Biclustering Algorithm for Embryonic Tumor Gene Expression Dataset: LAS Algorithm

Hamid Alavi Majd[1], Soodeh Shahsavari[1,*], Soheila Khodakarim[2], Seyyed Mohammad Tabatabaei[3], Bi bi Fatemeh Nobakht Motlagh Ghochani[4]

1 Biostatistics Department, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
2School of Public Health, Shahid Beheshti University of Medical Science, Tehran,Iran.
3Medical Informatics Department, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
4Proteomics Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran.

*Corresponding Author: email address: soodeh_shahsavari@sbmu.ac.ir (S. Shahsavari)

ABSTRACT

An important step in considering of gene expression data is obtained groups of genes that have similarity patterns. Biclustering methods was recently introduced for discovering subsets of genes that have coherent values across a subset of conditions. The LAS algorithm relies on a heuristic randomized search to find biclusters. In this paper, we introduce biclustering LAS algorithm and then apply this procedure for real value gene expression data. In this study after normalized data, LAS performed. 31 biclusters were discovered that 26 of them were for positive gene expression values and others were for negative. Biological validity for LAS procedure in biological process, in molecular function and in cellular component were 77.96% , 62.28% and 74.39% respictively. The result of biological validation of LAS algorithm in this study had shown LAS algorithm effectively convenient in discovering good biclusters.

Keywords: Biclustering algorithm; LAS procedure; Gene expression data

INTRODUCTION

Recently advancement in quick access to molecular sequences specially DNA microarray technologies have led to the development of studies in experimental biology[1]. An important step in considering of gene expression data is obtained groups of genes that have similarity patterns[2]. Gene expression data can be represented as a matrix of datasets that rows corressponding to genes (transactions) and columns corressponding to conditions (items) [3] such as different time points, different cells or different environmental conditions. This type of data depends on the context will be interpreted [4]. For analysis of these data clustering methods have been applied. Clustering of microarray expression data can lead to molecular classification of disease states, identification of co-fluctuation of functionally related genes, functional groupings of genes and logical descriptions of gene regulation, among others. Discovering the direct correlation patterns of genes with experimental conditions is a starting point for understanding the large-scale network that they comprise. The qualitative or quantitative correlation of genes with experimental conditions is a key to understand the biological interactions among genes in a better way[5]. Usually clustering approachs divided genes basis on their similarity of expression under all conditions. However, somtimes some genes behave similarly on a subset of condition and uncorrelated over the rest of the conditions[6]. So an important research problem of gene expression analysis is discovered submatrix patterns, therefore traditional clustering methods will fail to identify such gene groups. Biclustering methods was recently introduced that dicovered subsets of genes that have coherent values across a subset of conditions[7]. Biclustering was first described in the literature by Hartigan. It refers to a distinct class of clustering algorithms that perform simultaneous row-column clustering [8].

For the first time, Cheng and Church applied biclustering in gene expression data. They proposed a score for each candidate bicluster and used greedy algorithms for solved problem[9]. A lot of biclustering methods have been introduced for identifying subsets of gene expression such as ISA, CTWC, SAMBA, Spectral Biclustering, Plaids Models, OPSM and BiMax[10]. Different biclustering algorithms have been designed to discover different types of biclusters. There are some issuses with these algorithms. First, most of these algorithms are neglected the noise in data and required all values in bicluster to be coherent. In addition, they are top down greedy schemes that in each step iteratively eliminate some of rows and columns. Second, some of these algorithms do not find overlapping biclusters[11]. Recently, some algorithms have introduced that accounted the noise and performed exhaustive search for finding overlapping biclusters. The Large Submatrices in High Dimentional Data (LAS) algorithm relies on a heuristic randomized search to find biclusters. In this paper, we introduce biclustering LAS algorithm and then apply this procedure for real value gene expression data.

METHODS

The LAS score function rely on normal CDF and to departure from normality that usually happen in gene expression for heavy tails is sensitive. Then a first step in the algorithm is considered normal plot. Usually to solve heavy tail problem, applied transformation f(x)=sign(x) log(1+|x|) to each entry. Using a simple Gaussian null model obtain a score to each submatrix U using a Bonferroni-corrected p-value that let account multiple comparisons when searching many submatrices for finding biclusters[12].

Problem statement

The dataset is assumed to be a m×n data matrix D that

$$D = \{d_{i,j}; i \in [m], j \in [n]\}$$

$[m]=\{1,2,3,...,m\}$, $[n]=\{1,2,3,...,n\}$

Each row represents to gene and each columns represents to condition and Matrix element denoted by d(i,j) and a continuous variable. The purpose of this study has discovered submatrix (U) of D that co-regulated. Because of the random nature of the biological behavior of microarray data and uncertainty of appropriate criterion for converted continuous gene expression data into discrete, microarray data are subjected to the noise. So using a rules that data matrix D is denoted as sum of k constant, overlapping submatrix plus noise.

