OMICS A Journal of Integrative Biology
Volume 8, Number 2, 2004

Integrated Analysis of Microarray Data and Gene Function Information

YAN CUI,[1,2] MI ZHOU,[1,2] and WING HUNG WONG[3,4,5]

ABSTRACT

Microarray data should be interpreted in the context of existing biological knowledge. Here we present integrated analysis of microarray data and gene function classification data using homogeneity analysis. Homogeneity analysis is a graphical multivariate statistical method for analyzing categorical data. It converts categorical data into graphical display. By simultaneously quantifying the microarry-derived gene groups and gene function categories, it captures the complex relations between biological information derived from microarray data and the existing knowledge about the gene function. Thus, homogeneity analysis provides a mathematical framework for integrating the analysis of microarray data and the existing biological knowledge.

INTRODUCTION

MICROARRAY HAS BECOME a powerful tool for biomedical research. It detects the expression levels of thousands of genes simultaneously. Huge amount of genome-wide gene expression data have been generated using microarrays. However, microarray data by themselves tell us very little about the underlying biological processes. On the other hand, a lot of biological knowledge have been obtained by conventional biochemical or genetic methods and have been stored in public databases, such as MIPS Functional Classification Catalogue (Mewes et al., 2002), KEGG pathway database (Kanehisa et al., 2002) and Gene Ontology (The Gene Ontology Consortium, 2000). These functional classification systems represent well-organized knowledge about gene functions. In this paper, we use homogeneity analysis to integrate the analysis of microarray data and existing knowledge about gene function. Homogeneity analysis is a graphical multivariate method. It reveals the complex relations between microarray-derived gene groups and gene functional categories, and provides a global view of patterns of the correlations between gene groups derived from multiple types of data. It may help investigators to gain insights into the biological processes underlying microarray data by systematically connecting new data to existing biological knowledge.

Homogeneity analysis is mathematically equivalent to Multiple Correspondence Analysis under some conditions* (Michailidis and de Leeuw, 1998; Greenacre and Hastie, 1987), which is not satisfied in the in-

[1]Department of Molecular Sciences, University of Tennessee Health Science Center, Memphis, Tennessee.
[2]Center of Genomics and Bioinformatics, University of Tennessee Health Science Center, Memphis, Tennessee.
[3]Department of Statistics, Harvard University, Cambridge, Massachusetts.
[4]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts.
[5]Dana-Farber Cancer Institute, Boston, Massachusetts.

*Homogeneity Analysis is equivalent to Multiple Correspondence Analysis if all the row margins of the indicator table are equal.

tegrated analysis of microarray data and gene function information. Simple correspondence analysis (Benzecri, 1973; de Leeuw and van Rijckevorsel, 1980; Greenacre, 1993) has been applied to microarray data to analyze the associations between genes and samples (Waddell and Kishino, 2000; Kishino and Waddell, 2000; Fellenberg et al., 2001). The previous works focus only on microarray data. Gene function information and other biological knowledge have not been integrated into the analysis. Homogeneity analysis is a more general and flexible framework that can accommodate multiple types of data and utilize them in an integrated analysis. It allows us to analyze and visualize microarray data and gene function information simultaneously. This work is a new attempt to integrate the analysis of microarray data and existing biological knowledge in a single mathematical framework.

MATERIALS AND METHODS

Indicator table: unified coding of the microarray-derived gene groups and gene function categories

Microarrays are often used for identifying genes that are differentially expressed among different conditions. The groups of genes that are up-regulated or down-regulated in the testing sample (relative to the reference sample) can be selected. Thus, for each experimental condition, we can create two categories—one contains genes that are up-regulated under the condition and the other contains genes that are down-regulated under the condition.

Many computational methods have been developed for analyzing microarray data. Sophisticated analysis of large microarray dataset often results in overlapping gene groups such as transcriptional clusters (Wu et al., 2002; Lazzeroni and Owen, 2002; Lee and Batzoglou, 2003), biclique (Tanay et al., 2002), transcriptional modules (Ihmels et al., 2002; Segal et al., 2003), and genetic modules (Stuart et al., 2003). These gene groups are also microarray-derived categorical data.

Gene function classification systems assign genes to function categories. Gene classification data is also categorical data. We use an indicator table to code the different types of categorical data (Table 1). Each row contains the information of a gene—its membership to the gene groups and the function categories. Only 1 and 0 can occur in the indicator table. A "1" means a gene belongs to the corresponding category while a "0" means it does not.

