the matrix ($m \neq n$). Figure 1F is the corresponding matrix $D_{i,j}$ for the profiles shown in Fig. 1C. The match score is 20; and because the maximum value is from matrix $D_{i,j}$ rather than in matrix $E_{i,j}$ (not shown), these expression profiles are correlated in an inverted manner.

Cell Cycle Dataset and Generation of Similarity Matrix

We tested our algorithm on the yeast whole genome oligonucleotide expression array data generated by Cho et al (Cho *et al.*, 1998), which included over 6,000 ORFs. The data set consists of yeast cultures that were synchronized and sampled at intervals covering nearly two full cell cycles. This experiment was done using an Affymetrix oligonucleotide array (Lockhart *et al.*, 1996) containing oligos complementary to each of the yeast ORFs. The raw data was then scaled to account for the experimental differences between the four arrays used, and the scaled intensities are reported in the Cho data. (Of course, our algorithm can also be applied to a cDNA microarray (Shalon *et al.*, 1996), which measures changes relative to a reference state creating an expression ratio, rather than the measurement of mRNA expression levels as detected in oligonucleotide arrays.) After eliminating the negative expression levels in the Cho scaled measurements, 5,911 genes are included in our calculation.

We applied our local alignment procedure to all possible pairs of gene expression profiles. The match score and type of relationship (simultaneous, time-delayed or inverted) were calculated and assigned for each expression profile pair. This gave a matrix of all pairwise similarities that can be used as raw input of clustering algorithm.

Significance Statistics

In order to estimate the significance of a given match score in the distance matrix, a set of random profiles was generated by shuffling the expression ratio at different time points. The resulting profiles satisfied our earlier normalization conditions that the average ratio be zero, and the standard deviation, one. Using the local alignment procedure, the match scores of the random expression profiles and their distribution were calculated. The P-value is defined as P($S > C$), which is the probability of obtaining a match score larger than C from the random profiles. The smaller the P-value is, the more significant the match score.

Single-linkage Clustering

To define a network from the distance matrix, we used single-linkage neighbor joining clustering, with appropriate thresholds based on the significance statistics (discussed in

context in the text). Of course, based on the distance matrix, we could use other clustering methods, e.g. multiple linkage or K-means. However, as the focus of this paper is the determination of the distance matrix between genes rather than the clustering algorithm, we just choose a simple clustering method.

Overall Approach and Software Package

Throughout the paper we will refer to the results from clustering the local alignment of expression profiles as "local clustering". We will contrast this with results from "traditional, global clustering", which is based on computing a distance matrix only from simultaneous correlations between expression profiles (i.e. the traditional correlation coefficient). For instance, global clustering is the approach used in the original papers of Eisen et al. (1998) and Tamayo et al. (1999).

We have developed a distributed software package for clustering gene expression data sets with our local alignment algorithm. The package also incorporates global clustering and spectral analysis for comparison. It is available from http://bioinfo.mbb.yale.edu/expression/cluster.

Overall Network Topology

To provide a global view of the relationships detected by local clustering, a network is shown in Figure 2A. The threshold used to define connected genes is a match score of 16, which corresponds to a P-value of 10^{-6}. The network consists of 673 nodes (genes) and several large clusters. Dynamic navigation of the network can be obtained from our website. Figure 2B is a close up view of part of a large cluster in the rectangle outlined in Figure 2A. Different types of relationships can be seen in this plot. A gray solid line signifies the conventional simultaneous correlation relationship between two genes, an arrow denotes a time-delayed relationship with the arrow pointing to the delayed gene, and a dashed line denotes an inverted profile relationship. It is clear that by using our algorithm, additional nodes such as YMR320M and YKL177W are joined to this large cluster, making it even larger than if it were formed from simultaneous correlations alone. On the other hand, our method has generated many new clusters such as *SCH9*-YFL067W, as shown in the figure, which are very small. These two competing factors, growing a big clustering and forming new small clusters, can affect the overall connectivity and number of clusters in the network.

To quantitatively compare the network defined by local clustering to that defined by the traditional correlation coefficient, it is useful to compute some quantities. We calculated the average connections per node, i.e., the average number of nodes linked to a particular gene. It is obvious that this quantity depends on the size of the network, i.e., the number of nodes in the network, which in turn is controlled by the P-value threshold used to define the correlation. Figure 3A shows the average connection per node as a function of the number of nodes in the network (and P-value cutoff), generated both by local clustering and the traditional correlation coefficient. The overall pattern in the figure has

the following features: (i) The average number of connections increases with the size of the network. (ii) The two networks have approximately the same average connection below a network size of 200 nodes, suggesting that the highest ranked correlations detected by two algorithms are the same. (iii) However, the average connections diverge above this size. On average, the nodes have more connections in the network generated by the traditional correlation coefficient than that by our algorithm, which indicates that the configurations of these two networks are different. One possibility is that the network formed by our algorithm has more disjoint small clusters rather than large clusters. This is, in fact, the case as confirmed by looking at the number of clusters versus the size of the network, which is shown in Figure 3B. There are more clusters in the network formed by local clustering than by traditional methods. Furthermore, the configurations of the networks generated by our method differs from global clustering, as shown by the detection of more novel clusters in which the nodes have fewer connections.

Examples of Relationships Found by Local Clustering

Here we present some specific examples of profile relationships detected by our algorithm that have been classified as simultaneous, time-delayed or inverted. In addition to looking at how our procedure finds already known and well-documented relationships, we also explore some novel relationships as well as how they can shed light on the function of uncharacterized genes.

Simultaneous Relationships

Well-documented relationships: The majority of the correlated expression profiles have a simultaneous profile relationship, which is the same type of relationship detected by methods based on the simple correlation coefficient (Eisen *et al.*, 1998). Figures 4A and B show two examples that display a simultaneous relationship between gene pairs from the array data. The expression profiles of *RPS11A* and *RPS11B* are shown in Figure 4A. Both of the genes code for the ribosomal protein S11 and are 100 percent identical in sequence (Mewes *et al.*, 2000). *RPS11A* is located on yeast chromosome IV, and *RPS11B* is located on yeast chromosome II. Figure 4B contains the expression profiles of *HXT6* and *HXT7*, which are high-affinity hexose transporters nearly one hundred percent identical in sequence and have nearly identical functions (Boles & Hollenberg, 1997).

