Research on Microarray Dataset Mining

Miao Wang
School of Computer Science and Engineering
Northwestern Polytechnical University
Xi'an, China
Tel. (86)13909253548
riyushui@gmail.com

Xuequn Shang
School of Computer Science and Engineering
Northwestern Polytechnical University
Xi'an, China
Tel. (86)13319273686
shang@nwpu.edu.cn

Zhanhuai Li
School of Computer Science and Engineering
Northwestern Polytechnical University
Xi'an, China
Tel. (86)13709198435
lizhh@nwpu.edu.cn

ABSTRACT

With the rapid progress of bio-techniques of post genomic era, more and more bio-information needs to be analyzed. Using microarray data can reveal the structure of the transcriptional gene regulation processes. In this paper, we give an overview of recent research work of microarray data mining. We also introduce several woks which are developed, developing at present and future work, which are the focus of my PhD thesis.

1. INTRODUCTION

The execution of complex biological processes requires the precise interaction and regulation of thousands of molecules. The advent of high-throughput techniques has allowed the large-scale identification of components (genes, RNAs, and proteins), their expression patterns, and their biochemical and genetic interactions. Recent advances in microarray technologies have made it possible to measure the expression levels of genes simultaneously. Using microarray data can reveal the structure of the transcriptional gene regulation processes, which is called reverse engineering[1].

The microarray dataset can be seen as a matrix, denoted as *G* in real expression numbers which is shown in Table 1. The columns denote different experimental conditions or samples. The rows denote genes. In order to mine frequent patterns, microarray data would be converted each gene expression number into one of the three numbers: 1, -1 and 0, which denotes expressed, depressed and non-expressed, respectively, as shown in Table 2. Many obvious meaning of microarray data analysis are illustrated as follow. (1) Identify genes whose expression levels reflect biological processes of interest (such as development of aging). (2) Determine how the expression of any particular gene might affect the expression of other genes, e.g. several co-expressed genes may be composed to a protein. (3) It can provide clues for the function of genes or proteins of unknown role. (4) It can help biologist finding potential transcription factors.

However, unlike the traditional datasets, microarray dataset has its own characters. (1) The number of rows (genes) in microarray dataset far exceeds the number of columns (samples). For example, AGEMAP[2] is a highly standardized study of gene expression changes as a function of age in mice. AGEMAP has a total of 16,896 cDNA clones from only 16 tissues samples from each mouse. (2) The items (genes) in one sample are unique. Due to the large items in microarray datasets, mining procedure may be inherently costly in runtime and space usage. (3) According to biological interpretations, there are three types of regulation relations between genes: activation, inhibition and, and dependency. The definition of above three relations is different from traditional one. (4) There has large of noisy data in microarray data.

Table 1. An example of microarray dataset

	HiFA1	HiFA2	HiFA3	HiFA4
G1	2.39	-2.27	2.04	-1.6
G2	0.33	0.68	0.46	-0.06
G3	1.26	-1.70	1.58	-1.13
G4	0.77	1.01	0.73	-0.63
G5	-0.62	0.99	-0.87	0.42
G6	-1.03	0.02	-2.3	-0.31

Table 2. Converted microarray dataset

	HiFA1	HiFA2	HiFA3	HiFA4
G1	1	-1	1	-1
G2	0	1	0	0
G3	1	-1	1	-1
G4	1	1	1	-1
G5	-1	1	-1	0
G6	-1	0	-1	0

In a word, the research on microarray dataset mining aims to help biologist make use of the large volume information in microarray dataset effectively and efficiently. Due to the characters of microarray data, there have lots of challenging research issues in this area. My PhD thesis is focusing on designing some algorithms and addressing several challenging issues in this area. The following problems must be studied: Firstly, how to reduce the

VLDB 2010 PhD Workshop, September 13, 2010, Singapore

influence of inherent dimensionality problem in microarray data; Secondly, how to escape of noisy data; Thirdly, how to efficiently mining interesting patterns based on the character of microarray data; Finally, how to get patterns considering not only genes but also samples. In this paper, we will introduce our ideas, naïve solutions and some results in details.

