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, Haiyuan Yu, and Mark Gerstein*

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

Running title: Local Clustering of Gene Expression Profiles

*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 such 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 analogous to the way local sequence alignment (the Smith-Waterman algorithm) is derived from global alignment (Needleman-Wunsch). An integral part of our method is the use of random score distributions to assess the statistical significance of each cluster. 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 (as determined from the MIPS scheme) or their having known protein-protein interactions (as determined from the large-scale two-hybrid experiment); we found 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. We also found new relationships involving uncharacterized yeast genes and were able to suggest 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. We have also looked at other, less extensive datasets than the yeast cell-cycle and found further interesting relationships. Our clustering program and a detailed website of clustering results is available at bioinfo.mbb.yale.edu/expression/cluster or genecensus.org/expression/cluster .

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

Introduction

The massive datasets generated by microarray experiments present a challenge to those interested in studying the regulatory relationship between genes [1-5]. 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 [6-8].

There are further applications for expression clustering, especially in combination with other information about genes such as their subcellular localizations, metabolic functions, and intermolecular interactions [9-13]. In particular, microarray technology allows for studying the entire genome, while other types of gene annotation (e.g. biochemical functions) are often available only for a fraction of the genes. Therefore, researchers have attempted to predict protein function and interaction by expression clustering. This is based on 'guilt by association' [14], the premise that proteins with similar expression profiles (i.e., synexpression relationship) have similar functions [15-18].

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 sequences (e.g. BLAST [19]) and then grouping them into homologous families (e.g. Pfam [20] or Protomap [21]).

The most common algorithms for grouping genes with related profiles are hierarchical clustering [17,22], self-organizing maps [23,24], and K-means clustering [25]. 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 [26].

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, Spearman rank correlation, jackknife correlation coefficient, and Euclidean distance [22,27,28].

A 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, their 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 ignoring inhibition. Here, we propose a new algorithm; it is based on the dynamical programming method for local sequence alignment [29] and hence we call it local clustering. Its development from traditional gene expression clustering method [17] is strongly suggested by the way local sequence alignment [29] followed on the original global approach [30].

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 [16]. 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 (i.e. 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. These relationships have not been previously analyzed. However, they can be detected by the traditional correlation coefficient, if one looks at the correlation coefficients near –1.

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.[31] and a less extensive worm development dataset [32]. 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.

To make this comparison clear, throughout the paper we will refer to the results from our method as derived from "local clustering" and contrast these with results from "traditional, global clustering". The later approach, which is, for instance, used in Eisen et al. [17] and Tamayo et al. [23], is based on computing a distance matrix only from simultaneous correlations between expression profiles (i.e. the traditional correlation coefficient).

Algorithms and Datasets

Local Alignment between Pairs of Expression Profiles

We use a degenerate dynamical programming algorithm to find time-shifted and inverted correlations between expression profiles. 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 $M(x_i,y_i) = x_i y_j$. For simplification, it will be referred as $M_{i,j}$ for comparison of any two genes.

Then, two sum matrices **E** and **D** are calculated as $E_{i,j} = \max(E_{i-1,j-1}+M_{i,j}, 0)$ and $D_{i,j} = \max(D_{i-1,\ j-1}-M_{i,j}, 0)$. 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 $M_{i,j}$ in this segment. This can be accomplished by standard dynamic programming as in local sequence alignment [29] and results in an alignment of *l* aligned time points, where $l \leq n$.

Finally, an overall maximal value *S* is found by comparing the maximums for matrices **E** and **D**. This is the match score *S* for the two expression profiles. If the maximum is off diagonal in its corresponding matrix, the two expression profiles have a time-shifted relationship. This involves an alignment over a smaller number of time points *l* than the

total number *n*. A maximal value from matrix **D** 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," "inverted," or "inverted time-delayed". 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”.

Figure 1E is the corresponding matrix **E** for the expression profiles shown in Fig. 1B. The matrix **D** 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 S=19, is highlighted in the black 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. Figure 1F is the corresponding matrix **D** for the profiles shown in Fig. 1C. The match score is S=20; and because the maximum value is from matrix **D** rather than **E** (not shown), these expression profiles are correlated in an inverted fashion.

Cell Cycle Dataset and Generation of Similarity Matrix

We extensively tested our algorithm on the yeast whole genome oligonucleotide expression array data generated by Cho et al [31], which included over 6,000 ORFs and 17 time points. 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 [33] 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 [1], 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

If we divide the maximal match score by the number of time points (S/n), the resulting ratios are comparable with traditional correlation coefficients. This is strictly true for a global alignment resulting from a full-length simultaneous or inverted relationship. It is only approximately true, however, for local alignments, since these extend over a smaller number of matched positions l than n. This suggests that we could alternatively normalize the match by dividing by the total number aligned positions (S/l). Doing so will tend to

emphasize scores of the local time-shifted relationships in contrast to the global simultaneous relationships. Because of this normalization ambiguity we decide to simply report the unnormalized match score S and the number of aligned and total time points (l and n, where n is always 17 from the cell-cycle data). Then for further clarification of the significance of each match, we thought it better to calculate proper P-values from the distribution of scores (as is conventionally done in sequence and structural alignment [34-38]).

In order to estimate a P-value for a given match score, a set of random expression profiles was generated by shuffling the normalized expression levels at different time points (e.g. interchanging the expression level at time points 3 and 7, x_3 and x_7). The resulting profiles still satisfied our earlier normalization conditions with an average ratio of zero and a standard deviation of one. Using the local alignment procedure, we calculated optimal match scores S for each random expression profiles pair and then tabulated their distribution. This distribution is meant to approximate that of true negatives; through integration, we could calculate a conventional P-value, $P(s>S)$. This is defined as the probability of obtaining a match score s larger than S from the random profiles. The smaller the P-value is, the more significant the match score. Since we did not explicitly take into account length dependence, our P-value statistics are quite conservative, tending to de-emphasize local alignments in favor of global ones.

The distributions of random match scores in comparison the actual observed ones $P(S)$ for the cell-cycle are shown in Fig. 2A, and the relationship between the match score and P-value is shown in Fig. 2B.

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. 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.