Inverted Relationships

Well-documented relationship: Figure 4C shows the profiles of *YME1* and *YNT20*, displaying an inverted relationship. Yme1p (yeast mitochondrial escape) is a metal-dependent and ATP-dependent protease associated with the inner mitochondrial membrane as part of a larger complex of proteins, which is thought to control the assembly and degradation of multi-subunit protein complexes (Hanekamp & Thorsness, 1999). *YNT20* has been identified as a bypass suppressor of Yme1 and is believed to be a part of the Yme1-mediated mitochondrial DNA escape pathway by metabolizing RNA or mitochondrial DNA due to its 3'-5' exonuclease activity (Hanekamp & Thorsness, 1999).

This is a classic example of an inhibitor with an inverted relationship to what it inhibits, and it demonstrates the ability of our algorithm to find a known inverted relationship.

New, Suggested Relationship: Local clustering also detects a previously unknown but highly plausible relationship. Figure 4D displays the inverted gene expression profile relationship of *PUT2* and *SER3*, which are both enzymes of amino-acid metabolism. Put2p is a P5C dehydrogenase that carries out the second step in proline degradation to glutamate, allowing proline to be used as a nitrogen source (Brandriss, 1983). Ser3p is a 3-phosphoglyerate dehydrogenase that is involved in the synthesis of serine from glycolytic intermediates (Melcher & Entian, 1992). It was already found that Put2p could be inhibited by serine (and other amino acids) (Lundgren & Ogur, 1973). Therefore, even though it has not been directly shown that Ser3p inhibits Put2p, based on the related evidence between serine inhibition of Put2p, it is highly likely that this specific enzyme in serine synthesis could also inhibit Put2p as shown by our algorithm.

Time-delayed Relationships

Strongly Documented Suggested Relationship: The expression profiles of *ARC35* and *ARP3* are shown in Figure 4E. Arc35p is a subunit of the Arp2/3 complex in yeast. The Arp2/3 complex is involved in endocytosis and actin cytoskeleton organization (Schaerer-Brodbeck & Reizman, 2000a). Arc35p is required in late G1 actin cytoskeleton organization (Schaerer-Brodbeck & Reizman, 2000b). The expression profiles of *ARC35* and *ARP3* show a time-delayed relationship. Expression of *ARC35* is one time-point (20 minutes) delayed compared to *ARP3*.

New, Suggested Relationship: In addition to shedding light on known interactions, *l*ocal clustering can also suggest possible interactions or roles of proteins with unknown functions. *J0544* is a protein of unknown function -- based the documentation in the MIPS, YPD, and SGD databases (Ball *et al.*, 2000; Hodges, 1999; Mewes *et al.*, 2000). Analysis of the mRNA expression of this ORF with our algorithm showed that it has a time-delayed profile relationship with four ORFs associated with the mitochondria - *ATP11*, *MRPL17*, *MRPL19* and YDR116C. They are all time-delayed at approximately the same time shift compared to *J0544*. The expression profile relationships between *J0544* and these genes are shown in Figure 4F. Atp11p has been found in mitochondria, and is an F1-ATP synthase assembly protein (Ackerman & Tzagoloff, 1990). Mrpl17p and Mrpl19p are mitochondrial ribosomal proteins of the large ribosomal subunit (Kitakawa *et al.*, 1997). YDR116C has similarity to prokaryotic ribosomal protein L1 and is a probable component of mitochondrial ribosomes, as its mRNA abundance in DNA microarray analysis shows the same change patterns to a variety of drug treatments and mutations as do many mitochondrial proteins (Hughes *et al.*, 2000b). The profile relationship between *J0544* and these four mitochondrial ORFs suggests that *J0544* may be involved in mitochondrial processes, perhaps as an activator or some other type of component.

Additional Relationships

Our procedure can obviously uncover many more relationships than we have space to discuss in detail here. Additional time delayed and inverted relationships, with discussion of relevant documentation, for the cell-cycle dataset can be obtained from our web site.

Global Characterization of Local Clustering in Relation to Function Assignment

Early work has surveyed the ability of expression data to predict function or interaction (Altman & Raychaudhuri, 2001; Brown *et al.*, 2000; Gerstein & Jansen, 2000; Niehrs & Pollet, 1999); similar expression profiles may indicate similar cellular roles or physical interactions. In particular, it is quite plausible that tightly interacting proteins should have correlated patterns of gene expression. However, it is obviously the case (and demonstrated above) that genes with quite different (i.e. inverted or time-delayed) expression profiles may interact or have related cellular roles. It is interesting to evaluate how many additional new, functionally relevant relationships can be detected by local clustering compared to traditional, global clustering. Above, we have looked at specific examples identified by our method that were inverted or time-delayed, but it is also important to look at the percentage of newly detected relationships on a global level.

Odds Ratio Formulism for Assessing Interactions

For concreteness, we begin by focusing on protein-protein interactions and then generalize this to protein functions. We calculated the likelihood of finding a known interaction for a given score. This probability is normalized with the expected probability of finding a relationship for any score. This normalized probability is the odds ratio, which is defined as $R = P_o / P_e$, where P_o is the observed probability and P_e is the expected probability. The observed probability is the chance of finding two genes having a known interaction (i.e. from an interaction dataset such as the global yeast two-hybrid (Uetz, 2000)) for a given score, i.e. $P_o = P(i \mid S)$, where *i* means interaction and *S* is the match score. The expected probability is the probability of finding two genes having an interaction for any possible pairs in the genome, i.e. $P_e = \frac{N_{known}}{N_{whole}}$, where N_{known} is the number of detected interaction pairs by current techniques and N_{whole} is the number of all possible interactions.

A hypothetical example will illustrate the logic behind the odds-ratio calculation. Imagine an experiment where 2000 known interactions were detected among 6000 yeast genes. Then there are theoretically ~18 million ((6000^2-6000)/2), possible interactions among these 6000 genes. Therefore, the expected probability of finding an interaction if one randomly selects pairs from the 6000 genes is about 10^{-4} (=2000/18,000,000). To check whether this is related to expression profile relationships, we calculated the probability for the gene pairs with different expression profile match scores. Suppose 1000 gene pairs have a match score of 15, and 10 of these were found to have known interactions.

Therefore, the probability of finding an interaction with match score 15 is 10/1000=0.01, which corresponds to an odds ratio R 100 times higher ($0.01/10^{-4}$) than expected purely by chance. If the odds ratio is equal to 1, then the probability of finding an interaction is just as expected. By calculating this odds ratio, one can easily check if the probability of finding a biological relationship and the match score are correlated. The advantage of using the odds ratio instead of the probability itself is that when more interactions are found in future interaction studies, this quantity will remain stable while the probability will change significantly.

