Bioconductor Project

Bioconductor Project Working Papers

Year 2004 *Paper* 4

A graph theoretic approach to testing associations between disparate sources of functional genomic data

Raji Balasubramanian* Thomas LaFramboise†

Denise Scholtens‡ Robert Gentleman**

*Department of Biostatistics, Harvard School of Public Health, rbalasub@hsph.harvard.edu

†Department of Biostatistics, Harvard School of Public Health, tlaframb@hsph.harvard.edu

‡Department of Biostatistics, Harvard School of Public Health, dscholte@hsph.harvard.edu

**Department of Biostatistics, Harvard School of Public Health, rgentlem@hsph.harvard.edu

A graph theoretic approach to testing associations between disparate sources of functional genomic data

Raji Balasubramanian, Thomas LaFramboise, Denise Scholtens, and Robert Gentleman

Abstract

The last few years have seen the advent of high-throughput technologies to analyze various properties of the transcriptome and proteome of several organisms. The congruency of these different data sources, or lack thereof, can shed light on the mechanisms that govern cellular function. A central challenge for bioinformatics research is to develop a unified framework for combining the multiple sources of functional genomics information and testing associations between them, thus obtaining a robust and integrated view of the underlying biology.

We present a graph theoretic approach to test the significance of the association between multiple disparate sources of functional genomics data by proposing two statistical tests, namely edge permutation and node label permutation tests. We demonstrate the use of the proposed tests by finding significant association between a Gene Ontology-derived "predictome" and data obtained from mRNA expression and phenotypic experiments for Saccharomyces cerevisiae. Moreover, we employ the graph theoretic framework to recast a surprising discrepancy presented in Giaever et al. (2002) between gene expression and knockout phenotype, using expression data from a different set of experiments.

A graph theoretic approach to testing associations between disparate sources of functional genomics data

Raji Balasubramanian*, Thomas LaFramboise*, Denise Scholtens*, Robert Gentleman
Department of Biostatistics, Harvard School of Public Health
655 Huntington Avenue, Boston, MA 02115
*These authors contributed equally to this work

May 21, 2004

Abstract

Motivation: The last few years have seen the advent of high-throughput technologies to analyze various properties of the transcriptome and proteome of several organisms. The congruency of these different data sources, or lack thereof, can shed light on the mechanisms that govern cellular function. A central challenge for bioinformatics research is to develop a unified framework for combining the multiple sources of functional genomics information and testing associations between them, thus obtaining a robust and integrated view of the underlying biology.

Results: We present a graph theoretic approach to test the significance of the association between multiple disparate sources of functional genomics data by proposing two statistical tests, namely edge permutation and node label permutation tests. We demonstrate the use of the proposed tests by finding significant association between a Gene Ontology-derived *predictome* and data obtained from mRNA expression and phenotypic experiments for *Saccharomyces cerevisiae*. Moreover, we employ the graph theoretic framework to recast a surprising discrepancy presented in Giaever *et al.*, (2002) between gene expression (Causton *et al.*, 2001) and knockout phenotype, using expression data from a different set of experiments (Cho *et al.*, 1998).

Availability: An R software package, GraphAT, containing the data and statistical procedures is available from Bioconductor: http://www.bioconductor.org

Contact: tlaframb@hsph.harvard.edu

Introduction

High-throughput technologies are able to generate functional genomics data at an unprecedented rate. For example, microarray technology is used to provide quantitative information on expression levels for thousands of genes across time (Cho *et al.*, 1998) or multiple experimental conditions (Causton *et al.*, 2001). Other experiments focus on phenotypic outcomes, for example the work by Giaever *et al.*, (2002), in which a systematically constructed collection of gene-deletion mutants is analyzed with respect to the growth rate (fitness) of each knockout strain under varying growth conditions. Results from these types of gene expression and phenotypic outcome experiments are often clustered using standard algorithms such as k-means or represented in more general expression networks (Butte *et al.*, 2000) to yield connections between genes that have similar mRNA transcript or phenotypic profiles across various experimental conditions or time points. Other functional genomics experiments target protein-protein interactions, for example yeast two-hybrid (Y2H) experiments aimed at detecting binary physical interactions ((Uetz *et al.*, 2000); (Ito *et al.*, 2001)) and affinity purification-mass spectrometry experiments involving large scale analysis of purified protein complexes ((Gavin *et al.*, 2002); (Ho *et al.*, 2002)). Putative interactions found by such experiments can generate hypotheses regarding proteins that belong to a common protein complex and/or physically interact together in a metabolic pathway.

In addition to data from individual experiments, there exist several sources of meta-data, here defined as information compiled from different sources, including annotations procured by manual review of published research by expert biologists. Examples of meta-data include the Biological Process (BP), Molecular Function (MF) and Cellular Component (CC) networks of the Gene Ontology (GO) database (Ashburner *et al.*, 2000). In these GO networks, genes that share highly specific annotations can be inferred to perform closely related functions in similar metabolic processes (Zhou *et al.*, 2002). Additional sources of meta-data include Munich Information for Protein Sequences (MIPS) (Mewes *et al.*, 2004), *Saccharomyces* Genome Database (SGD) (Chervitz *et al.*, 1999), Yeast Proteome Database (YPD) (Hodges *et al.*, 1999), Biomolecular Interaction Network Database (BIND) (Bader *et al.*, 2001) and Database of Interacting Proteins (DIP) (Xenarios *et al.*, 2000).

Several previous studies have integrated multiple data sets from genomic experiments in *Saccharomyces cerevisiae*. The earliest such studies ((Grigoriev, 2001), (Ge *et al.*, 2001)) report significant association between data obtained from mRNA expression experiments and high-throughput experiments measuring protein-protein interactions. Similar results involving protein-protein interaction data and transcriptome data were also reported in other studies ((Kemmeren *et al.*, 2002), (Jansen *et al.*, 2003), (Deane *et al.*, 2002)). A study by Bader *et al.*, (2001) presents an analysis of overlap and biases comparing the four major sources of high-throughput data for protein-protein interactions (Gavin *et al.*, 2002; Ho *et al.*, 2002; Uetz *et al.*, 2000; Ito *et al.*, 2001) to meta-data on protein-protein interactions (i.e. *gold standard* interactions) compiled in the MIPS and YPD databases.