Homogeneity analysis

Homogeneity Analysis is a graphical multivariate method for analyzing categorical data. It has been used to display the main structures and regularities of complex data sets (de Leeuw and van Rijckevorsel, 1980; de Leeuw, 1984; Michailidis and de Leeuw, 1998). Points in p-dimensional space (p is the number of dimensions) are used to represent categories and genes. Let X be the $N \times p$ matrix containing the coordinates of the N genes, and Y the $M \times p$ matrix containing the coordinates of the M categories, a loss function is defined as:

$$\sigma(X;Y) = \sum_{i=1}^{N} \sum_{j=1}^{M} \left[G_{ij} \sum_{k=1}^{p} (X_{ik} - Y_{jk})^2 \right] \tag{1}$$

where G is indicator table. If edges are used to connect each category and the genes belonging to that category, the loss function is the total squared length of the edges. We used an alternating least squares (ALS) algorithm (Michailidis and de Leeuw, 1998) to minimize the loss function. The minimization is subject to two restrictions:

$$X'X = N\, I_p \tag{2}$$

$$u'X = 0 \tag{3}$$

where u is the vector of ones. The first restriction is for avoiding the trivial solution corresponding to $X = 0$ and $Y = 0$. The second one requires the points to be centered around the origin.

Table 1. Indicator Tables

A

	Sample1.up	*Sample1.down*	*Sample2.up*	*Sample2.down*	. . .	*Function1*	*Function 2*	. . .
Gene1	1	0	0	1	. . .	0	0	. . .
Gene2	0	1	0	0	. . .	0	0	. . .
Gene3	0	0	0	1	. . .	1	0	. . .
Gene4	1	0	1	0	. . .	0	1	. . .
Gene5	1	0	0	0	. . .	1	0	. . .
Gene6	1	0	1	0	. . .	0	1	. . .
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

B

	Module1	*Module2*	*Module3*	*Module4*	. . .	*Function1*	*Function2*	. . .
Gene1	1	1	0	1	. . .	1	1	. . .
Gene2	0	1	1	0	. . .	0	1	. . .
Gene3	0	0	0	1	. . .	1	0	. . .
Gene4	0	1	1	0	. . .	0	1	. . .
Gene5	1	0	0	0	. . .	1	0	. . .
Gene6	1	0	1	1	. . .	0	0	. . .
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

"SampleX.up" represents the group of genes that are up-regulated in sample X (comparing to the reference sample); "SampleX.down" denotes the groups of genes that are down regulated in sample X; "FunctionX" denotes gene function categories; ModuleX is the Xth transcriptional module. A "1" means a gene belongs to the corresponding category, while a "0" means it does not.

The ALS algorithm iterates the following steps until it converges. First, the loss function is minimized with respect to Y for fixed X. The normal equation is

$$CY = G'X \tag{4}$$

where G' is the transpose matrix of G, C is the diagonal matrix containing the column sums of G. The solution of Eq. 4 is

$$\hat{Y} = C^{-1}G'X \tag{5}$$

Second, the loss function is minimized with respect to X for fixed Y. The normal equation is

$$RX = GY \tag{6}$$

where R is the diagonal matrix containing the row sums of G. Therefore, we get that

$$\hat{X} = R^{-1}GY \tag{7}$$

Third, the coordinates of the genes are centered and orthonormalized by the modified Gram-Schmidt procedure (Golub and van Loan, 1989),

$$X = \sqrt{N}GRAM(W) \tag{8}$$

where

$$W = \hat{X} - u(u'\hat{X}/N) \tag{9}$$

This solution is called the HOMALS (homogeneity analysis by means of alternating least squares) solution. Here we list some basic properties of the HOMALS solution, which are useful for inter-

preting of the result of homogeneity analysis (Greenacre and Hastie, 1987; Michailidis and de Leeuw, 1998):

1. Category points and gene points are represented in a joint space.
2. A category point is the centroid of genes belonging to that category.
3. Genes with the same response pattern (i.e., identical rows in the indicator table) receive identical positions. In general, the distance between two genes points is related to the "similarity" of their profiles.
4. Genes with a "unique" profile will be located further away from the origin, whereas genes with a profile similar to the "average" one will be located closer to the origin.

RESULTS AND DISCUSSION

In this section, we will use two microarray datasets and two gene function classification systems to illustrate the applications of our method.

Rosetta Compendium dataset

We applied homogeneity analysis to the yeast gene expression data from the Rosetta Compendium (Hughes et al., 2000a), which includes 300 mutations and chemical treatment experiments. We excluded the mutant strains that are aneuploid for chromosomes or chromosomal segments because the aneuploidy often leads to chromosome-wide expression biases (Hughes et al., 2000b). The data was filtered to include only experiments with 20–100 genes up- or down-regulated greater than twofold, and significant at $p \leq 0.01$ (according to the error model described in Hughes et al., 2000a); and only genes that are up- or down-regulated at greater than twofold, and at $p \leq 0.01$, in two or more selected experiments. The filtered dataset includes 494 genes and 48 experiments.

