BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 18 2005, pages 3651–3657
doi:10.1093/bioinformatics/bti599

*Systems biology*

# Study of coordinative gene expression at the biological process level

Tianwei Yu, Wei Sun, Shinsheng Yuan and Ker-Chau Li*

Department of Statistics, University of California, Los Angeles, CA90095-1554, USA

Received on June 1, 2005; revised on July 21, 2005; accepted on July 24, 2005
Advance Access publication August 2, 2005

**ABSTRACT**

**Motivation:** Cellular processes are not isolated groups of events. Nevertheless, in most microarray analyses, they tend to be treated as standalone units. To shed light on how various parts of the interlocked biological processes are coordinated at the transcription level, there is a need to study the between-unit expressional relationship directly.

**Results:** We approach this issue by constructing an index of correlation function to convey the global pattern of coexpression between genes from one process and genes from the entire genome. Processes with similar signatures are then identified and projected to a process-to-process association graph. This top–down method allows for detailed gene-level analysis between linked processes to follow up. Using the cell-cycle gene-expression profiles for *Saccharomyces cerevisiae*, we report well-organized networks of biological processes that would be difficult to find otherwise. Using another dataset, we report a sharply different network structure featuring cellular responses under environmental stress.

**Contact:** kcli@stat.ucla.edu

**Supplementary information:** http://kiefer.stat.ucla.edu/lap2/download/KL_supplement.pdf

## INTRODUCTION

Microarray gene-expression profiling enables the assessment of transcript abundance at the full genome scale. A variety of methods has been proposed to process the microarray data for different purposes, such as annotating gene functions (Eisen *et al*., 1998; Zhou *et al*., 2002), finding transcription factor binding motifs (Conlon *et al*., 2003; Ihmels *et al*., 2002) and unraveling expression regulation circuitry (Li, 2002; Li *et al*., 2004; Qian *et al*., 2003; Qin *et al*., 2003; Spellman *et al*., 1998). Such studies are mainly at the gene-to-gene association level. Genes with similar expression patterns are thought to be more likely functionally associated. They may form structural complexes, participate in the same pathway or be regulated by a common mechanism.

Given that cellular processes are not isolated groups of events, the important issue of how various parts of the cellular system are coordinated needs to be addressed. Gene Ontology (GO) provides an excellent platform for dissecting the complex genetic circuitry into knowledge-based subunits (Ashburner *et al*., 2000). GO terms classify genes by the properties of their protein products. Because mRNA plays a critical role in regulating the linear flow of genetic information from DNA to protein, analysis of gene-expression data at the GO-term level can shed light on the cell's global management scheme in coordinating the uninterrupted supply of numerous protein products to meet the needs in various biological processes.

The term-to-term expressional relationship is more complex than the single gene-to-gene coexpression. As detailed in the Results section, the gene-to-gene correlations within most GO terms are not significant. Nevertheless, genes in each term have many strongly correlated genes from elsewhere of the genome. These correlated genes are not tightly correlated within themselves and they often come from diverse functional categories. Such results indicate that multiple intracellular and/or extracellular cues are utilized in regulating the mRNA sources.

We attempt to quantify the degree of expressional association between a pair of GO terms, A and B, by taking into account the aforementioned multiple cellular cues. Instead of considering only the coexpression pattern between the genes in A and B, we incorporate information from genes outside of the two GO terms. This idea is implemented by constructing a probability function, termed genome-wide index of correlation (GIOC) function, to convey the genome-wide coexpression pattern for genes in a GO term.

We first select a set of parallel terms, as outlined in Figure 1, from the gene ontology system (arrow a) to represent biological processes. Based on the gene-to-gene correlations obtained from microarray data (arrow b), we quantify the distance between every pair of terms (arrows c, d, e), using Kullback–Leibler (K–L) divergence between their GIOC functions. Following the measurement, a statistical hypothesis testing problem is formulated to obtain significant pairs of expressionally associated terms (arrow f). We convey the final results with a term-to-term association graph (arrow g). In this way, biologists can describe the global pattern of expressional association at the biological process level before going into the more detailed gene-to-gene level analysis. This strategy of portraying genetic circuitry at the GO-term level and the gene level, is in line with the increasingly more popular call for multiscale research when studying complex problems.

## METHODS

### Genome-wide index of correlation

For each GO term $H$, we create a probability function to serve as its GIOC. Denote the collection of all yeast genes present in the gene-expression dataset as $G$. For each gene profile $x_i$ in $G$, we first evaluate its correlation with every gene profile $y_j$ in $H$. The highest correlation, $c_i = \max_j \text{corr}(x_i, y_j)$, where the maximum is taken over all genes in $H$, indicates the level of interaction

*To whom correspondence should be addressed.