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 and is available from our website, http://bioinfo.mbb.yale.edu/expression/cluster or http://genecensus.org/expression/cluster.

Overall Network Topology

To provide a global view of the relationships detected by local clustering, we show in Figure 3A the network resulting from clustering the yeast cell-cycle data. In the diagram, the threshold used to define connected genes is a match score of 16, which corresponds to a P-value of 10^{-6} and correlation coefficient (S/n) of 0.94. The network consists of 673

nodes (genes) and several large clusters. Dynamic navigation of the network can be obtained from our website. Figure 3B is a close up view of part of a large cluster in the rectangle outlined in Figure 3A. 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, new relationships are found. For instance, additional nodes such as YMR320m and YKL177W are joined to a large central cluster, making it even larger than if it were formed from simultaneous correlations alone. On the other hand, our method also generates 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 one based on the traditional correlation coefficient, it is useful to compute some quantities. We calculated the average number of connections per node C (the average number of genes related to any particular gene). It is obvious that this quantity depends on the size of the network size N (number of nodes in the network), which in turn is controlled by the P-value threshold used to define the correlation. The top panel of Figure 4 shows how C varies as a function of N (and P-value cutoff), for networks generated both by local clustering and the traditional correlation coefficient. In both networks, the average number of connections per node C increases with network size N and has approximately the same value, for small networks (N<200). This suggests that the highest ranked correlations detected by two algorithms are the same. However, for large networks, the average connections per node C diverges, which suggests that the configurations of these two networks are topologically different. Overall, nodes have fewer connections in the local-clustering network. One way of understanding this difference is through plotting the number of clusters versus network size N, as shown in the bottom panel of Figure 4. For a given network size, there are slightly more clusters in the local-clustering network than the global-clustering one.

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, showing 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 [17]. Figures 5A and B show two examples. The expression profiles of *RPS11A* and *RPS11B* are shown in Figure 5A. Both

of the genes code for the ribosomal protein S11 and are 100 percent identical in sequence [39]. *RPS11A* is located on yeast chromosome IV, and *RPS11B* is located on yeast chromosome II. Figure 5B 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 [40].

Inverted Relationships

Well-documented relationship: Figure 5C shows the profiles of *YME1* and *YNT20*, which display an inverted relationship. Yme1p (yeast mitochondrial escape) is a metal and ATP dependent protease. It is 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 [41]. *YNT20* has been identified as a bypass suppressor of Yme1p; it 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 [41]. 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 5D 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 [42]. Ser3p is a 3-phosphoglyerate dehydrogenase that is involved in the synthesis of serine from glycolytic intermediates [43]. It has already been found that Put2p could be inhibited by serine (and other amino acids) [44]. 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 5E. Both these genes are part of the Arp2/3 complex in yeast and are thus clearly related. This complex, which comprises a total of 6 proteins, is involved in endocytosis and actin cytoskeleton organization [45]. The expression profiles of *ARC35* and *ARP3* show a time-delayed relationship, with the expression of *ARC35* being one time point (20 minutes) delayed compared to *ARP3*. This fits in well with Arc35p being required late in G1 for the cytoskeleton-organization functionality [46].

New, Suggested Relationship: In addition to shedding light on known interactions, local clustering can also suggest possible interactions or roles of proteins with unknown functions. *J0544* is a yeast protein of unknown function -- based the documentation in the MIPS, YPD, and SGD databases [39,47,48]. 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 by approximately the same phase as compared to *J0544*. The expression profile relationships between *J0544* and these genes are shown in Figure 5F. Atp11p has been found in mitochondria, and is an F1-ATP synthase assembly protein [49]. Mrpl17p and Mrpl19p are mitochondrial ribosomal proteins of the large ribosomal subunit [50]. 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 [51]. 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 publications, for the cell-cycle dataset can be obtained from our web site.

Overall Relationship of Local Clustering to Protein Function

Early work has surveyed the ability of expression data to predict functions, interaction, or localization [6,10,12-14,16,18]; 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 uncovered by local clustering as 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.

General Formalism

In general terms, we want to assess here whether there is a "global" relationship between expression profiles and a known biological association (e.g. similar functions or protein-protein interactions). A simple quantitative way to address this issue is to look at the conditional probability $P(k|S)$, the probability of that a pair of genes has a known biological association (k) given their expression profile match score (S). As diagrammed in figure 6A, $P(k|S)$ corresponds to the population density of known biological associations in all pairs with match score S. However, the number of known biological associations varies considerably depending on what type of associations one is focusing on. For example, there are relative few associations based on the two-hybrid data and other physical and genetic interactions [52,53] but many based on MIPS the function classes (5385 vs. 826,000). Therefore, it is useful to normalize $P(k|S)$ so it more generally

comparable between different types of associations. We normalize $P(k|S)$ by calculating the odds ratio

$$R = P(k|S)/P(k) \quad (1)$$

P(k) is the chance of having the known interaction, regardless of match score. It is essentially the number of known interactions divided by the number of all possible pairwise interactions, ~18 million in yeast. As shown in figure 6A, the odds ratio *R* is essentially the ratio of population density of biological association between the subgroup (with a given *S*) and whole genome (for any *S*).

To better understand the meaning of the odds ratio, we can rewrite it applying Bayes' rule: $R = P(k|S)/P(k) = P(S|k)/P(S)$. We can see that the right-hand side of the equation represents the distribution of match scores of the pairs with known biological interactions divided by the distribution of match scores of all possible pairs of genes in this genome (i.e. essentially the distribution in Fig. 2).

Likelihood of Local Clustering Finding Known Protein-Protein Interactions

Now we apply our formalism above explicitly to protein-protein interactions. Figure 6B shows the odd ratio 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 [52,53] and genetic and physical interaction data from MIPS [39], a similar combination to that used in other computational studies of protein-protein interactions [54]. There are 5385 total interactions in this dataset. One can observe that in the high match score region ($S > 14$, P-value better than 3.8e-4), 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 about 20 times more often than random expectation. On the other hand, in the low match score region ($S < 8$), 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 monotonically with the expression-profile match score.