Two groups of genes were selected from each experiment: (1) genes that are up-regulated at greater than twofold, and at $p \leq 0.01$; (2) genes that are down-regulated at greater than twofold, and at $p \leq 0.01$. The microarray-derived gene groups are encoded using an indicator table. Each experiment has two categories (up- and down-regulation). The selected genes are represented by "1"s in the indicator table. The categories (columns) with less than two "1"s and genes (rows) with less than two "1"s were deleted. Now we have 416 genes and 46 categories. We call these categories "expression categories." Seventeen MIPS functional categories (Fig. 1) were added to the indicator table. The indicator table contains 416 genes and 63 categories. We performed homogeneity analysis based on the indicator table. The result is shown in Figure 1. The red (green) category points represent the groups of genes that are up–(down)–regulated in the corresponding experiments and the blue points represent functional categories. A category point is located at the centroid of the genes that belong to it. The small gray points represent genes, each of them may represent one gene or a group of genes with same "response pattern," which means the genes have the same 0 and 1 strings in their rows in the indicator table. Because the total squared lengths of the edges are minimized, the categories that have large intersection set are likely to be pulled together by the common genes they share. The distances between the category points reflect the similarities between the gene contents of the categories. The plot shows the patterns of correlations between the groups of differentially expressed genes under various conditions and groups of genes with various functions.

The categories shown in Figure 1 approximately form four groups. Group A (left) contains ste12.down (40),[**] ste18.down (41), ste7.down (42), fus3_kss1.down[†] (32), rad6.down (35), hog1.up (10), dig1_dig2.up (7), sst2.up (20), pheromone response, mating-type determination, sex-specific proteins (47), cell differentiation (48), cell fate (50), chemoperception and response (52). Here we see the following functional categories: pheromone response, mating-type determination, sex-specific proteins (47) (a subcategory of cell

[**]"ste.down" denotes the group of genes that are down-regulated in the mutant in which ste12 is knocked out. In Figure 1, the category is labeled by the number in the parenthesis, see the legend for Figure 1.

[†]Double mutant in which both fus3 and kss1 are knocked out.

differentiation (48) and cell fate (50)) and chemoperception and response (52). This is consistent with the expression categories we observed in this region. Ste7, ste12, ste18, fus3 and kss1 belong to the pheromone signaling pathway (http://genome-www.stanford.edu/Saccharomyces/), removing these genes turns off the expression of pheromone-response genes. Ste7.down (42), ste12.down (40) and ste18.down (41) represent the groups of genes that are down-regulated when ste7, ste12 and ste18 are knocked out respectively. It is known that dig1 dig2 double mutants show constitutive mating pheromone specific gene expression and invasive growth and sst2 null mutants exhibit increased sensitivity to mating factors (http://genome-www.stanford.edu/Saccharomyces/). Consistently, we see dig1_dig2.up (7) and sst2.up (20) in this region. The expression of rad6 is induced early in meiosis and peaks at meiosis I, the mutant shows repression of retrotransposition, meiotic gene conversion and sporulation (http://genome-www.stanford.edu/Saccharomyces/). Hog1 is in the signaling pathway that responds to high osmolarity glycerol (Robberts et al., 2000), the presentation of hog1.up (10) in this region reflects the crosstalks between the HOG (high osmolarity glycerol) pathway and the pheromone pathway (Sprague, 1998). This method reveals positive correlations and negative correlations between the gene expression profiles of the samples simultaneously by displaying up- and down-regulation categories together. Clustering analysis failed to reveal the correlation between the dig1 dig2 double mutant and the mutants of the pheromone signaling pathway genes (ste7, ste12, ste18, fus3_kss1), the dig1 dig2 double mutant is located far away from the pheromone signaling pathway genes in the clustering dendrogram (Hughes et al., 2000a; http://download.cell.com/supplementarydata/cell/102/1/109/DC1/Tbl3ClnB.jpg). This is because the double knockout of dig1 and dig2 lead to constitutive mating pheromone specific gene expression (up-regulation) while the knockouts of pheromone signaling pathway genes turn off mating pheromone specific gene expression (down-regulation).

Group B (lower right) contains clb2.up (5), hda1.up (9), yhl029c.up (25), ckb2.down (30), gcn4.down (33), vps8.down (43), amino acid biosynthesis (46), amino acid metabolism (49), nitrogen and sulfur metabolism (56). Most of the genes involved in amino acid metabolism (the small light gray points in Fig. 1) are located in this region. The expression categories (clb2.up (5), hda1.up (9), yhl029c.up (25), ckb2.down (30), gcn4.down (33), vps8.down (43)) are enriched by the genes of two functional categories (amino acid biosynthesis (46), amino acid metabolism (49)) at very significant levels, ($p < 10^{-5}$).¶This means the knockouts of these genes (clb2, hda1, yhl029c, ckb2, gcn4, and vps8) impact many more genes involved in amino acid biosynthesis/metabolism than that could happen by chances. Gcn4 is a transcriptional activator of amino acid biosynthetic genes (http://genome-www.stanford.edu/Saccharomyces/). As far as we know, there is no literature describing the roles of the other five genes (clb2, hda1, yhl029c, ckb2, and vps8) in amino acid biosynthesis/metabolism. This result provides hints to some possible new functions of these genes.

