

Integrated Analysis of Microarray Data and Gene Function Information

YAN CUI^{1,2}, MI ZHOU^{1,2} and WING HUNG WONG^{3,4,5}

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 microarray-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.

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

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 conditions¹ (Michailidis and de Leeuw 1998; Greenacre and Hastie 1987), which is not satisfied in the integrated analysis of microarray data and gene function information. Simple Correspondence Analysis (Benzecri 1973; de Leeuw and van Rijkevorsel 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

¹ Homogeneity Analysis is equivalent to Multiple Correspondence Analysis if all the row margins of the indicator table are equal.

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 Rijkevorsel 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:

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

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:

$$X'X = NI_p, \quad (2)$$

$$u'X = 0, \quad (3)$$

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.

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

$$CY = G'X, \quad (4)$$

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

$$\hat{Y} = C^{-1}G'X \quad (5)$$

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

$$RX = GY \quad (6)$$

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

$$\hat{X} = R^{-1}GY \quad (7)$$

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

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

$$\text{where } W = \hat{X} - u(u'\hat{X}/N), \quad (9)$$

This solution is called HOMALS solution (Homogeneity Analysis by Means of Alternating Least Squares). Here we list some basic properties of the Homals solution, which are useful for interpreting 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 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 to 100 genes up- or down-regulated greater than 2 fold, 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 2 fold, and at $P \leq 0.01$, in 2 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 2 fold, and at $P \leq 0.01$; 2) genes that are down-regulated at greater than 2 fold, and at $P \leq 0.01$. The microarray-derived gene groups are encoded using an indicator table. Each experiment has two categories (up-regulation 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 (see the legend for Figure 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)², *ste18.down* (41), *ste7.down* (42), *fus3_kss1.down*³ (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 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

² “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.

³ Double mutant in which both *fus3* and *kss1* are knocked out.

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-regulation categories 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/Tb13ClnB.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 cyan points in Figure 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, ($P < 10^{-5}$)⁴. This means the knockouts

⁴ The P value is the probability of observing at least k genes in the intersection set of an expression category of size n

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.

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 – energy (60), cell rescue, defense and virulence (59) (which includes detoxification (62))

and a function category of size f , assuming there is no association between the expression category and the function category,

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{f}{i} \binom{g-f}{n-i}}{\binom{g}{n}}, \text{ where } g \text{ is the total number of}$$

genes in the indicator table.

and 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; <http://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 (Figure 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)⁵, 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.

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

⁵ In Figure 2, the module is labeled by the number in the parenthesis, see the legend for Figure 2.

according to the DNA-binding motifs occurring in their 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>.

Acknowledgements

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. Tome2: L'Analyse des Correspondances. (Dunod, Paris).
- DE LEEUW, J., AND VAN RIJCKEVORSEL, J. (1980). Homals and Princals. Some Generalizations of Principal Components Analysis. *Data Analysis and Informatics II*. (Amsterdam, North Holland).
- DE LEEUW, J. (1984). The Gifi-system of Nonlinear Multivariate Analysis. *Data Analysis and Informatics III*. (Amsterdam, North Holland).
- Dolinski, K., Balakrishnan, R., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., et al. Saccharomyces Genome Database <http://genome-www.stanford.edu/Saccharomyces/>
- FELLENBERG, K., HAUSER, N.C., BRORS, B., NEUTZNER, A., HOHEISEL, J.D., VINGRON M. (2001). Correspondence analysis applied to microarray data. *PNAS*, **98**: 10781-10786.
- GOLUB, G.H. AND VAN LOAN C.F. (1989). *Matrix Computations*, Baltimore: Johns Hopkins University Press.
- GREENACRE M. AND HASTIE T. (1987). The Geometric Interpretation of Correspondence Analysis. *Journal of the American Statistical Association*. **82**: 437-447.
- GREENACRE M.J. (1993). *Correspondence analysis in practice*. (Academic Press, London).
- 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., JONES, A.R., MEYER, M.R., SLADE, D., BURCHARD, J., DOW,