One advantage of the odds-ratio normalization is that it is not that sensitive to the number of associations currently known, a fact particularly important for the interaction data. Specially, as new known protein-protein interactions are uncovered by various experimental techniques, the probability $P(k|S)$ increases, but so does *P(k)*, keeping R relatively constant.

Likelihood of Local Clustering Finding Proteins with the Same Cellular Role

In figure 6C, we apply the odds-ratio formalism to protein function, i.e. we want to see whether genes clustered together by expression have a similar cellular role. We calculate probabilities that a pair of genes have the same cellular role based on the MIPS functional classification [39]. We use the second level of MIPS; for example, “amino-acid metabolism” is at this level whereas “metabolism” is at highest (most general) MIPS

level. Figure 6C shows the odds ratio for function versus match score. Very similar observations can be made to those above concerning interactions; the higher matched scores are definitely enriched in pairs of genes that have the same cellular role.

Composition of Different Relationships

As shown in Table 1, in the high match score region (P-value better than .01), there are a considerable number of time-delayed and inverted relationships found that would not be detected with global clustering. Even though the raw number 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. Moreover, we would like to emphasize that given our (conservative) statistical scoring scheme, all these new relationships are by definition significant.

Table 1 also shows that many of the significant time-delayed and inverted relationships uncovered by our procedure correspond to known interactions for similar cellular roles. Again, the number is obviously less than that for simultaneously clustering but one still uncovers many new statistically significant relationships.

Extension to Other Datasets Beyond the Yeast Cell-Cycle

Currently there are not that many long time course microarray experiments available in the public databases for analysis (see our website for the list of the available microarray time courses). The yeast cell cycle is by far the best of existing sets for local clustering, with the largest number of timepoints (16+), high-quality data (including Affymetrix), and multiple experimental repetitions. There are no other experiments with more than half this many timepoints; the next best set contains less than 7 points. Moreover, the time intervals in many of the other datasets are not uniform, which is not suitable for the current method without further extensions (see below).

However, it is anticipated that in the near future there will be a large number of long time courses available and being able to successfully deal with this type of data will be very important for expression analysis. This is especially true for development of multi-cellular organisms such as the worm and fly [55], and soon a fly developmental time course with more than 70 time points should be available (K White, personal communication).

For the present, to get some sense for how local clustering handles deal with a different data set we applied it in a preliminary fashion to a short time course from another organism: a seven-point *C. elegans* developmental time course [32]. Overall, we found about 12,885 significant inverted relationships and 677 shifted ones (with a P-value better than .001), corresponding to 0.5% and 0.03% of all the identified significant relationships, respectively. The corresponding numbers for the yeast cell cycle are ~72,000 inverted relationships and ~36,000 shifted ones, corresponding 32% and 16% of the identified relationships. While we found many significant non-simultaneous relationships for the worm, it seems we found proportionately fewer of them in this

organism than for yeast. This perhaps reflects the smaller size of the time course, which necessarily will give rise to fewer potential shifted relationships.

We also found that several of the time-shifted and inverted relationships represented documented or plausible biological associations. These tend to involve a transcription activator or repressor and their regulated genes. The results are available on our website, in terms of specific relationships and detailed network navigation.

Summary and Discussion

Microarray technology presents a new type of data for bioinformaticians to analyze, and given its large and growing scale, such analysis will clearly be centrally important in the near future. In order to detect relationships 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. It improves upon "traditional" gene-expression clustering in an analogous fashion to how for protein sequences local alignment [29] is derived from global alignment [30]. We related our newly found gene relationships to their similarity in function or known protein-protein interactions; we find 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 are justified by published biological experiments, local clustering was also 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 an overall comparison of the global clustering 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-delayed 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 [29], 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 [55]. 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 $M_{i,j} = (x_i y_j)^2$ or rank correlation coefficient, both of which 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 implemented this extension and present some 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. 6 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 meaningful 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 [51].

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 [56,57] and there is a problem with false positives in many of the protein-protein interaction studies, particularly the two-hybrid [52,53]. 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 may allow the investigation of gene networks or regulatory pathways.

Acknowledgements

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

References

1.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.

2.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.

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

4.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.

5.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.

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

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

8.Zhu, G., Spellman, P. T., Volpe, T., Brown, P. O., Botstein, D., Davis, T. N. & Futcher, B. (2000) Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth *Nature*, **406,** 90-94.

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

10.Drawid, A. & Gerstein, M. (2000) A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *J. Mol. Biol.*, **301,** 1059-1075.

11.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.

12.Jansen, R., Greenbaum, D., Qian, J. & Gerstein, M. Relating Whole-Genome Expression Data with Protein-Protein Interactions *Genome Res.*

13.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,** 262-7.

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

15.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.

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

17.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.

18.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.

19.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.

20.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.

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

22.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.

23.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.
24.Toronen, P., Kolehmainen, M., Wong, G. & Castren, E. (1999) Analysis of gene expression data using self-organizing maps. *FEBS Lett.*, **451,** 142-146.
25.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.
26.Friedman, N., Linial, M., Nachman, I. & Pe'er, D. (2000), Proceeding of the 4th Annual Conference on Research in Computational Molecular Biology. Universal Academy Press, Tokyo, Japan, pp. 127-135.
27.D'haeseleer, P., Wen, X., Fuhrman, S. & Somogyi, R. (1997) In Holcombe, M., Paton,R. (ed.), Information processing in cells and tissues. Plenum, pp. 203-212.
28.Heyer, L. J., Kruglyak, S. & Yooseph, S. (1999) Exploring expression data: identification and analysis of coexpressed genes *Genome Res*, **9,** 1106-15.
29.Smith, T. F. & Waterman, M. S. (1981) Identification of common molecular subsequences *J. Mol. Biol.*, **147,** 195-197.
30.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.
31.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.
32.Hill, A. A., Hunter, C. P., Tsung, B. T., Tucker-Kellogg, G. & Brown, E. L. (2000) Genomic Analysis of Gene Expression in *C. elegans Science*, **290,** 809-812.
33.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.
34.Pearson, W. R. (1998) Empirical statistical estimates for sequence similarity searches *J. Mol. Biol.*, **276,** 71-84.
35.Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs *Nucleic Acids Res.*, **25,** 3389-3402.
36.Gerstein, M. & Levitt, M. (1998) Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins *Protein Sci.*, **7,** 445-456.
37.Levitt, M. & Gerstein, M. (1998) A unified statistical framework for sequence comparison and structure comparison *Proc Natl Acad Sci U S A*, **95,** 5913-5920.
38.Wilson, C. A., Kreychman, J. & Gerstein, M. (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores *J. Mol. Biol.*, **297,** 233-249.
39.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.
40.Boles, E. & Hollenberg, C. P. (1997) The molecular genetics of hexose transport in yeasts *FEMS Microbiol Rev.*, **21,** 85-111.
41.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.
42.Brandriss, M. C. (1983) Proline utilization in Saccharomyces cerevisiae: analysis of the cloned PUT2 gene *Mol. Cell Biol.*, **3,** 1846-1856.
43.Melcher, K. & Entian, K. D. (1992) Genetic analysis of serine biosynthesis and glucose repression in yeast *Curr. Genet.*, **21,** 295-300.
44.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.

