

# Transitive functional annotation by shortest-path analysis of gene expression data

Xianghong Zhou\*, Ming-Chih J. Kao\*, and Wing Hung Wong\*†‡

\*Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115; and †Department of Statistics, Harvard University, Cambridge, MA 02138

Edited by Michael S. Waterman, University of Southern California, Los Angeles, CA, and approved July 19, 2002 (received for review March 18, 2002)

Current methods for the functional analysis of microarray gene expression data make the implicit assumption that genes with similar expression profiles have similar functions in cells. However, among genes involved in the same biological pathway, not all gene pairs show high expression similarity. Here, we propose that transitive expression similarity among genes can be used as an important attribute to link genes of the same biological pathway. Based on large-scale yeast microarray expression data, we use the shortest-path analysis to identify transitive genes between two given genes from the same biological process. We find that not only functionally related genes with correlated expression profiles are identified but also those without. In the latter case, we compare our method to hierarchical clustering, and show that our method can reveal functional relationships among genes in a more precise manner. Finally, we show that our method can be used to reliably predict the function of unknown genes from known genes lying on the same shortest path. We assigned functions for 146 yeast genes that are considered as unknown by the *Saccharomyces* Genome Database and by the Yeast Proteome Database. These genes constitute around 5% of the unknown yeast ORFome.

**D**NA microarrays simultaneously monitor the expression levels of thousands of genes. The massive gene expression data provide us with unique opportunities to analyze the functional and regulatory relationships among genes. One useful approach is to cluster genes with similar expression patterns. The most popular clustering methods include hierarchical clustering (1), *K*-means clustering (2), and self-organizing maps (3). On the assumption that genes with similar expression profiles have similar biological functions, functions of unknown genes can be predicted from their expression-similarity to known genes (1, 4).

However, do genes with similar functions always have similar expression profiles? The answer is, of course, no. First, genes with similar functions may not have been exposed to sufficient perturbations for their expression similarities to be revealed. Second, for some genes with similar functions, their product concentrations are partially or totally controlled at levels other than transcription. Third, measurements of expression similarity—e.g., Pearson's correlation or Euclidean distance—may not be able to completely capture the relationship between two expression profiles for such reasons as time-shift (5). Therefore, to identify the functional relationships between genes, we need to look beyond clustering methods. In this paper, we propose a method to group genes involved in the same biological process, even those without significant expression similarity.

First, we introduce an important characteristic of biological processes that we call *transitive co-expression*. Intuitively, this refers to situations where two genes are not strongly correlated in expression, but are both strongly correlated with the same set of other genes. In the simplest case, suppose genes *a* and *b* have strong expression correlation, as well as genes *b* and *c*. However, genes *a* and *c* do not have strong expression correlation, so we say that they are transitively co-expressed, with gene *b* serving as the *transitive gene*. It is clear that in a biological pathway, a gene is likely to show strong expression correlations with its neighbor genes, but not with genes that lie far apart in the pathway. This lack of correlation can be caused by various reasons. For

example, (i) some biological processes are protracted in time, e.g., the cell cycle, so that expression relationships are revealed at different time points along the process. Using transitive co-expression, we can establish linkages between these genes to unveil the complete pathway. (ii) genetic and biochemical networks of a cell must withstand substantial random perturbations. Based on the principle of efficiency, mechanisms such as negative feedback loops limit the number of genes that fluctuate in their expressions. Thus different sets of experiments perturb different segments of a biological pathway to different extents, so that the expression relationships among genes along the pathway are revealed in a compartmentalized fashion across experiments. Such compartmentalized pathway segments can be associated through overlapping genes by transitive co-expression so that the biological pathway can be revealed as a whole.

Given two genes known to be involved in the same biological pathway, identifying transitive genes between them may allow us to discover genes involved in the same biological process. Here, we propose a graph-theoretic scheme to identify such transitive genes. In our graph, vertices represent genes, and each gene pair that is highly correlated in expression is connected with an edge, where the edge length is a decreasing function of the expression correlation (Fig. 1A). Given such a graph, there can be multiple expression dependence paths between two genes—e.g., between genes *a* and *e*. The shortest of these paths (the shortest path, SP) would then be the most parsimonious explanation of dependence between *a* and *e*, given our expression data. If genes *a* and *e* are highly correlated, the shortest dependence path between them would just be the edge connecting them; if they are not significantly correlated but are involved in the same pathway, then we may still be able to discover their expression association by constructing the shortest dependence path between them. The transitive genes on the SP are likely to be important intermediate players between the two terminal genes in the same process.

In this paper, using yeast expression data, we first validate that the SP method is able to link functionally related genes, even without high expression correlation, and show the results to be highly statistically significant. In addition, we make comparisons to hierarchical clustering to show that our method can more precisely reveal functional relationships among genes. Finally, given that our method can group functionally related genes together, we make predictions for the functions of unknown genes based on those of known genes on the same SP. We made predictions for 146 unknown yeast genes, a significant number of which are supported by evidence other than the data we used.