- S., WARD, T.R., KIDD, M.J. et al. (2000b). Widespread aneuploidy revealed by DNA microarray expression profiling. *Nature Genetics*, **25**: 333-337.
- IHMELS J., FRIEDLANDER G., BERGMANN S., SARIG O., ZIV Y. & BARKAI N. (2002). Revealing modular organization in the yeast transcriptional network. *Nature Genetics* **31**: 370-377.
- KANEHISA, M., GOTO, S., KAWASHIMA, S., AND NAKAYA, A. (2002). The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30**: 42-46.
- KISHINO, H., WADDELL, P. (2000). Correspondence Analysis of Genes and Tissue Types and Finding Genetic Links from Microarray Data. *Gnome Informatics* **11**: 83-95.
- LAZZERONI, L., AND OWEN, A. (2002). Plaid Models for Gene Expression Data. *Statistica Sinica* **12**: 61-86.
- LEE S.I. AND BATZOGLOU S. (2003). APPLICATION OF INDEPENDENT COMPONENT ANALYSIS TO MICROARRAYS. *GENOME BIOLOGY* **4**: R76.
- MEWES HW, FRISHMAN D, GÜLDENER U, MANNHAUPT G, MAYER K, MOKREJS M, MORGENSTERN B, MÜNSTERKOETTER M, RUDD S, WEIL B. (2002). MIPS: a database for genomes and protein sequences. *Nucleic Acids Research* **30**: 31-34.
- MICHAILIDIS, G. & DE LEEUW J. (1998). The Gifi System of descriptive multivariate analysis. *Statistical Science* **13**: 307—336.
- ROBERTS, C.J., NELSON, B., MARTON, M.J., STOUGHTON, R., MEYER, M.R., BENNETT, H.A., HE, Y.D., DAI, H., WALKER, W.L., HUGHES, T.R. 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., PE'ER D., BOTSTEIN D., KOLLER D. & FRIEDMAN N. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, **34**: 166-176.
- SPRAGUE, G.F. JR. (1998) Control of MAP kinase signaling specificity or how not to go HOG wild. *Genes & Development* **12**: 2817-2820.
- STUART, J.M. SEGAL, E., KOLLER, D., KIM, S.K. 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. *Nature Genetics* **25**: 25-29.
- WADDELL, P.J. AND KISHINO, H. (2000). Cluster Inference Methods and Graphical Models Evaluated on NCI60 Microarray Gene Expression Data. *Genome Informatics* **11**: 129-140.
- WU, L.F., HUGHES T.R., DAVIERWALA A.P., ROBINSON M.D., STOUGHTON R. & ALTSCHULER S.J. (2002). Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nature Genetics* **31**: 255-265.

Address reprint request to:

Dr. Yan Cui

Department of Molecular Sciences

University of Tennessee Health Science Center

858 Madison Avenue

Memphis, TN 38163

Email: ycui2@utmem.edu

Or to:

Dr. Wing Hung Wong

Department of Statistics

Harvard University

Science Center, 1 Oxford Street

Cambridge, MA 02138

Email: wwong@stat.harvard.edu

Table 1 Indicator tables

(A)

	Sample1.up	Sample1.down	Sample2.up	Sample2.down	...	Function1	Function2	...
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.