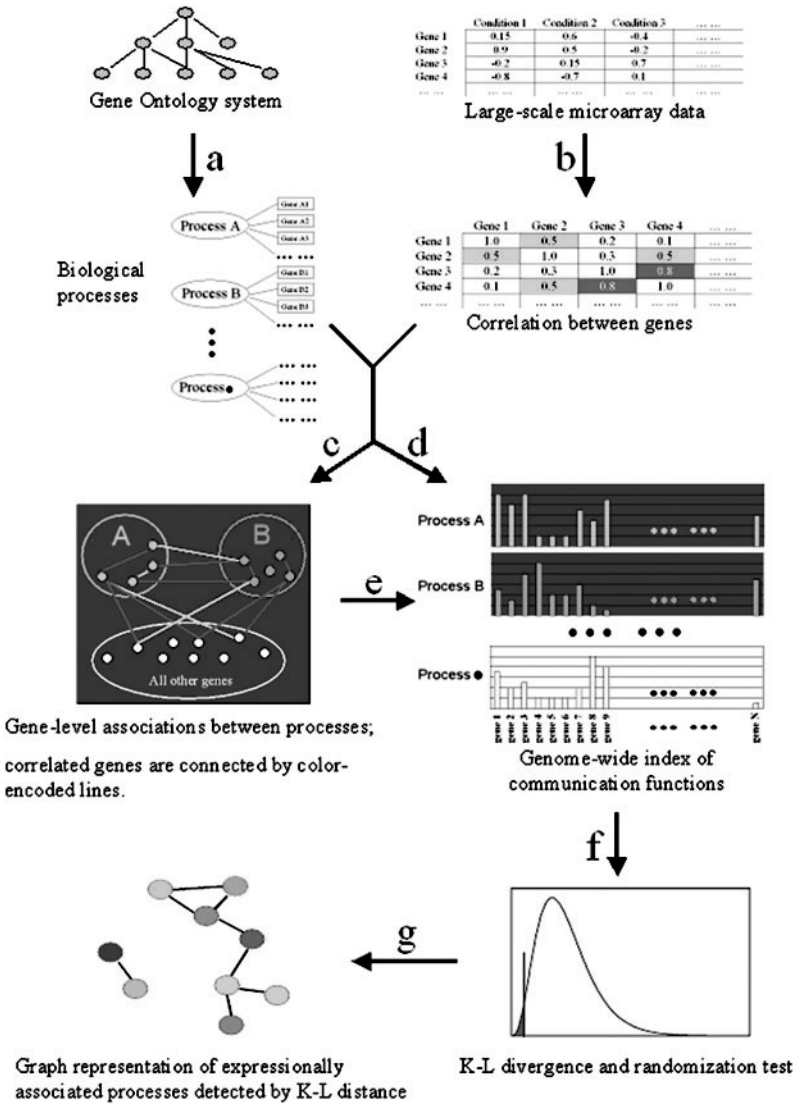


**Fig. 1.** Strategy of the study. Arrow **a**: select biological processes from the gene ontology system using a scheme described in Supplementary Figure 7. Arrow **b**: compute correlations from large scale microarray data. Arrow **c**: find gene level linkages between processes (this step may be skipped). Arrow **d**: GIOC functions are established for each process. Arrow **e**: use similarity between GIOC functions to measure the degree of expressional association between processes. Arrow **f**: determine the significance of process association by randomization test. Arrow **g**: connect associated processes and project the results as a graph.

between gene $i$ and term $H$. Using the clustering analysis terminology, this corresponds to the single linkage distance measure between $x_i$ and all genes in term $H$. We then convert $c_i$ into an index of correlation by a power function transformation. More specifically, we assign each gene $i$ in $G$ a probability mass $p_i \propto (1+c_i)^6$. Here the proportionality can be determined by setting the total probability mass equal to 1. The resulting probability function $P_H(x_i) = p_i, i = 1, \ldots, n$, is called the GIOC function for term $H$.

## GO term expressional association measure

The degree of expressional association between two GO terms $H_1$ and $H_2$ is determined by how similar their GIOC functions are. We use K–L divergence between probability measures to quantify the distance:

$$\mathrm{KL}(H_1, H_2) = \sum_{i=1}^{n} P_{H_1}(x_i) \log_2 \left[ \frac{P_{H_1}(x_i)}{P_{H_2}(x_i)} \right].$$

K–L divergence is not symmetric. A symmetrized version is to use the average $[\mathrm{KL}(H_1, H_2) + \mathrm{KL}(H_2, H_1)]/2$.

The next step is to determine if two GO terms can be called expressionally associated or not. This is achieved through a randomization test of significance to see if the observed K–L distance is shorter than what would be expected when two terms are not associated.