45.Schaerer-Brodbeck, C. & Reizman, H. (2000) Functional interactions between the p35 Subunit of the Arp 2/3 Complex and Calmodulin in Yeast *Molecular Biology of the Cell*, **11,** 1113-1127.
46.Schaerer-Brodbeck, C. & Reizman, H. (2000) *Saccharomyces cerevisiae* Arc35p works through two genetically separable calmodulin functions to regulate the actin and tubulin cytoskeletons. *J. Cell Sci.*, **113,** 521-532.
47.Hodges, P. E., McKee, A. H., Davis, B. P., Payne, W. E. & Garrels, J. I. (1999) The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucleic Acids Res.*, **27,** 69-73.
48.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.
49.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.
50.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.
51.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. (2000) Functional discovery via a compendium of expression profiles *Cell*, **102,** 109-126.
52.Uetz, P., et al. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae Nature*, **403,** 623-627.
53.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.
54.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.
55.White, K. P., Rifkin, S. A., Hurban, P. & Hogness, D. S. (1999) Microarray analysis of Drosophila development during metamorphosis *Science*, **286,** 2179-2184.
56.Gerstein, M. (2000) Integrative database analysis in structural genomics *Nat. Struct. Biol.*, **7 Suppl.,** 960-963.
57.Riley, M. (1998) Systems for categorizing functions of gene products *Curr. Opin. Struct. Biol.*, **8,** 388-392.

Figure captions:

Figure 1: Two examples showing simultaneous (A), time-delayed (B), and inverted (C) relationships in the expression profiles. Note there are only 8 time points for each profile, while in the real yeast cell-cycle 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. (D) The corresponding matrix **E** for the expression profile shown in (A). The corresponding matrix **D** is not shown because in this case the match score (the maximal score) is from **E** and not **D**. The numbers outside the border of the matrix are the expression ratio shown in (A). The black cell contains the overall match score *S* 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 the first encountered 0. (E) The corresponding matrix **E** for the expression profile shown in (B). Note the time-shifted relationship and how the length of the overall alignment can be shorter than 8 positions. (F) The corresponding matrix **D** for the expression profiles shown in (C). The matrix **E** is not shown because the best match score is not from this matrix in this case.

Figure 2: Relationship between the match score *S* and P-value. The top panel shows the distribution of match score for the cell-cycle expression dataset and a random dataset. Each random profile also has 17 time points and average 0 and standard deviation 1. The bottom panel shows how the P-value can be calculated by integrating the random distribution.

Figure 3: 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 4: Quantitative comparisons between networks generated by local clustering algorithm and the traditional correlation coefficient. The top panel shows the graph of the average connections per node C as a function of the number of nodes in the network N. The bottom panel shows the graph of the number of clusters as a function of the size of the network N. In both panels the indicated black and red dots highlight the thresholds used for different sizes of network. The numbers in the parentheses are the effective correlation coefficient for the match score.

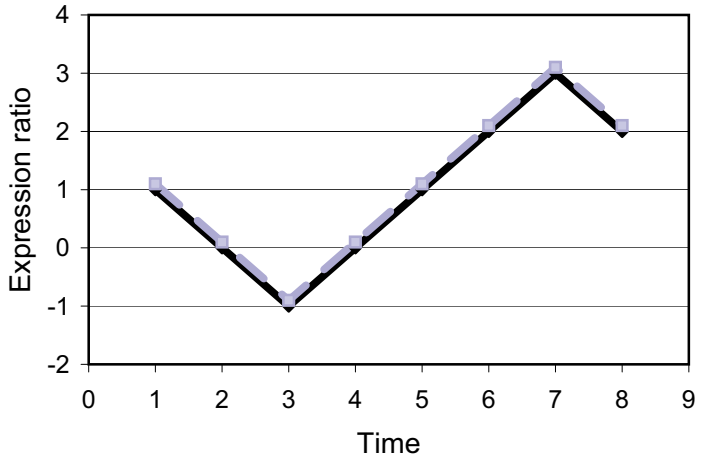
Figure 5: Examples of different profile relationships found by the algorithm. (A) Simultaneous expression profile relationship of *RPS11A* and *RPS11B*. (B) Simultaneous expression profile relationship of *HXT6* and *HXT7*. (C) Inverted expression profile relationship of *YME1* and *YNT20*. (D) Inverted 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 6: Odds ratio of having the same function or interaction between two genes. (A) A hypothetical example illustrating the logic behind the odds-ratio calculation. To check whether a biological interaction is related to expression profile relationships, we calculate the probability for finding the interaction between the gene pairs given a particular expression profile match score, say 16. A dot or a cross indicates the gene pairs, and the crosses indicate pairs with known biological association. The conditional probability $P(k|S)$ for finding an interaction for a given match score is the "density of crosses" in the different subgroup, e.g. the subgroup of match score 16. The odds ratio is the "density of crosses" in different subgroups normalized by the density for whole genome (big outer circle). Imagine an experiment where 2000 known interactions were detected among 6000 yeast genes. 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 16, and 10 of these were found to have known interactions. Therefore, the probability of finding an interaction with match score 15 is 16/1000=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. (B) Graph of the odds ratio 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. (C) Graph of the odds ratio that two genes have the same function for a given match score.

