[Frontiers in Bioscience 10, 2669-2675, September 1, 2005]

FUNCTIONAL GROUPING OF YEAST GENES VIA BICLUSTERING MICROARRAY DATA

Daqing Mao [1], Yi Luo[2], Maosheng Cheng [1] and Jinghai Zhang [1]

[1] Department of Bio-pharmaceutics, School of Pharmaceutical Engineering, Shenyang Pharmaceutical University, Shenyang 110016, China, [2] Faculty of Environmental Science, Liaoning University, Liaoning 110036, China

TABLE OF CONTENTS

1. Abstract
2. Introduction
3. Materials and methods
3.1. Microarray datasets of Yeast genome
3.2. MIPS (Munich information center for protein sequence)
3.3. Biclustering
3.3.1. Storing the datasets in database (SQL)
3.3.2. Preprocessing of biclustering in MATLAB
3.3.3. Running the biclustering scripts
3.3.4. Checking functions of biclusters through MIPS information
4. Results
4.1. Biclusters
4.2. Relevant functions of biclusters
4.3. Further analysis of bicluster VII
5. Discussion
5.1. "Compendium" Concept
5.2. Functional prediction of questionable ORF
5.3. Effectiveness and drawback of the biclustering algorithm
6. References

1. ABSTRACT

Biclustering algorithm on Gibbs sampling strategy is a recruit in the field of the analysis of gene expression data of microarray experiments. Its feasibility and validity still need to be researched not only for synthetic datasets but also for real datasets. Here we investigated a biclustering algorithm on a microarray dataset of Yeast genome through building a database for storing microarray datasets and MIPS data, and running the scripts on Matlab platform to discover gene patterns. In contrast with standard clusterings that reveal genes behaving similarly over all the conditions, biclustering groups genes over only a subset of conditions for which those genes have a sharp probability distribution. It has the key advantage of providing a transparent probabilistic interpretation of the biclusters. Its basic strategy of Gibbs sampling does not suffer from the problem of local minima that often characterizes expectation maximization, so that the patterns should be more global and accurate. Also we tested it with the known explanation of genes in MIPS, objectively to demonstrate the effectiveness and deficiencies of biclustering approach, and the functions of a few unknown ORFs in some bicluster can be deduced in the present research. In addition, the result of similarity searching in Blast-Search can be an assistant evidence for its effectivity.

2. INTRODUCTION

DNA microarray is an innovative technology that can measure the expression level of thousands of genes in parallel (1,2,3). Due to the huge throughput microarray experiments provide, computational methods that extract the knowledge from the large sets of experimental results become important. The knowledge of genes' functions and relationships between genes may provide possible drug target candidates or aide in the understanding of a disease process. Among the computational methods, clustering is one of the most popular approaches of analyzing gene expression data without prior knowledge. Several representative algorithmic techniques have been developed and experimented in clustering gene expression data, e.g. hierarchical clustering, k-means and a recruit of biclustering, etc. Although it is possible to obtain biologically meaningful results with these algorithms respectively, some characteristics often complicate their use for clustering expression data, e.g., the predefinition of the number of clusters in K-means (4), the finding of a suitable cleavage level on a big hierarchy tree of hierarchical clustering, the dispersedness of biologically similar genes on large dimensions in hierarchy tree (5), etc. These drawbacks induce even a mistaken result.