Group C (middle) contains cup5.up (6), fks1(haploid).up (8), med2(haploid).up (14), swi6(haploid).up (21), vma8.up (23), homeostasis of cations (51), ionic homeostasis (53), regulation of / interaction with cellular environment (54), cell wall (57), plasma membrane (61). Null mutant of cup5 is copper sensitive. Fks1 is involved in cell wall organization and biogenesis (http://genome-www.stanford.edu/Saccharomyces/). There are 57 and 61 genes in the expression categories cup5.up and vma8.up, respectively, the intersection set of these two categories contains 46 genes. The overlapping is very significant ($p = 2 \times 10^{-37}$). The knockout of cup5 or vma8 makes largely the same group of genes over-express. Med2(haploid).up (14) and swi6(haploid).up (21) do not significantly overlap with other categories in this region. This may reflect the limitation of the two-dimensional visualization of high dimension data.

¶The P value is the probability of observing at least k genes in the intersection set of an expression category of size n and a function category of size f, assuming there is no association between the expression category and the function category,

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{f}{i}\binom{g-f}{n-i}}{\binom{g}{n}}$$

where g is the total number of genes in the indicator table.

Group D (upper right) contains ade2(haploid).up (0), aep2.up (1), afg3(haploid).up (2), cem1.up (3), msu1.up (15), top3(haploid).up (22), ymr293c.up (26), lovastatin.up (28), dot4.down (31), c-compound and carbohydrate metabolism (55), lipid, fatty-acid and isofenoid metabolism (58), cell rescue, defense and virulence (59), energy (60), detoxification (62). All the function categories in this region belong to three super-categories: (a) energy (60), (b) cell rescue, defense, and virulence (59), which includes detoxification (62), and (c) metabolism, which includes c-compound and carbohydrate metabolism (55) and lipid, fatty-acid and isofenoid metabolism (58). Ade2 is a purine-base metabolism gene (http://genome-www.stanford.edu/Saccharomyces/). Aep2 mutant is non-conditional respiratory mutant and unable to express the mitochondrial OLI1 gene afg3. Cem1, msu1, ymr293c are mitochondrial genes (http://genome-www.stanford.edu/Saccharomyces/) and are involved in energy generation and processing.

Yeast transcription modules

Ihmels et al. identified 86 context-dependent and potentially overlapping transcription modules by mining yeast microarray data of more than 1,000 experiments (Ihmels et al., 2002; www.weizmann.ac.il/home/jan/NG/MainFrames.html). The genes in a module are co-regulated under some experimental conditions. The modules reflect the modular organization of the yeast transcription network. Here we use Homogeneity Analysis to present a global view of the relations between the modules and their connections to the underlying biological processes.

We selected 72 modules that contain more than 20 genes and overlap with at least one other selected modules. Altogether, the 72 modules contain 2,159 genes. The modules and 18 biological processes defined by Gene Ontology (The Gene Ontology Consortium 2000) are quantified using Homogeneity Analysis and displayed in two-dimensional space (Fig. 2). The graph reveals the relations between the genes (small gray dots), modules (big black dots) and the biological processes (big blue dots). The modules related to nitrogen and sulfur metabolism (78, 84) are in the lower left corner of the plot; modules related to cellular fusion (74), conjugation with cell proliferation (76), sporulation (77), response to DNA damage stimulus (81), nucleobase, nucleoside, nucleotide and nucleic acid metabolism (82), signal transduction (89) are in the lower right corner; the upper area of the plot is related to electron transport (80), oxidative phosphorylation (73), and aldehyde metabolism (85); the middle area are related to carbohydrate metabolism (86), response to oxidative stress (87), oxygen and reactive oxygen species metabolism (88), alcohol metabolism (79), transport (83), lipid metabolism (75), protein metabolism (72).

The function categories that are closely located show strong associations. For example, electron transport (80) and oxidative phosphorylation (73) contain 17 and 25 genes respectively, the intersection set of these two categories contains 12 genes. The p-value associated with the overlapping is 1.5×10^{-21}. It is well known that electron transport and oxidative phosphorylation are closely related biological processes. Similar examples include response to oxidative stress (87) and oxygen and reactive oxygen species metabolism (88) ($p = 1.4 \times 10]^{-41}$), cell proliferation (76) and response to DNA damage stimulus (81) ($p = 7.0 \times 10^{-14}$). This indicates that arrangement of the genes and categories is biologically meaningful.