The rest of this paper is organized as follows. The related research work is illustrated in Section 2. Section 3 presents the works we are focusing now and will research in the futures. We will make the conclusion in Section 4.

2. RELATED RESEARCH WORK

Until now, large numbers of efforts are worked in microarray analysis. Due to the paper space limit, the related work can not be discussed in detail. Recent advances in microarray technologies have generated larger volume of datasets, using which can reveal the functions and biological processes of genes. These datasets pose a great challenge for frequent pattern discovery algorithms. Therefore, efficient mining methods have been studied extensively to extract meaningful biological information from huge microarray datasets.

-Clustering. One of the widely used method to reveal the relationship among genes is clustering[3,4]. However, using cluster analysis to infer regulatory modules or biological function has several inherent limitations[5]. First, genes that are biologically related often are not related in their expression profiles[6]. Second, one gene may participate in more than one biological process or function. Third, the relationship between clusters can not be discovered.

-Association rule. Many researchers have been also focused on using association rules mining method to construct the gene regulatory network. An association rule has the form LHS=>RHS, where LHS and RHS are sets of items, the LHS occurs may result in the occurrence of RHS. Biologically speaking, the expression of one gene of gene sets may induce the expression of another gene or genes. For example, GENE1=>GENE2 (support 80%, confidence 100%), which means, when GENE1 is expressed, 100 percent of the time GENE2 is also expressed and both genes are expressed in 80 percent of the microarray experiments. Traditional method to generate association rules is to mine frequent items firstly. Many frequent pattern mining algorithms can be used to generate frequent items, such as [7,8]. Then high confident association rules can be obtained using frequent items. For example, ABC is a frequent item. Using it, we can obtain six potential rules: A=>BC, AB=>C, B=>AC, C=>AB, AC=>B, BC=>A. Based on our experiment, compared to generate frequent items, the obtaining high confident association rules is very time-consuming.

-Frequent (closed) pattern. The existing approaches of mining frequent closed pattern in microarray dataset are item enumeration and sample enumeration. Item enumeration approach generates frequent patterns by combination of items. As discussed above, the microarray dataset has larger items. As a result, item enumeration approach is very time consuming. Therefore, [9] proposed to mine frequent patterns by using sample enumeration, which explore the enumeration space by constructing projected transposed tables recursively[10], is another algorithm on the row enumeration tree, which utilizes set intersection operations on horizontal layout data. However, there are still several limitations with the existing microarray mining algorithms: (1) when the microarray dataset was dense or the number of samples is quite same as the number of genes, the existing algorithms are time-consuming. (2) The existing algorithms need to keep the historical frequent closed patterns in memory, which limits the scalability. (3) The patterns generated by the existing algorithm can not reveal the complex regulation relations from microarray data. According to biological interpretations, there are three types of expression relations between genes: positive expressed, negative expressed and unexpressed. However, previous algorithms neglect the negative expressed relations.

-Biclustering. Biclustering[11] has proved of great value of finding interesting patterns in microarray dataset, which records the expression data of each gene in different biological samples. Biclustering can identify the co-expression patterns of a subset of genes which may be relevant to a subset of the interesting samples[12]. An important example of the utility of biclustering is the discovery of transcription modules from microarray data, which denote groups of genes that show coherent activity only across a subset of all the conditions constituting the data set, and may reveal important information about the regulatory mechanisms operating in a cell[13]. The advantage of biclustering for microarray analysis can be shown as follows. First, the discretization techniques used in traditional frequent pattern mining paradigm results in loss of information. Biclutering will address the problem. Second, biclustering method can mine not only the co-express gene pattern, but also the details about their corresponding samples. The details are with remarkable biology significance.