In response to the multiplicity of functional genomics data sources and previous interest in studying their overlap and integration, the goals of the present paper are: (a) To present a unified graph theoretic approach for testing the association between multiple sources of functional genomics data by proposing two tests, namely the *edge permutation* and *node label permutation* tests ; (b) To illustrate the proposed methods by integrating data from gene expression and phenotypic fitness experiments with the GO-derived data. We note that the goal of the analyses presented in this paper is not to present a comprehensive interaction network for yeast but rather to illustrate our proposed statistical methods using a few data examples.

Data

In this section, we describe in detail the three primary data sets used for demonstration of our methods, namely (a) transcriptome data from the study by Causton *et al.*, (2001), (b) phe-

notypic fitness data from the study by Giaever *et al.*, (2002), and (c) the GO-derived dataset. These three data sets measure different aspects of gene behavior in *Saccharomyces cerevisiae*. It is of interest to elucidate the complementarity of the different information sources for further insight into their biological relatedness.

The transcriptome data are obtained from experiments by Causton *et al.*, (2001), in which the authors measure the abundance of 6191 mRNA species in *S. cerevisiae* gathered under a variety of experimental conditions. Only the subset of the 3000 most variable (as measured by the ratio of the standard deviation to the mean expression) of these genes was considered. Following a methodology similar to that of Butte *et al.*, (2000), we first computed Spearman correlations of the expression profiles for all pairs of the 3000 genes. All expression values were then permuted independently within each gene 100 times. For each permutation, all pairwise correlations were again computed, resulting in a permutation distribution of over 440 million correlation values. Of these, there were no values below -0.76 and only four above 0.76. We note that Spearman correlation was used both to avoid parametric assumptions, and because many correlations in the original data were clearly outside of the range of values of the permutation distribution. Thus, pairs of genes whose expression profiles had correlations above 0.76 or below -0.76 were considered to be putatively functionally linked in the transcriptome. The 108,352 pairs putatively linked in this way have significantly correlated or anti-correlated transcriptional behavior across the experimental conditions of interest.

The phenotypic fitness data was obtained from a study performed by Giaever *et al.*, (2002). In this study the authors systematically created a unique *S. cerevisiae* gene-deletion mutant for each of nearly 6000 genes. Each strain was then grown in a variety of media similar to those in Causton *et al.*, (2001). A fitness score was computed for each gene-knockout strain in each medium based on the decrease in growth rate as compared to the wild-type strain in the same medium. The authors compare their results with the Causton *et al.*, (2001) data, and surprisingly find little correspondence between genes crucial for growth in a given medium and genes highly expressed in the same medium. We applied to the fitness data the same approach as described above for the transcriptome data set. Again by using the Spearman correlation as a statistic and comparing to the observed pairwise correlations to permutation distribution, we obtained 133,828 significantly correlated or anti-correlated pairs for our phenotype data.

We refer to each dataset derived from the BP, MF and CC ontologies of the GO database as a *predictome*, the nomenclature following a previous study of a predictome of putative functional links for pairs of genes (Mellor *et al.*, 2002). See Ashburner *et al.*, (2000) for details regarding the GO database. Previous studies have implemented two broad classes of semantic similarity measures based on *information content* and *distance* respectively, to quantify similarities between pairs of genes. Information content measures are based on the belief that terms that occur less frequently are more informative. Several such semantic measures have been explored in previous studies (Lord *et al.*, 2003). On the other hand, distance based similarity measures are based on the topological dissimilarity between shared terms in a graph, such as GO. An example of this class of measures is that implemented by Jansen *et al.* (2003), where the similarity between pairs of genes is based on the depth of the terms shared in common in the GO graph. We implement a similar measure to derive the predictome datasets, which we then use to predict whether pairs of genes are functionally linked and test whether these functional links are reflected in the transcriptional and/or phenotypic fitness data.

The algorithm for constructing the GO-based predictome is as follows. Suppose there are M yeast genes under consideration. For each pair of genes (G_i, G_j), $i, j = 1, \ldots M$ and $i \neq j$ and for each of the three networks/ontologies, we assign a unique measure of similarity, denoted D_{ij}.

1. Find all the terms to which gene G_i is annotated, and similarly for gene G_j. These are the GO graphs induced by G_i and G_j, respectively.

2. Find the set of terms that both genes G_i and G_j share in common. Denote this set S_{ij}.

3. Define the *depth* of each term in S_{ij} to be the length of the shortest path between the term and the root node of the ontology (here length refers to number of edges transversed).

4. Find the maximum depth of terms in the set S_{ij}. We refer to this value as D_{ij}.

We now choose a threshold C, based on quantiles of D_{ij} values for all pairs of genes i and j, for each ontology. The choice of quantile cutoff depends on the desired number of predicted functional links, with less stringent thresholds yielding a higher number of predicted relationships. From the three thresholds, we derive two groups of predictome data sets, based on integrating the BP and CC dissimilarities and the MF and CC respectively. That is, pairs of genes are predicted to be functionally linked if they share highly specific BP or MF annotations, respectively, and their products are localized in close proximity in the cell, where the level of specificity is determined by the chosen threshold C. The rationale for including the CC ontology in both predictomes is that the opportunity for two gene products to have coordinated activity is severely restricted by their physical proximity in the cell.

Table 1 presents the sizes of different predictome data sets obtained by choosing different values of the threshold C for each of the three GO ontologies. We considered only the yeast genes analyzed in Causton *et al.*, (2001) and Giaever *et al.*, (2002), where genes showing little variability in both mRNA expression and fitness over different conditions are excluded. Genes with unknown BP (MF) or CC functions are excluded in the BP-CC (MF-CC) derived predictome data sets.