The similar modules are grouped together. Module 26 (22),[‖] Module 35 (29), Module 48 (40), Module 54 (45), Module 70 (59) and Module 75 (63) are clustered together near the origin. The sizes of these modules are 60, 73, 88, 66, 69, and 72, respectively. The six modules share 45 common genes, more than 50% of the largest module.

The associations between modules and biological processes are also readily to be found in Figure 2. We can see that Module 5 (4), Module 55 (46) and Module 74 (62) are closely related to the biological process "oxidative phosphorylation" (73). The p-value associated with the overlapping between "oxidative phosphorylation" and the three modules are 2.0×10^{-41}, 2.9×10^{-33}, and 2.2×10^{-5} respectively. Module 1 (0), Module 51 (42) and Module 57 (48) are grouped with "protein metabolism" (72). The p-value associated with the overlapping between "protein metabolism" and the three modules are 1.9×10^{-72}, 4.0×10^{-4}, and 5.7×10^{-51}, respectively.

[‖]In Figure 2, the module is labeled by the number in the parenthesis.

FIG. 1. Homogeneity analysis for the Rosetta Compendium data and MIPS functional catalogue. In this bipartite plot, the small gray dots represent genes; the red (up-regulation) and green (down-regulation) dots represent expression categories, and the blue dots represent MIPS gene function categories. The categories are labeled by numbers:

0: ade2 (haploid).up
1: aep2.up
2: afg3 (haploid).up
3: cem1.up
4: cka2.up
5: clb2.up
6: cup5.up
7: dig1_dig2 (haploid).up
8: fks1 (haploid).up
9: hda1.up
10: hog1(haploid).up
11: isw1_isw2.up
12: kim4.up
13: kin3.up
14: med2 (haploid).up
15: msu1.up
16: qcr2 (haploid).up
17: rrp6.up
18: rtg1.up
19: spf1.up
20: sst2 (haploid).up
21: swi6 (haploid).up
22: top3 (haploid).up
23: vma8.up
24: yar014c.up
25: yhl029c.up
26: ymr293c.up
27: HU.up
28: Lovastatin.up
29: Terbinafine.up
30: ckb2.down
31: dot4.down
32: fus3,kss1 (haploid).down
33: gcn4.down
34: med2 (haploid).down
35: rad6 (haploid).down
36: rpl12a.down
37: rtg1.down
38: sir4.down
39: sod1 (haploid).down
40: ste12 (haploid).down
41: ste18 (haploid).down
42: ste7 (haploid).down
43: vps8.down
44: yel033w.down
45: ymr014w.down
46: AMINO ACID BIOSYNTHESIS
47: PHEROMONE RESPONSE, MATING-TYPE DETERMINATION, SEX-SPECIFIC PROTEINS
48: CELL DIFFERENTIATION
49: AMINO ACID METABOLISM
50: CELL FATE
51: HOMEOSTASIS OF CATIONS
52: CHEMOPERCEPTION AND RESPONSE
53: IONIC HOMEOSTASIS
54: REGULATION OF / INTERACTION WITH CELLULAR ENVIRONMENT
55: C-COMPOUND AND CARBOHYDRATE METABOLISM
56: NITROGEN AND SULFUR METABOLISM
57: CELL WALL
58: LIPID, FATTY-ACID AND ISOPRENOID METABOLISM
59: ENERGY
60: CELL RESCUE, DEFENSE AND VIRULENCE
61: PLASMA MEMBRANE

CONCLUSION

Homogeneity analysis is a powerful method that is capable of integrating the analysis of microarray-derived gene groups and categorical gene function information. It is a useful mathematical framework for interpreting microarray data in the context of existing biological knowledge.

Homogeneity analysis can be used for analyzing the relations between any gene groups regardless how they are derived. For example, we can group genes according to the DNA-binding motifs occurring in their