However, the precondition of most bicluster algorithms is to know the specific characters of the dataset. [14] classified different types of biclusters into four categories, which is shown in Figure 1. Therefore, one biclustering mining method may only suitable to one category of bicluster. In order to get more accuracy, the category of microarray data need to be analyzed before microarray data mining. For instance, the MicroCluster adapt to the dataset with a scaling feature. The precondition required huge data analysis workload.

3. SEVERAL RESEARCH WORKS

In this section, several research works are proposed for discussion, which are developed, developing at present and future work.

3.1 Strong Association Rules Mining

Most of existing association rule mining algorithms consisted of two steps: discovery of frequent itemsets which satisfy with the user-defined support threshold and then generation association rules that satisfy with the confidence threshold from generated frequent itemsets. As discussed in Section 1, this two steps method is not very efficiently.

We present an algorithm, SAW[15], to generate high confident association rules without using frequent items. First, all the paired rules are generated. Second, larger strong association rules can be obtained by combing the above rules. The combined strategy is composed by two parts: forward combined, backward combined, which are show in Fig.2. Fig. 3 shows the process of SAW dealing with an example dataset. Using our method can avoid lots of unnecessary computing. For example, if A=>B is not strong rule,

according to the Apriori property, A=>BC must not be the strong one. Therefore, using frequent item ABC to produce strong rules A=>BC and AC=>B must be failed. However, using our method can avoid the unnecessary computing, so the efficiency can be improved.

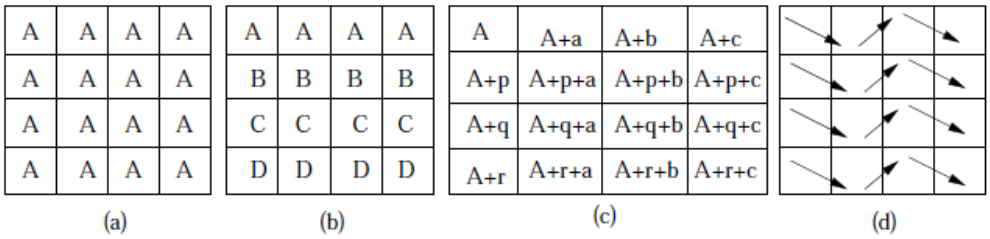

Figure 1: Types of biclusters: (a) Constant value biclusters; (b) Constant row biclusters; (c) Coherent value biclusters; (d) Coherent evolution biclusters.