## Randomization test of significance

We first specify the null hypothesis. Suppose there are $n$ genes in term $H_1$, and $m$ genes in term $H_2$. To incorporate the case that there may be genes that are annotated to both terms, we further assume that there are $r$ overlap genes. Under the null hypothesis of no association between two terms, the $m+n-r$ gene-expression profiles for these two terms should behave as if they were randomly drawn from the entire gene-expression database. To find the null distribution of the K–L distance, we use the Monte Carlo method. We first

draw $n + m - r$ profiles randomly from the collection of all gene profiles. We use the first $n$ of them to form one term and the last $m$ of them to form the second term. This naturally leads to $r$ overlaps between the two terms. We then compute the K–L distance between these two artificially created terms. This procedure is iterated many times to yield an approximation of the distribution of K–L distance. Once the null distribution is available, we can call a pair of GO terms significantly associated if their K–L distance is shorter than a cutoff percentile.

## RESULTS

### Selecting GO terms to represent biological processes

We use the 'biological process' ontology for *Saccharomyces cerevisiae*. The GO system forms a directed acyclic graph. In this report, we restrict the study to a representative set of GO terms that do not have ancestor–descendent relationships. This is because the analysis of a full size GO, which contains both ancestor–descendent and sibling relationships, involves too much complexity and redundancy to yield easily interpretable results.

Selecting representative terms is not a simple issue to address. Instead of relying solely on expert opinions, we use computer search to gain objectiveness. Our program traverses the entire 'biological process' branch of GO from top to bottom (Supplementary Figure 5). A couple of parameters are optimized to reach the dual aim of choosing terms as close to the bottom level as possible, and covering as many genes as possible. The result is a collection $C$ of 214 parallel terms. This representative list is at a scale finer than 'GO slims' (Ashburner *et al*., 2000; Dwight *et al*., 2002). The distribution of the number of genes in the selected terms is shown in Supplementary Figure 6.

### Within-GO term and between-GO term correlation structures

In order to find a proper measure of the expression association between two GO terms, we first study how gene-expression profiles within a GO term are correlated. We created an on-line GO term computation page (a module in http://kiefer.stat.ucla.edu/lap2) to facilitate the investigation. Given a pair of terms $X$ and $Y$, the system computes gene-level correlations within each term and between the two terms. Subject to a user-specified size limit, the system also searches the entire genome for two lists of highest co-expressed genes, one for each term. These two lists are then linked to the GO Term Finder of SGD to identify enriched functional groups.

Our preliminary study shows that not all genes from the same GO term are tightly coexpressed. To the contrary, the correlations within the majority of the terms we investigate are low (Supplementary Figure 7); e.g. the range is between $-0.50$ and 0.47 for 'actin cortical patch assembly' (14 genes), between $-0.59$ and 0.80 (median 0.03) for 'axial budding' (21 genes), and between $-0.18$ and 0.43 (median 0.19) for 'NAD biosynthesis' (6 genes). The correlations are much higher for terms involving translation mechanism, e.g. from $-0.16$ to 0.85 (median 0.53) for 'ribosomal large subunit biogenesis' (14 genes).

Our preliminary study also suggests that yeast uses multiple intracellular or extracellular cues in regulating the resources devoted to a functional module. Despite the low average correlation within a GO term, each term has many strongly correlated genes from elsewhere of the genome; but these genes are not highly correlated within themselves, and their cellular roles are diverse. For instance, when we submit the top 200 genes which have the best correlations (all $>0.57$) with 'NAD biosynthesis' to GO Term Finder, no more than one-quarter of them fall into functionally enriched groups, the most visible ones being 'catabolism' (27 genes), 'protein folding' (10 genes) and 'regulation of protein metabolism' (5 genes).

These preliminary findings argue for the merit of considering GIOC function. Our aim is to find a higher order organization among a diverse list of biological processes. Therefore, in quantifying the degree of expressional association between a pair of GO terms, we should not isolate the genes in the term pair from the rest of the genome. On the contrary, the information from genes outside of the two GO terms must be integrated first.

### Expressional association in cell cycle

Using the cell-cycle dataset (Spellman *et al*., 1998), we compute the GIOC function for each term. Furthermore, based on the K–L distance and $P$-values from randomization tests, we find a total of 202 GO-term associations significant at level 0.025. The output is displayed as a graph by Cytoscape (Shannon *et al*., 2003) (Fig. 2). Four large sections are visible topologically; together they show a very clear higher order functional organization between GO terms that would be hard to detect using standard bottom-up analyses.

Component A features cell-cycle mechanisms, which have often been discussed in the literature for this dataset. Component B exhibits a coherent operation within the translation mechanism, which is also well-anticipated. Component C features the protein transport mechanism. This component is further enriched by two actin-related terms: 'actin cortical patch assembly' and 'actin polymerization and/or depolymerization' (14 and 7 genes, respectively; no overlap), which agrees with the role actin plays in the cell (Palmgren *et al*., 2002). Component D shows an extensively connected network of metabolic processes including four major categories: coenzyme metabolism, amino acid/lipid metabolism, small molecule transport/homeostasis and polysaccharide metabolism/energy generation. We further discuss the importance of coenzyme terms and calcium homeostasis in Supplementary information Text 1.