Biclustering is a recruit, but a promising methodology in the field of clustering microarray data. The original biclusters of gene expression datasets were based on uniformity criteria, and were discovered by applying the greedy algorithm developed by Cheng and Church (6). The approximate uniformity in a submatrix in gene expression data can be detected by another model Plaid developed by Lazzeroni and Owen (7), in which they used a form of overlapping two-sided clustering with an embedded ANNOVA in each other. Patterns in which genes differ in their expression levels by a constant vector can be detected by Plaid model. Ben-Dor *et al.* discussed approaches for unsupervised identification of patterns in expression data that distinguish two subclasses of a tissue on the basis of a supporting set of genes that can offer accurate classification (8). Ben-Dor *et al.* also introduced the model of Order preserving submatrix (9). Tanay *et al.* defined a bicluster as a subset of genes that jointly respond across a subset of conditions for reducing the biclustering problem (10). A biclustering algorithm based on Gibbs sampling has been successfully developed and implemented by Sheng *et al.* (11), and applied on microarray datasets. With discretizing the expression datasets into fixed number of bins, Sheng *et al.* detected the motif subsequences in sequence data. Caliafano *et al.* also previously observed this analogy, and they applied a pattern-discovery algorithm SPLASH for finding patterns in strings to gene expression data (12). Also with Gibbs sampling, Wu *et al.* developed a running scheme and expand its application to biclustering continuous gene expression data (13).

However, biclustering method as a recruit in this field still need to be tested and improved in practice for a wide application. In the present research, a biclustering algorithm is applied on a microarray dataset of Yeast genome, and the result of biclusters is discussed detailedly for finding functional information of genes.

3. MATERIALS AND METHODS

3.1. Microarray datasets of Yeast genome

Whole-genome expression profiling, facilitated by the development of DNA microarrays (14), represents a major advance in genome-wide functional analysis. Because the relative abundance of transcripts is often tailored to specific cellular needs, most expression profiling studies conducted on microarray have focused on the genes that respond to conditions or treatments of interests. Not only can we directly apply a single assay to measure the interaction items of unknown or known genes in identifying functions, but also the idea of "compendium" can be used for the purpose of predicting or diagnosing etc (14,15). Hughes datasets as a comprehensive datasets of reference profiles were created for the aim of analyzing functions, testing drug target, etc (15). The reference datasets of three-hundred full-genome expression profiles in *S. cerevisiae* corresponding to mutations and chemical treatments in both characterized genes and uncharacterized open reading frames (ORFs), as well as treatments with compounds with known molecular targets were developed. A gene-specific error model was built for compensating for differences in variation of transcript abundance among different yeast genes. Hughes datasets contain totally 6316 genes corresponding to 300 conditions related to *S. cerevisiae*. For each experiment (condition), five values were calculated: logIntensity, logRatio, errors of error model, errors of measurements and P value. The values of logIntensity and P value have been investigated in present research.

3.2. MIPS (Munich information center for protein sequence)

The MIPS Comprehensive Yeast Genome Database (CYGD, http://mips.gsf.de/genre/proj/yeast/) presents the information on the functional network and molecular structure of the entirely sequenced and well-studied model eukaryote, the budding yeast *S. cerevisiae*. In addition, the data of various projects on related yeasts has been used for comparative analysis. Nearly seven thousands genes and ORFs documented in MIPS, and being categorized into root main 19 functional groups. These information known as for checking exact functions related to each gene of each pattern was stored into a database in local server.

3.3. Biclustering

3.3.1 Storing the datasets in database (SQL)

Some tables are built in local database to store Hughes datasets. Then the values of logIntensity and P_value in the original dataset are elicited for clustering analysis. Also some tables built in local database are for the purpose of storing MIPS datasets and explanation which are useful for information searching of index of clusters and genes' annotation in chunks.

3.3.2. Preprocessing of biclustering in MATLAB

The main objective of preprocessing of biclustering is to reduce noise. In the previous step, we have elicited the values of "logIntensity" and "P_value" of gene expression respectively, transfered and stored it in text file. The following is to load them in Matlab in matrix form. The data showing less standard deviation along column vectors (genes) is deleted with the way of putting a certain threshold on the original data in Matlab. Also the variations of each gene along all of the experiments are examined for filtering the ORFs holding a certain P_value in quantitative experiments ($P \leq 0.01$, experiments≤ 100). Then discretizing the filtered expression data into fixed number of bins in the last of this step.

3.3.3. Running the biclustering scripts

The biclustering algorithm is based on the Gibbs sampling strategy. In this method, a greedily iterative searching is applied to find interesting patterns in the matrices, and probabilistic models are proposed in which matrix rows (genes in this case) and columns (experimental conditions) are divided into clusters, and there are linking probabilities between these clusters. These linking probabilities can describe the association between a gene cluster and an experimental condition cluster, and can be found by using iterative Gibbs sampling and approximated Expectation Maximization algorithms (11).