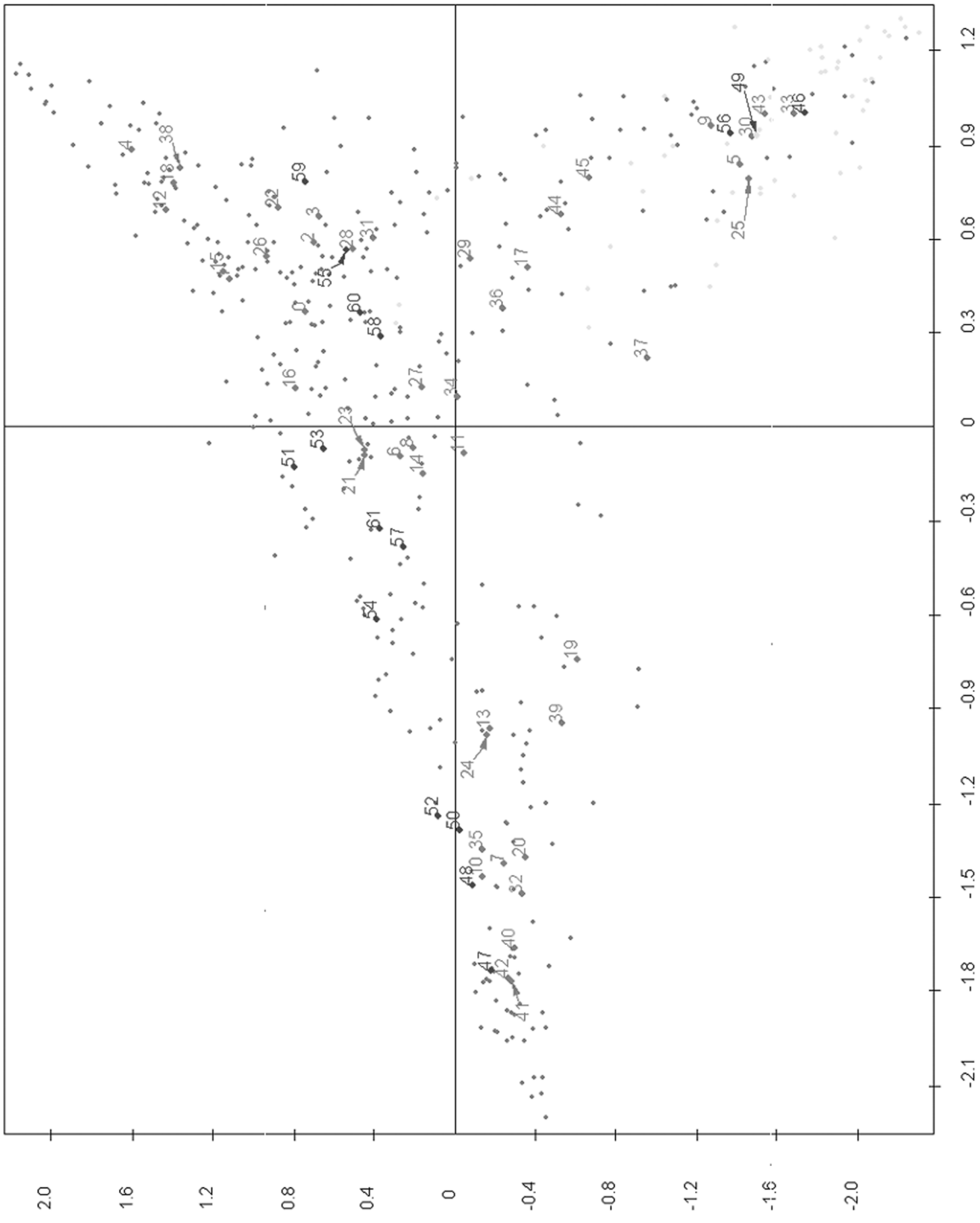
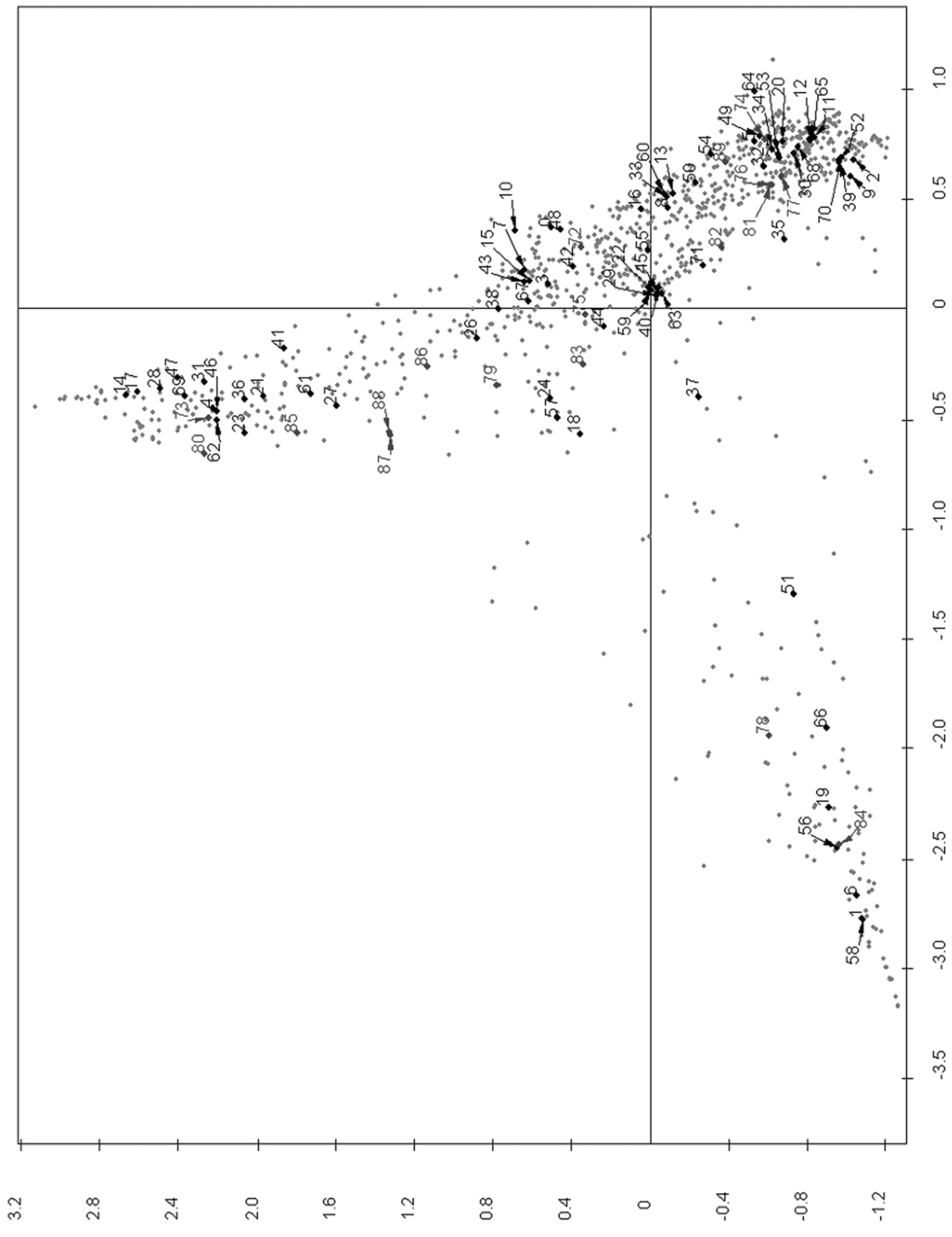