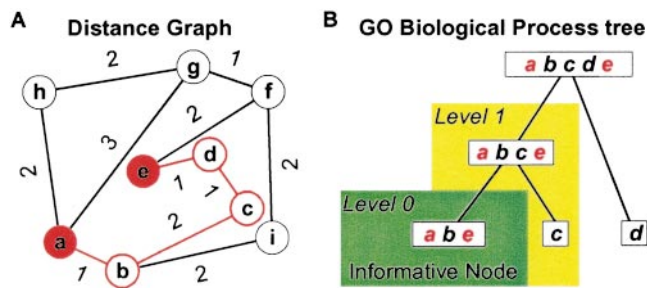
## Methods

**Data Processing.** We used the *Saccharomyces cerevisiae* gene expression profiles from the Rosetta Compendium (6), which includes 300 deletion and drug treatment experiments. Genes

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: SP, shortest path; L0, level 0; L1, level 1; GO, Gene Ontology; SGD, *Saccharomyces* Genome Database; YPD, Yeast Proteome Database.

†To whom reprint requests should be addressed at: Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115. E-mail: wwong@hsph.harvard.edu.



**Fig. 1.** (A) Application of the shortest-path (SP) algorithm to gene expression data. Nine genes are depicted in the graph. The distance between two genes is a decreasing function of their correlation. For example, there are multiple expression dependence paths leading from gene *a* to gene *e*. Among them, the shortest dependence path is *a–b–c–d–e*, with genes *b*, *c*, and *d* serving as the *transitive genes*. This is the most parsimonious summary of the expression relationship between the terminal genes *a* and *e*. (B) Level 0 (L0) and level 1 (L1) matches of genes on the SP *a–b–c–d–e* defined according to their relationships in the Gene Ontology (GO) classification tree. With respect to the terminal genes *a* and *e*, the transitive gene *b* is a L0 match because it is annotated in the informative node where *a* and *e* are annotated; the transitive gene *c* is a L1 match because it shares the same direct parent as the two terminal genes; the transitive gene *d* is neither a L0 nor a L1 match.

were annotated by using the biological process ontology of Gene Ontology (GO) (7) provided by the *Saccharomyces* Genome Database (SGD) (8). To verify that genes on the same SP are likely to be involved in the same biological process, we applied our method to the Rosetta dataset and checked the results against GO-annotated biological processes in the three major cellular compartments: mitochondria, cytoplasm, and nucleus. Genes are separated according to their subcellular localizations because sometimes a GO process category may actually encompass distinct processes. Although similar in nature, these processes may occur in different cellular compartments and thus are not necessarily tightly controlled in regulation. For example, “protein biosynthesis” can take place in the mitochondria and the cytoplasm, and “membrane transport” processes in different compartments are distinct. We refine the process categories by sorting genes into the three major cellular compartments.

After removing the genes without GO process annotation and the 20 genes for which there are less than 80 experimental measurements in the Rosetta Compendium, we were left with 266 mitochondrial, 398 cytoplasmic, and 659 nuclear GO-annotated genes. For each of the three sets of genes, we calculated the expression similarities of all gene pairs  $\{a, b\}$  using  $C_{a,b}$ , the minimum of the absolute value of leave-one-out Pearson correlation coefficient estimates. This estimate is a measurement robust against single experiment outliers and sensitive to overall similarities in expression patterns.

**Graph Construction and SP Computation.** We constructed three graphs, one for each set of the 266 mitochondrial genes, the 398 cytoplasmic genes, and the 659 nuclear genes. In each graph, two genes were assigned an edge if their absolute expression correlation  $C_{a,b}$  was higher than  $\tau = 0.6$ . This cut-off, while conservative, nonetheless retains a sufficient number of connected gene pairs in the graph. The edge length between vertices *a* and *b* is  $d_{a,b} = f(C_{a,b}) = (1 - C_{a,b})^k$ . The powering factor *k* is used to enhance the differences between low and high correlations. Because the length of a path is the sum of the individual edge lengths, by exaggerating the differences between edge lengths, the SPs will be more likely to cover more transitive genes. Thus by increasing *k* we gain more power to reveal transitive co-expression. We set  $k = 6$  because for  $k \geq 6$ , the numbers of transitive genes stabilizes (detailed results at [www.biostat.harvard.edu/complab/SP/](http://www.biostat.harvard.edu/complab/SP/)).

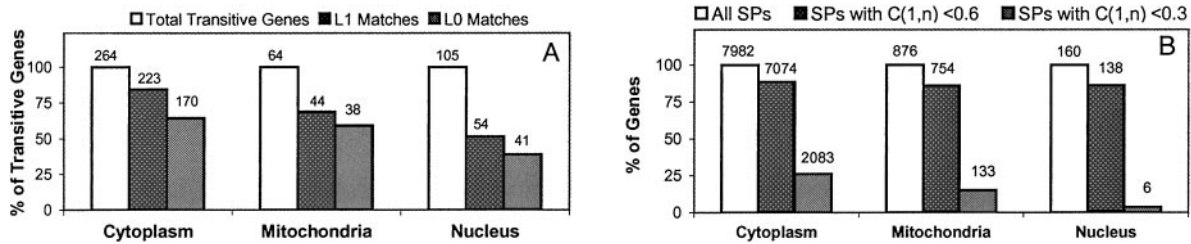
To ensure the quality of SPs, we consider only SPs with total path lengths  $< 0.008$ .