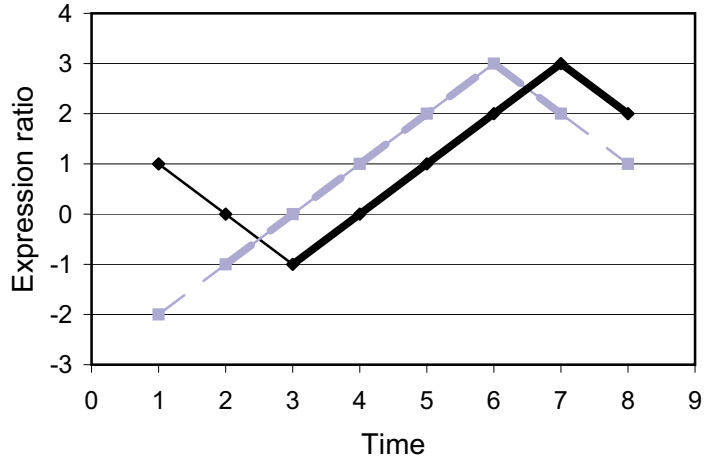
Table Caption

The actual number of new types of relationships found by local clustering (time-shifted and inverted) for a given match score. The table also gives a breakdown into the various types of non-simultaneous relationships by association. Note that the division of non-simultaneous relationships by associations does not sum up to the total number of non-simultaneous relationships since it is possible to have a relationship with both a known function and a known interaction.

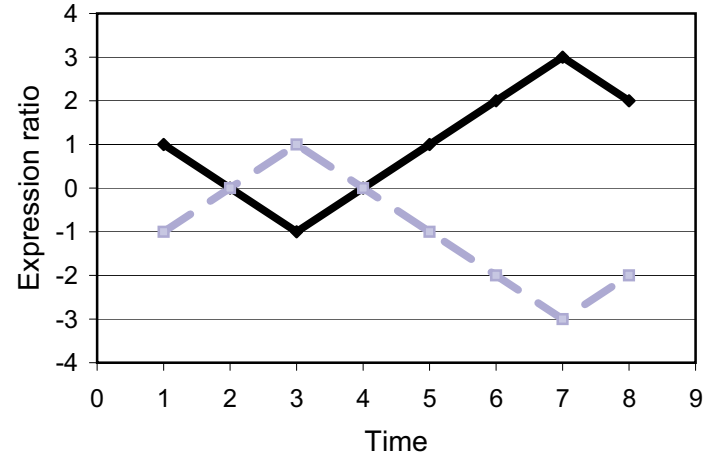
A



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

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

E

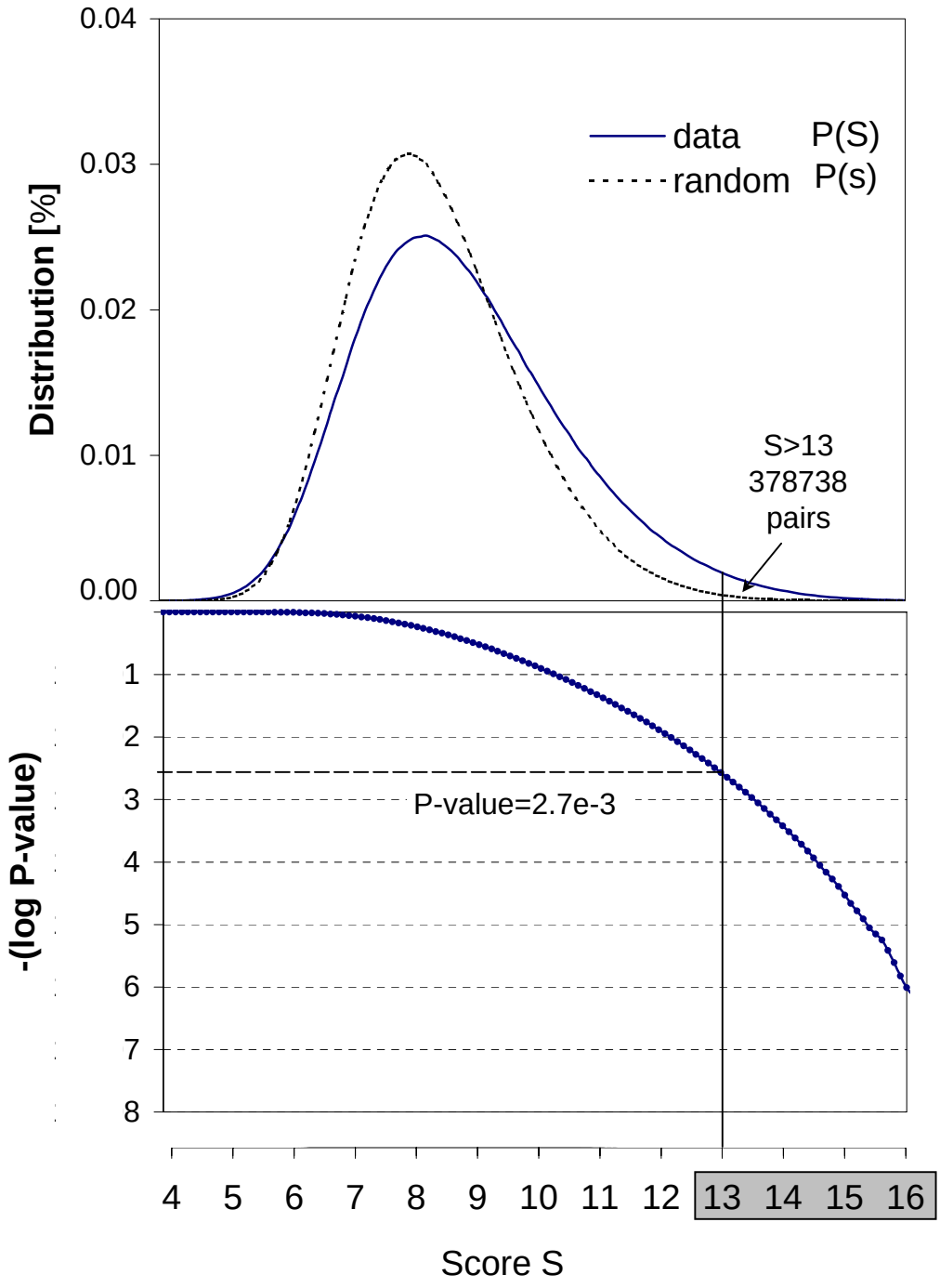
		-2	-1	0	1	2	3	2	1
	0	0	0	0	0	0	0	0	0
1	0	0	0	0	1	2	3	2	1
0	0	0	0	0	0	1	2	3	2
-1	0	2	1	0	0	0	0	0	2
0	0	0	2	1	0	0	0	0	0
1	0	0	0	2	2	2	3	2	1
2	0	0	0	0	4	6	8	7	4
3	0	0	0	0	3	10	15	14	10
2	0	0	0	0	2	7	16	19	16