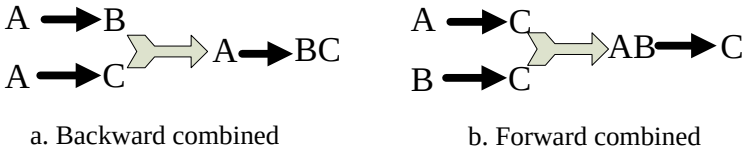

Figure 2. The combined strategy

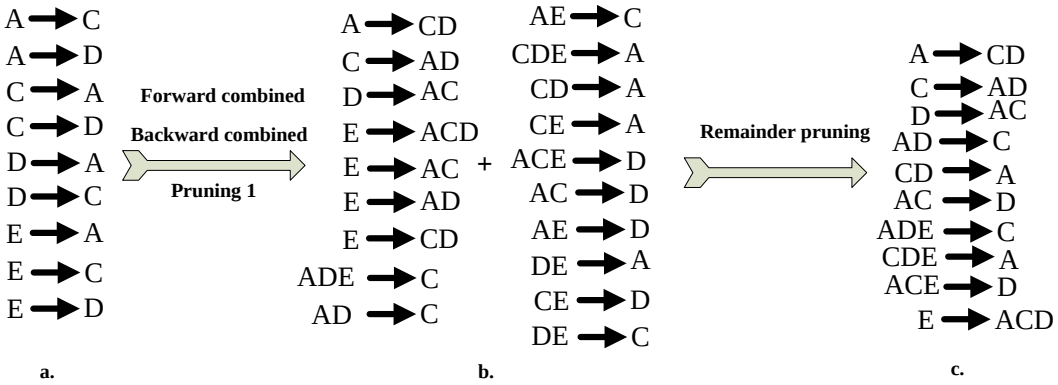

Figure 3. An example of rule combined method

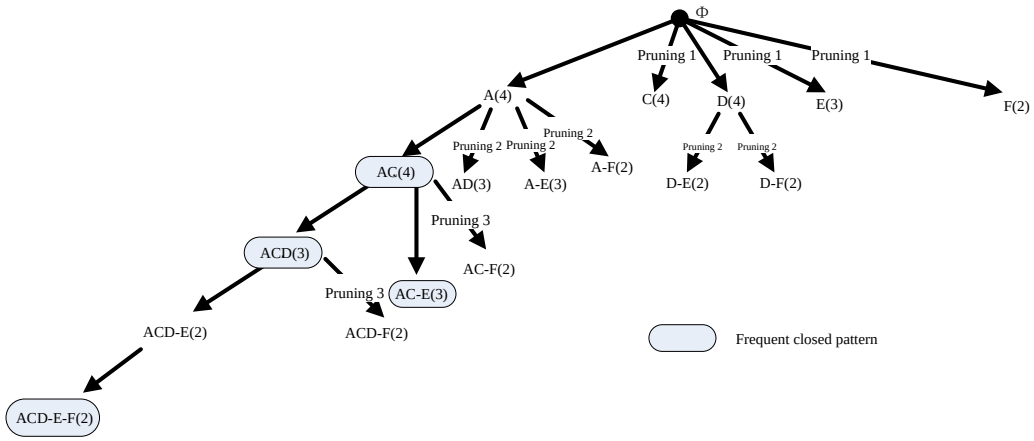

Figure 4. The process of WIBE dealing with example dataset.

3.2 Frequent Closed Pattern Mining

As discussed in Section 2, according to biological interpretations, there are three types of expression relations between genes: positive expressed, negative expressed and unexpressed. However, previous algorithms neglect the negative expressed relations. The three types of relations between genes *X* and *Y* can be respectively defined as follows:

1. X and Y is positive expressed (*XY*) if *X* and *Y* are all expressed; or if *X* and *Y* are depressed;
2. X and Y is negative expressed ($X\bar{Y}$) if *X* is expressed and *Y* is depressed; or if *X* is depressed and *Y* is expressed;
3. *X* and *Y* is unexpressed if *X* is expressed or depressed, then *Y* shows both expressed and depressed state, or remains unchanged.

Traditional frequent patterns mining algorithms in microarray dataset do not distinguish positive and negative ones. The reason may be the complicated computing. Each gene has two kind of expression, if there are *N* genes, the number of patterns they generate would be $C_N^1 + C_N^2 + \cdots + C_N^{N-1} = 2^N - 2$. Therefore, the efficiency is poor. Based on the relation among genes, the gene pattern *{A-BC}* has the same biological significance as *{-AB-C}*. If both of them can be mined as one, it would be easy to generate. The approach for this procedure in our method is described as the following examinations:

1) The first item of extending pattern is set to positive one;
2) If the first item of extending pattern is negative in some transactions, the candidate in these transactions is set to opposite one.

Based on the above definitions and examinations, we develop an algorithm, WIBE[16], to mine frequent closed patterns without candidate maintenance and backward checking. In our method, each item in the generated closed patterns would be set a weight. If the weight of each item in candidate is the same, the candidate pattern can be pruned. We also consider both positive and negative relations between genes, which can deliver more valuable potential regulation information of gene network construction. The giving example is illustrated as shown in Fig. 4.

3.3 Differential Frequent Bicluster Mining

A broad range of solutions to the clustering problem use cancerous dataset as their experimental data, based on our observation, the proceeding progress of their method can not represent the key advantage of biclustering method. These cancerous dataset take cancerous person as irrelevant individuals corresponding to health persons. For example, given Table 3 be an interesting bicluster. The dataset present that G1 and G4 are co-expressed in Sample 1 and Sample 2. However, as S1 and S2 are irrelevant individuals, the outcome can not explain the biology significance.