We used Dijkstra’s algorithm to identify the SPs between a source vertex to all other vertices in a graph. The central idea of Dijkstra’s algorithm is the “relaxation” operation, which is based on the fact that every subpath of a SP is also a SP. For instance, in Fig. 1A, suppose that *a* is the source vertex and that we already have the SP from *a* to *d* via *b* and *c*. By relaxing each edge leaving *d* we attempt to see what is the SP from *a* to all other vertices that can be directly reached from *d*, in particular vertex *e*. If the path from *a* to *e* going through *d* is shorter than the current best SP estimate for *e*, then this estimate is updated to the new path. A detailed introduction to Dijkstra’s algorithm can be found in ref. 9 and other algorithm textbooks. Exploiting the sparse nature of our graph, we implemented a priority queue of vertices with a Fibonacci heap to achieve  $O[n \log(n) + m]$  time complexity, where *n* is the number of vertices and *m* is the number of edges in the graph. To determine the SPs starting from *v* source vertices, we applied Dijkstra’s algorithm to each of the *v* vertices. Thus the overall time complexity is  $O[nv \log(n) + mv]$ .

**Analyzing the SP Between Two Genes Involved in the Same Biological Process.** GO is a set of controlled biological vocabularies organized in a rooted directed acyclic graph. For our purposes it can be treated as a tree. Nodes in the tree refer to biological process categories. A parent node refers to a more general annotation than that of its children. SGD annotates each known yeast gene with one or more nodes in the GO tree. From all available annotations, we want to select those process categories that do not include too many genes to guarantee the functional homogeneity of member genes, and that do not include too few genes to provide sufficient numbers of genes for the purpose of validation. Using an approach similar to that proposed by Hvidsten *et al.* (10), we retrieved such process categories from GO by traversing the tree breadth-first from the root and selecting nodes that satisfy the properties that (i) the node contains more than  $\gamma = 30$  genes and (ii) each of the node’s children contains less than  $\gamma$  genes. We define such GO nodes as *informative nodes*, and the biological process categories they represent as the *informative categories*.

To test the validity of the SP method, we need to see whether genes on the same SP share the same GO process annotations. Given any two genes from the same informative process category (*terminal genes*), we determine whether there is a SP connecting them. If there is a SP including one or more transitive genes, we check the GO process annotations of these genes. A transitive gene is termed a *level 0* (L0) match if it is annotated in the informative node in the GO tree from which the terminal genes are selected; it is termed a *level 1* (L1) match if it shares the same direct parent node with the terminal genes (Fig. 1B). For all SPs connecting all gene pairs in each informative category, we count the total number of transitive genes as well as the numbers of L0 and L1 matches. For each cellular compartment we sum over the results of its informative categories and calculate the L0 and L1 match ratios relative to the total number of transitive genes.

**Using a Permutation Test to Assess the Statistical Significance of the SPs.** For the purpose of validating the SP method, we need to evaluate the numbers of L0 and L1 matches, taking into account the numbers of such matches expected under the null hypothesis. Keeping the graph structure constant, we randomly permute the gene labels over the vertices to decouple gene annotations from their expression profiles. We then perform the SP method over these graphs, and calculate the L0 and L1 match ratios. This is done for 1,000 iterations. The distribution of L0 and L1 match ratios generated under the null hypothesis is compared with the observed quantities. The *P* values so derived give us an assess-



**Fig. 2.** Summary of the performance of the SP method. (A) The percentages of L0- and L1-matched transitive genes in the three cellular compartments. Values shown above the bars are the numbers of genes. All match ratios at L0 and L1 are statistically significant at  $P < 0.001$  by permutation test. (B) The percentages of SPs with at least one transitive gene in which terminal genes show weak ( $<0.6$ ) and very weak ( $<0.3$ ) expression correlations.  $C(1, n)$  denotes the expression correlation between the terminal genes. Values shown above the bars are the numbers of SPs.

ment of the ability of the SP method to reveal biological relationships between genes based on microarray data.

**Predicting the Functions of Unknown Genes.** We use the SP method to classify previously unannotated yeast genes by adding the 3,255 ORFs unknown to SGD into the graphs of known genes in the mitochondrial, cytoplasmic, and nuclear compartments. As before, an edge is constructed between two genes if their absolute expression correlation is higher than 0.6. For all pairs of known genes, we determine the SPs connecting them. For the purpose of functional prediction, we would like to assign a putative function that is as specific as possible to the gene. Given all known genes on a SP, we achieve this by tracing back their annotations along the GO process tree and finding their lowest common ancestor. If the lowest ancestral node is at least 4 levels below the root of the GO tree, that is, it defines a sufficiently specific gene function, we then assign this function to the unknown genes on the SP. Analogous to the L0 and L1 matches, here the L0 prediction then corresponds to the lowest common ancestor, and the L1 prediction to its direct parent. In this way, the function represented by the lowest common ancestor can be more specific than that defined by the informative nodes. Under two circumstances an unknown gene may be assigned with multiple functions: (i) Because known genes on a SP may each have multiple functions, they may share several lowest common ancestors in the GO tree. (ii) An unknown gene may reside in different SPs with different lowest common ancestors. For each predicted gene function, we provide both the number of *support SPs* from which the prediction was derived and the number of unique known genes on those support SPs (*support genes*). The more support genes there are, the more confidence we have in the corresponding prediction. Note that a gene can be assigned putative functions in multiple graphs, because many genes are known to function in multiple cellular compartments.

## Results

### SP Method Clusters Genes Involved in the Same Biological Process.

We constructed the graphs for all yeast genes with GO process annotations in mitochondria (266 genes), cytoplasm (398 genes), and nucleus (659 genes). Using the procedure defined in *Methods*, we obtained 4, 8, and 22 informative GO categories for genes in the graphs of mitochondria, cytoplasm, and nucleus, respectively. The numbers of genes in those informative categories range from 31 to 174. In each graph, given any two genes belonging to the same informative GO process category, we identify the SP connecting them. We check the GO process classification of the transitive genes on the SP to identify the L0 and L1 matches, as discussed in *Methods*.