### Expressional association and other distance measures of GO

Terms with short GO-graph distances tend to have shorter K–L distances (Fig. 3a), and higher chance of being connected (Fig. 3b). However, the trend is weak. Terms located far apart according to the node-to-node distance induced by the graph of the GO system can still have strong expressional association. For example, the term 'translation initiation' is located 10 steps away from 'ribosomal large subunit assembly and maintenance' (35 and 34 genes, respectively; no overlap). Biologically, they have tight relationship and we do observe strong expressional association between them. Although the GO graph distance is simple to compute, the implicit equal path weight for the entire GO graph may not be appropriate. Upon the advice of a reviewer for this paper, we provide further discussion using a modified version of Lord *et al*.'s semantic similarity measure (2003) in Supplementary information Text 2.

### Environmental stress data

One major purpose of our method is to allow the comparison of cellular regulation strategy under different conditions. For a comparison, we apply the same method to the environmental stress dataset (Gasch *et al*., 2000). The significant term-to-term associations are

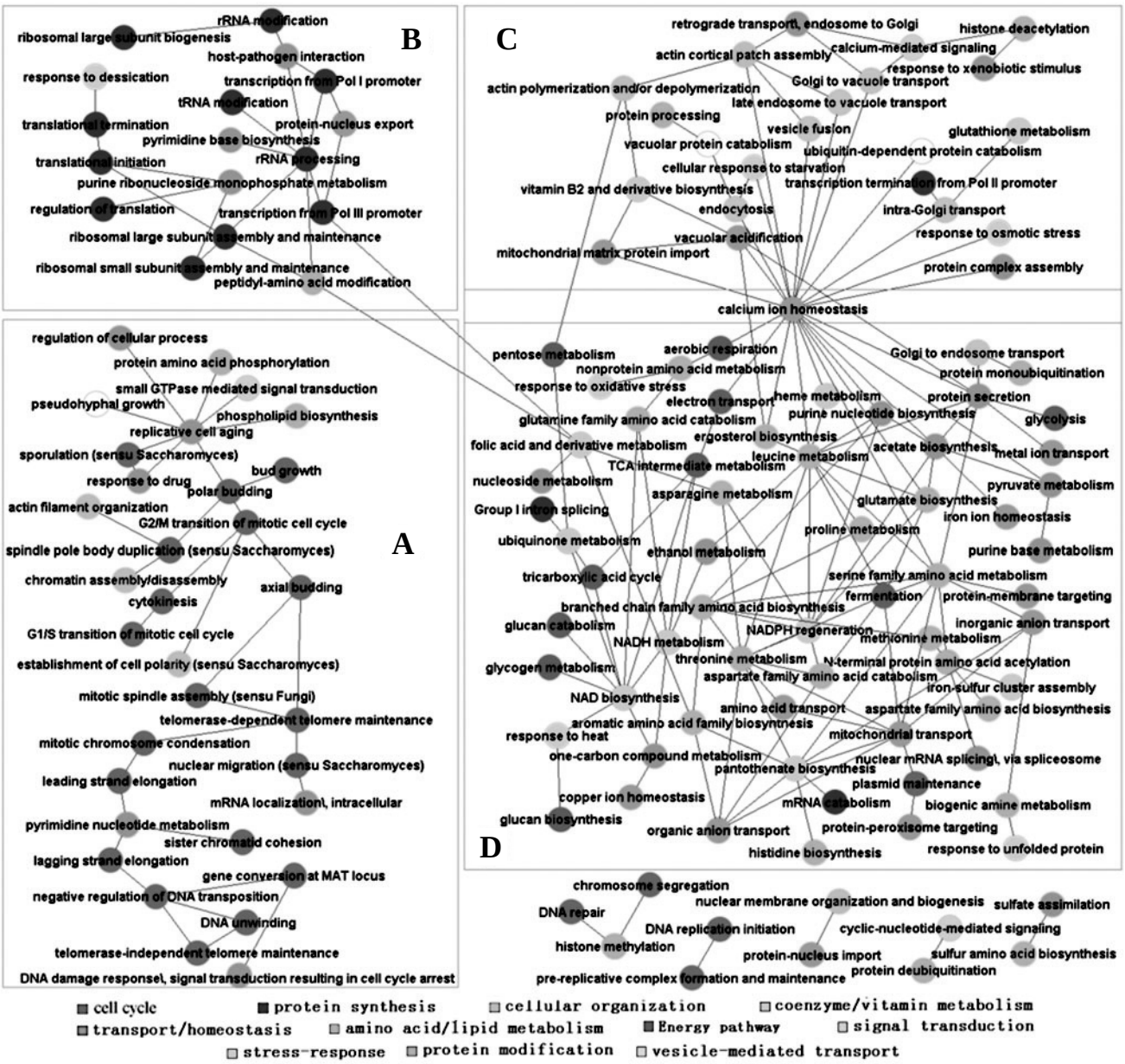


**Fig. 2.** Term-to-term expressional association based on cell-cycle data. This figure gives the final graph output following steps outlined in Figure 1. A representative list of 214 GO terms were used. Four sections, **A**, **B**, **C** and **D**, are easily identified.