By contrast, using the Aging Mice dataset[2], the advantage of biclustering can be presented. There are 12 conditional samples, respectively are: HiFA, HiMA, HiFC, HiMC, HFA, HMA, HFC, HMC, GFA, GMA, GFC, GMC. Our purpose is to find the correlative genes and the critical factor about the mice aging progress. Analysis of the data can highlight the advance of biclustering method. Assuming our result is in Table 4, which shows diet is a key factor on gene expression in the Hippocampus. In the interest of accuracy, there are five samples in each gender, to get more notable biology significance. The original data are divided to five modified dataset, as we have dataset in five age period, so there are totally 4*5=20 expression datasets. We plan to get interesting information by mining frequent differential bicluster, which can reveal not only aging-related genes, but also the aging-related samples.

Table 3. An example of bicluster

	S1	S2
G1	23	46
G2	23	46
G3	23	46
G4	23	46

Table 4. A bicluster of mice aging

	HiFA1	HiFA2
G1	12	24
G2	12	24
G3	12	24
G4	12	24

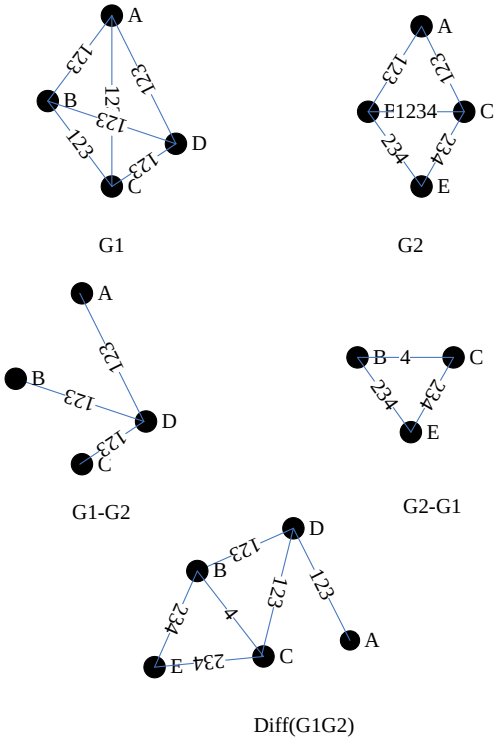

Figure 5. The process of differential weighted sample graph

Each original microarray dataset would convert to be a weighted sample relational graph, which can be modeled using a undirected and weighted graph $G=\{E_i=(U_i, V_i, W_i)\}$ of a set of edges E_i which contains two vertices U_i and V_i and the weighed items W_i between U_i and V_i, where each vertex represents a unique sample, edge is a pair of samples if they have co-expressed genes and weighted item

set is the genes which co-express in both vertices. Traditional method to generate differential bicluster or frequent pattern need to check in both datasets[17], which is time-consuming. In order to get differential frequent bicluster, the adjacent age weighted sample relational graph should be converted to be a differential graph. The process is illustrated in Fig. 5. Then we can mine frequent differential bicluster in the proceeded weighted graph with sample growth. The extension of our method can generate maximal or closed frequent differential bicluster. In the interest of noise elimination, we can use quasi or dense degree to get error tolerant bicluster.

Discriminative coexpression patterns can be considered as biclusters that exist mostly in one class but not in the other. Some studies[18,19,11] use two-step approach to discovery discriminative coexpression patterns. A set of biclusters are discovered in the first step and then the ones that unique to a single class are selected in the second step. Such two-step approach is very time-consuming. The approach proposed in [17] can also be used to generate discriminative biclusters. However, according to the discussion in the last paragraph, the double checking procedure is less efficient. In this study, we propose to mine discriminative coexpression biclusters in differential data set, which is shown in Fig. 6. Using differential data set to analyze has a premise, which is the sample must be an unique one. It may not be suitable to mine cancer microarray data set, since we cannot get the expression value of a sample of health and cancerous. However, the AGEMAP[2] data set provides a unique opportunity to create difference matrix as the data set from the 8-month old mice is matched to the data set from the 16-month old mice. Since the mice sample (five males, five females) is unique one during aging, so the difference matrix does not lose information comparing to original data sets.