For each compartment, we summarize the results in Fig. 2A. In the cytoplasm and mitochondria graphs the SP method achieved surprisingly accurate results, where it successfully called 84%/64% and 69%/59% of the transitive genes at the L1/L0 levels, respectively. The nuclear graph shows relatively lower match ratios of 51%/39% at the L1/L0 levels. This result

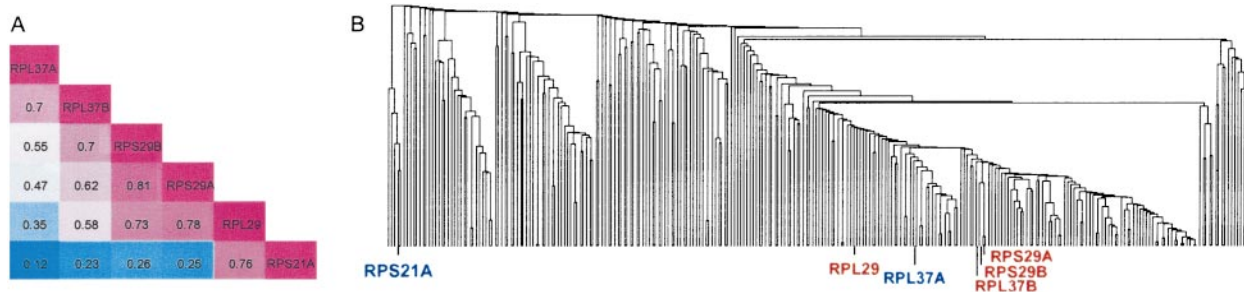
can be attributed to regulatory mechanisms exerted at levels other than transcription in the nucleus, for example regulated import/export of nuclear proteins. For the L1/L0 match ratios in all compartments, our method showed better performance than the 1,000 iterations of the permutation test, giving us  $P$  values less than 0.001. This demonstrates the significant power of the SP method in revealing biological relationships between genes based on microarray data.

It should be noted that, first, the match ratios we compute are in fact conservative. Because GO annotation is based only on positive biological evidence and is sparse in SGD, genes that are not classified as L1 or L0 matches may still be functionally related to the terminal genes on the SP. Second, we believe one major source of errors is the heterogeneity of perturbations in the Rosetta Compendium, where experimental conditions overall were not specifically designed for any particular biological process. Thus genes on a SP may be involved in several distinct biological processes that are nonetheless affected simultaneously by nonspecific perturbations. With experiments specifically designed for particular biological processes, we believe our approach will achieve higher accuracy.

### SP Method Reveals Functional Relationships Between Genes with Low Expression Similarity.

Among the identified SPs, particularly interesting are those with low expression similarities between the terminal genes. In fact, over all three graphs, for more than 85% of the SPs that include at least one transitive gene, their terminal genes have weak expression correlations ( $<0.6$ ) (Fig. 2B). This finding means that, in the graph, there is no edge between those terminal gene pairs; their relationships are revealed only by transitive co-expression by transitive genes. In a considerable subset of the SPs, transitive co-expression is even more pronounced. In fact, there are significant numbers of SPs with very weak expression correlations between the terminal gene pairs ( $<0.3$ ). In the cytoplasm graph, there are 2,083 such SPs, which make up 26% of the total SPs. Even for transitive genes in these SPs, we achieved a L1 match ratio of 74%. This result confirms that our method can successfully reveal the relationship between functionally related gene pairs through transitive co-expression, even if their expression correlation is not significantly high.

An example is the SP we identified in the graph of cytoplasm, RPL37A–RPL37B–RPS29B–RPS29A–RPL29–RPS21A from the GO process “protein biosynthesis.” The correlation matrix of the 6 genes is shown in Fig. 3A. In this SP, the expression correlation between the terminal genes is only 0.12. In addition, it is apparent that genes farther apart on the SP have lower expression correlations. All 6 genes encode ribosomal proteins and participate in protein biosynthesis. However, it is clear that they are not tightly coregulated as a group. We compare our method to standard hierarchical clustering methods based on expression correlations from the same Rosetta Compendium dataset. Among different linkages of hierarchical clustering, single linkage is the closest to the SP method in its greedy



**Fig. 3.** (A) The correlation matrix for the shortest path RPL37A–RPL37B–RPS29B–RPS29A–RPL29–RPS21A. The gene names and their correlations in the Rosetta dataset are indicated. The magnitudes of the correlations are represented by the colors: red, high; blue, low. (B) The minimum subtree covering the two terminal genes in the SP (RPL37A and RPS21A) contains 276 genes.

approach to building clusters by incrementally joining the closest, most correlated genes. For the SP mentioned above, using the 398 genes in the cytoplasm graph, the minimum hierarchical clustering subtree covering the 2 SP terminal genes (RPL37A and RPS21A) contains 276 genes (Fig. 3B). Among them, 144 are not ribosomal genes. Even genes involved in carbohydrate metabolism, osmosensory signaling, and fatty acid biosynthesis are included. Such remarkable contrasts between hierarchical clustering and the SP method are common in our results. As another example, the genes on the SP *COR1–RIP1–SDH2–SDH4–STF1* in the mitochondria graph are all involved in oxidative phosphorylation. The correlation between the two terminal genes is 0.16. By hierarchical clustering, the minimum subtree covering the 2 SP terminal genes contains 241 genes, which constitutes 91% of the genes in the mitochondria graph. The weakness of the hierarchical clustering method is that once a gene is assigned to a cluster, comparisons with other genes are no longer done at the gene-to-gene level but at the cluster-to-

gene or cluster-to-cluster level. Thus it does not offer us as parsimonious a description of the expression dependence as does the SP method. The same drawback also exists in other commonly used clustering methods, such as *K*-means clustering. Examples of such comparisons are available at our web site ([www.biostat.harvard.edu/complab/SP/](http://www.biostat.harvard.edu/complab/SP/)).