Likelihood of Local Clustering Finding Pairs with Known Protein-Protein Interactions

Figure 5A shows the likelihood that two genes interact genetically or physically for a given match score. The interaction data is based on the union of the yeast two-hybrid data (Ito, 2000; Uetz, 2000) and genetic and physical interaction data from MIPS (Mewes *et al.*, 2000), a similar combination to that used in other computational studies of protein-protein interactions (Park *et al.*, 2001). One can observe that in the high match score region (score > 14), the overall likelihood of having interactions for two genes is much higher than expected because their odds ratios are much larger than 1. For instance, gene pairs with a match score of 16 are found to interact with each other more than 20 times more often than random expectation. On the other hand, in the low match score region (score < 8), the likelihood of finding interactions is either close to or lower than expected according to their odds ratios. The likelihood of finding an interaction increases with the expression profiles' match score.

Likelihood of Local Clustering Finding Pairs with the Same Cellular Role

The odds ratio formalism can easily be generalized to assess whether genes clustered together by expression have a similar function. Here we just calculate probabilities that two genes are known to have the same cellular role, based on a functional classification, rather than a known interaction. For a functional classification we use the second highest level of MIPS functional catalogue (Mewes *et al.*, 2000). For example, “amino-acid metabolism” is the second highest MIPS functional catalogue, while “metabolism” is the highest MIPS functional catalogue.

Figure 5B shows the likelihood that two genes have the same function for a given match score of their expression profiles. Very similar observations can be made to those above concerning interactions. That is, the higher matched scores are definitely enriched in pairs of genes that have the same cellular role.

Composition of Different Relationships

In the high match score region, there are some time-delayed and inverted relationships detected besides the simultaneous relationships. The majority of gene pairs with known interactions or the same cellular role have very similar simultaneously correlated expression profiles. However, one can still see that some new gene pairs are detected to

have a functional similarity or interaction according to their high match score, even though their actual expression profiles are very different from each other (either inverted or time-delayed). Even though the composition of time-delayed or inverted relationships is smaller than that from simultaneous relationships, we believe that each additional relationship is important in thoroughly understanding biological systems.

Summary and Discussion

Microarray technology presents a big challenge of how to analyze the resulting data sets. In order to detect correlations other than simultaneous ones, we developed an alternative similarity measure distinct from the traditional correlation coefficient. Our approach, which we call local clustering, can be used to identify new relationships between genes that have time-delayed or inverted expression profiles, as well as to detect conventional simultaneous profile relationships. We related these new relationships between genes to their similarity in function or their having known protein-protein interactions, finding that genes strongly related by local clustering were considerably more likely than random expectation to have a known interaction or a similar cellular role.

On a reasonable level, one would not expect all relationships in gene expression data to be simple correlations, so there is an obvious justification for many of the new relationships turned up by our procedure. While some of time-delayed and inverted relationships found by our method have published biological relationships, local clustering was able to identify many additional pairs of genes whose functions and relationships need to be further explored. We described a number of examples in detail and provide others on our website.

In addition, in a quantitative comparison of the correlation coefficient to our method, it is clear that different network configurations result. For the gene pairs with the highest match score based on our algorithm, the percentage of time-delay and inverted relationships are low because most gene pairs with the same function also have very similar simultaneous correlated expression profiles. However, we believe that the new relationships are important for the understanding of a whole biological system.

Possible Extensions to Algorithm

In analogy with local sequence alignment (Smith & Waterman, 1981), we could easily extend our local clustering method to handle "gaps" in the aligned expression profiles. These would be useful if time points are not uniformly sampled, as often happens in the long time series such as during the development of Drosophila or other organisms (White *et al.*, 1999). The inclusion of gaps into the alignment effectively adds some pseudo-time points to the real expression profile, making the time points uniformly sampled.

As for score schema, similarity functions other than direct multiplication could be defined; these might include $S(x_i, y_j) = (x_i \bullet y_j)^2$ or rank correlation coefficient. These might be a useful way to handle particularly noisy expression data.

Finally, the similarity of two expression profiles could be measured in the frequency space. In other words, we would compare the spectra of the expression profiles generated by Fourier transformation. We actually did implement this extension and present some limited results on our website. However, we found that for the cell-cycle dataset spectral comparisons did not reveal as many new but well documented relationships as local clustering -- i.e. the odds ratio plots as in Fig. 5 showed fewer known relationships at high match scores. Hence, we decided not to emphasize them here. However, the spectral methods may have suffered comparatively from the relatively few time points in the cell-cycle dataset (which gives rise to poor Fourier transformations) and may be more successful on longer time series that will be available in the future.

Limitations and Future Directions

Local clustering can most usefully be applied to time series. It may not apply under other conditions, especially for the detection of time-delay relationships that would only be relative in a time-dependent array study. It would be better to use normal clustering methods for non-time series data -- e.g. for the yeast knockout study (Hughes *et al.*, 2000b).

In addition, while the analysis of the highly scored pairs found by local clustering can shed light on novel biological relationships, it is limited by the quality of the information available on protein function and protein-protein interactions. There are many ambiguities in the current functional classifications (Gerstein, 2000; Riley, 1998) and there is a problem with false positives in many of the protein-protein interaction studies, particularly the two-hybrid (Ito, 2000; Uetz, 2000). Thus, the novel relationships we uncovered should be viewed as potential hypotheses until they are validated by appropriate biological experiments. In order to more accurately predict gene interactions and relationships, it is important to combine the clustering results with other experimental information. As a future direction, this type of hybrid computational and experimental analysis will allow the investigation of gene networks or regulatory pathways.

Acknowledgements

The authors are grateful to Dr. Nicholas Luscombe, Dov Greenbaum for comments on the manuscript and useful discussion. We also thank the Keck Foundation for support.

References

Ackerman, S. H. & Tzagoloff, A. (1990). Identification of two nuclear genes (ATP11, ATP12) required for assembly of the yeast F1-ATPase. *Proc Natl Acad Sci U S A* **87**, 4986-4990.

Altman, R. B. & Raychaudhuri, S. (2001). Whole-genome expression analysis: challenges beyond clustering. *Curr. Opin. Struct. Biol.* **11**, 340-347.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.

