[Frontiers in Bioscience 10, 1619-1627, May 1, 2005]

A NEW STRATEGY OF COOPERATIVITY OF BICLUSTERING AND HIERARCHICAL CLUSTERING: A CASE OF ANALYZING YEAST GENOMIC MICROARRAY DATASETS

Daqing Mao[1], Yi Luo[2], Jinghai Zhang[1] and Jun Zhu[1]

[1] Department of Biochemistry, School of Pharmaceutical Engineering, Shenyang Pharmaceutical University, Shenyang 110016, China, [2] State Key Laboratory of Pollution Control and Resource, School of Environment, Nanjing University, Nanjing 210093, China

TABLE OF CONTENTS

1. ABSTRACT

Hierarchical clustering is difficult to be deployed effectively in finding meaningful subtrees since genes rarely exhibit similar expression pattern across a wide range of conditions. It is also difficult to find a suitable level in cleaving a big hierarchy tree. Biclustering is a promising methodology in the field of the analysis of gene expression data of genechip. Generally it can be employed in identification of gene groups, which show a coherent expression profile across a subset of conditions. But in some cases of biclustering analysis of gene expressions, the genes in one bicluster are involved in more than one functional group, or all genes in one bicluster are involved in unknown functional group (e.g. pattern VI and VIII in our studies). Then, how to predict the function of genes in these patterns? In the present research, we developed a new strategy of combining both of the clustering methods, hierarchical clustering and biclustering. The reserved conditions in datasets for hierarchical clustering were elicited according to the conditions in biclusters, and after hierarchical clustering, more detailed results in predicting unknown genes in certain patterns were obtained. This strategy of cooperating both of the methods during clustering procedure should be an effective guideline for functional predictions.

2. INTRODUCTION

Recent advances of microarray technology of high throughput profiling of gene expression have catalyzed an explosive growth in functional genomics for the aim of the elucidation of genes that are differentially expressed in various tissues or cell types across a range of experimental conditions, and the data analysis techniques have been intensively studied. 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, which include but not limited to hierarchical clustering (1, 2), self-organizing maps (3, 4), and have been widely applied in public website, e.g. http://ep.ebi.ac.uk/EP/EPCLUST/ etc. The applicability of clustering in prediction of gene

functions is based on the hypothesis that similar expression profiles imply a functional relation in biological activities (5). As a result, the quality of the clusters has often been evaluated by their correlations to the known genes' function groups. Although these studies have successfully shown that genes participating in the same biological process have similar expression profiles, there are still some deficiencies in preventing these methods from solving a large dataset analysis, e.g. It is difficult to detect a certain level of cleavage in hierarchy tree, and 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.* (6), clinical databases by Hirano *et al.* (7).

Biclustering is a promising methodology in this field, and might be a powerful measurement to solve the above all problems. 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 (8). The approximate uniformity in a submatrix in gene expression data can be detected by another model Plaid developed by Lazzeroni and Owen (9), which they use 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 (10). Ben-Dor *et al.* also introduced the model of Order preserving submatrix (11). Tanay *et al.* defined a bicluster as a subset of genes that jointly respond across a subset of conditions for reducing the biclustering problem (12). A biclustering algorithm based on Gibbs sampling has been successfully developed and implemented by Sheng *et al.* (13), 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 (14). Also with Gibbs sampling, Wu *et al.* developed a running scheme and expand its application to biclustering continuous gene expression data (15). Liu *et al.* presented a technique named Smart Hierarchical Tendency Preserving clustering, based on a bicluster model, Tendency Preserving clusters. They incorporated Gene Ontology information to subtrees directly (16).

But most biclustering strategy may meet several baffles, e.g. all genes in one bicluster are involved in unknown functional group. How to judge their functions? In another case, genes included in one bicluster relate to more than one functional group. 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? In order to solve these problems, our strategy to this situation is a cooperation of biclustering and hierarchical clustering. The preferable results were obtained.

3. MATERIALS AND METHODS

3.1. Microarray gene datasets

Whole-genome expression profiling, facilitated by the development of DNA microarrays (Lockhart *et al.* 1996) (17), 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 we can directly apply a single assay to measure the interaction items of unknown or known genes in finding functions, but also the idea of "compendium" can be used for the purpose of predicting or diagnosing etc (17, 18). Hughes datasets as a comprehensive datasets of reference profiles were created for the aim of analyzing functions, testing drug target, etc (18). The reference datasets of three-hundred full-genome expression profiles in *S.cerevidiae* 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) 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 informations known as for checking the exact function related to each gene of each pattern were stored into tables of local database.