**Prediction of the GO Process Category for Unknown Genes.** After validating that the SP method can link functionally related genes, we used it to classify previously unannotated yeast genes. We used two rules to predict gene function. (i) A *general rule*: given all known genes on a SP, if their lowest common ancestor on the GO tree is more than 3 levels below the root, we then assign this function to the unknown genes on the SP. Over all SPs in the graphs of cytoplasm, mitochondria, and nucleus, we have made predictions for 80, 54, 115 genes, respectively, which are considered as unknown by SGD. (ii) Using a *conservative rule*, we consider those SPs that satisfy the general rule and that contain only one unknown gene.

**Table 1. Validation of the predictions made by the conservative rule for 24 genes without SGD GO process annotation, against their YPD cellular role annotation (as of March 2002)**

Gene	GO process category prediction (L0)	YPD annotation
<u>ADH1</u>	Main pathways of carbohydrate metabolism	Carbohydrate metabolism [E]
<u>ADH5</u>	Amino acid metabolism	Other metabolism [E]; carbohydrate metabolism [E]
<u>BMS1</u>	Nucleobase, nucleoside, nucleotide metabolism	RNA processing/modification [E]
<u>BUD28</u>	Protein biosynthesis	Cell polarity [E]
<u>CLC1</u>	Actin cytoskeleton organization	Vesicular transport [E]; cell polarity [E]
<u>COR1</u>	ATP synthesis coupled proton transport	Energy generation [E]; small molecule transport [P]
<u>CPR7</u>	Cell cycle	Protein folding [P] [protein folding, G <sub>1</sub> phase of cell cycle (19)]
<u>ELP3</u>	Transcription, DNA-dependent	Pol II transcription [E, P]; protein modification [P]; chromatin/chromosome structure [P]
<u>ERB1</u>	35S primary transcript processing	RNA processing/modification [E] [involved in 35S primary transcript processing (11)]
<u>GPH1</u>	Carbohydrate metabolism	Carbohydrate metabolism [E]; cell stress [E]
<u>GSP1/Ran</u>	Protein biosynthesis; ribosome biogenesis; 35S primary transcript processing; G <sub>2</sub> /M transition of mitotic cell cycle	Cell cycle control [E]; nuclear–cytoplasmic transport [E]; RNA processing/modification [E]
<u>MAK16</u>	Ribosome biogenesis; RNA processing	RNA processing/modification [E]
<u>NFI1/SIZ2</u>	Protein metabolism and modification; protein degradation	Cell cycle control [E]; protein modification [E]
<u>NOC3</u>	Ribosome biogenesis; 35S primary transcript processing	Protein synthesis [E] [involved in the biogenesis of the 60S ribosomal subunit (13)]
<u>NUG1</u>	RNA metabolism	Nuclear–cytoplasmic transport [E] [export of ribosomal subunit] (20)
<u>PRO2</u>	Biosynthesis	Amino acid metabolism [E]
<u>QCR8</u>	ATP synthesis coupled proton transport	Energy generation [E]; small molecule transport [P]
<u>RIB5</u>	Amino acid and derivative metabolism	Other metabolism [E]
<u>RIP1</u>	ATP synthesis coupled proton transport	Energy generation [E]; small molecule transport [P]
<u>RSM26</u>	Protein biosynthesis	Energy generation [E]; cell cycle control [E]; protein synthesis [P] [protein of the mitochondrial ribosome small subunit (21)]; cell stress [E]
<u>THR1</u>	RNA processing	Amino acid metabolism [E]
<u>UGP1</u>	Carbohydrate metabolism	Carbohydrate metabolism [E]
<u>XKS1</u>	Protein metabolism and modification	Carbohydrate metabolism [E]
<u>YGL068W</u>	Protein biosynthesis	Energy generation [P]; cell cycle control [E]; protein synthesis [P] [homolog of <i>Escherichia coli</i> L7/L12 ribosomal protein (22)]

The gene with underlined names have YPD annotations that either agree or are closely coupled to our predictions. “[E]” denotes experimental evidence; “[P]” denotes computational prediction. We provide additional references in parentheses if (i) the YPD cellular role annotation is not sufficiently specific (*ERB1*, *NUG1*), (ii) YPD has misclassified (*NOC3*), or (iii) computational cellular role predictions included by YPD match our prediction (*CPR7*, *RSM26*, *YGL068W*).