Ball, C. A., Dolinski, K., Dwight, S. S., Harris, M. A., Issel-Tarver, L., Kasarskis, A., Scafe, C. R., Sherlock, G., Binkley, G., Jin, H., Kaloper, M., Orr, S. D., Schroeder, M., Weng, S., Zhu, Y., Botstein, D. & Cherry, J. M. (2000). Intergrating functional genomic information into the saccharomyces genome database. *Nucleic Acids Res.* **28**, 77-80.

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. & Sonnhammer, E. L. (2000). The Pfam Protein Families Database. *Nucleic Acids Res.* **28**, 263-266.

Boles, E. & Hollenberg, C. P. (1997). The molecular genetics of hexose transport in yeasts. *FEMS Microbiol Rev.* **21**, 85-111.

Brandriss, M. C. (1983). Proline utilization in Saccharomyces cerevisiae: analysis of the cloned PUT2 gene. *Mol. Cell Biol.* **3**, 1846-1856.

Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., Jr. & Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A* **97**(1), 262-7.

Bussemaker, H. J., Li, H. & Siggia, E. D. (2001). Regulatory element detection using correlation with expression. *Nat. Genet.* **27**, 167-171.

Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J. & Davis, R. W. (1998). A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle. *Molecular Cell* **2**, 65-73.

D'haeseleer, P., Wen, X., Fuhrman, S. & Somogyi, R. (1997). Mining the gene expression matrix: inferring gene relationships from large scale gene expression data. In *Information processing in cells and tissues* (Holcombe, M., Paton,R., ed.), pp. 203-212. Plenum.

Drawid, A., Jansen, R. & Gerstein, M. (2000). Genome-wide analysis relating expression level with protein subcellular localization. *Trends Genet.* **16**, 426-430.

Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863-14868.

Ermolaeva, O., Rastogi, M., Pruitt, K. D., Schuler, G. D., Bittner, M. L., Chen, Y., Simon, R., Meltzer, P., Trent, J. M. & Boguski, M. S. (1998). Data management and analysis for gene expression arrays. *Nat. Genet.* **20**, 19-23.

Friedman, N., Linial, M., Nachman, I. & Pe'er, D. (2000). *Proceeding of the 4th Annual Conference on Research in Computational Molecular Biology, Tokyo, Japan*.

Gaasterland, T. & Bekiranov, S. (2000). Making the most of microarray data. *Nat. Genet.* **24**, 204-206.

Gerstein, M. (2000). Integrative database analysis in structural genomics. *Nat. Struct. Biol.* **7 Suppl.**, 960-963.

Gerstein, M. & Jansen, R. (2000). The current excitement in bioinformatics analysis of whole-genome expression data: how does it relate to protein structure and function? *Curr. Opin. Struct. Biol.* **10**, 574-584.

Hanekamp, T. & Thorsness, P. E. (1999). YNT20, a bypass suppressor of yme1 yme2, encodes a putative 3'-5' exonuclease located in mitochondria of Saccharomyces cerevisiae. *Curr. Genet.* **34**, 438-448.

Hegde, P., Qi, R., Abernathy, K., Gay, C., Dharap, S., Gaspard, R., Hughes, J. E., Snesrud, E., Lee, N. & Quackenbush, J. (2000). A concise guide to cDNA microarray analysis. *Biotechniques* **29**, 548-550.

Heyer, L. J., Kruglyak, S. & Yooseph, S. (1999). Exploring expression data: identification and analysis of coexpressed genes. *Genome Res* **9**(11), 1106-15.

Hodges, P. E., et al. (1999). The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucleic Acids Res.* **27**, 69-73.

Hughes, J. D., Estep, P. W., Tavazoie, S. & Church, G. M. (2000a). Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J. Mol. Biol.* **296**, 1205-1214.

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

Ito, T., et al. (2000). Toward a protein-protein intearaction map for the budding yeast: A comprehenisve system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl. Acad. Sci. USA* **97**, 1143-1147.

Jansen, R. & Gerstein, M. (2000). Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins. *Nucleic Acids Res.* **28**, 1481-1488.

Kim, S., Dougherty, E. R., Bittner, M. L., Chen, Y., Sivakumar, K., Meltzer, P. & Trent, J. M. (2000). General nonlinear framework for the analysis of gene interaction via multivariate expression arrays. *J. Biomed. Opt.* **5**, 411-424.

Kitakawa, M., Graack, H. R., Grohmann, L., Goldschmidt-Reisin, S., Herfurth, E., Wittmann-Liebold, B., Nishimura, T. & Isono, K. (1997). Identification and characterization of the genes for mitochondrial ribosomal proteins of Saccharomyces cerevisiae. *Euro. J. Biochem.* **245**, 449-456.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M. & Horton, H., et al. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **14**, 1675-1680.

Lundgren, D. W. & Ogur, M. (1973). Inhibition of yeast 1 -pyrroline-5-carboxylate dehydrogenase by common amino acids and the regulation of proline catabolism. *Biochim. Biophys Acta* **297**, 246-257.

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

Melcher, K. & Entian, K. D. (1992). Genetic analysis of serine biosynthesis and glucose repression in yeast. *Curr. Genet.* **21**, 295-300.

Mewes, H. W., Heumann, K., Kaps, A., Mayer, K., Pfeiffer, F., Stocker, S. & Frishman, D. (2000). MIPS: a database for protein sequences and complete genomes. *Nucleic Acids Res.* **27**, 44-48.

Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443-453.

Niehrs, C. & Pollet, N. (1999). Synexpression groups in eukaryotes. *Nature* **402**(6761), 483-7.

Park, J., Lappe, M. & Teichmann, S. A. (2001). Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *J. Mol. Biol.* **307**, 929-938.

Riley, M. (1998). Systems for categorizing functions of gene products. *Curr. Opin. Struct. Biol.* **8**, 388-392.

Schaerer-Brodbeck, C. & Reizman, H. (2000a). Functional interactions between the p35 Subunit of the Arp 2/3 Complex and Calmodulin in Yeast. *Molecular Biology of the Cell* **11**, 1113-1127.

Schaerer-Brodbeck, C. & Reizman, H. (2000b). *Saccharomyces cerevisiae* Arc35p works through two genetically separable calmodulin functions to regulate the actin and tubulin cytoskeletons. *J. Cell Sci.* **113**, 521-532.

Shalon, D., Smith, S. J. & Brown, P. O. (1996). A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res* **6**, 639-645.

Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195-197.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S. & Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* **96**, 2907-2912.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. (1999). Systematic determination of genetic network architecture. *Nat. Genet.* **22**, 281-285.