In the following section, we discuss testing the association between multiple datasets, such as those described in this section, using a graph theoretic framework. In subsequent sections, we apply these ideas to the transcriptional data, phenotypic fitness data, and GO-based predictomes and discuss the biological interpretability of the results.

Graph theoretic framework and statistical tests

In this section, we demonstrate how to use graphs for statistical inference on the level of association between disparate data sources. Each data source is represented by a single graph. A graph in this context is a set of nodes, each representing a gene or its protein product, along with edges connecting some pairs of these nodes. Each edge represents a functional link, as asserted by or inferred from the data source, between the nodes/genes it connects.

Figure 1 outlines both of our proposed algorithms using a simple example of one cluster graph and one non-cluster graph. By cluster graph, we mean here a graph derived from a clustered data set (for example, from hierarchical or k-means clustering of expression data). We demonstrate how the common features of multiple graphs can be recorded in a separate *intersection graph*. Various features of the intersection graph can be used to measure association between multiple data sets, and we describe how to assess their statistical significance using an *edge permutation* scheme. We also discuss an alternate *node label permutation* test, which may be more biologically valid. Both the edge and node label permutation schemes can be applied to any collection of graphs with a common set of nodes, but one test may be more appropriate than the other depending on the topology of the graphs under consideration. This question warrants further investigation.

First consider a clustered data set with clusters of sizes 1, 2, 3, and 4. In a graph of such clustered data, each pair of nodes (often representing genes) that shares common cluster membership is connected by an edge. The cluster graph is then a disjoint collection of completely connected subgraphs, representing the partition of the data induced by cluster membership. For our simple illustrative example, the cluster graph contains ten nodes and ten edges with the edges organized as in observed graph A in Figure 1.

For the non-cluster graph, we retain the same set of nodes, but now two nodes are connected by an edge based on some other criterion. For our illustrative example, the edge criterion

is not specifically defined, but they could be GO-based predictome edges, for example. In observed graph B in Figure 1, the ten nodes are connected by ten edges in a fairly unstructured fashion. Data from a real biological experiment could very well result in a more highly organized set of relationships between the genes.

The common features of the observed graphs A and B can be recorded in an intersection graph containing the same nodes, but only those edges common to both graphs A and B (see the observed intersection graph in Figure 1). The intersection graph is used for quantifying the association between the two data types. In our examples, we count the number of edges in the intersection graph to measure association, but we could also use other features of the intersection graph, for example the number of connected components, the number of triads (three nodes all connected to one another), and any other biologically relevant characteristics.

The two tests proposed in this paper statistically analyze whether the edges in observed graph A are over-represented in observed graph B; that is, whether there is an association between the different observations recorded in the two graphs. The edge permutation test is constructed as follows: Let X denote the number of edges in the intersection graph. To obtain realizations of intersection graphs under the null hypothesis of no association between the graphs A and B, we randomly permute cluster graph A's edges multiple times, say N. For each permutation, we intersect the permuted graph A with the observed graph B and count the number of edges in the new intersection graph. The proportion of permutations for which the number of edges in the intersection graph is at least as large as the observed value of X is an approximate P value for testing the null hypothesis of no association between the two graphs. In this edge permutation scheme, each count of intersection graph edges (i.e. X) is a realization of a Hypergeometric(m, n, k) random variable, where m is the number of edges in observed graph A, $n = [N(N-1)/2] - m$ is the number of missing edges in observed graph A, and k is the number of edges in observed graph B (in our case, $m = 10$, $n = 35$, and $k = 10$, see Appendix for details). Thus, for the special case of 2 datasets, our edge permutation scheme is actually a simulation of Fisher's exact test (Fisher, 1925) obtained by conditioning on the total number of edges in the two observed graphs. The first column of Figure 1 depicts three iterations of the edge permutation test for observed graphs A and B. It is recommended that several thousand iterations be used for P value calculation.

As suggested previously, some functional genomics data may result in highly structured graphs. Given a collection of graphs with apparent non-random connectivity, it may be argued that conditioning only on the number of edges in the observed graphs for a permutation test may not provide the correct reference distribution for assessing statistical significance. An alternative approach is to condition on the entire structure of both graphs and permute the node labels rather than the edges, yielding the node label permutation test. In this case, the P value represents the probability of observing at least the observed number of intersecting edges given no association between graphs A and B, conditional on the edge structures of both graphs. Among other things, the node degree distribution — the degree of a node being the number of nodes to which it is connected — present in both graphs is preserved through node label permutation. If we consider the graphical structure in the observed data to be our best approximation of the underlying biology of the relationships under investigation, then the node label permutation test may be more appropriate than the edge permutation test. The second column of Figure 1 depicts three iterations of the node label permutation scheme for testing association between graphs A and B.

Thus far we have considered the setting where there is a single clustered data set to be integrated with a non-cluster data set. In the more general setting, we consider the integration of K data sets, where each can be represented as a graph, with nodes representing genes and edges representing any of several functional genomics relationships. The K graphs may have varying degrees of structure among the edges. We denote the K graphs by $G_1, \cdots, G_K$. The intersection of the K graphs results in a graph where a pair of nodes is connected by an edge if and only if the pair of nodes is connected in all the graphs $G_1, \cdots, G_K$. If X

denotes the number of edges in the intersection graph, the null hypothesis corresponds to the statement: the probability that a pair of nodes in graph G_k is connected is independent of the set of graphs $\{G_i : i \neq k, i = 1, \cdots, K\}$, for all $k = 1, \cdots, K$. The edge permutation test described earlier easily generalizes to this setting, where realizations under the null hypothesis are generated by randomly permuting the edges of each of the graphs $G_1, \cdots, G_K$ multiple times. As before, the proportion of permutations for which the number of edges in the intersection graph is at least as large as observed value of X is an approximate P value for testing the null hypothesis. In this general case, the number of edges in the intersection now follows a multidimensional Hypergeometric distribution. We note that the null distribution of the number of edges in the intersection graph after random permutation of edges in the graphs $G_1, \cdots, G_K$ is identical to that resulting from random permutation of edges in any subset of $K-1$ out of the original K graphs. The node label permutation test can also be generalized in a similar manner, based on random permutations of the gene node labels in each of K graphs, $G_1, \cdots, G_K$.

