Finding Similar Patterns in Microarray Data

Xiangsheng Chen[1], *Jiuyong Li*[1], *Grant Daggard*[2], *and Xiaodi Huang*[3]

1 Department of Mathematics and Computing,
2 Department of Biological and Physical Sciences,
The University of Southern Queensland, Australia

3 Department of Mathematics, Statistics and Computer Science,
The University of New England, Armidale, NSW, 2350

Abstract. In this paper we propose a clustering algorithm called s-Cluster for analysis of gene expression data based on pattern-similarity. The algorithm captures the tight clusters exhibiting strong similar expression patterns in Microarray data, and allows a high level of overlap among discovered clusters without completely grouping all genes like other algorithms. This reflects the biological fact that not all functions are turned on in an experiment, and that many genes are co-expressed in multiple groups in response to different stimuli. The experiments have demonstrated that the proposed algorithm successfully groups the genes with strong similar expression patterns and that the found clusters are interpretable.

Keywords: data mining, bioinformatics, Microarray data analysis, clustering

1. Introduction

Many clustering techniques in bioinformatics have been applied to analyze gene expression data. Most clustering models [4, 1, 8, 10, 7, 9] are distance based clusterings such as Euclidean distance and cosine distance. However, these similarity functions are not always sufficient in capturing correlations among genes or conditions.

To remedy this problem, the bicluster model [2] uses a similarity score to measure the coherence of genes and conditions in a sub matrix of Microarray data. Wang *et al.* [11] proposed an algorithm to find all (maximum) submatrices such that they are δ-pClusters.Liu *et al.* [5] introduced a u-Cluster model to capture the general tendency of objects across a subset of dimensions in a high dimensional space. In reality, errors are unavoidable in biological experiments and perfect pattern matching in Microarray data may not occur even among known coordinately regulated genes. In this paper, we will present a model which tolerates such possible errors in the data. Our proposed algorithm is simple, interpretable, and deterministic. The proposed algorithm is distinct from δ-p Clustering model in that it is a full space clustering model and allows dissimilarities, possibly caused by experimental errors, in clusters while δ-pClustering does not.

2. s-Clusters

We define *s*-clusters by a threshold as the minimum proportion of conditions in which genes have the similar express. Our model does not cluster all genes and allows clusters to overlap. The resulting clusters are tight. A tight cluster is better for refining a hypothesis.

2.1 Model

The original gene data matrix is first normalized. Gene-condition expression data is represented as a *n*-by-*p* matrix where each entry *xij* denotes the expression level of the *i*th gene in the *j* th condition (where i = 1,...,n and j = 1,...,p).

The new standardized data matrix *Z* is obtained by converting the raw values to *z-scores*, and it will be used for the following clustering analysis. The mean of *z-scores* in each row is zero.

Definition 1. *Let N be the set of genes and P be the set of conditions in a standardized data set Z. Given x, y ∈N, Zx and Zy denote the vectors of the xth gene and yth gene, respectively. We define the sScore of two genes under the jth condition as*

$$sScorex,y,j = |zxj - zyj| \quad (1)$$

With two given thresholds $0 < \alpha \leq 1$ *and* $\delta > 0$, *we say two genes x and y are similar, if at least in a α fraction of conditions, sScore ≤ δ for the two genes.*

Definition 2. *Let S = {Z*1*, Z*2*, ...,Zk} be a set of genes, S ⊂N. Zk denotes a vector of a gene. We say* S *forms an s-Cluster if every pair of genes in* S *is similar by definition 1.*

In the s-Cluster model, one gene can be in several different clusters. In other words, the clusters are not exclusive. This is very meaningful in the underlying biological processes in which many individual genes are co-expressed in multiple function groups in response to different stimuli.

2.2 Algorithm

The algorithm contains three phases: (1) preprocess the data into a normalized data matrix. The mean and mean absolute deviation are calculated for each row, and are then converted the raw data into *z-scores*;(2) find similar gene pairs. We go through the *z-scores* data and identify all similar gene pairs according to Definition 1; (3) form all s-Clusters. construct a graph where every gene is represented as a vertex, and two similar genes as an edge. s-Clusters can be viewed as the cliques in this graph according to Definition 2. We design an algorithm similar to Bierstone's algorithm [6] to generate all maximum cliques, interesting s-Clusters.

In general, finding all maximal cliques in a graph is NP-complete. The algorithm can enumerate all maximal cliques efficiently only when the equivalent graph is sparse, i.e. edge density is low. Edge density of a gene graph is usually very low since there are not many genes expressing similarly across most conditions. Therefore, this method produces good results with high efficiency in Microarray data.

A simple heuristic to set *δ* is outlined as follows. It is set high initially, and then is reduced gradually. When the visual inspection of similarity of gene expression patterns in clusters is unacceptable, the process stops. The setting of *α* is straightforward since its meanings is clear.

The definition of similarity in this model is more strict than that in most other clustering models. As a result, the clusters of this model are usually very tight, including much fewer genes than clusters from other models. We do not intend to find regular clusters to group all genes, but to find small groups of genes that exhibit strong similar expression patterns. We find that these clusters are very interpretable.

3 Experiments

We apply the s-Cluster algorithm to yeast *Saccharomyces cerevisiae* cell cycle expression data from Cho *et al.* [3]. The yeast data contains expression levels of 2,884 genes under 17 conditions. The data set is organized in a matrix where each row corresponds to a gene and each column represents a condition.