Toronen, P., Kolehmainen, M., Wong, G. & Castren, E. (1999). Analysis of gene expression data using self-organizing maps. *FEBS Lett.* **451**, 142-146.

Uetz, P., et al. (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623-627.

Wen, X., Fuhrman, S., Michaels, G. S., Carr, D. B., Smith, S., Barker, J. L. & Somogyi, R. (1998). Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci. USA* **95**, 334-339.

White, K. P., Rifkin, S. A., Hurban, P. & Hogness, D. S. (1999). Microarray analysis of Drosophila development during metamorphosis. *Science* **286**, 2179-2184.

Yona, G., Linial, N. & Linial, M. (2000). Protomap: automatic classification of protein sequences and hierarchy of protein families. *Nucleic Acids Res.* **28**, 49-55.

Figure captions:

Figure 1: Two examples showing the time delayed (A) and inverted (B) relationships in the expression profiles. Note there are only 8 time points for each profile, while in the real data there are 17 time points. Also, the expression ratio is not normalized, whereas in the real data each profile is normalized so that the averaged expression ratio is 0 and the standard deviation is 1. The thick segments of the expression profiles are the matched part. (C) The corresponding matrix $E_{i,j}$ for the expression profiles shown in (A). The corresponding matrix of $D_{i,j}$ is not shown because in this case the match score (the maximal score) is from $E_{i,j}$, not from $D_{i,j}$. The numbers outside the border of the matrix are the expression ratio shown in (A). The dark gray cell contains the match score for these two expression profiles, and the light gray cells indicate the path of the optimal alignment between the expression profiles. The path starts from the match score and ends at first encountered 0. (D) The corresponding matrix $D_{i,j}$ for the expression profiles shown in (B). The matrix of $E_{i,j}$ is not shown because the match score is not from this matrix in this case.

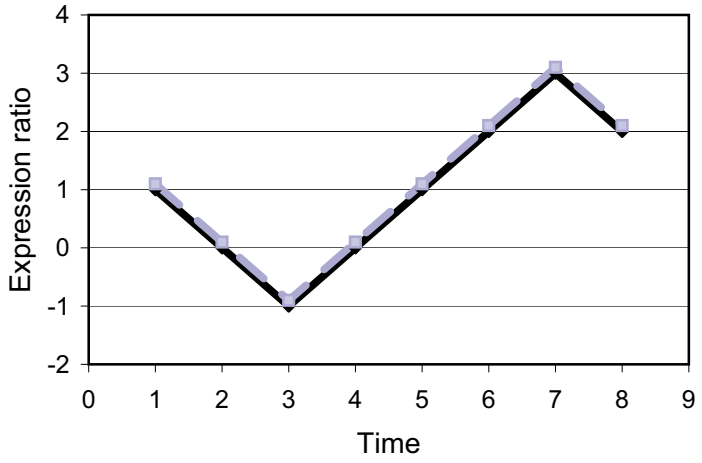
Figure 2: Network view of relationships defined by the algorithm. This figure was prepared using a software program based on the graph-drawing library "AGD" (http://www.mpi-sb.mpg.de/AGD). (A) A global view of the network formed by relationships detected by the algorithm. The threshold used for this network is a match score of 16 (P-value of 10^{-6}). (B) A close-up view of the rectangle outlined in Fig 2A. A solid line signifies a simultaneous profile relationship, an arrow denotes a time delay in the relationship with the arrow pointing to the delayed gene, and a dashed line denotes an inverted profile relationship.

Figure 3: Quantitative comparisons between networks generated by local clustering algorithm and the traditional correlation coefficient. (A) Graph of the average connection per node as a function of the number of nodes in the network (B) Graph of the number of clusters as a function of the size of the network. The black and red dots highlight the thresholds used for different sizes of network.

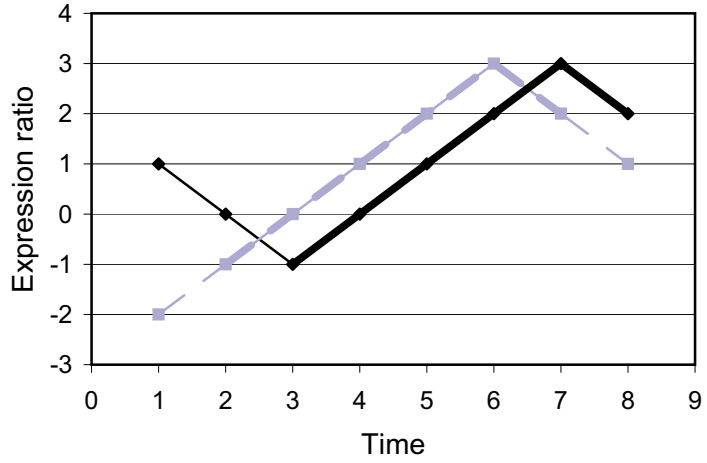
Figure 4: Examples of different profile relationships found by the algorithm. (A) Simultaneous expression profile relationship of *RPS11A* and *RPS11B*. (B) Simultaneous expression profile relationship of *HXT6* and *HXT7*. (C) Inverted expression profile relationship of *YME1* and *YNT20*. (D) Inverted gene expression profile relationship of *PUT2* and *SER3*. (E) Time-delayed profile relationship between *ARC35* and *ARP3*. The arrow indicates the time shift between two profiles. (F) Time-delayed relationship between *J0544* and *ATP11*, *MRPL17*, *MRPL19* and YDR116C. The arrow indicates the time shift between two profiles.

Figure 5: Conditional probability of having the same function or interaction between two genes. (A) Graph of the conditional probability that two genes interact genetically or physically for a given match score of their expression profiles. The inverted relationships and the inverted time-delayed relationships are pooled into "inverted" in conditional probability analysis. For the very low match score, the profile relationship, such as simultaneous, time-delay or inverse, is not relevant, since their profiles are so disparate that their grouping into a particular type of relationship does not have significance. These are marked as “unmatched”. (B) Graph of the conditional probability that two genes have the same function for a given match score.