Simulation study

In this section, we present the results of a simulation study to compare the behavior of the edge and node label permutation tests under a range of alternative hypotheses. We use the same simple cluster membership data set discussed in Figure 1 and integrate it with a non-cluster data set on the same ten nodes with ten edges obtained through simulation (similar to graph B in Figure 1). The results of our simulation study indicate a difference in the computed probabilities of the observed number of intersecting edges between the edge and node label permutation schemes.

The simulation study was carried out as follows: Given the cluster graph A, we randomly select ten edges to appear in graph B, with X of the ten taken from graph A. The intersection graph for graphs A and B would then have ten nodes and X edges. We take X to be distributed according to a non-central Hypergeometric distribution with parameters $(m = 10, n = 35, k = 10, \psi)$, where ψ is equal to the ratio of the odds of a random edge connecting intracluster nodes from graph A to that of it connecting intercluster nodes (see Appendix for details). Thus, $\psi > 1$ indicates preferential connection of intracluster nodes in graph B using the cluster information in graph A. For each of $\psi = 1, 2, 3, 4, 5$, we generated 1000 realizations of X drawn from the non-central Hypergeometric distribution with parameters $(m = 10, n = 35, k = 10, \psi)$. For each realization of X, we created graph B connecting X intracluster edges and $10 - X$ intercluster edges with no particular structure for the observed edges. Both permutation tests were performed, with 10,000 permutations each. For each value of ψ and each statistical test, we computed the proportion of the 1000 graphs for which the P value was less than 0.05. As Figure 2 demonstrates, the node label permutation approach results in a consistently lower probability of detecting preferential connection of intracluster nodes compared to the edge label permutation approach in this setting. This difference in probability is reduced as the graphs become larger and richer in structure, as will be seen in the real data examples.

Given the fact that the edge permutation scheme more readily detects preferential connection in the above example, it might be tempting to conclude that the edge permutation test is more powerful than the node label permutation test. Rather than power, what is really of interest is the ability of these two permutation schemes to cover the sample space of the correct reference distribution for use in statistical inference. In the above simulation study, graph B followed a random edge allocation pattern with preferential connection of intracluster edges under the alternative hypothesis. In this case, the edge permutation scheme quickly covers the correct null sample space, and is a suitable algorithm. In other cases where the graphs are more highly structured, the edge permutation scheme includes samples in the reference distribution that may be outside of the sample space of the graphs of interest.

Several papers (Maslov & Sneppen, 2002), (Jeong *et al.*, 2001) have demonstrated that protein-interaction networks do not have the structure of a random edge graph, but rather have a scale-free, small-world structure. In a

scale-free graph, the number of nodes of degree d is inversely proportional to a positive power of d. Small-world networks are characterized by *hub* nodes having very high degree, giving the graph a local order but a global disorder. If one of our graphs represented a protein network, the node label permutation approach seems to be more appropriate since it preserves the appropriate structure of all graphs involved when sampling the reference distribution.

We note here that when using the number of edges as the feature of interest of the intersection graph, the edge and node label permutation schemes frequently yield very similar results for graphs with a large number of nodes and a relatively low proportion of node pairs connected by edges. This will be evident in the examples we discuss in the next section. When other features, such as the number of connected components or the number of triads of the intersection graph are considered, the two permutation algorithms typically yield very different results (unpublished data). Investigators are encouraged to examine several features of intersection graphs to completely explore all relevant biology, and then carefully select the edge or node label permutation scheme based on the structure of the graphs involved.

Data analysis

Association between transcriptome (Causton et al.*, 2001) and GO-derived predictome*

As previously discussed, GO annotation is the result of manual review by expert biologists of current research on MF, BP and CC for genes and their products. Given the availability of high-throughput microarray gene expression data, it is of interest to know whether pairs of genes with similar known annotations also share gene expression profiles. If such a relation does exist, this sheds light on at least one of the mechanisms by which the cell coordinates genes with similar functions.

Table 2 presents the number of nodes with positive degree and the number of edges in the intersection graphs obtained by integrating each GO-derived predictome graph with the transcriptome graph. The nodes with positive degree in Table 2 refer to genes connected to at least one other gene in these intersections of the transcriptome graph and the GO-based predictome graphs. The P values for the number of edges in the intersection graph resulting from edge and node label permutation tests were highly significant ($P < 0.007$) for the association between transcriptome and each predictome derived using a combination of BP and CC ontologies as well as a combination of MF and CC ontologies respectively. This suggests that there is a significant relationship between gene expression and functional annotation, confirming previous suggestions that transcription is one way in which the cell regulates cofunctioning genes. We note that in all but the smallest graphs, both the edge and node label permutation tests yield similar P values. A secondary analysis done by excluding annotations that were inferred from expression profile data also yielded similar results. Details are available upon request.

Association between phenotypic fitness data (Giaever et al.*, 2002) and GO-derived predictome*

In addition to gene expression data, it is also of interest to know whether the phenotypic fitness data significantly overlaps with the GO-based predictomes. If genes with similar phenotypic fitness outcomes share functional annotation, this could suggest future experiments for testing and constructing specific gene networks that govern cellular fitness.