3.3.4. Checking functions of biclusters through MIPS information

Each gene's function in each bicluster was checked in those tables of MIPS local database with SQL queries for inducing the functions of biclusters.

Table 1. Number and Proportion of ORFs and experiments in each bicluster

Bicluster Composition	I	II	III	IV	V	VI	VII	VIII	IX	X	XI
Experiments	13 (4.3%)	19 (6.3%)	23 (7.7%)	7 (2.3%)	5 (1.7%)	23 (7.7%)	21 (7.0%)	15 (5.0%)	18 (6.0%)	20 (6.7%)	5 (1.7%)
ORFs	537 (28.3%)	155 (8.2%)	85 (4.5%)	24 (1.3%)	20 (1.1%)	35 (1.8%)	21 (1.1%)	19 (1.0%)	19 (1.0%)	24 (1.3%)	15 (0.8%)

4 RESULTS

4.1. Biclusters

The biclustering algorithm enables detection of multiple biclusters, through the way of masking the genes selected for found biclusters and perform the algorithm on the rest of data. 11 biclusters were found in the original datasets in the present research. Different bicluster consists of various genes and experiments, also different in their quantity. The compositions and proportions of ORFs and experiments are shown in table 1. The patterns of those biclusters are shown in figure 1. For the purpose of clear vision of those figures, some patterns of biclusters are partly shown.

4.2. Relevant functions of biclusters

Through checking in MIPS database which were restored in local server, the information of relevant functions and proportions of genes holding the function in each bicluster was reckoned and enumerated, which is shown in table 2 as below. The most other parts in each pattern are those "open reading frames" with unknown or unclassified function, e.g., in bicluster I, 37% of the genes (199/537) participate in cell metabolism, 62% (333/537) with unknown functions, and less than 1% (5/537) are classified into other function groups.

4.3. Further analysis of bicluster VII

The bicluster VII being a small subgroup with a higher percent of genes' homogeneity is shown and detected detailedly in the present research. The pattern of bicluster VII is shown in figure 2 as below.

By checking in MIPS database, the details of functions involving bicluster VII have been uncovered. 86% of the genes in this bicluster involves the functional group of 'protein synthesis', concretely involveing 'ribosome biogenesis'. The details of gene functions can be checked in figure 2. The ORF YGL064c plays 'RNA helicase activity' during the process of ribosome biogenesis (16). Thereinto, 3 questionable ORFs (14%) locate in this bicluster.

According the results of Blast sequence-searching program in NCBI (http://www.ncbi.nlm.nih.gov/BLAST/), YPL197c is full overlapped with gene YL8B (alias: RPL7B) (Score=283 bits (723), Expect=2e-75, Identities=137/137(100%), Positives=137/137 (100%)), partially overlapped with gene YL8A (alias: RPL7A) (Score=238 bits (607), Expect=5e-62, Identities=114/119 (95%), Positives=117/119(98%)). The both genes participate in the process of ribosomal protein synthesis (17). According to index searching in Yeast GRID (http://biodata.mshri.on.ca/yeast_grid/servlet), YPL197c and YCL047c have no interactions report, YDR154c has 2 interaction proteins (YHRO41c RNA polymerase II transcription mediator activity, YCR009c cytoskeletal protein binding). The result of bicluster VII and those searching information in public database may be a guide for functional detecting of the questionable ORF in experimental biology.

