Mining Coherent Patterns from Heterogeneous Microarray Data

Xiang Zhang
Department of Computer Science
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599, USA

xiang@cs.unc.edu

Wei Wang
Department of Computer Science
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599, USA

weiwang@cs.unc.edu

ABSTRACT

Microarray technology is a powerful tool for geneticists to monitor interactions among tens of thousands of genes simultaneously. There has been extensive research on coherent subspace clustering of gene expressions measured under consistent experimental settings. However, these methods assume that all experiments are run using the same batch of microarray chips with similar characteristics of noise. Algorithms developed under this assumption may not be applicable for analyzing data collected from heterogeneous settings, where the set of genes being monitored may be different and expression levels may be not directly comparable even for the same gene. In this paper, we propose a model, F-cluster, for mining subspace coherent patterns from heterogeneous gene expression data. We compare our model with previously proposed models. We analyze the search space of the problem and give a naïve solution for it.

Categories and Subject Descriptors: H.2.8 [Database Management]: Database Applications - Data Mining; J.3 [Life and Medical Sciences]: Biology and Genetics

General Terms: Algorithm, Experimentation

Keywords: Subspace Clustering, Microarray Analysis, Heterogenous Data

1. INTRODUCTION

Recent advances in microarray technology have made large amounts of gene expression data available from a variety of different experimental settings. Clustering is a popular method used to analyze microarrays. A special class of subspace clustering algorithms called bi-clustering can effectively identify meaningful gene groups [12, 13]. By finding shifting and/or scaling patterns, bi-clustering algorithms can identify co-regulated genes whose expressions differ in value but are highly correlated. Tricluster [14] extends the model to the temporal domain. It finds clusters in gene-sample-time space.

Although previous methods have demonstrated their usefulness, all of these methods focus on analyzing expression data generated by a single microarray technique. There have been several microarray techniques using fundamentally different mechanisms to measure gene expression levels, including Affymetrix oligonucleotide microarrays [5], cDNA microarrays [8], and serial analysis of gene expression (SAGE) [11]. The expression levels reported by different techniques are not necessarily comparable with each other. Since previously proposed subspace clustering methods were developed under the assumption that the gene expressions are measured under consistent experimental settings, they can not be readily applied to analyze data generated by different techniques.

In the biology community, there has been wide interest in the cross-platform comparisons of gene expression values [1, 2, 3, 6, 7, 9]. However, these studies have focused on the expression values for individual genes. Statistical and probabilistic methods [4, 10] have been proposed for finding the pair-wise interactions of genes or proteins in heterogenous datasets. They are not applicable to finding subspace coherent patterns in heterogenous microarray datasets.

To overcome the limitation of previous methods, we propose a more general model for mining coherent subspace clusters from multiple microarrays that may be generated by different techniques.

2. PROPOSED MODEL

We propose the model of *F-cluster* (Frequent cluster), which mines genes co-expressed in subsets of samples within multiple microarrays. Informally speaking, an F-cluster is a group of genes that show similar expression patterns in some subspaces (of samples) in at least δ matrices, where δ is a user specified number. For example, as shown in Figure 1, the gene group $\{g_1, g_2, g_3\}$ is an F-cluster which is preserved in matrices $M_1, M_2,$ and M_3 if $\delta = 3$.

2.1 Comparison with Previous Methods

Previous coherent subspace clustering methods focus on mining data generated under homogeneous settings. Our model, on the other hand, can be applied to microarray data collected from heterogeneous settings. Since experimental settings used in different microarray techniques and the samples may be different and incompatible, we do not require the genes in an F-cluster have same expression patterns in different microarrays. As shown in Figure 1, the F-cluster $\{g_1, g_2, g_3\}$ is preserved in M_1, M_2, and M_3 sepa-

	M_1			M_2			M_3		
	a	b	c	α	β	γ	x	y	z
g_1	30	40	30	45	35	45	8	18	8
g_2	25	35	25	15	5	15	28	38	28
g_3	20	30	20	28	18	28	20	30	20

(a) Expression values of g_1, g_2, g_3
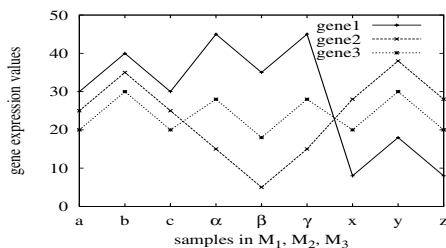
(b) Expression patterns of g_1, g_2, g_3

Figure 1: Example 1

rately. But the expression patterns are not preserved across matrices. (There are cross-overs between matrices as shown in Figure 1.)