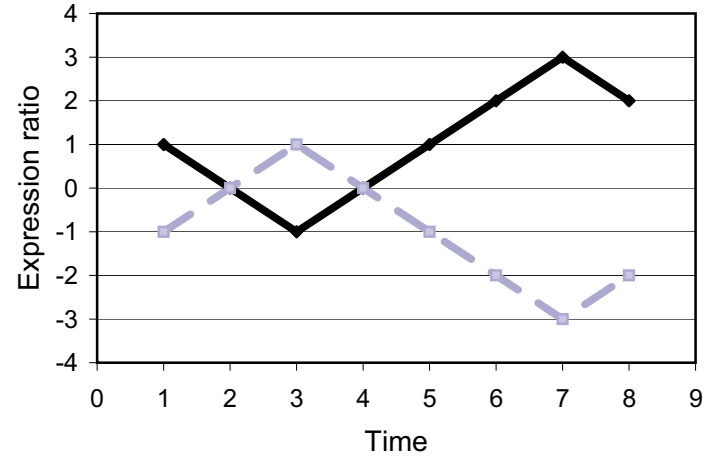
A



B



C



D

		1	0	-1	0	1	2	3	2
	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	1	2	3	2
0	0	0	1	0	0	0	1	2	3
-1	0	0	0	2	0	0	0	0	0
0	0	0	0	0	2	0	0	0	0
1	0	1	0	0	0	3	2	3	2
2	0	2	1	0	0	2	7	8	7
3	0	3	2	0	0	3	8	16	14
2	0	2	3	0	0	2	7	14	20

$$E_{i,j} = \max(E_{i-1,j-1} + x_i \bullet y_j, 0)$$

E

		-2	-1	0	1	2	3	2	1
	0	0	0	0	0	0	0	0	0
1	0	0	0	0	1	2	3	2	1
0	0	0	0	0	0	1	2	3	2
-1	0	2	1	0	0	0	0	0	2
0	0	0	2	1	0	0	0	0	0
1	0	0	0	2	2	2	3	2	1
2	0	0	0	0	4	6	8	7	4
3	0	0	0	0	3	10	15	14	10
2	0	0	0	0	2	7	16	19	16

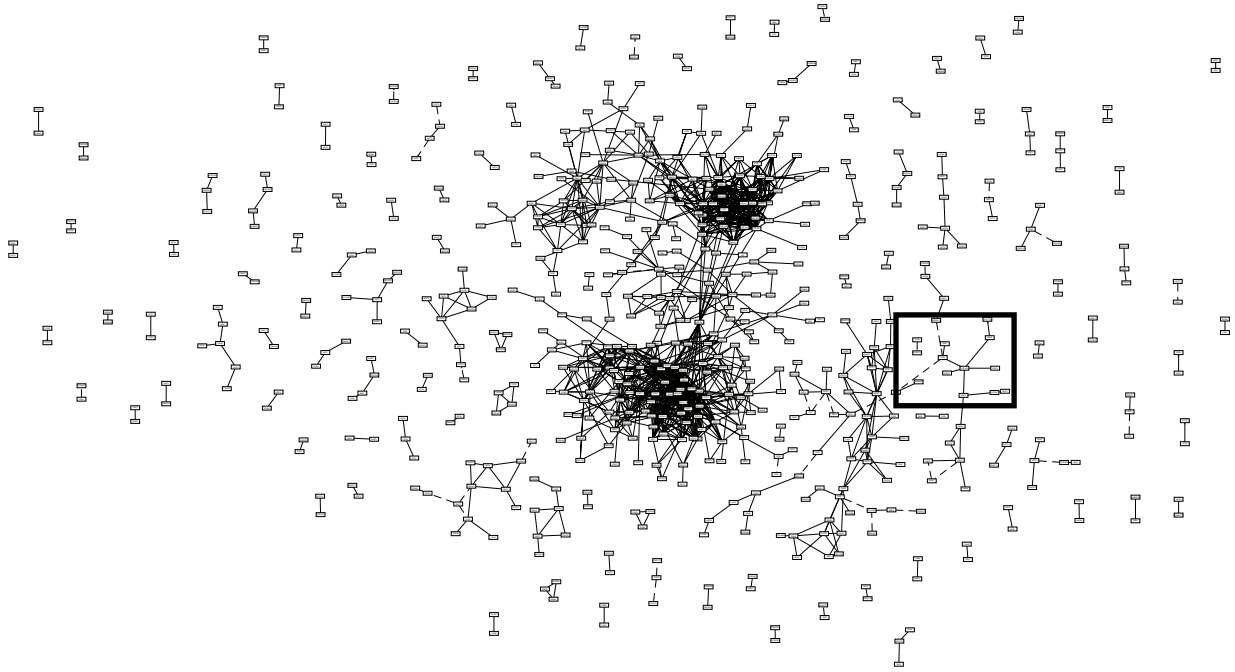
$$E_{i,j} = \max(E_{i-1,j-1} + x_i \bullet y_j, 0)$$

F

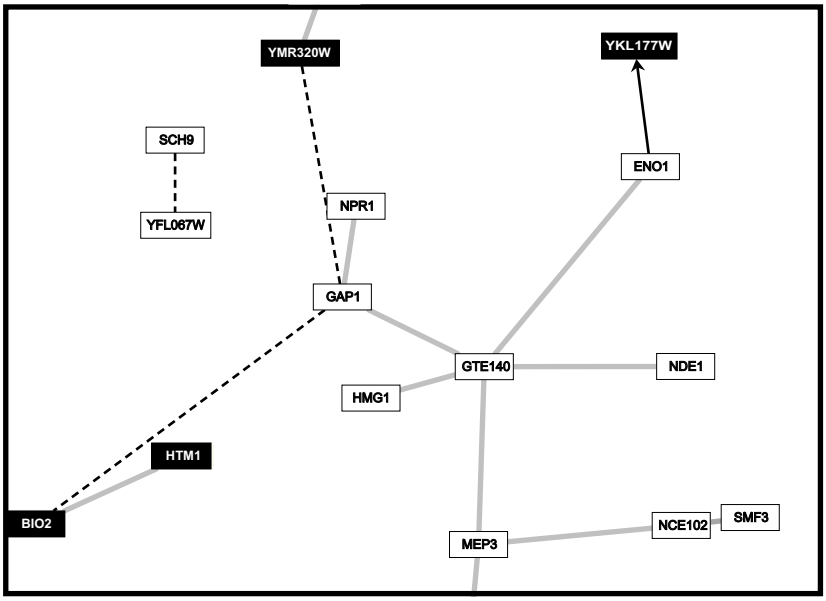
		-1	0	1	0	-1	-2	-3	-2
	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	1	2	3	2
0	0	0	1	0	0	0	1	2	3
-1	0	0	0	2	0	0	0	0	2
0	0	0	0	0	2	0	0	0	0
1	0	1	0	0	0	3	2	3	2
2	0	2	1	0	0	2	7	8	7
3	0	3	2	0	0	3	8	16	14
2	0	2	3	0	0	2	7	14	20