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

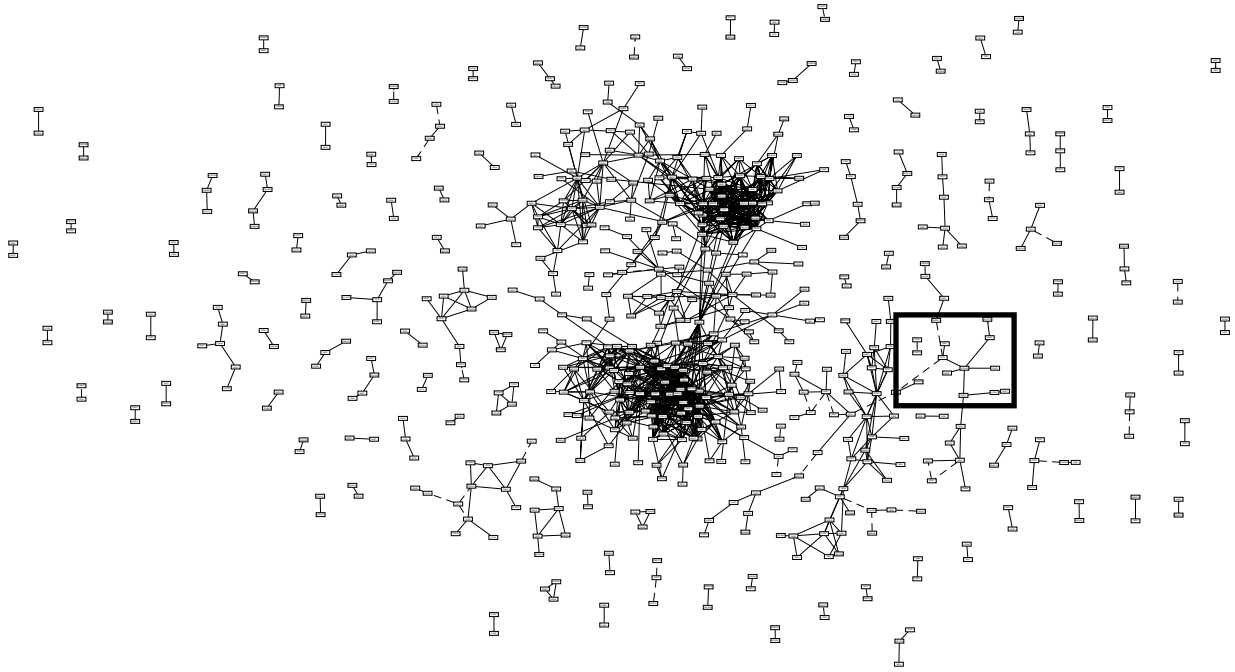
F

		-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

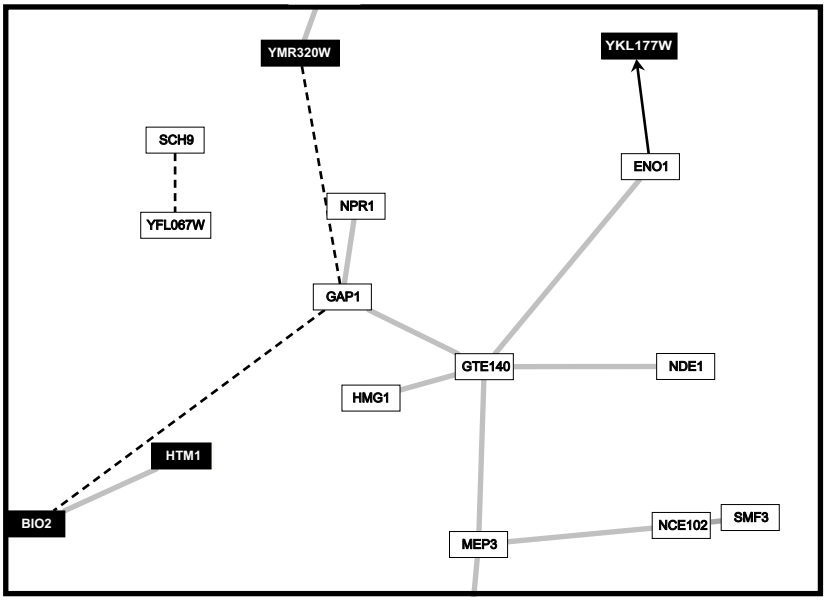
$$D_{i,j} = \max(D_{i-1,j-1} - x_i \;\; y_j, 0)$$