displayed in Figure 4. In total, 91 connections are found. Comparison with Figure 2, shows that the two graphs have only a small portion of overlap. Of the 12 preserved connections (Supplementary Table 8), 6 are ribosome related. This agrees with the notion that cellular protein synthesis machinery is constantly under tight control irrespective of the growth conditions (Gasch *et al*., 2000). Connections in the component A of Figure 2, which feature the cell-cycle mechanism, are no longer present in Figure 4. In the stress–response experiments, mRNA samples were collected from unsynchronized cell cultures, which make it difficult to distill cell-cycle information from the data. For further discussion about the network, see Supplementary information Text 1.

## DISCUSSION

GO term relationships have been studied using protein interaction (Giot *et al*., 2003) and genetic interaction data (Tong *et al*., 2003). For microarray data, differentially expressed genes can be mapped to GO or other knowledge sources to identify enriched functional group (Berriz *et al*., 2003; Cheng *et al*., 2004; Draghici *et al*., 2003; Dwight *et al*., 2002; Mootha *et al*., 2003; Robinson *et al*., 2002). Alternatively, some authors argue that whether a group is enriched or not should be determined based on the average expression for all genes within the group (Pavlidis *et al*., 2004; Smid and Dorssers, 2004; Volinia *et al*., 2004).

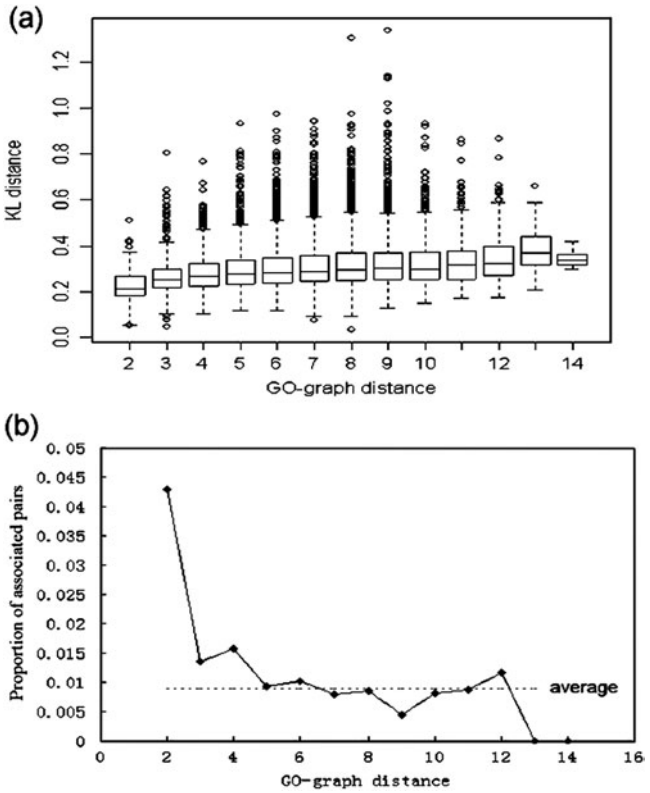


**Fig. 3.** GO-graph distance and expressional association. (**a**) Boxplots showing the relationship between GO-graph distances and K–L distances. (**b**) Proportion of expressionally associated pairs versus GO-graph distance. The GO-graph distance between two terms is the length of the shortest path between them, considering all edges as bi-directional. The K–L distances were computed from cell-cycle data.

Our method is different from those used in comparative gene-expression studies. It systematically finds biological processes that are more tightly coordinated under the cell's mRNA allocation program. At the core of our approach is the creation of a GIOC function that conveys the extent of correlation between genes annotated to a GO term and genes in the entire genome.

## Power transformation in constructing the GIOC function

We chose to use power 6 for the transformation in constructing the GIOC function. A similar idea has been used to give weight to distances (Zhou *et al*., 2002). The purpose is to assign more weights to higher correlations without enforcing a hard threshold. For example, using a power of 6, a correlation of 0.5 will receive $\sim$20% of the weight that the perfect correlation of 1 will receive, whereas a correlation of 0.8 will increase the weight to 50%. Our method is insensitive to the choice of power. Supplementary Table 9 shows the correlation between term-to-term distances induced by different powers. Even if a power of 3 or 9 is used, we still find the correlation exceeding 0.96 across the board.