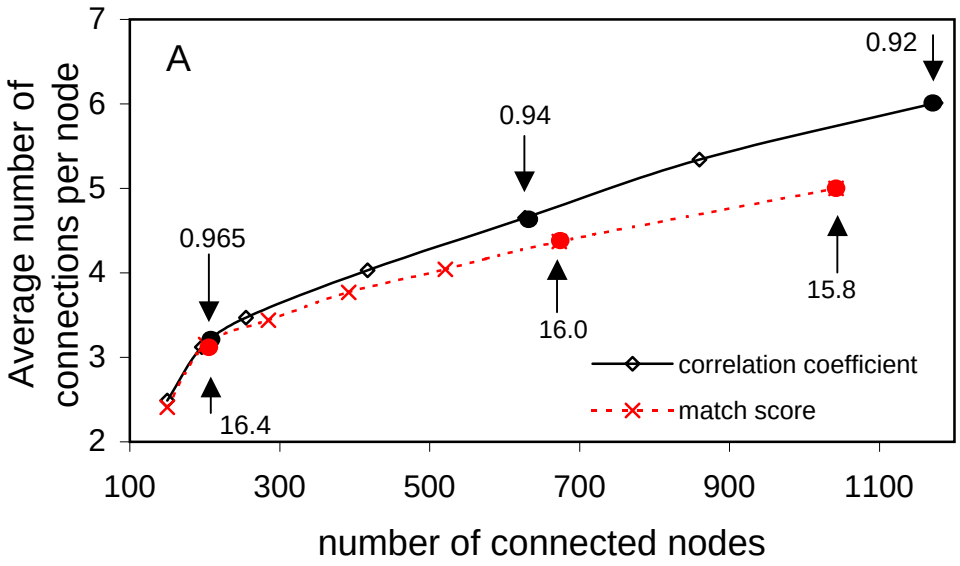
$$D_{i,j} = \max(D_{i-1,j-1} - x_i \bullet y_j, 0)$$