2.0
1.6
1.2
0.8
0.4
0
-0.4
-0.8
-1.2
-1.6
-2.0
1.2
0.9
0.6
0.3
0
-0.3
-0.6
-0.9
-1.2
-1.5
-1.8
-2.1

FIG. 2. Homogeneity analysis for yeast transcription modules and the biological processes defined by gene ontology. In this bipartite plot, the small gray dots represent genes; the black dots represent modules, and the blue dots represent biological processes defined by gene ontology. The categories are labeled by numbers:

0: Module 1
1: Module 2
2: Module 3
3: Module 4
4: Module 5
5: Module 6
6: Module 7
7: Module 8
8: Module 10
9: Module 11
10: Module 12
11: Module 13
12: Module 15
13: Module 16
14: Module 17
15: Module 18
16: Module 19
17: Module 20
18: Module 21
19: Module 22
20: Module 24
21: Module 25
22: Module 26
23: Module 27
24: Module 28
25: Module 29
26: Module 30
27: Module 32
28: Module 34
29: Module 35
30: Module 36
31: Module 37
32: Module 40
33: Module 41
34: Module 42
35: Module 43
36: Module 44
37: Module 45
38: Module 46
39: Module 47
40: Module 48
41: Module 50
42: Module 51
43: Module 52
44: Module 53
45: Module 54
46: Module 55
47: Module 56
48: Module 57
49: Module 58
50: Module 59
51: Module 61
52: Module 62
53: Module 63
54: Module 64
55: Module 65
56: Module 66
57: Module 67
58: Module 68
59: Module 70
60: Module 71
61: Module 73
62: Module 74
63: Module 75
64: Module 76
65: Module 77
66: Module 80
67: Module 81
68: Module 82
69: Module 84
70: Module 85
71: Module 86
72: protein metabolism
73: oxidative phosphorylation
74: conjugation with cellular fusion
75: lipid metabolism
76: cell proliferation
77: sporulation
78: sulfur metabolism
79: alcohol metabolism
80: electron transport
81: response to DNA damage stimulus
82: nucleobase, nucleoside, nucleotide and nucleic acid metabolism
83: transport
84: nitrogen metabolism
85: aldehyde metabolism
86: carbohydrate metabolism
87: response to oxidative stress
88: oxygen and reactive oxygen species metabolism
89: signal transduction


3.2
2.8
2.4
2.0
1.6
1.2
0.8
0.4
0
-0.4
-0.8
-1.2
1.0
0.5
0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
-3.5

up-stream regions, the protein domains they encode or the sub-cellular locations of the products of the genes. The relations between various classifications of genes can be revealed using this method.