## Gene sharing

Gene sharing between GO terms is a positive factor in shortening the K–L distance between their GIOC functions. This factor is automatically adjusted for by the randomization test. For the same term sizes, the reference curve is shifted to the left as the number of overlapping genes increases (Supplementary Figure 10). Thus, if two terms have more overlapping genes, a shorter distance has to be observed in order to establish a significant connection. Consequently, we find no systematic selection bias toward pairs of terms with more overlaps in our results (Supplementary Table 11).

## From term level association to gene level association

Our global sketch of the term-to-term relationship sets tone for conducting more elaborate gene-level investigations. Using the on-line system described in the preliminary study section, we can compute the gene-level correlations for any linked GO term pairs. The system also searches the entire genome for two lists of highest coexpressed genes, one for each term. The coexpressed genes shared between the two lists are likely to be the source of intracellular cues that bridge the connection between the two terms. As proven by Zhou *et al*. (2002), 'transitive expression similarity' is an important attribute for discovering more functionally associated genes.

We find two possible scenarios for a pair of terms to be linked by our expressional measure: (1) by tight coexpression between their genes directly; (2) by their shared co-expressed genes elsewhere in the genome. Ribosome and translation related genes are known to be under tight cellular control. As expected, both the within-term and the between-term correlations in component B of Figure 2 are high. In contrast, we find both the within-term and the between-term correlations in component D are much lower (Supplementary Figure 12). This indicates that multiple intracellular cues have been utilized to ensure the proper flow of metabolites across a variety of metabolic processes.

As an example of the first scenario, in Supplementary Figure 13a, the expression profiles for genes in the pair 'rRNA modification' and 'ribosomal large subunit biogenesis' (both 14 genes; no overlap) are compared by hierarchical clustering. Many cross-term neighbors are observed.

As an example of the second scenario, we revisit the term 'NAD biosynthesis' in component D of Figure 2. As one of the key coenzymes involved in multiple metabolic pathways, the level of NAD and NAD/NADH ratio is crucial for maintaining well-regulated metabolism. Reflecting this important physiological relationship, our method finds a direct link between 'NAD biosynthesis' and 'NADH metabolism' (6 and 7 genes respectively; no overlap). In order to identify the source of the link, we find the coexpressed genes for each term. There are 463 genes that have correlations of $>0.5$ with 'NAD biosynthesis', and 363 genes with 'NADH metabolism'. The two groups share 117 genes. These 117 genes serve as the bridges that link the two terms. However, there are only two cross-term correlations $>0.5$. We note that the two terms share an ancestor 'nicotinamide metabolism'. Among the 13 genes that are annotated to this ancestor but not to the two NAD terms, 11 are in the descendent term 'NADPH regeneration' (no overlap with the two NAD terms). However, 'NADPH regeneration' is connected to neither of the two terms, and none of its 11 genes serve as a bridge for the two terms.

Another example is the pair 'NAD biosynthesis' and 'tricarboxylic acid cycle' (6 and 14 genes respectively; no overlap). It is well-known that multiple steps in the TCA cycle require NAD (Alberts *et al*., 2002) and our method does find the link between these two terms. There are 463 genes that have correlations of $>0.5$ with 'NAD