Figure 1

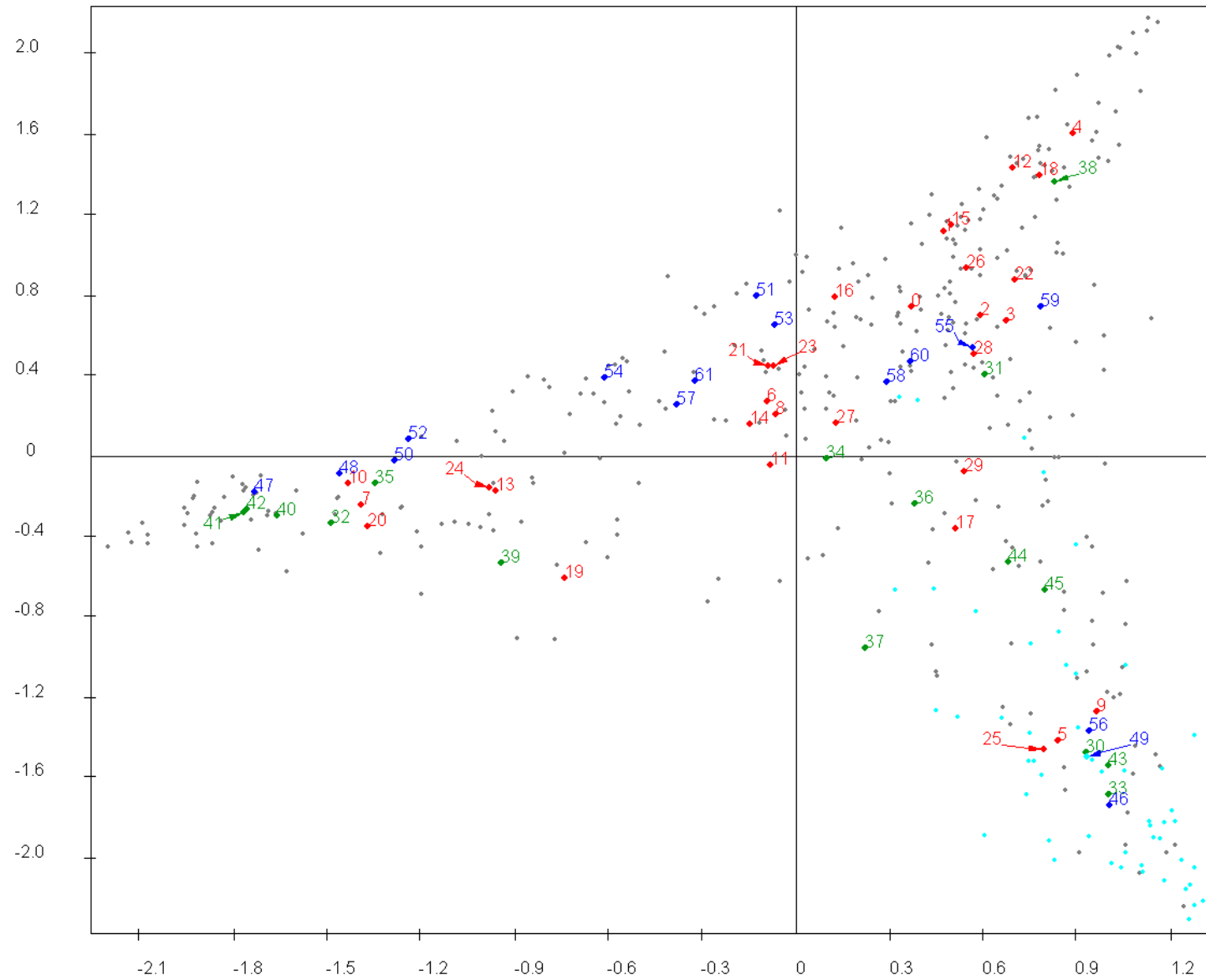


Figure 1. Homogeneity Analysis for Rosetta Compendium data and MIPS functional catalogue. In this bipartite plot, the small gray dots represent genes; the red (up-regulation) and green (down-regulation) dots represent expression categories, and the blue dots represent MIPS gene function categories. The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