**Table 2. Predictions of GO process category made by the conservative rule for 51 genes with neither SGD GO process annotation nor YPD cellular role annotation**

GO process category prediction (L1 ⇒ L0)	Gene [no. unique support genes, no. support SPs, graph]
Amino acid metabolism ⇒ amino acid biosynthesis	<i>YHR029C</i> [9, 14, C]
Cell cycle ⇒ DNA replication and chromosome cycle	<i>TOS4</i> [5, 5, N]
Cell cycle ⇒ M phase	<i>YPL267W</i> [2, 1, N]
Cell cycle ⇒ mitotic cell cycle	<i>TOS4</i> [6, 8, N]
Cell organization and biogenesis ⇒ cytoplasm organization and biogenesis	<i>YNR046W*</i> [5, 4, N]
Cytoplasm organization and biogenesis ⇒ organelle organization and biogenesis	<i>RIM21</i> [2, 1, M]
Cytoplasm organization and biogenesis ⇒ ribosome biogenesis	<i>PUF6</i> [14, 46, N], <i>RRP12*</i> [3, 2, N], <i>YDR324C*</i> [6, 6, N], <i>YGR128C*</i> [2, 1, C], <i>YIL019W*</i> [2, 1, C], <i>YJL010C*</i> [3, 2, N], <i>YLR132C*</i> [3, 2, N], <i>YLR287C</i> [4, 4, N], <i>YML093W*</i> [2, 1, M] (23), <i>YML093W*</i> [3, 2, N], <i>YPL146C*</i> [4, 3, N]
DNA metabolism ⇒ DNA repair	<i>YBR089W</i> [2, 1, N]
Carbohydrate metabolism ⇒ catabolic carbohydrate metabolism	<i>PST2</i> [2, 1, M] (24)
Metabolism of energy reserves ⇒ trehalose metabolism	<i>TFS1</i> [2, 1, C]
Metabolism ⇒ biosynthesis	<i>YDR165W</i> [2, 1, N], <i>YLR132C*</i> [118, 117, C], <i>YPL246C</i> [3, 2, C] (16)
Metabolism ⇒ coenzymes and prosthetic group metabolism	<i>YLR356W</i> [2, 1, N]
Metabolism ⇒ nucleic acid metabolism	<i>DAT1</i> [3, 2, N], <i>YBR267W</i> [7, 6, N], <i>YDL063C</i> [5, 4, N], <i>YDR165W</i> [2, 1, N], <i>YGR128C*</i> [4, 4, N], <i>YGR145W</i> [2, 1, M], <i>YLR132C*</i> [3, 2, N], <i>YLR287C</i> [4, 4, N], <i>YNL311C</i> [2, 1, N], <i>YOR004W*</i> [4, 3, N], <i>YOR042W</i> [2, 1, N], <i>YPR045C</i> [5, 8, N]
Metabolism ⇒ protein metabolism and modification	<i>RNY1</i> [3, 3, C], <i>TOS5</i> [2, 1, C] (16), <i>YDR165W</i> [114, 113, C] (23), <i>YKL053C-A</i> [16, 62, M], <i>YKL195W*</i> [25, 37, M] (16), <i>YLR356W</i> [46, 45, C], <i>YLR434C</i> [119, 124, C]
Mitotic cell cycle ⇒ S phase of mitotic cell cycle	<i>YBR089W</i> [2, 1, N] (25)
Monosaccharide metabolism ⇒ hexose metabolism	<i>YCRO13C*</i> [3, 2, C] (18)
Nucleic acid metabolism ⇒ RNA metabolism	<i>YDR324C*</i> [3, 1, N], <i>YML093W*</i> [3, 2, C]
Protein biosynthesis ⇒ protein synthesis initiation	<i>LSG1</i> [3, 2, C]
Protein complex assembly ⇒ cytochrome c oxidase biogenesis	<i>YKL053C-A</i> [3, 3, M]
Protein metabolism and modification ⇒ protein biosynthesis	<i>RNQ1</i> [9, 8, M] (16), <i>YGL069C*</i> [24, 44, M] (26), <i>YGL102C*</i> [4, 4, C] (27)
Protein-mitochondrial targeting ⇒ mitochondrial translocation	<i>RNQ1</i> [3, 2, M]
Ribosome biogenesis ⇒ ribosomal large subunit assembly	<i>YDR496C</i> [2, 1, M], <i>YML093W*</i> [2, 1, N]
Ribosome biogenesis ⇒ rRNA processing	<i>YDR496C</i> [12, 38, N]
RNA metabolism ⇒ RNA processing	<i>BCP1*</i> [8, 14, N], <i>TCI1</i> [2, 1, M], <i>TCI1</i> [2, 1, N], <i>YGR145W*</i> [5, 5, N]
rRNA processing ⇒ 35S primary transcript processing	<i>BCP1*</i> [5, 5, N], <i>PUF6</i> [2, 1, M], <i>YDR101C</i> [9, 24, N], <i>YHR085W*</i> [4, 5, N], <i>YJL010C*</i> [2, 1, N], <i>YML093W*</i> [2, 1, M]
Transcription ⇒ transcription, DNA-dependent	<i>RIO2*</i> [5, 6, N], <i>YNL114C*</i> [4, 3, N]
Transcription, DNA-dependent ⇒ transcription, from Pol I promoter	<i>SAS10*</i> [7, 6, N], <i>YDR101C</i> [13, 16, N], <i>YGR210C</i> [3, 2, N], <i>YMR310C</i> [3, 2, N], <i>YNL182C*</i> [8, 9, N]

The last two levels of the predicted GO processes are shown. Essential genes—i.e., genes with lethal null phenotypes—are marked with \*. For each prediction, we show in square brackets the number of support SPs, the number of unique support genes, and the graph in which the prediction is made. References to experimental supports are noted in parentheses. “C” denotes cytoplasm, “M” denotes mitochondria, and “N” denotes nucleus.