4. CONCLUSION

The recent development of high-throughput bio-techniques for post genomics has generated a large volume of gene expression data. Microarray data has made the new challenges which make many traditional data mining methods infeasible for mining the hidden knowledge. In this paper, we have given an overview of the recent research techniques in microarray data analysis. We propose some methods to generate interesting information based on the character of microarray. Based on our recent experimental dataset, we plan to get frequent differential bicluster from multiple microarray data sources. The focuses of my PhD thesis are mining interesting information from multiple microarray datasets and addressing several issues from it.

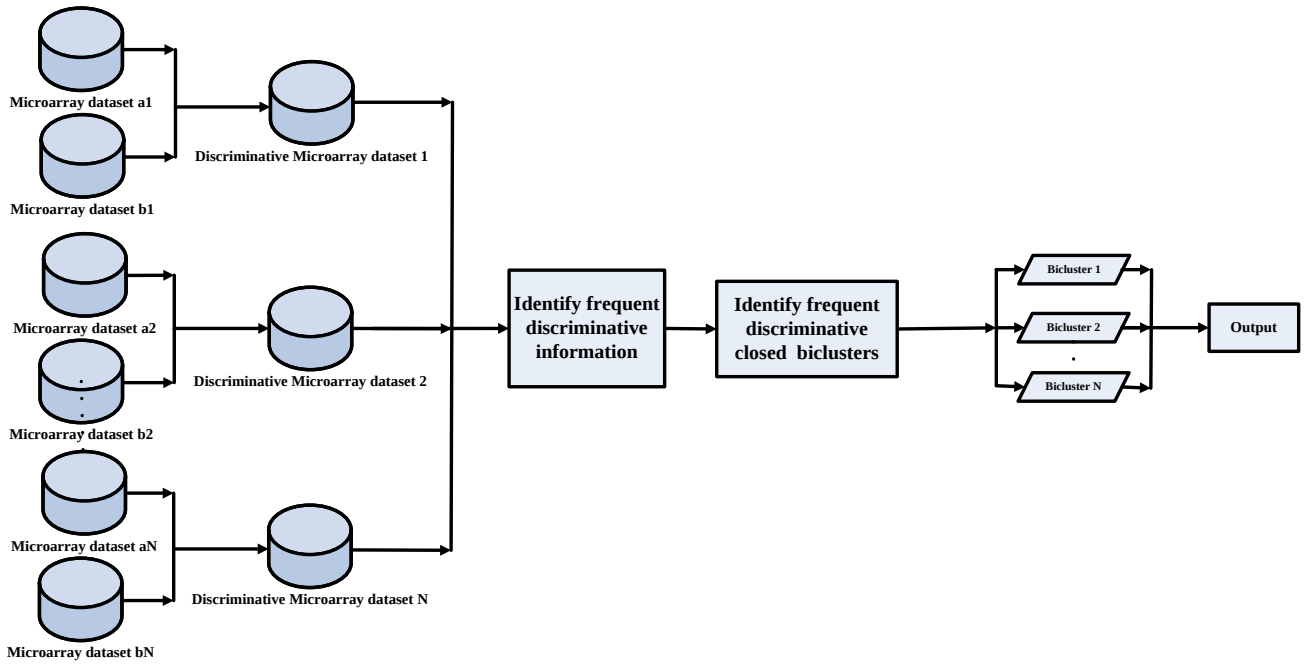

Figure 6. The overview of mining frequent discriminative bicluster approach.