3.3. Biclustering

3.3.1. Storing the datasets in database (SQL)

Several tables were built for storing Hughes datasets, and SQL tools were used to elicit the values (logInternsity and P value) for biclustering. For the purpose of information searching index of patterns, some tables were built to store functional description of genes in MIPS classes. These tables contain functional groups of Metabolism, Energy, Cell cycle and DNA processing, Transcription, Protein synthesis, Protein fate (folding, modification, destination), Cellular transport and transport mechanisms, Cellular communication/signal transduction mechanisms, Cell rescue, Defense and

Table 1. Number and proportion of ORFs and experiments (conditions) 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%)

virulence, Regulation of interaction with cellular environment, Cell fate, Transposable elements, Viral and plasmid proteins, Control of cellular organization, Subcellular localization, Protein activity regulation, Protein with binding function or cofactor requirement (structural or catalytic), Transport facilitation, etc.

3.3.2. Preprocessing of biclustering in matlab

The main objective of preprocessing of biclustering was to reduce noise. In the previous step, the values of "logIntensity" and "P value" of genes' expression were elicited respectively. After that, the dataset was transfered and stored into text file. The following was to load them in Matlab in matrix form, then filtered datasets in MATLAB to delete the data which show less standard deviation. In this step, a certain filtering ratio was used for leaching. Afterward, variations of each gene along all of experiments were examined to delete those ORFs which hold a certain range of P value in less than 100 experiments ($P \leq 0.01$, experiments$\leq$100). After that, the filtered expression dataset was discretized into fixed number of bins in the last of this step (13).

3.3.3. Running the biclustering scripts

The biclustering algorithm was based on the Gibbs sampling strategy. In this method, a greedily iterative searching was applied to find interesting patterns in the matrices, and probabilistic models were proposed in which matrix rows (genes in this case) and columns (experimental conditions) were divided into clusters, and there were linked probabilities between these clusters. These linked 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 (13).

3.3.4. Checking functions of patterns through MIPS database

This was a time-consuming work. In this step, each table was checked in the MIPS database for annotating function of patterns.

3.4. Hierarchical clustering

3.4.1. Eliciting conditions according to biclusters

The datasets were restocked in a local database after preprocessing of biclustering and filtering on both of LogIntersity and P value levels. Then the conditions (experiments) involved in the interesting patterns of biclusters were elicited, e.g. pattern VI, VIII, in present experiment, etc. Total 75 conditions were reserved in the new basal datasets.

3.4.2. Hierarchical clustering and subtree obtaining

The datasets were analyzed by hierarchical clustering (http://ep.ebi.ac.uk/EP/EPCLUST/). Then the output of a big hierarchy tree was cleaved according to the patterns of interesting bicluster. In this step, firstly, the ID and accession number of ORFs were redeposited in database, simultaneously the sequence of ID was ordered according to the hierarchy tree. Afterwards, the interesting ORFs can be detected by SQL queries. Secondly, subtrees were obtained by cleaving on certain level of linkage according to the genes position in the sequence. The pattern VI and VIII from hierarchical clustering were obtained by this way.

4. RESULTS

4.1. Biclustering

4.1.1. Biclusters

Biclustering algorithm enables the detection of multiple biclusters, through the way of masking the genes or the experiments selected for the found biclusters and performing the algorithm on the rest of the data. 11 biclusters were found in the original datasets. Table 1 lists the compositions of biclusters which consist of various genes and experiments.

4.1.2. Relevant functions of biclusters

Through checking in MIPS database restocked in tables of local SQL database, the information of relevant gene functions in each bicluster was obtained. E.g. in bicluster I, 37% of the genes (199/537) participate in cell metabolism, 62% of the genes (333/537) involve in unknown functions, and less than 1% of the genes (5/537) is classified into other function groups. Most parts in this pattern are those of "open reading frames" with unknown or unclassified function, also in other several patterns. See table 2 below for the details of gene functions involved in each bicluster.

Pattern I comprises nearly 2 parts, the known ORFs with the same function and the unknown ORFs, the same for pattern V and VII. It is facilitated to predict function of the unknown ORFs in these patterns on the postulate that the same function exerts the same behavior. But how to predict the 100% unknown ORFs in the pattern VIII and the other patterns in which includes more than one kind of functions? These tangles can be settled down by a combination of biclustering and hierarchical clustering. The analysis of pattern VI and pattern VIII are shown below as examples.