These are predictions for which we have higher levels of confidence, because these unknown genes are each functionally bounded by all of the other genes on the SP. Using the conservative rule, we made predictions for a total of 75 unique genes.

While some genes do not have GO biological process annotations in SGD, the biological processes that they are involved in are annotated in the “Cellular Role” category in the Yeast Proteome Database (YPD; [www.incyte.com/sequence/proteome/databases/YPD.shtml](http://www.incyte.com/sequence/proteome/databases/YPD.shtml)). We obtained the YPD cellular role annotations for 24 genes (Table 1) of the total 75 genes predicted with the conservative rule, and we used them as a positive internal control for our prediction. Among the 24 genes, the predictions for 16 matched the experimentally derived annotations in YPD; 3 matched computationally derived annotations in YPD. By “match” we mean that our prediction and the YPD annotation are either identical or closely coupled. In addition, we identified a case (*NOC3*) in which the YPD cellular role annotation did not correspond to its cited experimental reference, whereas our prediction matched the experimental conclusions perfectly. This amounts to a successful prediction ratio of 83%, again validating our SP method.

For example, *COR1*, *RIP1*, and *QCR8* are all predicted to be involved in “ATP synthesis coupled proton transport.” YPD annotates them as “energy generation” and “small molecule transport.” *ERB1* is predicted to be involved in “35S primary transcript processing.” In a recent study it was found to be essential for 35S primary transcript processing (11). The nuclear

trafficking protein GSP1/Ran was assigned with a set of diverse functions in both the cytoplasm and the nucleus graphs, including “protein biosynthesis,” “ribosome biogenesis,” and “G<sub>2</sub>/M transition of mitotic cell cycle.” Evidence in the literature suggests that it plays a central role in a large number of biological processes and is a master regulator of cell cycle and proliferation (12), which explains our putative functional assignments. The fact that we made predictions about GSP1/Ran from both nucleus and cytoplasm graphs agrees with the multiple subcellular localizations of this nuclear trafficking protein.

We predicted *NOC3* to be in the GO category “ribosome biogenesis” and “35S primary transcript processing.” While this did not agree with its YPD annotation “protein biosynthesis,” a close inspection of the YPD-cited reference (13) revealed that *NOC3* is in fact involved in the biogenesis of the 60S ribosomal subunit, which exactly matched our prediction.

The 51 novel genes predicted with the conservative rule for which no biological process annotation is available in either database are listed in Table 2. The additional 95 novel genes predicted by using the general rule that are unknown to both databases are listed on our web site ([www.biostat.harvard.edu/complab/SP/](http://www.biostat.harvard.edu/complab/SP/)). Together, we have assigned functions to around 5% of the unknown yeast ORFome. While some unknown genes have been assigned three-letter gene symbols, their cellular roles remain uncharacterized. For a significant portion of the genes listed we found partial experimental support, each to a different extent. We present some examples below.

One intriguing gene for which we made predictions is *RNQ1*. It is known to form the [PIN+] prion, and it possesses a glutamine- and asparagine-rich domain characteristic of other yeast prions (14). The function of *RNQ1*, however, is so far unknown. We predicted it to be involved in both “mitochondrial translocation” and “protein biosynthesis.” Interestingly, the prion responsible for the [PSI+] determinant, SUP35, not only is an important factor for protein translation termination but also is implicated in the system of cotranslational translocation into the mitochondria (15). In addition, a large-scale transcription profiling study found *RNQ1* to be clustered with genes involved in protein biosynthesis (16).

There are numerous other interesting cases. *BRX1* is predicted to be “ribosome biogenesis.” In fact, a very recent experimental study showed that it is involved in ribosomal large subunit assembly (17). YCR013C was predicted to be involved in “hexose metabolism.” There is evidence based on SAGE data (18) that supports our prediction. In total, for 11 of the 51 genes we have collected suggestive experimental evidences corroborating our predictions (Table 2). Most of the evidence comes from gene expression studies. To determine the biological function and verify our predictions of those genes, further experimental work is required.

## Discussion

Systematic approaches to the functional annotation of genes identified in the genome sequencing projects are urgently needed in the postgenomic era. The rapid increase in large-scale gene expression data provides us unique opportunities to meet this need. However, the development of methodology to discover novel gene functions by using expression data has so far been slow. With its 300 deletion and drug-treatment experiments, the Rosetta Compendium (6) is to date the most systematic expression profiling of the yeast genome published. In its original publication, by clustering approaches the authors identified and experimentally confirmed 8 novel gene functions. The Compendium has since remained mostly unexplored. In this paper we have proposed a systematic approach to predict gene functions based on such large-scale data. Applying it to the Rosetta Compendium, we assigned putative functions to 75 unknown yeast genes in SGD by using the conservative rule. We have shown that a positive internal control based on 24 genes gives a successful prediction ratio of 83%. Based on the general

prediction rule, we assigned functions to an additional 95 unknown yeast genes. Together these assignments make up 5% of the unknown yeast ORFome. While impressive, the Compendium contains only a tiny fraction of possible perturbations. We believe the method can achieve even better results with more comprehensive expression datasets.