Table 3 presents the number of nodes with positive degree and the number of edges in the intersection graphs of each GO-derived predictome graph with the phenotypic fitness graph. For each test of association of phenotypic fitness data with the predictome derived from the MF-CC ontologies, the P values for the number of intersecting edges were highly significant ($P < 0.001$). This suggests that phenotypic fitness and molecular function are indicative of one another, and further investigations along these lines may prove fruitful. The associations between phenotypic fitness data and predictomes derived from the BP-CC ontologies were weaker, especially for the predictomes derived using a CC cut-off equal to the 99th percentile of D_{ij} values. It is

somewhat surprising to find a correspondence between phenotypic fitness data and MF yet not between phenotypic fitness data and BP. This may be explained to some degree by the fact that the phenotypic fitness data in Giaever *et al.*, (2002) were collected from yeast grown in suboptimal conditions. Alternatively, this result may simply be an artifact of incompleteness of the MF ontology due to lack of data sources. Under such conditions, normal biological processes may have been repressed, and genes may have been more involved in stress response. A secondary analysis done by excluding annotations that were inferred from expression profile data (GO evidence code IEP) also yielded similar results. Details are available upon request.

Assocation between transcriptome (Causton et al.*, 2001) and phenotypic fitness data (Giaever* et al.*, 2002)*

We also tested the association between the transcriptional data and phenotypic fitness data to see whether genes that affect cellular fitness under similar experimental conditions are also transcriptionally coregulated. We obtained non-significant P values of 0.96 and 0.74 for the association between transcriptome data from Causton *et al.*, (2001) and the phenotypic fitness data from the gene knock out experiments (Giaever *et al.*, 2002) using the edge and node label permutation tests, respectively. This result is in accordance with the findings reported in the Giaever *et al.*, (2002) paper. It is interesting to note that we obtained highly significant P values (< 0.001) for testing a similar association using the k-means transcriptome clusters generated by Tavazoie *et al.*, (1999) from a different set of mRNA expression experiments (Cho *et al.*, 1998). The data used for clustering in the Tavazoie study involved profiling the mRNA expression of the yeast genome at multiple time points across the duration of two cell cycles. The phenotypic data's strong association with one gene expression data set but not another is curious. A potential explanation for this phenomenon could be that a gene's role in phenotypic fitness has (at least) two components — stress response and cell cycle — and that these components roughly correspond to the expression data in Causton *et al.*, (2001) and Cho *et al.*, (1998), respectively. Given the strong association between phenotypic fitness and the cell cycle expression data set but weak association between phenotypic fitness and the stress response expression data set, we hypothesize that the signal due to the cell cycle role , in the phenotypic fitness data, overwhelmed their stress response role signals. Clearly, this idea merits closer examination.

Association between transcriptome (Causton et al.*, 2001), phenotypic fitness data (Giaever* et al.*, 2002) and GO-derived predictome*

All three data sets can also be combined to assess the joint correspondence of transcriptional and phenotypic profiles with functional annotation. Pairs of genes sharing all three features may be members of transcriptionally regulated functional pathways that contribute to cellular fitness under certain conditions.

Table 4 presents the number of nodes with positive degree and the number of edges in the intersection graphs obtained by integrating each GO-derived predictome graph with the transcriptional and phenotypic fitness graphs. In order to obtain rich graphs for the intersection of all three datasets, we used less stringent cut-offs for creating the predictomes. All four three-way associations yielded statistically significant P values from both edge and node label permutation tests, confirming the significant overlap of transcriptionally and phenotypically similar genes that are also known to be functionally related. A secondary analysis done by excluding annotations that were inferred from expression profile data also yielded similar results. Details are available upon request.

Discussion

In this paper, we have demonstrated that the congruency of disparate data types can be assessed using a graph theoretic framework. For a more general setting involving the association of an arbitrary number of data sets, this framework allows for easy representation of the intersection graph and facilitates the development of statistical tests of association.

We also saw that the edge and node label permutation tests may yield different results, and that the appropriate algorithm should be chosen based on the structure of the graphs being used. We then explored the use of the Gene Ontology database as a source for deriving predictome data sets, based on the fact that genes with closely related biological process or molecular function annotation should be more likely to generate protein products that are functionally linked. Our permutation tests resulted in highly significant P values for the association between transcriptome and GO-derived predictome data sets, as well as for the association between phenotypic fitness profiles in gene-deletion mutants and the GO-derived predictomes. These integrated data sets could be used to investigate high-throughput data sets exploring protein-protein interactions, such as Gavin *et al.*, (2002), Ho *et al.*, (2002), Uetz *et al.*, (2000) and Ito *et al.*, (2001), and can also be used to suggest targets for further investigation via direct experiments assessing protein interactions.

We see many directions for extending the approach we have explored in this paper. While we have emphasized tests of association for multiple data types, these methods could be extended to combine imperfect predictors of functional links from multiple independent sources of data to form a stronger predictor. In a recent paper by (Jansen *et al.*, 2003), the authors propose a Bayesian networks approach to the integration of several sources of data. In our graph theoretic setting, we could weight the edges connecting the nodes to account for the false positive and/or false negative rates in the original data set. The graphs developed in this paper are a special case where the weights are either 0 or 1 (i.e. an edge is either present of absent between any two nodes). Moreover, graphs with weighted edges could be used to represent putative functional links obtained from several GO-derived predictome graphs using a range of cut-off values. In this case, putative links present in predictome datasets obtained using high cut-off values can be represented by edges having proportionately higher weights compared to putative links only present in less stringent predictome datasets. Finally, by statistically combining several weighted graphs, one could propose a more general method to both integrate information from multiple experimental sources and assess their correspondence. Such an approach would not only be useful in generating more robust hypotheses regarding functional relationships between genes, but could also serve to probe and possibly validate new experimental data sets.