5. DISCUSSION

5.1. "Compendium" Concept

"Compendium Approach" indicates determination of functions of genes and ORFs affected by uncharacterized perturbation through comparisons of expression profile with a large and diverse set of reference profiles, and the profiles and the functions are consistent to a known ORF or gene (15). In the present research, 11 biclusters were gained by the biclustering algorithm which base on Gibbs sampling strategy. Each bicluster contains a subgroup of experiments and a subgroup of ORFs, and shows a homogeneity respectively. So the compendium approach may be improved for determining the function of genes by performing the subgroup of experiments of a bicluster which shows a higher homogeneity of genes and experiments respectively and holds known ORFs. The bicluster V and VII express a high homogeneity of unknown gene functions involving 'protein fate' or 'protein synthesis' respectively. Those experiments in the bicluster may be a 'compendium' or an indicator for detecting this kind of gene function involving the bicluster. Especially, bicluster V, which contains only 5 experiments, deserves to be further detected for the possibility of being an indicator for functional analysis of genes.

5.2. Functional prediction of questionable ORF

Model organisms such as Saccharomyces cerevisiae have also proven to be powerful tools for mechanistic studies of clinically relevant compounds (18,19), which are made possible by the fact that many human disease-associated genes have highly conserved yeast counterparts (20,21). The questionable ORF YPL197c was involved in analyzing molecular mechanisms of drugs, as a biological target for elucidating biosynthetic pathway of effective compounds (22,23). In the present research, the YPL197c was classified in bicluster VII in which all known genes hold the function relating ribosome biosynthesis. And after Blast searching in translated protein databases, we found that the sequence of YPL197c (411bp) is a part of gene YL8B (1919bp), and 87% (357) isoverlapped with the third exon (630bp) in the gene YL8B. Being compared to gene YL8A (Score = 238 bits (607), Expect = 5e-62, Identities = 114/119 (95%), Positives = 117/119 (98%)), it is full overlapped by an exon (630bp) in gene YL8A. Dose it individually makes roles in the process of translation or there are other regulative mechanism existing simultaneously in transcription or translation process? And the above analysis of biclustering gave us a