4.2. Pattern VI

4.2.1. Pattern VI from biclustering

The details of pattern VI are shown in figure 1, containing 35 ORFs in total. In figure 3, the proportions of three parts in this pattern are shown (according to MIPS),

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%

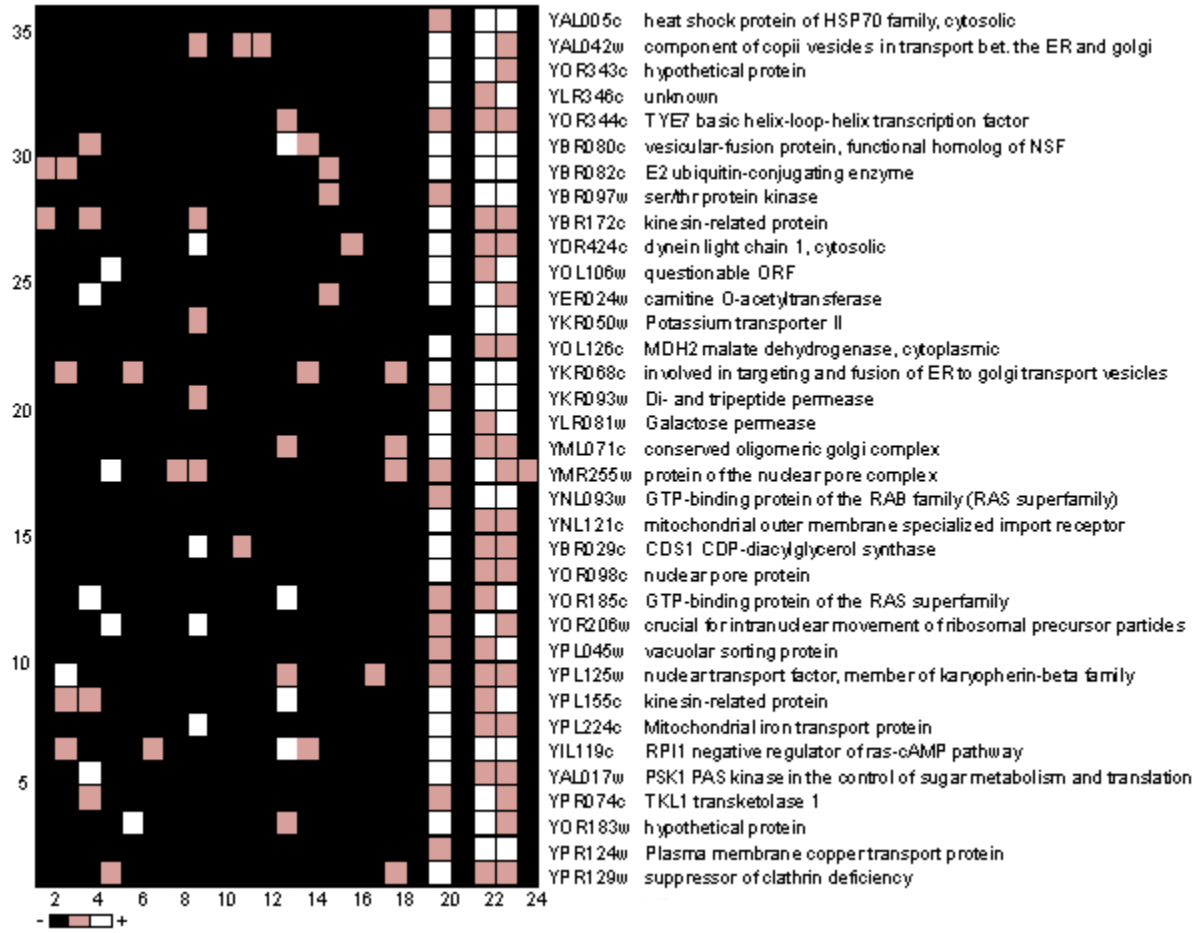

Figure 1. The pattern of bicluster VI comprises 35 ORFs and 23 experiments. 35 ORFs are involved in 3 functional groups, cellular transport and transport mechanism, metabolism and unknown function. The experiments include "anp1", "bim1", "bul1", "fpr1", "gln3", "ost3", "pep12", "pfd2", "rpl12a", "rpl20a", "rpl34a", "sap30", "sbh2", "spf1", "vac8", "vma8", "yar014c", "yel033w", "yel067c", "yer084w", "yml005w", "ymr293c" and "yor009w", the details can be seen in experiment_list of Hughes datasets (18).

72% represents the functional group I of cellular transport and transport mechanism, 17% represents the functional group II of metabolism and 11% represents unknown function. Obviously, it seems less evident to assign the unknown ORFs to the functional group I or II. So we deployed the hierarchical clustering for the elicited datasets which contain the same conditions with interesting patterns of biclusters after preprocessing of biclustering.