Gene	System	Name	Description
58	YAR007C		69 kDa subunit of the heterotrimeric RPA (RF-A) singlestranded DNA binding protein, binds URS1 and CAR1
216	YBR088C		Profilerating cell nuclear antigen (PCNA) accessory factor for DNA polymerase delta, mRNA increases in G1, peaks in S in mitosis, and increases prior to DNA synthesis in meiosis"
217	YBR089W		Unknown
448	YDL003W		Unknown
526	YDL164C		DNA ligase
616	YDR097C		Homolog of the human GTBP protein, forms a complex with Msh2p to repair both single-base and insertion-deletion mispairs, redundant with Msh3p in repair of insertion-deletion mispairs"
1022	YFL008W		Coiled-coil protein involved in chromosome structure or segregation
1184	YGR152C		GTP-binding protein of the ras superfamily involved in bud site selection
1286	YHR154W		Establishes Silent omatin
1795	YLR103C		Omosomal DNA replication initiation protein
1836	YLR183C		Unknown
2278	YNL102W		DNA polymerase I alpha subunit, p180
2375	YNL312W		1-7, 116-930" subunit 2 of replication factor RF-A 29% identical to the human p34 subunit of RF-A
2538	YOR074C		Thymidylate synthase
2725	YPL153C		Protein kinase, Mec1p and Tel1p regulate rad53p phosphorylation"

Fig. 1. A list of genes in s-Cluster #111. 12 genes are related to DNA synthesis and replication and 3 are unknown. This raises the possibility that the 3 genes are also DNA synthesis and replication related.

Each entry represents the relative abundance values (percentage of the mRNA for the gene in all mRNA) of the mRNA of a gene under a specific condition, which is scaled into an integer in the range of 0 and 600. We conducted the experiment with the parameters of $\delta = 0.8$ and $\alpha = 0.8$. A total of 1764 s-Clusters with a minimum size of 5 was generated by the algorithm. Clusters of four or fewer genes were ignored. The 1764 s-Clusters covered 453 genes, or 15.7% of the 2884 genes. This method only groups some interesting genes, which express coherently with other genes. All clusters are highly overlapping, and this captures a biological fact that some genes participate in a number of functions.

There are 15 members in the s-cluster #111 in Figure 1, 12 genes of which are related to DNA synthesis and replication, and 3 genes (YBR089W, YDL003W, ULR183C) are unknown. This raises the possibility that the 3 genes are also related to DNA synthesis and replication. Figure 1 shows genes in this s-Cluster in details.

Our findings are interesting when compared with those of Tavazoie *et al.* [8]. Our 15 members in s-Cluster #111 are all in the cluster #2 discovered by Tavazoie *et al.*. Their cluster #2 contains 186 genes which are related to four functions: DNA synthesis and replication, cell cycle control and mitosis, recombination and DNA repair, and nuclear organization. Our approach successfully subcategorized Tavazoie's cluster #2 into several smaller sized s-Clusters containing genes which are clearly related to one of the four functional categories. This indicates that the s-Clusters are more tightly grouped and more interpretable than the clusters from the alternative analysis approach.

4 Conclusions

We have proposed a new pattern-similarity clustering model called s-Cluster to capture some tight clusters containing groups of genes with strong coherent expression patterns.Our experimental results show that the proposed algorithm can successfully group genes with similar expression patterns. When compared with the clustering results from a conventional method [8], the clusters found by our algorithm are tighter and more interpretable.

References

1. U. Alon, N. Barlai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide array. In *Proc. Natl. Acad. Sci. USA*, volume 96(12), pages 6745–6750, 1999.

2. Y. Cheng and G. Church. Biclustering of expression data. In *Proc Int Conf Intell Syst Mol Biol*, pages 93–103, 2000.

3. R. Cho and Team. A genome wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2:65–73, 1998.

4. M. Eisen, P. Spellman, P. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. In *Proc. Natl. Acad. Sci. USA*, volume 95, pages 14863–14868, 1998.

5. J. Liu and W.Wang. Op-cluster: Clustering by tendency in high dimensional space. In *Proc of IEEE International Conference on Data Mining (ICDM)*, pages 19 – 22, 2003.

6. G. D. Mulligan and D. G. Corneil. Corrections to Bierstone's algorithm for generating cliques. *Journal of the ACM*, 19(2):244–247, 1972.

7. P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. lander, and T. Golub. Interpreting patterns of gene expression with selforganizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci., USA*, 96:2907–2912, 1999.

8. S. Tavazoie, J. Hughes, M. Campbell, R. Cho, and G. Church. Systematic determination of genetic network architecture. *Natrue Cenetics*, 22:281–285, 1999.

9. P. Toronen, M. Kolehmainen, G.Wong, and E. Castren. Analysis of gene expression data using self-organizing maps. *Federation of European Biochemical Societies FEBS Lett*, 451(2):142–146, 1999.

10. J. Vilo, A. Brazma, I. Jonassen, A. Robinson, and E. Ukkonen. Mining for putative regulatory elements in the yeast genome using gene expression data. In *Proc. 8th Int Conf on Intelligent Systems for Molecular Biology. AAAI Press*, pages 384–394, 2000.

11. H. Wang, W. Wang, J. Yang, and P. Yu. Clustering by pattern similarity in large data sets. In *Proc of the ACM SIGMOD International Conference on Management of Data SIGMOD*, pages 394–405, 2002.