Appendix

The Hypergeometric distribution arises naturally in the context of a 2×2 contingency table. In our setting, the observations represented in the table are *pairs* of nodes present in the graph. The rows of the table classify each pair as either connected by an edge in graph A or not, and the columns classify pairs as connected in graph B or not. The situation with N nodes, m edges in graph A, $n = N(N-1)/2 - m$ missing edges in graph A, and k edges in graph B is shown in Table 5. The odds ratio here is defined as

$$\psi = \frac{\pi_{\mathrm{B|A}}(1-\pi_{\mathrm{B|A^c}})}{\pi_{\mathrm{B|A^c}}(1-\pi_{\mathrm{B|A}})},$$

where $\pi_{\mathrm{B|A}}$ denotes the probability that a pair of nodes will be connected in graph B conditional on being connected in graph A, and $\pi_{\mathrm{B|A^c}}$ denotes the probability that a pair of nodes will be connected in graph B conditional on not being connected in graph A. If we condition only on the number of edges in each graph, we may think of X, the number of edges in the intersection of graphs A and B, as a random variable with distribution given by

$$P(X = x|m,n,k,\psi) = \frac{\binom{k}{x}\binom{m+n-k}{m-x}\psi^x}{\sum_{l=0}^{\min(k,m)}\binom{k}{l}\binom{m+n-k}{m-l}\psi^l}.$$

This is a noncentral Hypergeometric distribution with noncentrality parameter ψ. The hypothesis of no association between edge connections in the two graphs is equivalent to $\psi = 1$, resulting in the standard Hypergeometric distribution. Each permutation in the edge permutation test then yields a realization of X under this distribution, and thus this test may be thought of as a simulation of Fisher's exact test (Fisher, 1925). In order to simulate graphs, for power calculations, from the alternative hypothesis of preferential connection in graph B of nodes connected in graph A, we generate realizations x_i ($i = 1, \ldots,$ number of realizations)

of a noncentral Hypergeometric variable X with $\psi > 1$ and connect x_i edges in graph B that are also in graph A and $k - x_i$ edges in graph B that are not in graph A. The results of these simulations are presented in Section 4. Integrating three or more, say K, graphs would generalize the above 2×2 contingency table to a 2^K contingency table, with the Hypergeometric distribution becoming the multidimensional Hypergeometric distribution. The node label permutation test, on the other hand, conditions on the topological structure of all graphs under consideration, and thus the null distribution of the number of common edges is more challenging to determine analytically.

Acknowledgments

We acknowledge helpful comments by two reviewers. This work was supported by grants AI024643, CA09647, T32-AI-07358, and 5R33HG002708-02.

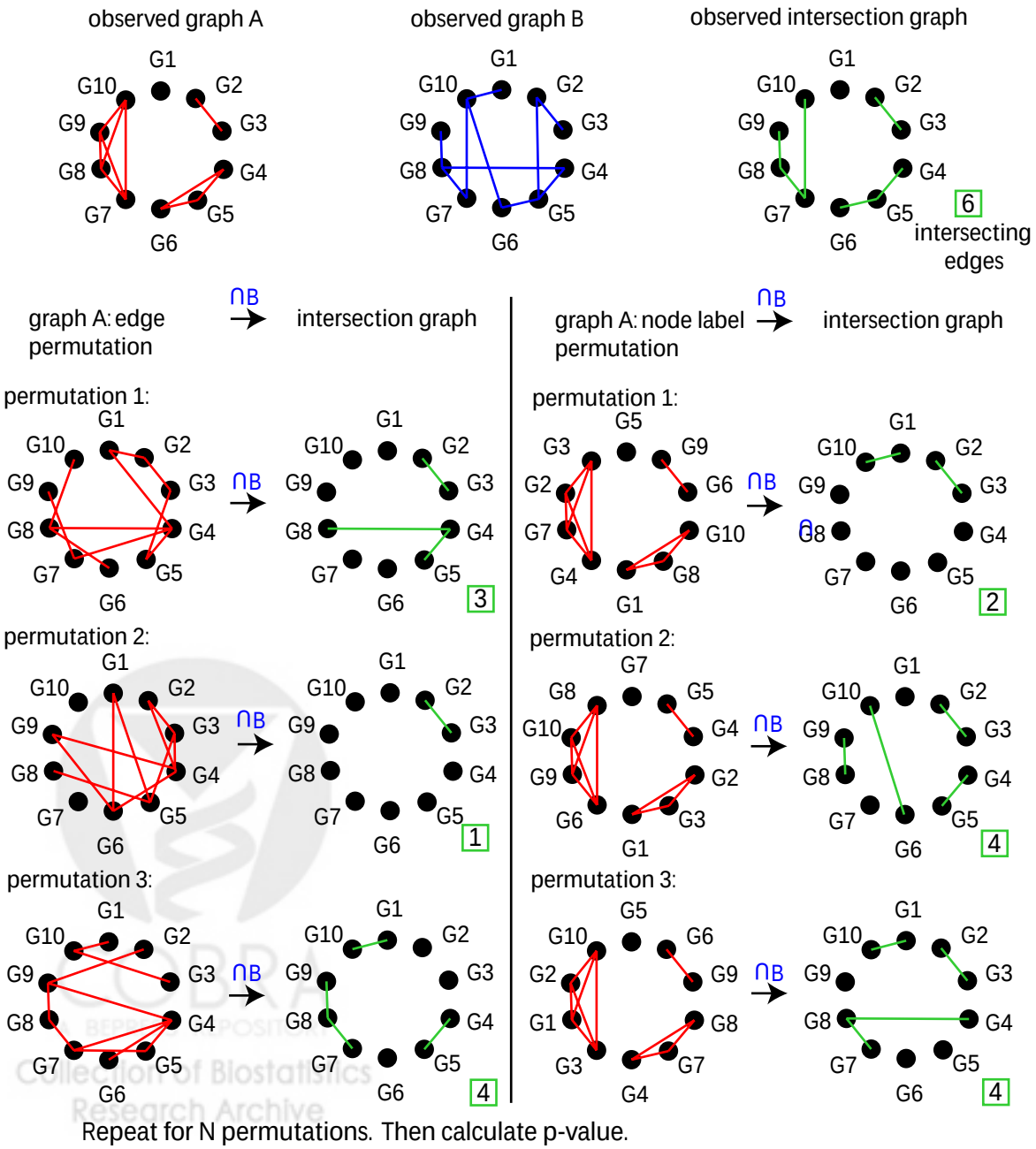

Figure 1: Edge and node label permutation schemes.