4.2.2. Pattern VI from hierarchical clustering

On the basis of the preprocessing datasets of biclustering and the interesting patterns of biclusters, the datasets including all rows (genes) and 75 columns (conditions) involved in the interesting patterns were restocked. A hierarchical clustering with an average linkage strategy was deployed on the datasets in the website http://ep.ebi.ac.uk/EP/EPCLUST/. After depositing the

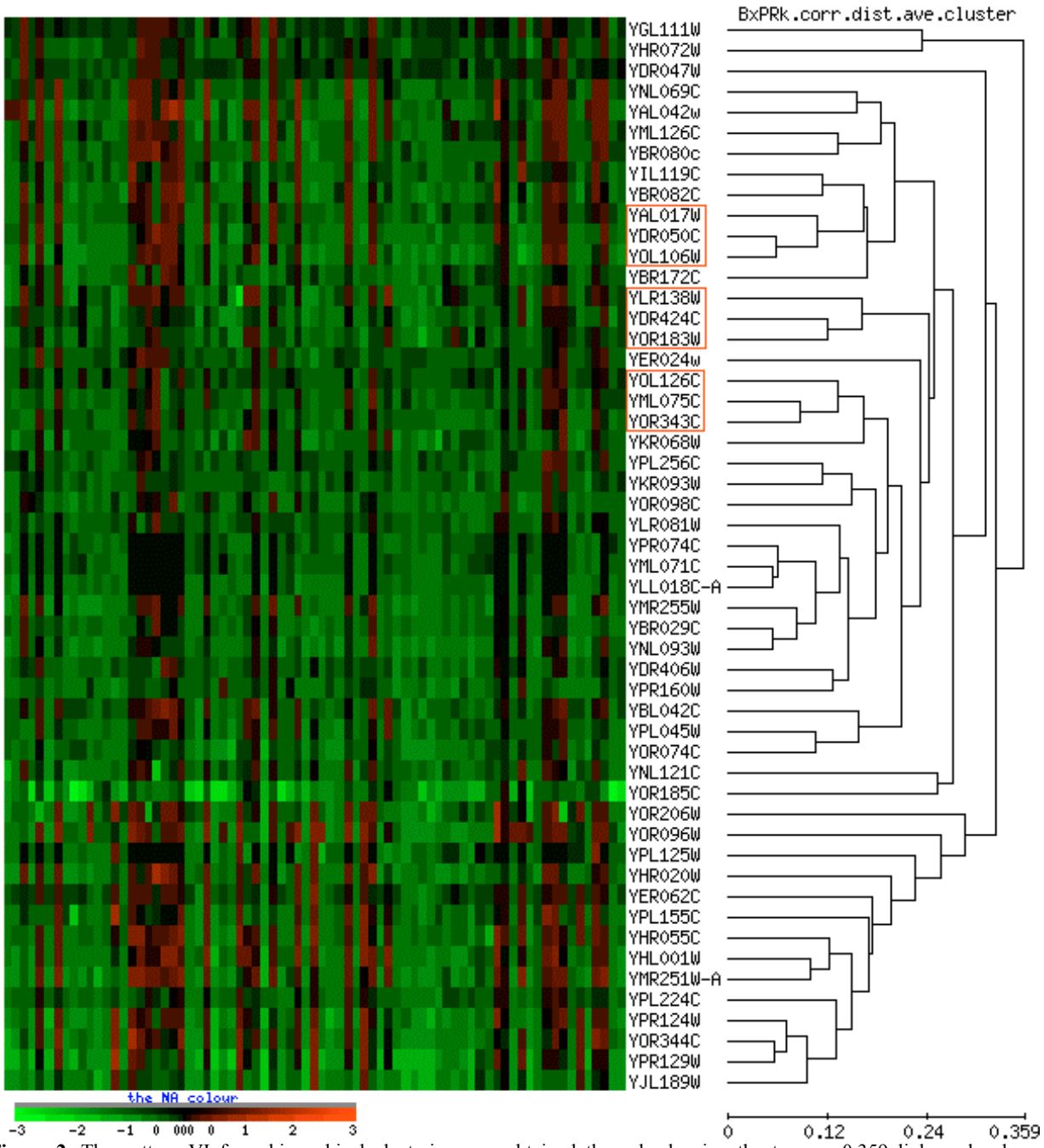

Figure 2. The pattern VI from hierarchical clustering was obtained through cleaving the tree on 0.359 linkage level, and comprises 52 ORFs. It holds the higher similarity with pattern VI from biclustering. The details of ORF functions are shown in figure 3. The three red rectangles are shown for 3 unknown ORFs and linked genes (ORFs).