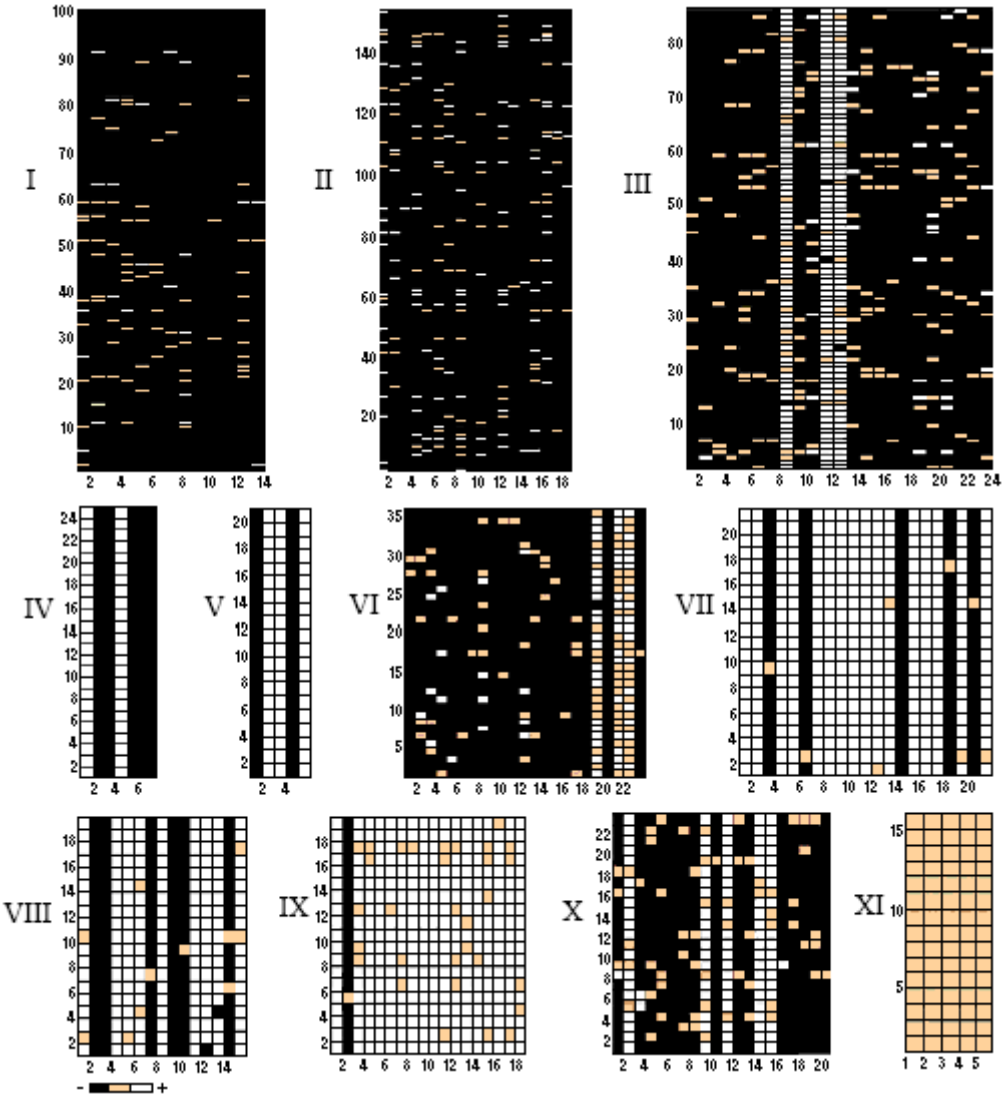

Figure 1. 11 patterns of biclusters. 11 biclusters were found in the original datasets through the way of masking the genes selected for the found biclusters and perform the algorithm on the rest of data. Each bicluster contains a subset of genes responding similarly in a subset of experiments. Distinct patterns of biclusters possess different number of experiments, and the behavior of genes in the pattern is consistent in induced (white color) or repressed (black color) response, or both (yellow color). Also characterized genes in each pattern are categorized and reckoned according to their involvement in MIPS Yeast Function List, which is shown in table 2.

direction for a further research. Apparently this is a further topic of our research in which combination between bioinformatics and biological experiments is necessary.

5.3. Effectiveness and drawback of the biclustering algorithm

Biclustering on the expression profiles of genes is an approach to detect groups of genes with similar expression profiles, even a small group, that can further identify their functions and putative regulatory elements in their promoter or coding sequence region (24). These studies help to understand gene co-regulation on the transcription level. Functional categorization also provides information on protein interaction and pathways. Gene expression and regulation are complex biological processes, which rouse the complexity of microarray data. But those genes involving same functions probably have similar expression patterns, which is a basic of clustering analysis. The clustering analysis on microarray data has been a research focus, and several of clustering approaches are developed, nevertheless some drawbacks or characters embarrass its further application in this field. E.g., the first drawback of standard k-means algorithm is an iterative procedure and requires to predetermine the number of clusters k as a given priory which will be not known ahead of time (4). Although several programs have been developed to estimate the cluster number, the accuracy of the estimation and the accuracy of results still depend on original datasets. The second, the problem of dead units locates into the procedure of k-means steps (4), i.e. if some

Table 2 Relevant functions of biclusters

Bicluster	Function and Proportion of genes holding the function in the Cluster	Unknown ORFs
I	Lipid, cofactors, prosthetic groups, fatty-acid and isoprenoid metabolism (37%)	62%
II	Protein synthesis (61%)	32%
III	C-compound and carbohydrate metabolism (56%)	38%
IV	Transcription (62%)	33%
V	Protein fate (folding, modification, destination) (55%)	45%
VI	Cellular transport and transport mechanisms (72%)	11%
VII	Protein synthesis (86%)	14%
VIII	Unknown (100%)	100
IX	Energy (47%)	42%
X	Cell Cycle and DNA Processing (38%)	50%
XI	Amino acid, nucleotide metabolism (27%)	33%