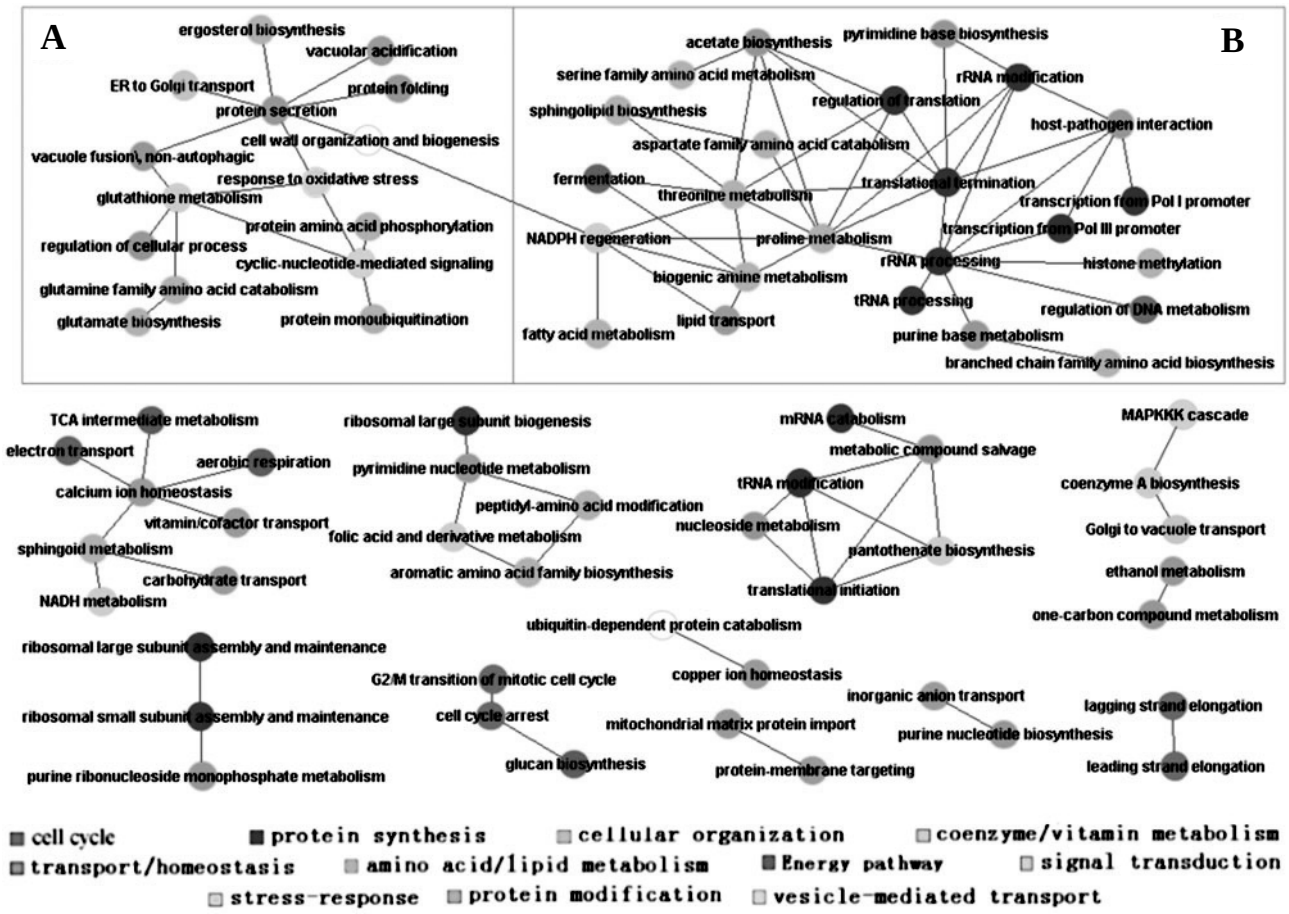


**Fig. 4.** Term-to-term expressional association based on environmental stress gene expression data. This figure is generated in the same way as in Figure 2. Less connections are found. Section A features yeast's characteristic responses under stress. Section B features a cluster of ribosome/protein synthesis terms, together with a group of closely related metabolic terms.

biosynthesis', and 566 genes with 'tricarboxylic acid cycle'. The two groups share 207 genes. However, there is only one cross-term correlation $>0.5$. Supplementary Figure 13b and c show how the clustering patterns in these two examples are different from what is seen in Supplementary Figure 13a.

## Network display and inference

We view our method mainly as an exploratory tool. Although we have presented our results as networks of biological processes, the method itself can serve the general purpose of comparing the behavior of two gene categories in microarray data. Unlike correlation coefficient which is bounded between $-1$ and 1, K–L distance is not upper-bounded. We make its interpretation easier by comparing with a reference distribution. The reported $P$-value for a pair of terms serves as a Monte Carlo estimate of the relative standing of their K–L distance as compared to all possible K–L distances. A small $P$-value indicates a short distance.

The $P$-value cutoff of 0.025 used in this study is only suggestive. In fact, as one referee points out, there are other ways of presenting the results. In Supplementary information Text 3, we present the results using heat-map and single-linkage hierarchical clustering trees. Areas in the heat map that correspond to the components in Figures 2 and 4 are labeled. We further collect all term associations with $P$-value $<0.1$ in our website http://kiefer.stat.ucla.edu/lap2.

We did not adjust $P$-value for multiple testing. Statistical inference on network configuration is a very complex issue and it requires balanced attention on both the false positive rate and false negative rate. The classical adjustment by the Bonferroni method is too conservative. More recently there has been a growing interest on false discovery rate (FDR) (Reiner *et al.*, 2003). However, in view of the complex dependence in the network structure, the independence or the special dependence requirement on the testing statistics (Benjamini and Yekutieli, 2001) makes such FDR procedures inadequate for our problem.

In order to elaborate more on the network structure issue, consider the typical question about the number of false positives. With a fixed unadjusted $P$-value cutoff at 0.025, the standard answer is $0.025M$, where $M$ is the number of true null hypotheses. However, this straightforward answer can be misleading because it does not take into account the network configuration. For instance, consider the case that $M = 3$. If the three true null hypotheses are (1) term A and term B randomly assembled; (2) term B and term C randomly assembled; (3) term C and term D randomly assembled, then implicitly all four terms A, B, C, D are randomly assembled. Thus

by accounting for the three additional implicated hypotheses, the answer should be $0.025 \times 6$ instead of $0.025 \times 3$. In real application, $M$ is unknown. If we were to assume $M$ equals the upper bound, $\binom{214}{2}$, we should expect 570 false-positives—a number far greater than 202, the number of significant connections found in the cell-cycle data. This shows that we need a better proposal to estimate $M$. Unfortunately, the proposals in the literature generally require independence assumption which is clearly violated in our setting.