ordered accession number of ORFs of hierarchy tree in a table of local database (SQL), the interesting ORFs were checked by SQL queries, and also the tree was cleaved on a certain level of linkage according to the elements in the pattern VI of biclustering. Therefore a correlated subtree similar to a certain bicluster was found. The subtree shown in figure 2 is pattern VI from hierarchical clustering (The name is originated according to the corresponding bicluster). In figure 3, the compositions of pattern VI from biclustering and pattern VI from hierarchical clustering are shown. Although they do not fully match each other, more details are complemented from the hierarchy subtree to the bicluster. These are important to predict the function of the unknown ORFs in the pattern. Furthermore, both of the clustering methods show a higher similarity, i.e. 89% of bicluster VI overlaps with 60% of hierarchy subtree VI. The similarity mostly depends on the selected conditions. If the elicited conditions from the original dataset for

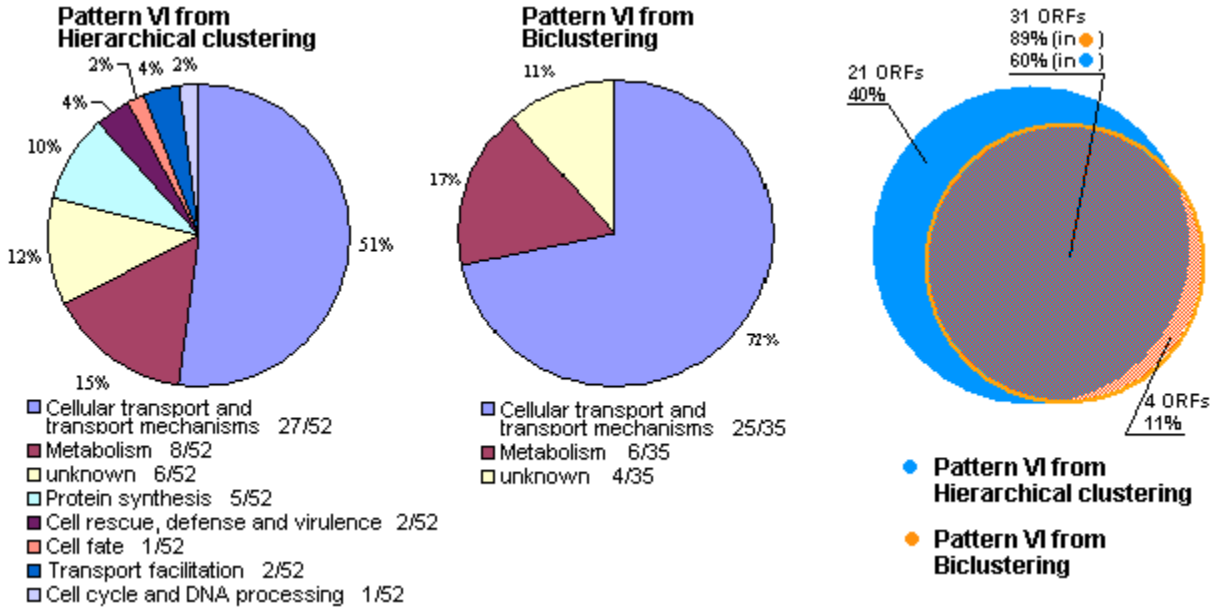

Figure 3. Functional compositions of pattern VI. The same color shows the same function group in both patterns. The pattern from hierarchical clustering comprises 52 ORFs, and the details of functions and proportions are shown in this graph. The pattern from biclustering comprises 35 ORFs categorized to 2 function groups and unknown category, and a higher similarity is shown between the two patterns from both clusterings. The overlap-part contains 31 ORFs, it is 89% in pattern from biclustering and 60% in pattern from hierarchical clustering respectively.

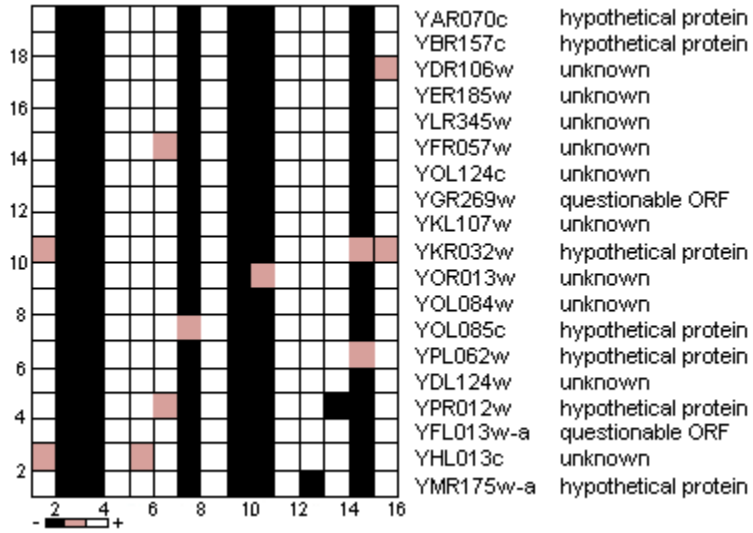

Figure 4. The pattern of bicluster VIII. It comprises 19 unknown ORFs and 15 experiments (conditions). The experiments include “ard1”, “ecm31”, “eft2”, “erg2”, “erg6”, “hdf1”, “mrt4”, “rnr1 (haploid)”, “rpl27a”, “rpl6b”, “ste11 (haploid)”, “ste5 (haploid)”, “yhl029c”, “ymr025w”, “erg11 (tet promoter)”, the details can be seen in experiment_list of Hughes datasets (18).