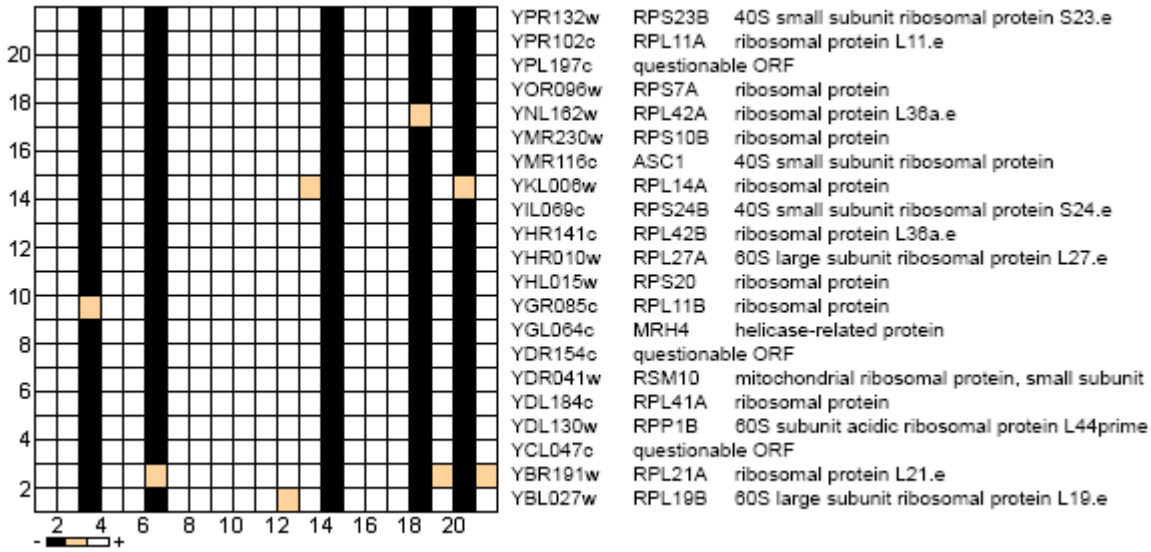

Figure 2. The pattern of bicluster VII. The pattern of bicluster VII comprises 21 genes and 21 experiments, experiments include "anp1", "cem1", "cmk2", "dig1, dig2 (haploid)", "fus3, kss1 (haploid)", "pet127", "pfd2", "rpl20a", "sbh2", "spf1", "sst2 (haploid)", "ste18 (haploid)", "ste7 (haploid)", "utr4", "yap1", "yar014c", "yel033w", "yhr039c", "yil037c (haploid)", "yor006c", "terbinafine", the details of experiments can be seen in the experiment_list of Hughes datasets (15).

Units are initialized far away from the input dataset in comparison with other units, they then immediately become dead without learning chance any more in the whole learning process. The third, k-means produces fairly round clusters, resulting in inaccurate identification of close or geometrically shaped clusters. But the conventional k-means clustering algorithm has been studies very well, and some improved k-means algorithms which can be applicable to ellipse-shaped data clusters are developed. The similar research locates in fuzzy k-means clustering methods (25,26,27). The fourth, k-means algorithm is very expensive in clustering a massive data because much iteration is needed in obtaining a good cluster. More recently, Nittel et. al. (28) proposes to apply k-means algorithm to cluster massive datasets, scanning the dataset only once. Their algorithm splits the entire dataset into chunks, and each chunk can fit into the main memory. Then, it applies k-means on each chunk of data, and merges the clustering results by another k-means type algorithm. Good results are shown on a real dataset, however, no theoretical bounds on the results are established. Conclusively, one can overcome one drawback but not all.

Hierarchical clustering is another popular approach on analyzing microarray data (5,29). Although it is possible to obtain biologically meaningful results, some of its characteristics often complicate the use for clustering expression data (30), and some even prevent the hierarchical clustering from obtaining a meaningful subtree because of the uncertainty in cleaving a big hierarchy tree on a suitable level. Also clustering over all dimensions (conditions) may separate the biologically related genes from each other. These have been observed by comparison of several clustering methods which have been deployed in diverse datasets, e.g. cancer classification by Romualdi *et al.* (31), clinical databases by Hirano *et al.* (32).