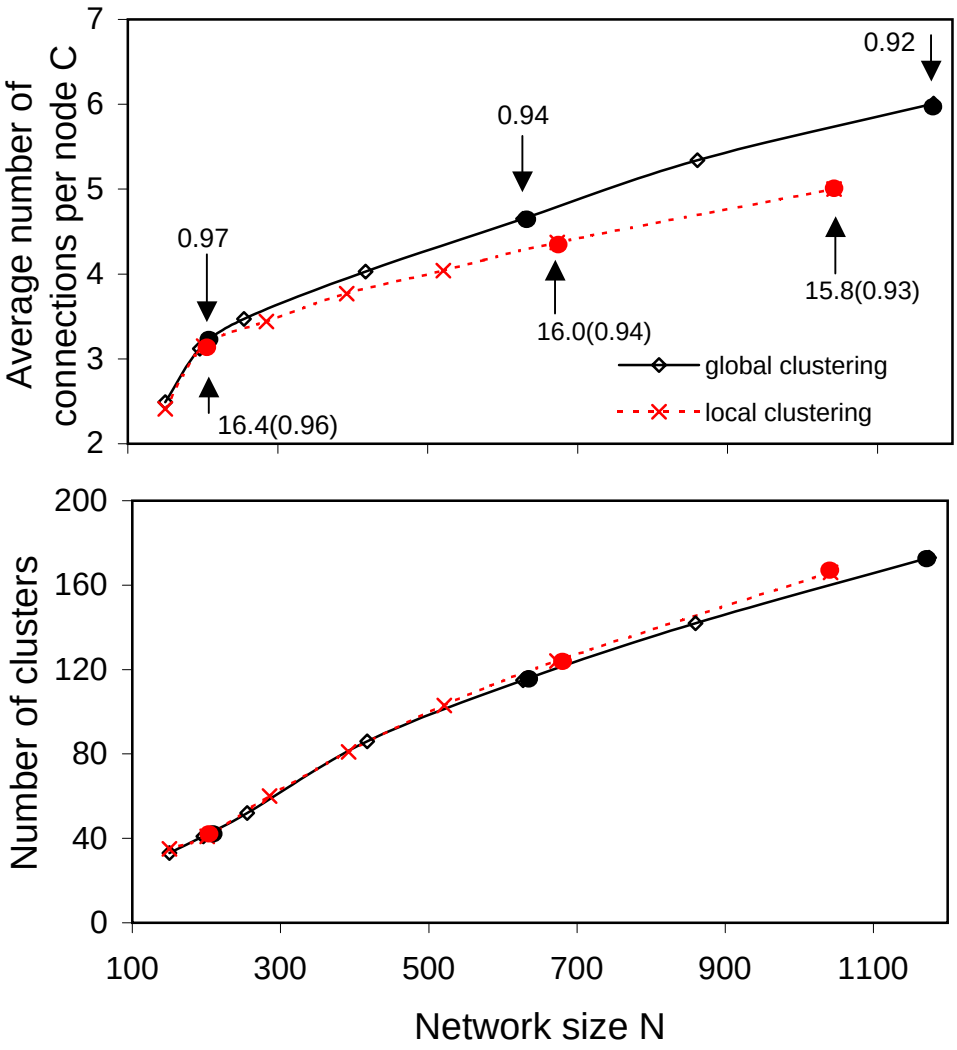

0.04
0.03
0.02
0.01
0.00
Distribution [%]
data P(S)
random P(s)
S>13
378738
pairs
1
2
3
4
5
6
7
8
-(log P-value)
P-value=2.7e-3
4 5 6 7 8 9 10 11 12 13 14 15 16
Score S