hierarchical clustering resemble mostly to the conditions in the pattern of bicluter, the proportion of overlap-part between the subtree and the bicluster should be higher.

In the pattern VI from hierarchical clustering, the unknown ORF “YOL106w” is extremely correlated to “YDR050c” and “YAL017w” which both hold the function involved in “C-compound and carbohydrate metabolism”. Therefore the function of “YOL106w” can be predicted. Same procedure for “YOR343c”, it is correlated to “YML075c” and “YOL126c”, and is predicted with the same function as the both latter ORFs, i.e. “Carbohydrate or lipid metabolism”. The unknown ORF “YOR183w” is correlated to “YDR424c”, and is predicted with the same function as the latter, i.e. “Cellular transport and transport mechanisms”, and so on.

The comparability between the pattern VI from

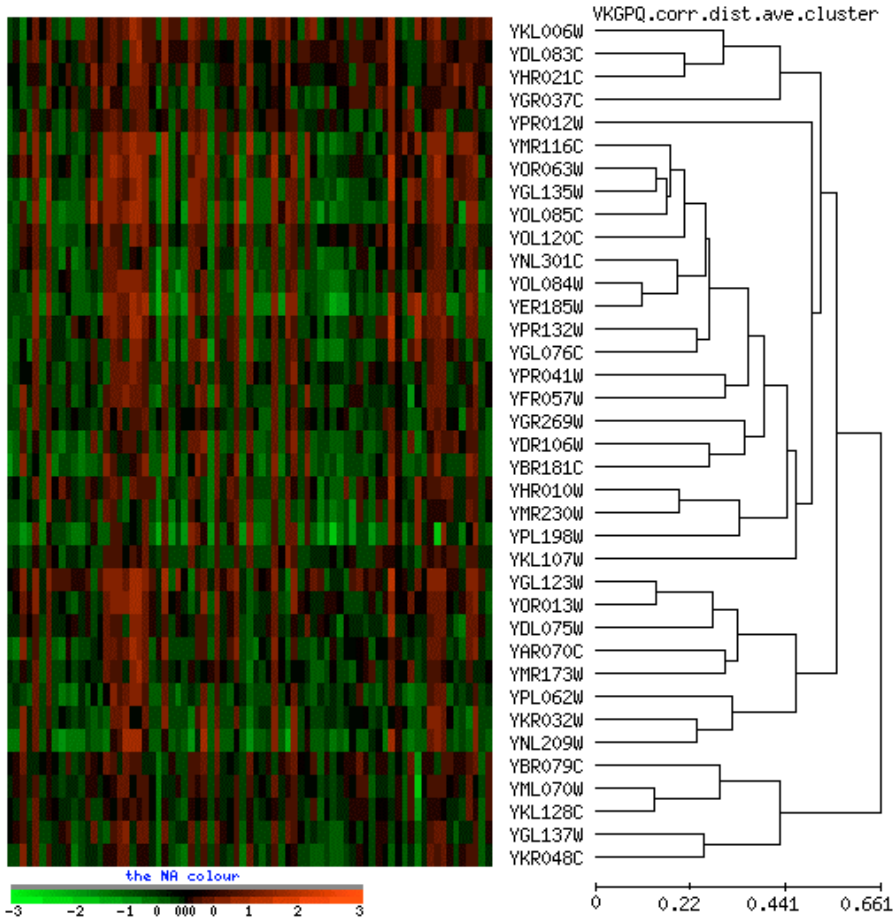

Figure 5. The pattern VIII from hierarchical clustering was obtained through cleaving the tree on 0.661 linkage level, and comprises 37 ORFs. It holds a certain similarity with pattern VIII from biclustering. The details of these ORF functions and proportions are shown in figure 6.