The strength of our method is its utilization of transitive co-expression, an important feature among genes of the same biological process. To link such genes, we applied the SP algorithm. In contrast with traditional clustering approaches, ours is able to group not only functionally related genes with correlated expression profiles but also those without. Moreover, the expression dependence relationships between individual genes are given by the SPs. Using transitive co-expression, we are able to capture the functional relationships between genes beyond “synexpression” (6). While clustering is a useful approach to group genes with similar expression patterns, it is not sensitive to other types of expression relationships such as transitive co-expression. As we have seen, the SP method offers a complementary and informative tool for large-scale expression data analysis.

Another advantage of the SP method over traditional clustering methods is that it transparently and actively uses available biological knowledge as a guide to discover additional relevant genes. Clustering methods first group genes according to their expression profiles, then make inferences on the functions of genes within clusters. In contrast, our method starts with two genes with a known biological function, then uses them as a bound to identify intermediate genes related to them. Such active incorporation of biological annotation into the knowledge discovery process is one of the current challenges in microarray data analyses.

The SP method is scalable to larger graphs. For our prediction based on the nuclear graph with 3,914 vertices (genes) and 20,815 edges, determination of the SP from the 659 known genes as the source vertices took only 7 min on a 700-MHz Pentium III processor running Linux. Because for a given graph the computation needs to be done only once and is easily distributed over multiple processors, the SP method is applicable to the transcriptomes of higher eukaryotes such as the human and mouse.

The work of X.Z. and W.H.W. is supported by National Institutes of Health Grant 1R01HG02341. The work of M.-C.J.K. is supported by a Howard Hughes Medical Institute Predoctoral Fellowship.

- Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868.
- Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. (1999) *Nat. Genet.* **22**, 281–285.
- Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., et al. (1999) *Science* **286**, 531–537.
- Niehrs, C. & Pollet, N. (1999) *Nature (London)* **402**, 483–487.
- Qian, J., Dolled-Filhart, M., Lin, J., Yu, H. & Gerstein, M. (2001) *J. Mol. Biol.* **314**, 1053–1066.
- Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., et al. (2000) *Cell* **102**, 109–126.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000) *Nat. Genet.* **25**, 25–29.
- Dwight, S. S., Harris, M. A., Dolinski, K., Ball, C. A., Binkley, G., Christie, K. R., Fisk, D. G., Issel-Tarver, L., Schroeder, M., Sherlock, G., et al. (2002) *Nucleic Acids Res.* **30**, 69–72.
- Cormen, T. H. (2001) *Introduction to Algorithms* (MIT Press, Cambridge, MA).
- Hvidsten, T. R., Komorowski, J., Sandvik, A. K. & Laegreid, A. (2001) *Pac. Symp. Biocomput.*, 299–310.
- Pestov, D. G., Stockelman, M. G., Strezoska, Z. & Lau, L. F. (2001) *Nucleic Acids Res.* **29**, 3621–3630.
- Clarke, P. R. & Zhang, C. (2001) *Trends Cell Biol.* **11**, 366–371.
- Milkereit, P., Gadal, O., Podtelejnikov, A., Trumtel, S., Gas, N., Petfalski, E., Tollervy, D., Mann, M., Hurt, E. & Tschochner, H. (2001) *Cell* **105**, 499–509.
- Oshervovich, L. Z. & Weissman, J. S. (2001) *Cell* **106**, 183–194.
- Shumov, N. N., Volkov, K. V. & Mironova, L. N. (2000) *Genetika* **36**, 644–650.
- Jelinsky, S. A., Estep, P., Church, G. M. & Samson, L. D. (2000) *Mol. Cell Biol.* **20**, 8157–8167.
- Kaser, A., Bogengruber, E., Halleger, M., Doppler, E., Lepperdinger, G., Jantsch, M., Breitenbach, M. & Kreil, G. (2001) *Biol. Chem.* **382**, 1637–1647.
- Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., Jr., Hieter, P., Vogelstein, B. & Kinzler, K. W. (1997) *Cell* **88**, 243–251.
- Fujimori, F., Gunji, W., Kikuchi, J., Mogi, T., Ohashi, Y., Makino, T., Oyama, A., Okuhara, K., Uchida, T. & Murakami, Y. (2001) *Biochem. Biophys. Res. Commun.* **289**, 181–190.
- Bassler, J., Grandi, P., Gadal, O., Lessmann, T., Petfalski, E., Tollervy, D., Lechner, J. & Hurt, E. (2001) *Mol. Cell* **8**, 517–529.
- Saveanu, C., Fromont-Racine, M., Harington, A., Ricard, F., Namane, A. & Jacquier, A. (2001) *J. Biol. Chem.* **276**, 15861–15867.
- Mewes, H. W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S. & Weil, B. (2002) *Nucleic Acids Res.* **30**, 31–34.
- Causton, H. C., Ren, B., Koh, S. S., Harbison, C. T., Kanin, E., Jennings, E. G., Lee, T. I., True, H. L., Lander, E. S. & Young, R. A. (2001) *Mol. Biol. Cell* **12**, 323–337.
- Lee, J., Godon, C., Lagniel, G., Spector, D., Garin, J., Labarre, J. & Toledano, M. B. (1999) *J. Biol. Chem.* **274**, 16040–16046.
- Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. & Futcher, B. (1998) *Mol. Biol. Cell* **9**, 3273–3297.
- Traven, A., Wong, J. M., Xu, D., Sopta, M. & Ingles, C. J. (2001) *J. Biol. Chem.* **276**, 4020–4027.
- Rep, M., Krantz, M., Thevelein, J. M. & Hohmann, S. (2000) *J. Biol. Chem.* **275**, 8290–8300.