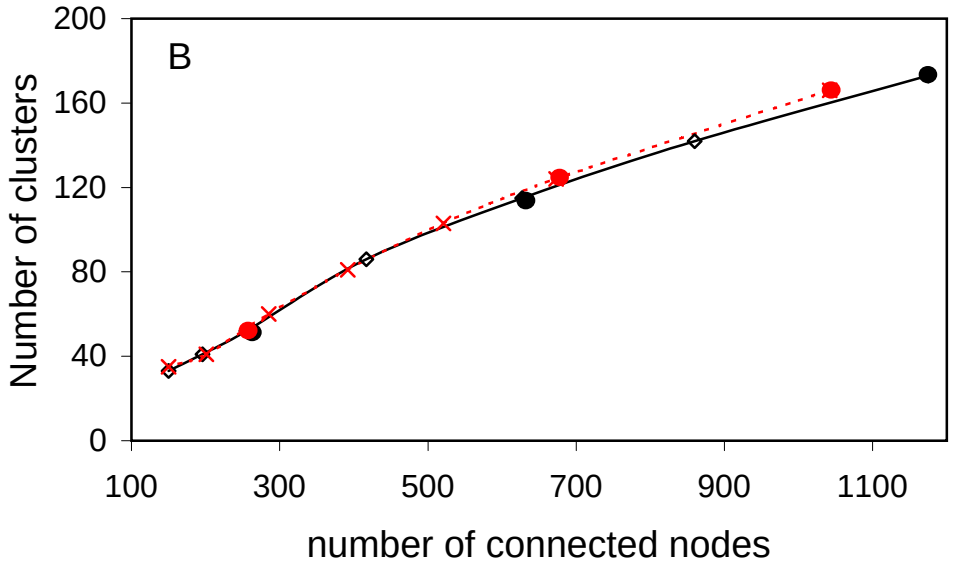
A

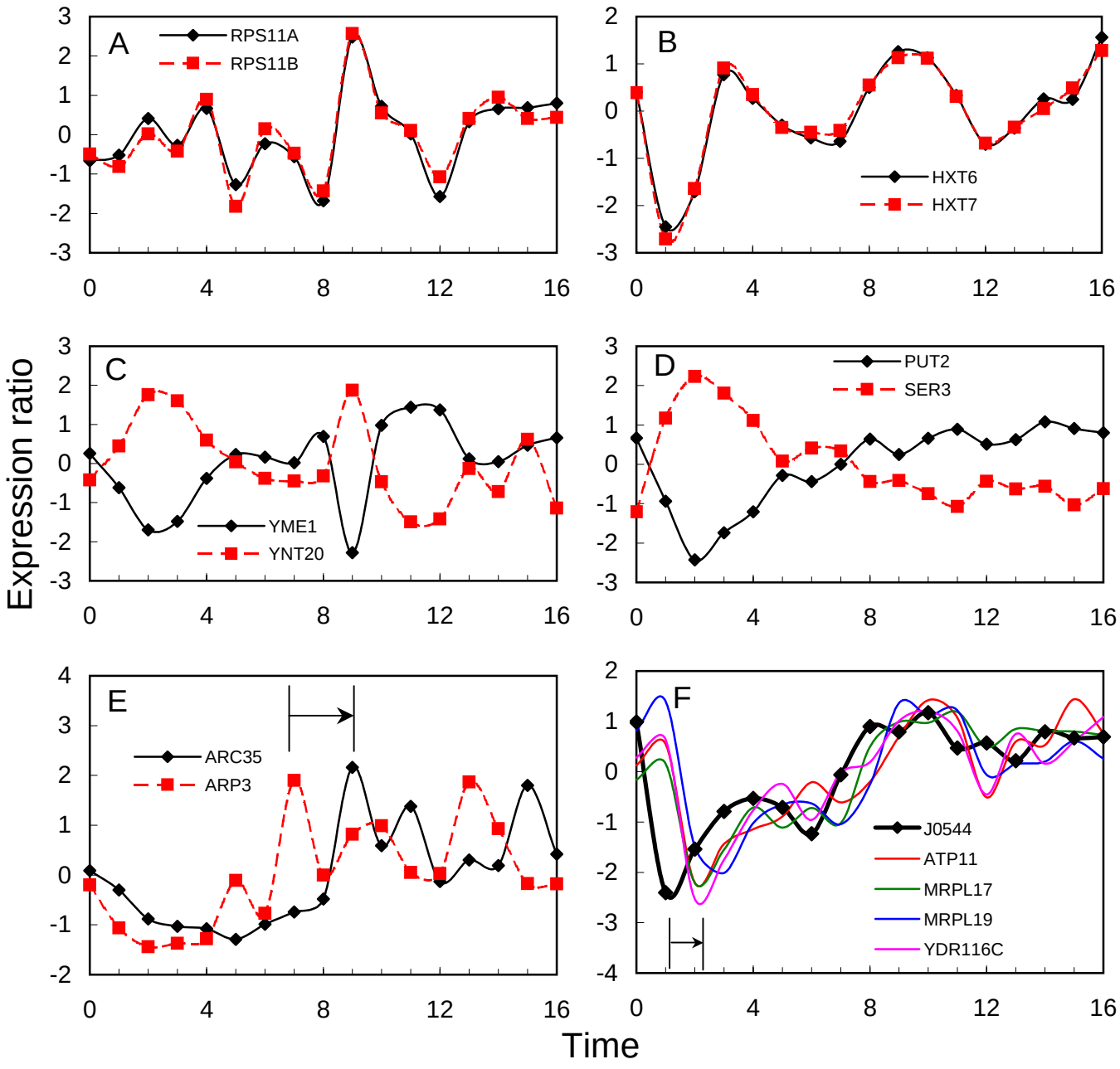


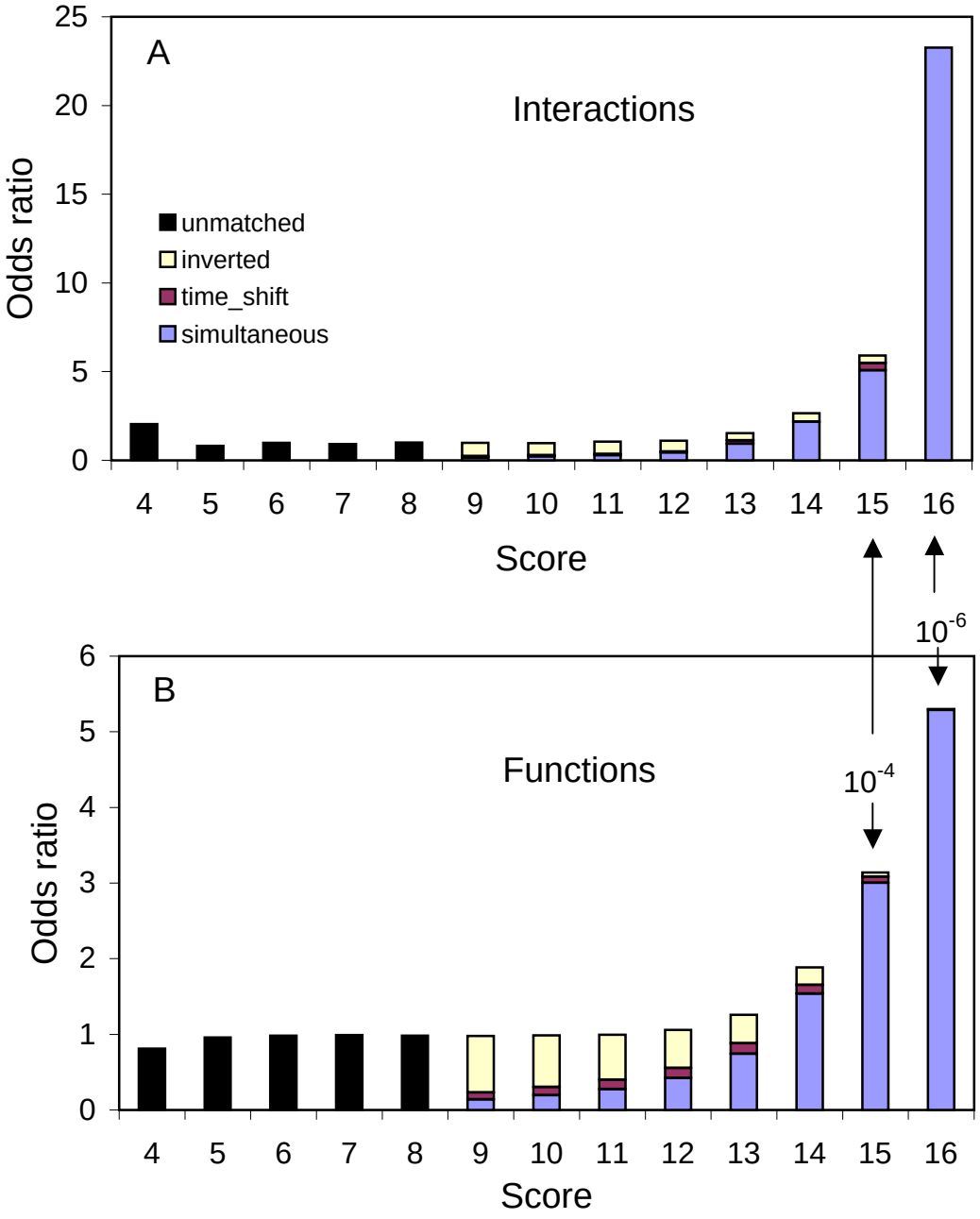
B


YMR320W
YKL177W
SCH9
ENO1
NPR1
YFL067W
GAP1
GTE140
NDE1
HMG1
HTM1
BIO2
NCE102
SMF3
MEP3


A
Average number of connections per node
number of connected nodes
0.92
0.94
0.965
15.8
16.0
16.4
correlation coefficient
match score
2
3
4
5
6
7
100
300
500
700
900
1100


B
Number of clusters
number of connected nodes
0
40
80
120
160
200
100
300
500
700
900
1100


A
RPS11A
RPS11B
B
HXT6
HXT7
C
YME1
YNT20
D
PUT2
SER3
E
ARC35
ARP3
F
J0544
ATP11
MRPL17
MRPL19
YDR116C
Expression ratio
Time


A
Interactions
Odds ratio
25
20
15
10
5
0
unmatched
inverted
time_shift
simultaneous
4 5 6 7 8 9 10 11 12 13 14 15 16
Score
B
Functions
Odds ratio
6
5
4
3
2
1
0
10^{-4}
10^{-6}
4 5 6 7 8 9 10 11 12 13 14 15 16
Score