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

Figure 2

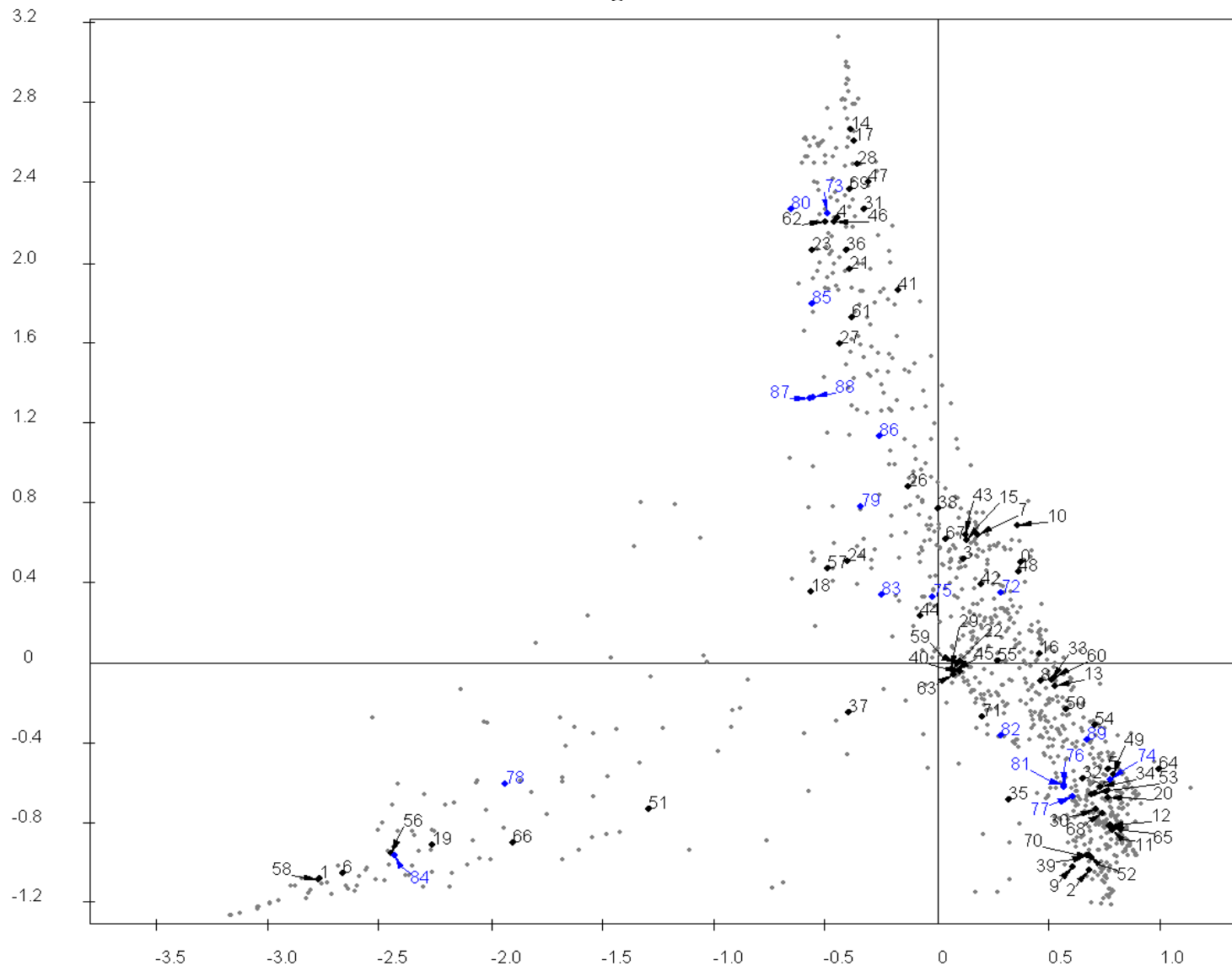


Figure 2. Homogeneity Analysis for yeast transcription modules and the biological processes defined by Gene Ontology. In this bipartite plot, the small gray dots represent genes; the black dots represent modules, and the blue dots represent biological processes defined by Gene Ontology. The categories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

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