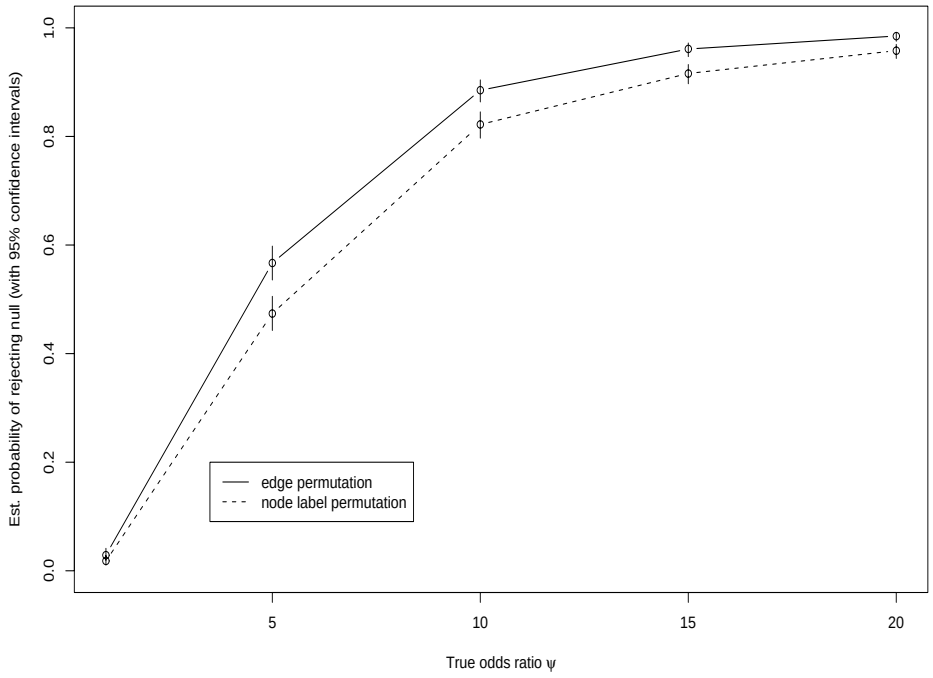

Figure 2: Estimated probability of detecting enrichment of intracluster edges for node label permutation and edge permutation tests (with 95% confidence intervals) as a function of the non-centrality parameter ψ of the non-central Hypergeometric distribution. As ψ increases away from 1, the simulated data is more likely to have preferential connection of nodes from the same cluster.

Table 1: Predictomes obtained by combining BP with CC and MF with CC networks respectively

	Biological Process: Cut-off Percentile (value)	
Cellular Component: Cut-off Percentile (value)	95 % (6)	99 % (7)
95 % (5)	No. of nodes (edges): 363 (2070) (predicted interactions)	No. of nodes (edges): 220 (1163) (predicted interactions)
99 % (6)	No. of nodes (edges): 151 (664) (predicted interactions)	No. of nodes (edges): 106 (444) (predicted interactions)

	Molecular Function: Cut-off Percentile (value)	
Cellular Component: Cut-off Percentile (value)	95% (6)	99 % (7)
95% (5)	No. of nodes (edges): 268 (743) (predicted interactions)	No. of nodes (edges): 110 (210) (predicted interactions)
99% (6)	No. of nodes (edges): 107 (176) (predicted interactions)	No. of nodes (edges): 42 (64) (predicted interactions)

Note: Only nodes with at least one predicted interaction are included.

Table 2: Intersection graph between transcriptome and each predictome derived by combining the GO ontologies — BP with CC and MF with CC networks respectively. P values are reported from edge and node label permutation tests respectively

	Biological Process: Cut-off Percentile (value)	
Cellular Component: Cut-off Percentile (value)	95% (6)	99 % (7)
95% (5)	No. of nodes (edges): 148 (204) P values: < 0.001, < 0.001	No. of nodes (edges): 79 (99) P values:< 0.001, < 0.001
99% (6)	No. of nodes (edges): 60 (63) P values: < 0.001, < 0.001	No. of nodes (edges): 35 (39) P values: < 0.001, 0.005

	Molecular Function: Cut-off Percentile (value)	
Cellular Component Cut-off Percentile (value)	95% (3)	99 % (5)
95% (5)	No. of nodes (edges) : 95 (114) P values: < 0.001, < 0.001	No. of nodes (edges): 49 (57) P values: < 0.001, < 0.001
99% (6)	No. of nodes (edges) : 36 (128) P values: < 0.001, < 0.001	No. of nodes (edges): 12 (8) P values: < 0.001, 0.007

Note: Only nodes with degree > 0 are included.

Table 3: Intersection graph between phenotypic fitness data and each predictome derived by combining the GO ontologies — BP with CC and MF with CC networks respectively. P values are reported from edge and node label permutation tests respectively

	Biological Process: Cut-off Percentile (value)	
Cellular Component Cut-off Percentile (value)	95% (6)	99 % (7)
95% (5)	No. of nodes (edges): 243 (297) P values: < 0.001, < 0.001	No. of nodes (edges): 154 (183) P values: < 0.001, < 0.001
99% (6)	No. of nodes (edges): 87 (79) P values: < 0.001, 0.017	No. of nodes (edges): 54 (46) P values: 0.044, 0.133

	Molecular Function: Cut-off Percentile (value)	
Cellular Component Cut-off Percentile (value)	95% (3)	99 % (5)
95% (5)	No. of nodes (edges) : 149 (171) P values: < 0.001, < 0.001	No. of nodes (edges): 55 (55) P values: < 0.001, < 0.001
99% (6)	No. of nodes (edges) : 54 (45) P values: < 0.001, < 0.001	No. of nodes (edges): 28 (23) P values: < 0.001, < 0.001

Note: Only nodes with degree > 0 are included.