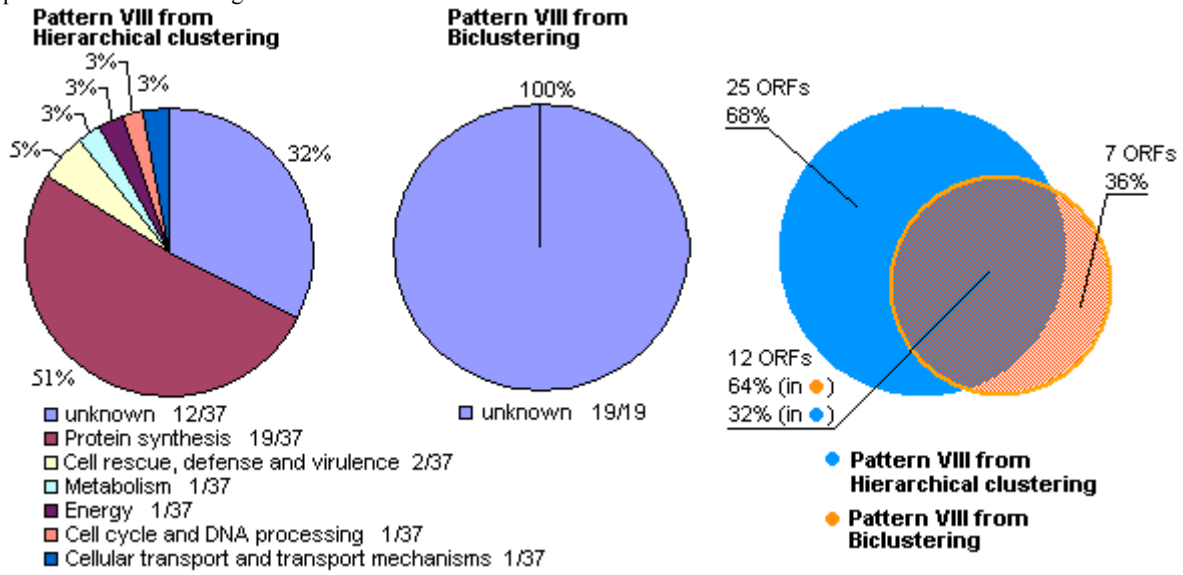

Figure 6. Functional composition of pattern VIII. The pattern VIII from hierarchical clustering comprises 37 ORFs. The proportions of functions are shown in this graph. The dominant function in pattern VIII from hierarchical clustering is "protein synthesis" (51%). The pattern VIII from biclustering comprises 19 unknown ORFs, and shows a certain similarity with the pattern from hierarchical clustering. The overlap-part contains 12 unknown ORFs which are 64% in pattern VIII from biclustering and 32% in pattern VIII from hierarchical clustering.

biclustering and from hierarchical clustering was also analyzed, as shown in figure 3. The intersection part is dominant in the patterns from both of the clusterings. And the most important is to detect the information relating to functional prediction of genes, regardless whether the intersection part is dominant or not.

4.3. Pattern VIII

The details of pattern VIII from biclustering are shown in figure 4**,** and all 19 ORFs in the pattern fall into group of questionable proteins or group with unknown function (according to MIPS). We cannot make a functional prediction of the pattern only on the basis of this bicluster.

The pattern VIII from the hierarchical clustering was obtained according to the pattern of bicluster VIII by cleaving on a certain level of linkage. The corresponding pattern VIII from hierarchical clustering is shown in figure 5 which consists of 37 ORFs that mainly relate to the functional group of "protein synthesis" (51%) according to MIPS. 12 out of the 37 ORFs are unknown ORFs (32%) and they are also elements in pattern VIII of biclustering (12/19=64%). These results shown in figure 5 provide more details of information about functional analysis than only depending on biclustering. So the function of those unknown genes may be predicted relating to "protein synthesis" according to neighbor-joining in hierarchy subtree. For example, YOL085c as an unknown ORF, which is allocated in an unknown function pattern of bicluster is linked directly to YGL135w and YDR063w in hierarchy subtree. The latter two genes participate in the functional group of "protein synthesis". Not only this ORF, but also total 12 ORFs in pattern VIII from biclustering are linked to the genes with the function of "protein synthesis" in this hierarchy subtree. So we can make a functional prediction for pattern VIII on the basis of not only biclustering but also hierarchical clustering. Certainly, these are only from prediction, and still need to be validated in biological experiments. The details of proportions in the two patterns from both of the clusterings can be seen in figure 6.

In figure 6, the third graph is on the show of an interaction between the patterns from biclustering and hierarchical clustering. The overlap-part contains 12 unknown ORFs which are 64% in the pattern of biclustering and 32% in the pattern of hierarchical clustering respectively.