5. ACKNOWLEDGMENTS

The work is supported by the National Natural Science Foundation of China under Grant No.60703105 and the Doctorate Foundation of Northwestern Polytechnical University under Grant No.CX200913. It is also partly supported by the Natural Science Foundation of Shaanxi Province under Grant No.2007F27.

6. REFERENCES

[1] D'haeseleer, P., Liang, S., and Somogyi, R. (2000) Genetic network inference: from co- expression clustering to reverse engineering. Bioinformatics, 16, 707-726.

[2] J.M. Zahn, S. Poosala, etc. AGEMAP: A gene expression database for aging in mice. PLOS Genetics, 3(11):2326-2337, 2007.

[3] Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999). Systematic determination of genetic network architecture. Nature Genetics, 22, 281-285.

[4] Ramoni, M., Sebastiani, P., and Kohane, I. (2002) Cluster analysis of gene expression dynamics. PNAS, 99, 9121-9126.

[5] M. Wang, X.Q.Shang, Z.H.Li. Strong association rules mining without using frequent items for microarray analysis. The 3rd International Conference on Bioinformatics and Biomedical Engineering (iCBBE 2009). Beijing, China:IEEE pp 978-1-4244-2902-8.

[6] Torgeir, R.H., Astrid, L. and Jan, K. (2002). Learning rule-based models of biological process from gene expression time profiles using Gene Ontology. Bioinformatics, 19, 1116-1123.

[7] Pei Jian, Han Jiawei. Mining Sequential Patterns by Pattern-growth: The PrefixSpan Approach[J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 6(10): 1-17.

[8] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M.-C. Hsu, FreeSpan: Frequent Pattern-Projected Sequential Pattern Mining, Proc. 2000 ACM SIGKDD Int'l Conf. Knowledge Discovery in Databases (KDD '00), pp. 355-359, Aug. 2000.

[9] Pan, F., Cong, G., Tung, K., Yang, J., Zaki, M.: Carpenter: Finding closed patterns in long biological datasets. In: Proc. ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining (KDD),2004, pp. 637–642.

[10] Cong, G., Tan, K., Tung, A. et al.: Mining Frequent Closed Patterns in Microarray Data. ICDM'04. IEEE Press, 2004, 363–366.

[11] Y. Cheng and G.M. Church, "Biclustering of Expression Data," Proc. 8th Int'l Conf. Intelligent Systems for Molecular Biology (ISMB00),ACM Press, 2000, pp. 93–103.

[12] L. Zhao and M. Zaki. Microcluster: Efficient deterministic biclustering of microarray data. IEEE Intelligent Systems, 20(6):40–49, 2005.

[13] J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai. Revealing modular organization in the yeast transcriptional network. Nat. Genet., 31:370–377, 2002.

[14] S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: a survey. IEEE/ACM TCBB, 1(1):24–45,2004.

[15] M. Wang, X.Q.Shang, Z.H.Li. Strong association rules mining without using frequent items for microarray analysis. The 3rd International Conference on Bioinformatics and Biomedical Engineering (iCBBE 2009). Beijing, China:IEEE pp 978-1-4244-2902-8.

[16] M.Wang, X.Q.Shang, Z.H.Li. WIBE: Mining frequent closed patterns without candidate maintenance in microarray dataset. In Proc. The 6th International Conference on Data Mining, 2010. To be appeared.

[17] Gang Fang, Rui Kuang, Gaurav Pandey, Michael Steinbach, Chad L. Myers and Vipin Kumar, Subspace Differential Coexpression Analysis: Problem Definition and A General Approach, Proceedings of the 15th Pacific Symposium on Biocomputing (PSB), 15:145-156, 2010.

[18] T. Murali and S. Kasif. RankGene: identification of diagnostic genes based on expression data. In Proc. Pacific Symposium on Biocomputing 8:77-88, 2003.

[19] A. Tanay, R. Sharan, M. Kupiec, and R. Shamir. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. PNAS, 101(9):2981–2986, 2004.