Table 4: Intersection graph between transcriptome, phenotypic fitness data and each predictome derived by combining the GO ontologies — BP with CC and MF with CC networks respectively. P values are reported from edge and node label permutation tests respectively.

	Biological Process: Cut-off Percentile (value)	
Cellular Component Cut-off Percentile (value)	75% (4)	90 % (5)
75% (4)	No. of nodes (edges): 60 (49) P values: $< 0.001, < 0.001$	No. of nodes (edges): 41 (29) P values: $< 0.001, < 0.001$

	Molecular Function: Cut-off Percentile (value)	
Cellular Component Cut-off Percentile (value)	75% (2)	90 % (3)
75% (4)	No. of nodes (edges) : 44 (35) P values: $< 0.001, < 0.001$	No. of nodes (edges): 24 (15) P values: < 0.001 , < 0.001

Note: Only nodes with degree > 0 are included

Table 5: Tabular representation of intersecting edges of a two graphs A and B. Here N =total number of nodes, m =number of edges in graph A, $n = N(N-1)/2 - m$ =number of missing edges in graph A, and k =number of edges in graph B. If there are K graphs with $K > 2$, the integration of the graphs can instead be represented by a K-dimensional 2^K table.

	No. edges in graph B	No. missing edges in graph B	
No. edges in graph A	X	$m - X$	m
No. missing edges in graph A	$k - X$	$n - k + X$	n
	k	$m + n - k$	$\binom{N}{2} = m + n$

References

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S. & Eppig, J. *et al.* (2000) Gene ontology: tool for unification of biology. The Gene Ontology Consortium. *Nat. Genet.,* **25**, 25–29.

Bader, G., Donaldson, I., Wolting, C., Ouellette, B., Pawson, T. & Hogue, C. (2001) BIND-The Biomolecular Interaction Network Database. *Nucleic Acids Res.,* **29**, 242–245.

Butte, A., Tamayo, P., Slonim, D., Golub, T. & Kohane, I. (2000) Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl. Acad. Sci. U. S. A.,* **97**, 12182–12186.

Causton, H., Ren, B., Koh, S., Harbison, C., Kanin, E., Jennings, E., Lee, T., True, H., Lander, E. & Young, R. (2001) Remodeling of yeast genome expression in response to environmental changes. *Mol. Biol. Cell,* **12**, 323 – 337.

Chervitz, S., Hester, E., Ball, C., Dolinski, K., Dwight, S., Harris, M., Juvik, G., Malekian, A., Roberts, S. & Roe, T. *et al.* (1999) Using the Saccharomyces Genome Database (SGD) for analysis of protein similarities and structure. *Nucleic Acids Res.,* **27(1)**, 74–78.

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

Deane, C., Salwinski, L., Xenarios, I. & Eisenberg, D. (2002) Protein interactions: two methods for assesment of the reliability of high throughput observations. *Mol. Cell. Proteomics,* **1**, 349 – 356.

Fisher, R. (1925) *Statistical Methods for Research Workers.* Oliver and Boyd, Edinburgh.

Gavin, A., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J., Michon, A. M. & Cruciat, C. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature,* **415**, 141–147.

Ge, H., Liu, Z., Church, G. & Vidal, M. (2001) Correlation between transcriptome and interactome mapping data from *saccharomyces cerevisiae*. *Nat. Genet.,* **29**, 482–486.

Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K. & Andre, B. *et al.* (2002) Functional profiling of the *saccharomyces cerevisiae* genome. *Nature,* **418**, 387–391.

Grigoriev, A. (2001) A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast *saccharamyces cerevisiae*. *Nucleic Acids Res.,* **29**, 3513–3519.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G., Moore, L., Adams, S., Millar, A., Taylor, P., Bennett, K. & Boutilier, K. *et al.* (2002) Systematic identification of protein complexes in *saccharomyces cerevisiae* by mass spectometry. *Nature,* **415**, 180–183.

Hodges, P., McKee, A., David, B., Payne, W. & Garrels, J. (1999) The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucleic Acids Res.,* **27**, 69–73.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. & Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U.S.A.,* **98**, 4569 – 4574.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. & Gerstein, M. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science,* **302**, 449–453.

Jeong, H., Mason, S., Barabasi, A.-L. & Oltvai, Z. (2001) Lethality and centrality in protein networks. *Nature,* **411**, 41–42.

Kemmeren, P., van Berkum, N., Vilo, J., Bijma, T., Donders, R., Brazma, A. & Holstege, F. (2002) Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol. Cell,* **9**, 1133–1143.

Lord, P., Stevens, R., Brass, A. & Goble, C. (2003) Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics,* **19**, 1275–1283.

Maslov, S. & Sneppen, K. (2002) Specificity and stability in topology of protein networks. *Science,* **296**, 910–913.

Mellor, J., Yanai, I., Clodfelter, K., Mintseris, J. & DeLisi, C. (2002) Predictome: a database of putative functional links between proteins. *Nucleic Acids Res.,* **30 (1)**, 306 – 309.

Mewes, H., Amid, C., Arnold, R., Frishman, D., Guldener, U., Mannhaupt, G., Munsterkotter, M., Pagel, P., Strack, N. & Stumpflen, V. *et al.* (2004) MIPS: analysis and annotations of proteins from whole genomes. *Nucleic Acids Research,* **32**, D41–D44.

Uetz, P., Giot, L., Cagney, G., Mansfield, T., Judson, R., Knight, J., Lockshon, D., Narayan, V., Srinivasan, M. & Pochart, P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in *saccharomyces cerevisiae. Nature,* **403**, 623–627.

Xenarios, I., Rice, D., Salwinski, L., Baron, M. K., Marcotte, E. & Eisenberg, D. (2000) DIP: the Database of Interacting Proteins. *Nucleic Acids Res.,* **28**, 289 – 291.

Zhou, X., Kao, M. & Wong, W. (2002) Transitive functional annotation by shortest path analysis of gene expression data. *Proc Natl Acad Sci USA,* **99**, 12783 – 12788.