The biclustering approach, which is applied in the present research, can overcome some drawbacks mentioned above. Biclustering is a local technique by nature, i.e., the algorithm try to find local, significant signals in dataset for finding biclusters which contain subset of rows and columns (genes and experiments in the present research). So the results of biclusters are transparent. In the present

research, the biclustering approach has gained several meaningful biclusters which show a higher homogeneity of gene functions, e.g. bicluster V and VII. Functions of some questionable ORFs in these biclusters can be deduced, although they are just putative. The biclustering approach has exhibited a promising application. But the biclustering strategy may meet several baffles, e.g. all genes in one bicluster are involved in unknown functional group, e.g. bicluster VIII in the present research. How to judge their functions? In another case, genes included in one bicluster relate to more than one functional group, e.g. bicluster VI. How to judge the unknown genes in this cluster even if there is a dominant functional group within the pattern? Can we ignore the minor functional group? These are our further focuses on the field of clustering analysis of microarray data.

6. REFERENCES

1. Lin S.M., & Johnson K.F.: Methods of Microarray Data Analysis. *Kluwer Academic*, pp 137-150 (2002)

2. Slonim D., Tamayo P., Mesirov J.P., Golub T.R. & Lander E.S.: Class Prediction and discovery using gene expression data. *Fourth Annual Inter. Conf. on Computational Molecular Biology (RECOMB)* (2000)

3. Huels C., Muellner S., Meyer H.E. & Cahill D.J.: The impact of protein biochips and microarrays on the drug development process. *Drug Discov Today* 7, S119-S124 (2002)

4. Cheung Y.M.: K*-Means: Anew generalized k-means clustering algorithm. *Pattern Recog. Lett.* 24, 2883-2893 (2003)

5. Eisen M., Spellman P., Brown P. & Botstein D.: Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 95, 14863-14868 (1998)

6. Cheng Y. & Church G.M.: Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol* 8, 93-103 (2000)

7. Lazzeroni L. & Owen A.: Plain models for gene expression data. *Statistica Sinica* 12, 61-86 (2002)

8. Ben-Dor A., Friedman N. & Yakhini Z.: Class discovery in gene expression data. *Proc. Fifth Annual Inter. Conf. on Computational Molecular Biology (RECOMB)* (2001)

9. Ben-Dor A. & Chor B.: Discovering local structure in gene expression data: The order-Preserving Submatrix Problem. *J Comput Biol* 10, 373-384 (2003)

10. Tanay A., Sharan R. & Shamir R.: Discovering statistically significant biclusters in gene expression data. *Bioinformatics* 18, Suppl 1, 136-144 (2002)

11. Sheng Q., Moreau Y. & Moor B.D.: Biclustering Microarray data by Gibbs sampling. *Bioinformatics* 19, Supl 2, 196-205 (2003)

12. Caligano A., Stolovitzky G. & Y. Tu: Analysis of gene expression microarrays for phenotype classification. *Proc Intell Syst Mol Biol* 8, 75-85 (2000)

13. Wu C.J., Fu Y.T., Murali T.M. & Simon K.: Gene Expression Module Discovery Using Gibbs Sampling. *Genome Informatics* 15, 239-248 (2004)

14. Lockhart D., Dong H., Byrne M., Follettie M., Gallo M., Chee M., Mittmann M., Wang C., Kobayashi M., Horton H. & Brown E.: Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 14, 1675-1680 (1996)

15. Hughes T.R., Marton M.J., Jones A.R., Roberts C.J., Stoughton R., Armour C.D., Bennett H.A., Dai H., He Y.D., Kidd M.J., King A.M., Meyer M.R., Slade D., Lum P.Y., Stepaniants S.B., Shoemaker D.D., Gachotte D., Chakraburtty K., Simon J., Bard M. & Friend S.H.: Functional Discovery Via a Compendium of Expression Profiles. *Cell* 102, 109-126 (2000)

16. Schmidt U., Lehmann K. & Stahl U.: A novel mitochondrial DEAD box protein (Mrh4) required for maintenance of mtDNA in Saccharomyces cerevisiae. *FEMS Yeast Res* 2(3), 267-76(2002)