5. DISCUSSION

5.1. Comparability of biclustering and hierarchical clustering

Biclustering and hierarchical clustering show a certain level of similarity in corresponding patterns (e.g. pattern VI). If the exact conditions of biclusters are elicited for hierarchical clustering analysis, the similarity between the corresponding patterns from biclustering and hierarchical clustering can be much higher. And the most important information is the linkage between each gene in hierarchy subtree. As we have known in normal strategy, the hierarchical clustering is difficult to be deployed effectively in finding meaningful subtrees since genes rarely exhibit similar expression pattern across a wide range of conditions, and it is also difficult to find a suitable level in cleaving a big hierarchy tree. So the strategy in our research is a significant supplement from hierarchical clustering to biclustering in detailed information of linked genes, and it is also a felicitous methodology to overcome the drawback of hierarchical clustering in normal strategy.

5.2. Cooperativity of biclustering and hierarchical clustering

For biclustering, it can generally be employed in identification of gene groups that show a coherent expression profiles across a subset of conditions. Genes that exhibit similar expression profiles may imply strong correlations between their functions in the biological activities and the soundness of clustering in the analysis of gene expression profiles. Genes' functional prediction are based on this hypothesis. But in some cases of biclustering analysis of gene expression, the genes involve in one pattern of bicluster referring to more than one functional group, or all genes involve in one bicluster referring to unknown functional group, e.g. pattern VI and VIII mentioned above. How to predict the function of these patterns? In the present research, we combined both of the two clustering methods, biclustering and hierarchical clustering, and gained better results. These mainly depend on the features of both methods. First, biclustering of gene expression data is a promising methodology, which can be deployed in identification of gene groups on subset of conditions. These subsets of conditions show us an important range of datasets for hierarchical clustering. Therefore the accuracy of hierarchy tree from hierarchical clustering can be enhanced, and the two corresponding patterns from two methods behave a strong similarity. Second, hierarchical Clustering links each gene one by one on the average or complete linkage strategy finally to form one big tree. Consequently, more details of genes are related to the final results in our new strategy, these can be used for gene predictions. Therefore, we conclude that the cooperativity between biclustering and hierarchical clustering is distinct.

5.3. Robustness for functional predictions and clues for interactomic analysis

A perfect pattern of bicluster is involved in a general group of function, e.g. "protein synthesis" or "metabolism" etc. Hierarchy subtrees display the details of interaction and similarity of each branch through the way of linked genes. However, it is deficient not only in pathway and interactome analysis of genes or proteins in biclustering, but also in preprocessing of analyzing genes in hierarchical clustering. The combination of both methods can overcome these drawbacks. Most unknown genes can be predicted by this new strategy of combining both clustering methods, although the validating experiments are still needed.

Interactomic analysis *in silico* generally bases on several methods, i.e. genes fusion, genes neighborhood, phylogenetic profiles, *etc*. This new strategy of

combination of the two methods can provide some clues in interactomic analysis on transcriptional level by the neighbor-joining of genes in hierarchy trees. This will be focused on our research in future.

In fact, no one method is suitable for any kind of datasets. Considering the accuracy, it is better to make a combination of different methods even for a single kind of datasets. However, the strategy in our study is a combination in essence of both of the methods, not in a simple form, and it has put up a promising application.

6. ACKNOWLEDGEMENTS

The author wishes to thank Dr. SHENG for her kindly help in analyzing algorithm, including resetting parameters and discussing the results of biclusters. These are active factors for achievements of this experiment.

7. REFERENCES

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

2. M. Eisen, P. Spellman, P. Brown & D. Botstein: Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 95, 14863-148678 (1998)

3. Tamayo P., Slonim D., Mesirov J., Zhu Q., Kitareewan S., Dmitrovsky E., Lander E.S. & Golub T.R.: Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation *Proc Natl Acad Sci USA* 96, 2907-2912 (1999)

4. Kaski S., Nikkila J., Toronen P., Castren E. & Wong G.: Analysis and visualization of gene expression data using self-organizing maps, *Proceedings of NSIP-01, IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing* (2001)

5. Brown M.P., Grundy W.N., Lin D., Cristianini N., Sugnet C.W., Furey T.S., Ares M. & Haussler D.: Knowledge-based analysis of microarray gene expression data by using support vector machine. *Proc Natl Acad Sci USA* 97, 262-267 (2000)

6. 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)

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

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

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

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

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

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

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

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

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

16. Liu J., Yang J. & Wang W.: Gene Ontology Friendly Biclustering of Expression Profiles. *Proceeding of the IEEE Computational System Bioinformatics Conference* (2004)

17. 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)

18. 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)

Key Words: Biclustering, Hierarchical Clustering, Unknown ORF, Pattern

Send correspondence to: Dr Jinghai Zhang, Department of Biochemistry, School of Pharmaceutical Engineering, Shenyang Pharmaceutical University, Shenyang 110016, China, Tel: 0086-24-23843711-3512, Fax: 0086-24-23843711-3641, E-mail: zhang.jinghai@tom.com

http://www.bioscience.org/current/vol10.htm