Up to now, our discussion reflects only the statistical point of view. It is possible to approach the FDR issue by incorporating some biological perspectives. A rough estimate of FDR can be obtained by considering two distant categories of GO terms that are less likely to be expressionally associated. Denote the number of GO terms that fall under the two categories by $a$ and $b$, respectively. There are a total of $ab$ possible cross-category connections. Suppose among them our method finds $x$ significant connections. Denote the number of true connections by $c (c < x)$. Then the false positive rate can be estimated by $(x - c)/(ab - c) \leq x/ab$. For example, if we take one category to be cell-cycle/reproduction/DNA metabolism related terms ($a = 38$) and the other to be small molecule metabolism/transport terms ($b = 112$), then the estimated false positive rate is bounded by $1.7 \times 10^3$ ($x = 5$). This translates into 27 false positives among our 202 reported connections.

### Other technical issues

In Supplementary information Text 4, we show that term-level connection can be mediated by multiple transcription factors. In Supplementary information Text 5, we conduct simulations to evaluate how well our method can tolerate aberrations in microarray measurement.

## ACKNOWLEDGEMENTS

We thank Dr Feng Qiao for discussions about the biological aspects of this work, and Rana Bahhady for correcting grammar and spelling. We would like to thank the two anonymous referees, whose constructive criticisms led to additional work that strengthened the results. The term ‘genome-wide index of correlation’ was suggested by one referee to replace our original wording ‘genome-wide index of communication’. This work is supported by NSF grants DMS-0201005, DMS-0104038 and DMS-0406091.

*Conflict of Interest:* none declared.

## REFERENCES

Alberts,B., Johnson,A., Lewis,J., Raff,M., Roberts,T. and Walter,P. (2002) *Mol. Biol. Cell*, 4th edn. Garland Publishing, NY.

Ashburner,M. *et al*. (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, **29**, 1165–1188.

Berriz,G.F. *et al*. (2003) Characterizing gene sets with FuncAssociate. *Bioinformatics*, **19**, 2502–2504.

Cheng,J. *et al*. (2004) NetAffx Gene Ontology Mining Tool: a visual approach for microarray data analysis. *Bioinformatics*, **20**, 1462–1463.

Conlon,E.M. *et al*. (2003) Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl Acad. Sci. USA*, **100**, 3339–3344.

Draghici,S. *et al*. (2003) Global functional profiling of gene expression. *Genomics*, **81**, 98–104.

Dwight,S.S. *et al*. (2002) Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.*, **30**, 69–72.

Eisen,M.B. *et al*. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

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

Giot,L. *et al*. (2003) A protein interaction map of *Drosophila melanogaster*. *Science*, **302**, 1727–1736.

Ihmels,J. *et al*. (2002) Revealing modular organization in the yeast transcriptional network. *Nat. Genet.*, **31**, 370–377.

Li,K.C. (2002) Genome-wide coexpression dynamics: theory and application. *Proc. Natl Acad. Sci. USA*, **99**, 16875–16880.

Li,K.C. *et al*. (2004) A system for enhancing genome-wide coexpression dynamics study. *Proc. Natl Acad. Sci. USA*, **101**, 15561–15566.

Lord,P.W. *et al*. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275–1283.

Mootha,V.K. *et al*. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**, 267–273.

Palmgren,S. *et al*. (2002) Twinfilin, a molecular mailman for actin monomers. *J. Cell Sci.*, **115**, 881–886.

Pavlidis,P. *et al*. (2004) Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochem. Res.*, **29**, 1213–1222.

Qian,J. *et al*. (2003) Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics*, **19**, 1917–1926.

Qin,Z.S. *et al*. (2003) Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. *Nat. Biotechnol.*, **21**, 435–439.

Reiner,A. *et al*. (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**, 368–375.

Robinson,M.D. *et al*. (2002) FunSpec: a web-based cluster interpreter for yeast. *BMC Bioinformatics*, **3**, 35.

Shannon,P. *et al*. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Smid,M. and Dorssers,L.C. (2004) GO-Mapper: functional analysis of gene expression data using the expression level as a score to evaluate Gene Ontology terms. *Bioinformatics*, **20**, 2618–2625.

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

Tong,A.H. *et al*. (2003) Global mapping of the yeast genetic interaction network. *Science*, **303**, 808–813.

Volinia,S. *et al*. (2004) GOAL: automated Gene Ontology analysis of expression profiles. *Nucleic Acids Res.*, **32**, 492–499.

Zhou,X. *et al*. (2002) Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl Acad. Sci. USA*, **99**, 12783–12788.