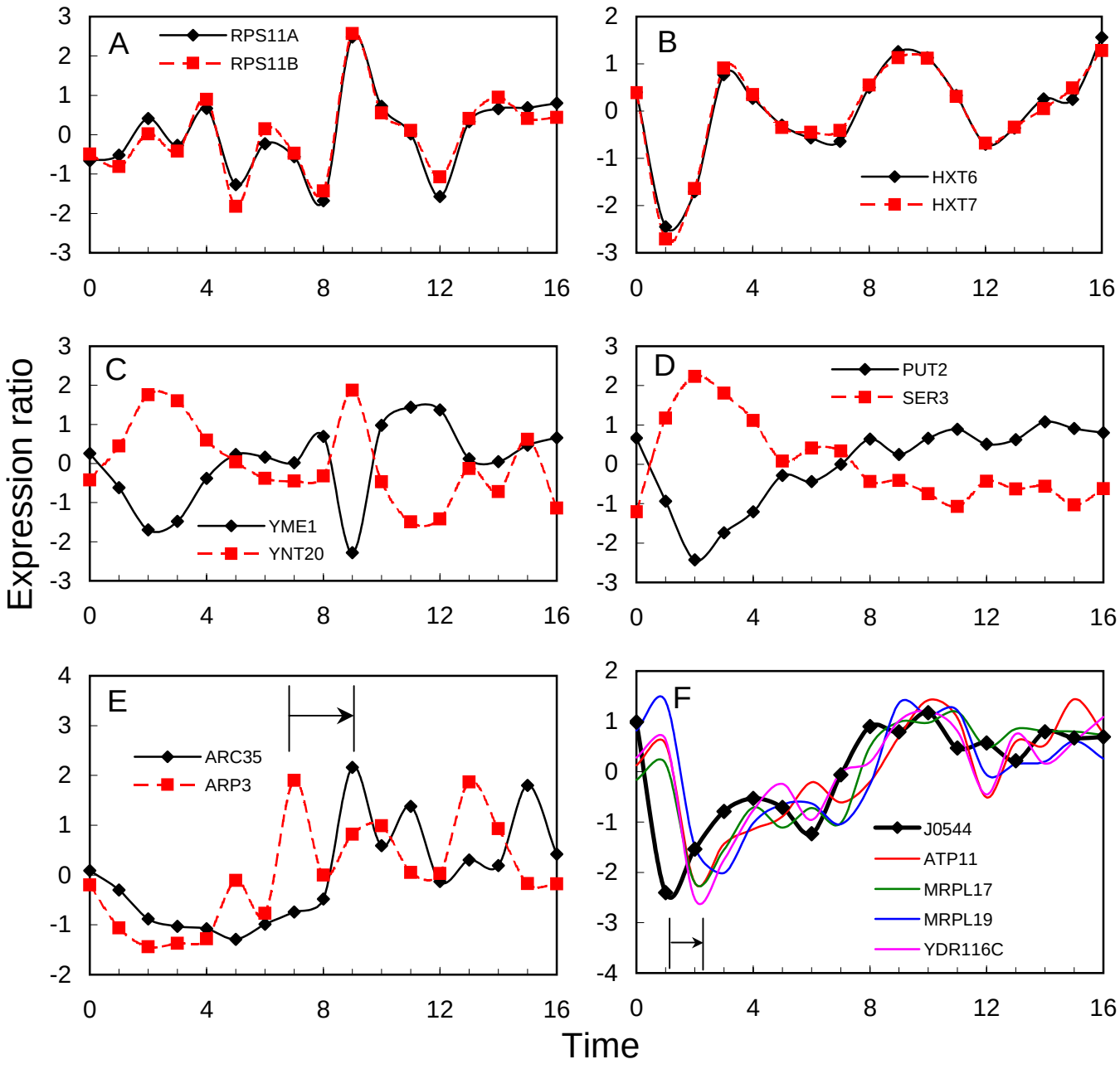
A

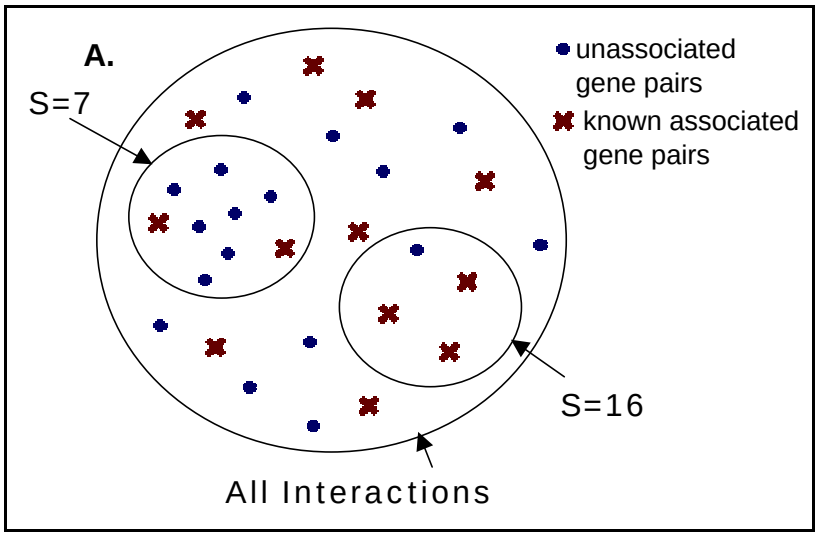


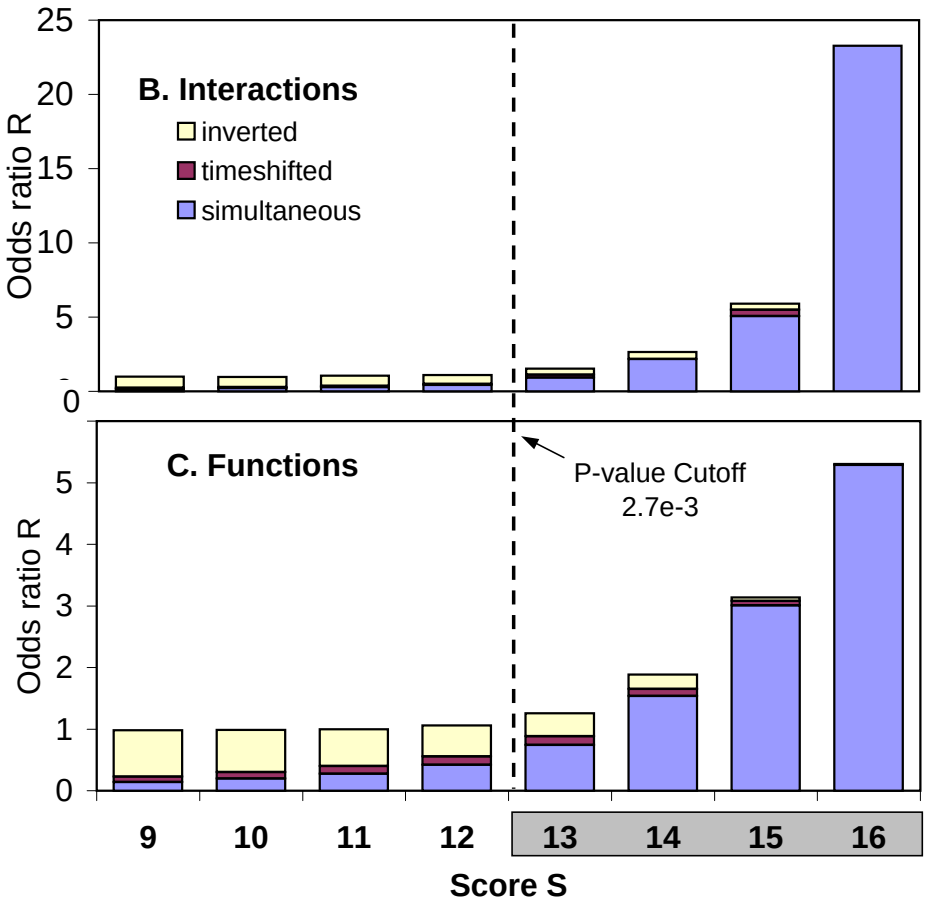
B


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


Average number of connections per node C
7
6
5
4
3
2
0.92
0.94
0.97
15.8(0.93)
16.0(0.94)
16.4(0.96)
global clustering
local clustering
Number of clusters
200
160
120
80
40
0
100
300
500
700
900
1100
Network size N


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.
S=7
unassociated gene pairs
known associated gene pairs
S=16
All Interactions


B. Interactions
inverted
timeshifted
simultaneous
Odds ratio R
25
20
15
10
5
0
C. Functions
P-value Cutoff
2.7e-3
Odds ratio R
5
4
3
2
1
0
9
10
11
12
13
14
15
16
Score S

score of pair relationship			number of simultaneous relationships with score S	number of nonsimultaneous relationships with score S						
				divided by expression relationship				divided by association		
score S	approx. correlation	P-value		total	timeshifted only	inverted only	timeshifted & inverted	same known function	known interaction	new relationships
13	0.76	2.7E-03	81863	250393	92607	71835	85951	23722	12	120695
14	0.82	3.8E-04	37408	74253	24854	27373	22026	5692	13	81111
15	0.88	3.0E-05	11580	12997	3657	6244	3096	626	2	13809
16	0.94	1.0E-06	1406	775	183	476	116	10	1	788
total			132257	338418	121301	105928	111189	30050	28	216403