We developed a computer program to implement the method. It is free for nonprofit research and is downloadable at http://compbio.utmem.edu/Gifi.php.

ACKNOWLEDGMENTS

We thank Drs. Jan de Leeuw and George Michailides for their help in the implement of homogeneity analysis. This work is partly supported by NIH grants P20 CA96470 and R01 GM67250 to W.H.W.

REFERENCES

BENZECRI, J.P. (1973). *L'Analysis des Donnees. Tome 1: La Taxinomie. Tone2: L'Analyse des Correspondances* (Dunod, Paris).

DE LEEUW, J., and VAN RIJCKEVORSEL, J. (1980). Homals and princals. Some generalizations of principal components analysis. In *Data Analysis and Informatics II* (Amsterdam, North-Holland).

DE LEEUW, J. (1984). The Gifi-system of nonlinear multivariate analysis. In *Data Analysis and Informatics III* (Amsterdam, North-Holland).

DOLINSKI, K., BALAKRISHNAN, R., CHRISTIE, K.R., et al. (2003). Saccharomyces genome database. Available: http://genome-www.stanford.edu/Saccharomyces/.

FELLENBERG, K., HAUSER, N.C., BRORS, B., et al. (2001). Correspondence analysis applied to microarray data. Proc Natl Acad Sci USA, **98,** 10781–10786.

GOLUB, G.H., and VAN LOAN, C.F. (1989). *Matrix computations* (Johns Hopkins University Press, Baltimore).

GREENACRE, M.J. (1993). *Correspondence Analysis in Practice* (Academic Press, London).

GREENACRE, M., and HASTIE, T. (1987). The geometric interpretation of correspondence analysis. J Am Statist Assoc **82,** 437–447.

HUGHES, T.R., MARTON, M.J., JONES, A.R., et al. (2000a). Functional discovery via a compendium of expression profiles. Cell **102,** 109–126.

HUGHES, T.R., ROBERTS, C.J., DAI, H., et al. (2000b). Widespread aneuploidy revealed by DNA microarray expression profiling. Nat Genet **25,** 333–337.

IHMELS, J., FRIEDLANDER G., BERGMANN S., et al. (2002). Revealing modular organization in the yeast transcriptional network. Nat Genet **31,** 370–377.

KANEHISA, M., GOTO, S., KAWASHIMA, S., et al. (2002). The KEGG databases at GenomeNet. Nucleic Acids Res **30,** 42–46.

KISHINO, H., and WADDELL, P. (2000). Correspondence analysis of genes and tissue types and finding genetic links from microarray data. Genome Inform **11,** 83–95.

LAZZERONI, L., and OWEN, A. (2002). Plaid models for gene expression data. Statist Sin **12,** 61–86.

LEE, S.I., and BATZOGLOU, S. (2003). Application of independent component analysis to microarrays. Genome Biol **4,** R76.

MEWES, H.W., FRISHMAN, D., GÜLDENER, U., et al. (2002). MIPS: a database for genomes and protein sequences. Nucleic Acids Res **30,** 31–34.

MICHAILIDIS, G., and DE LEEUW, J. (1998). The Gifi System of descriptive multivariate analysis. Statist Sci **13,** 307—336.

ROBERTS, C.J., NELSON, B., MARTON, M.J., et al. (2000). Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. Science **287,** 873–880.

SEGAL, E., SHAPIRA, M., REGEV, A., et al. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet **34,** 166–176.

SPRAGUE, G.F., Jr. (1998). Control of MAP kinase signaling specificity or how not to go HOG wild. Genes and Dev **12,** 2817–2820.

STUART, J.M., SEGAL, E., KOLLER, D., et al. A gene-coexpression network for global discovery of conserved genetic modules. Science **302,** 249–255.

TANAY, A., SHARAN, R., and SHAMIR, R. (2002). Discovering statistically significant biclusters in gene expression data. Bioinformatics **18,** s136–s144.

THE GENE ONTOLOGY CONSORTIUM. (2000). Gene ontology: tool for the unification of biology. Nat Genet **25,** 25–29.

WADDELL, P.J., and KISHINO, H. (2000). Cluster inference methods and graphical models evaluated on NCI60 microarray gene expression data. Genome Inform **11,** 129–140.

WU, L.F., HUGHES T.R., DAVIERWALA A.P., et al. (2002). Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. Nat Genet **31,** 255–265.

Address reprint requests to:
Dr. Yan Cui
Department of Molecular Sciences
University of Tennessee Health Science Center
858 Madison Avenue
Memphis, TN 38103

E-mail: ycui2@utmem.edu

or

Dr. Wing Hung Wong
Department of Statistics
Harvard University Science Center
1 Oxford Street
Cambridge, MA 02138

E-mail: wwong@stat.harvard.edu