2.2 Exponential Search Space

Let I be the number of genes in the heterogenous microarray dataset, K be the number of matrices, and δ be the user specified threshold. There can be as many as 2^I potential gene clusters in each matrix, where I typically ranges from thousands to tens of thousands. Identifying those supported by at least δ out of K matrices requires intelligent pruning strategies to confine the search space.

3. A NAÏVE ALGORITHM

F-clusters can be viewed as *consensus* subspace clusters in multiple microarray matrices. To mine the F-clusters, in a straightforward manner, we can adopt the following 2-step approach.

- In the first step, we mine and maintain subspace clusters in each matrix separately. Existing bi-clustering algorithms [12, 13] can be applied to find these clusters.

- In the second step, we exam the combinations of the subspace clusters found in previous step to see if any sub-cluster is preserved in at least δ matrices. This step can be accomplished by using either breadth first search or depth first search strategy.

This naïve approach is inefficient, since there are usually thousands to millions of subspace clusters (even if we only keep maximal subspace clusters) in a single matrix.

4. CONCLUSION AND FUTURE WORK

In this paper, we propose F-cluster model for mining coherent subspace clusters. This model can find the clusters that are preserved in expression data generated by different microarray techniques. We compare our model with previous models and propose a straightforward method to mine

the F-clusters. In our future work, we shall develop more efficient algorithm for this problem. The algorithm should adopt efficient search strategy and incorporate effective punning methods. We will apply our mining algorithm to real life gene expression datasets and validate the biological significance of the resulting gene clusters.

5. REFERENCES

[1] V. Detours, J.E. Dumont, H. Bersini and C. Maenhaut. Integration and cross-validation of high-throughput gene expression data: comparing heterogenous data sets. *BEBS Lett.*, 546: 98-102, 2003.

[2] C.A. Iacobuzio-Donahue and et al. Highly expressed genes in pancreatic ductal adenocarcinomas: a comparison of the transcription profiles obtained from three major technologies. *Cancer Research*, 63: 8614-22, 2003.

[3] A.K. Jarvinen and et al. Are data from different gene expression microarray platforms comparable? *Genomics*, 83: 1164-68, 2004.

[4] I. Lee and et al. A probabilistic functional network of yeast genes. *Science*, 306: 1555-58, 2004.

[5] D.J. Lockhart and et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, 14: 1675-80, 1996.

[6] J. Lu and et al. A comparison of gene expression profiles produced by SAGE, long SAGE, and oligonucleotide chips. *Genomics*, 84: 631-36, 2004.

[7] A.T. Rogojina and et al. Comparing the use of Affymetrix to spotted oligonucleotide microarrays using two retinal pigment epithelium cell lines. *Mol. Visioin*, 9: 482-96, 2003.

[8] M. Schena, D. Shalon, R.W. Davis and P.O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270: 467-70, 1995.

[9] P.K. Tan and et al. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Research*, 31: 5676-84, 2003.

[10] O.G. Troyanskaya, K. Dolinski, A.B. Owen, R.B. Altman and D. Botstein. A bayesian framework for combining heterogeneous data sources for gene function prediction in Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. USA*, 100: 8348-53, 2003.

[11] V.E. Velculescu, L. Zhang, B. Vogelstein and K.W. Kinzler. Serial analysis of gene expression. *Science*, 270, 1995.

[12] H. Wang, W. Wang, J. Yang and P. Yu. Clustering by pattern similarity in large data sets. *SIGMOD*, 2002.

[13] X. Xu, Y. Lu, A. Tung and W. Wang. Mining shifting-and-scaling co-regulation patterns on gene expression profiles. *ICDE*, 2006.

[14] L. Zhang and M. Zaki. Tricluster: An effective algorithm for mining coherent clusters in 3d microarray data. *SIGMOD*, 2005.