17. Planta R.J. & Mager W.H.: The list of cytoplasmic ribosomal proteins of *Saccharomyces cerevisiae*. *Yeast* 14(5), 471-7 (1998)

18. Heitman J., Movva N.R. & Hall M.N.: Targets for cell cycle arrest by the immunosuppressant rapamycin in yeast. *Science* 253, 905-909(1991)

19. Schreiber S.L. & Crabtree G.R.: The mechanism of action of cyclosporin A and FK506. *Immunol Today* 13, 136-142(1992)

20. Foury F.: Human genetic diseases: a cross- talk between man and yeast. *Gene* 195, 1-10(1997)

21. Sternmetz L.M., Scharfe C., Deutschbauer A.M., Nokranjac D., Herman Z.S., Jones T., Chu A.M., Giaever G., Prokisch H., Oefner P.J. & Davis R.W.: Systematic screen for human disease geens in yeast. *Nat Genet* 31, 400-404(2002)

22. Lum P.Y., Armour C.D., Stepaniants S.B., Cavet G., Wolf M.K., Butler J.S., Hinshaw J.C., Garnier P., Prestwich G.D., Leonardson A., Garrett-Engele P., Rush C.M., Bard M., Schimmack G., Phillips J.W., Roberts C.J. & Shoemaker D.D.: Discovering Modes of Action for Therapeutic Compounds Using a Genome-Wide Screen of Yeast Heterozygotes. *Cell* 116, 121–137(2004)

23. Rine J., Hansen W., Hardeman E. & Davis,R.W.: Targeted selection of recombinant clones through gene dosage effects. *Proc Natl Acad Sci USA* 80, 6750–6754(1983)

24. Dettling M. & Bühlmann P.: Supervised clustering of genes. *Genome Biol* 3(12), 0069.1–0069.15(2002)

25. Dembele D. & Kastner P.: Fuzzy C-means method for clustering microarray data. *Bioinformatics* 19, 973–980(2003)

26. Gasch A. & Eisen M.: Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology* 3(11), 0059.1–0059.22(2002)

27. Arima C. & Hanai T.: Gene Expression Analysis Using Fuzzy K-Means Clustering. *Genome Informatics* 14, 334-335(2003)

28. Nittel S., Leung K.T. & Braverman A.: Scaling clustering algorithms for massive data sets using data stream. In Umeshwar Dayal, *Proceedings of the 19th International Conference on Data Engineering, Bangalore, India. IEEE Computer Society*(2003)

29. Jeremy T., Alejandro M. & Werner S.: Hierarchical model-based clustering of large datasets through fractionation and refractionation. *Inf Systems* 29, 315-326 (2004)

30. Sherlock G.: Analysis of large-scale gene expression data. *Curr. Opin. Immunol.* 12, 201-205(2000)

31. Romualdi C., Campanaro S., Campagna D., Celegato B., Cannata N., Toppo S., Valle G. & Lanfranchi G.: Pattern recognition in gene expression profiling using DNA array: a comparative study of different statistical methods applied to cancer classification. *Hum Mol Genet* 12, 823-836 (2003)

32. Hirano S., Sun X. & Tsumoto S.: Comparison of clustering methods for clinical databases. *Inform Sciences* 159, 155-165 (2004)

Key Words: Biclustering, Yeast, Gene, Pattern

Send correspondence to: Dr Maosheng Cheng, Department of Bio-pharmaceutics, School of Pharmaceutical Engineering Shenyang Pharmaceutical University, Shenyang 110016, China, Tel: 0086-024-23894287, Fax: 0086-024-23995043, E-mail: maoshengcheng@tom.com; or Dr Jinghai Zhang, Department of Bio-pharmaceutics, School of Pharmaceutical Engineering Shenyang Pharmaceutical University, Shenyang 110016, China, Tel: 0086-024-23843711-3512, Fax: 0086-024-23843711-3641, E-mail: zhang.jinghai@tom.com