Definition of LAS Algorithm

This model is expressed as

$$d_{i,j} = \sum_{k=1}^{K} \alpha_k \, I(i \in A_k, j \in B_k) + \varepsilon_{ij} \qquad i\epsilon[m], j\epsilon[n] \qquad (1)$$

that $A_k \subseteq [m]$, $B_k \subseteq [n]$ are kth row and column submatrix and $\alpha_k \in R$ is kth level of submatrix. In addition $\{\varepsilon_{i,j}\}$ are independent normal random variables N(0,1).

If k=0 model (1) converted to null simple model that

$$\{d_{i,j}; i \in [m], j \in [n]\} \qquad d_{i,j} \sim iid\, N(0,1)$$

that D is a random normal matrix m×n. This model is provided score function for submatrices. For submatrix U with dim k×l with average AVG(U)=τ score function is

$$S(U) = -\log\left[\binom{m}{k}\binom{n}{l}\, \Phi(-\tau\sqrt{kl})\right]$$

that k=|A| , l=|B| are number of rows and columns for U and Φ is CDF distribution of standard normal. Term of $\binom{m}{k}\binom{n}{l}$ is as a Bonferroni correction for number of possible submatrices in D with dimensions of k×l. Further the term of Φ(-$\tau\sqrt{kl}$) show significant average bicluster under null model. Score function can be computed deviations from null model that accounted dimensions and average values in submatrices. This function is sum of two terms. First, -log Φ(-$\tau\sqrt{kl}$) is positive and for finding the submatrix k×l with average τ named "reward". Second, $-\log\binom{m}{k}\binom{n}{l}$ is negative and for multiple comparison is penalty. LAS algorithm search for biclusters that have large score. This algorithm iteratively maximize score fore set of rows and columns and repeated several times. After discovered a bicluster, τ (average) subtracts to entries in U and repeat process. This procedure can not all of submatrices for D and iteratively performed greedy search set of rows and columns of candidate submatrices until obtained local maximum for score function[12].

Heat Map

The heatmap is popular and widely used graphs in the biological sciences which compacts large amounts of information into a small space to finding coherent patterns in the data[13]. A heatmap is a graphical matrix representation that entries represented as color. In gene expression data large values were represented by dark and smaller value by lighter squares[14].

Gene Expression Dataset

In this study a central nervous system (CNS) embryonic tumor gene expression dataset was analyzed. This dataset is constituted of 40 tumor samples (including 10 medulloblastomas, 10 malignant gliomas, 10 atypical teratoid/rhabdoid tumours (AT/RT, 5 brain, 3 renal and 2 extrarenal), 4 normal cerebellums and 6 supratentorial primitive neuroectodermal tumours (PNETs)) analyzed on Affymetric HuGeneFLFor. This dataset was used in research that was performed in 2011 for DNA microarray gene expression data matrix[4]. Dimensions of this matrix was 7129×40. The dataset is standardized (mean 0 and variance 1) and then discretized into 12 different levels (with values from 0 to 11). Each value corresponds to a gene expression value range. The purpose of this analysis discovered biclusters that show relation between genes and type of CNS tumor.

Gene Enrichment Analysis

Knowledge of the biological role of genes and proteins, usually create inference about organisms functional. For this goal Gene Ontology (GO) was formed that had nomenclature systems for genes and their products. Such information can be used to recovered human health. Biological process, molecular function and cellular component are all attributes of genes, gene products or gene-product groups. Biological process refers to a biological objective to which the gene product contributes. Molecular function is defined as the biochemical activity of a gene product. Cellular component refers to the place in the cell where a gene product is active[1]. Genes included in a bicluster are expected to be involved in similar biological processes. The web application Babelomics was used to discover gene ontology.

RESULTS

The aim of this section is to evaluate the usefulness of LAS procedure in relation to real data. To attain this, a central nervous system embryonic tumor gene expression dataset was used. Figure 1 shows heatmap for 7129 genes and 40 conditions that lighter color represents smaller values and dark color represent larger values of expression. Furthermore, coexpressed genes almost were given near each other.

Table1: performance information of LAS biclustering algorithms

Tumor Type	Bicluster Type	Num Bicluster	Num Gene	Num Conditions	Average of Bicluster	Score of bicluster
Total	**Positive**	26	729	39	2.072	58677.011
	Negative	5	456	39	-0.450	13384.382
Brain	**Positive**	8	499	4	2.979	6301.285
	Negative	3	1	4	-2.177	1.440
Extra Renal	**Positive**	17	465	9	2.904	15936.333
	Negative	4	14	9	-1.746	94.574
Renal	**Positive**	6	246	2	4.028	2929.376
	Negative	3	1	2	-4.117	9.691
Gliblastoma	**Positive**	10	481	9	2.681	13806.451
	Negative	4	294	9	-1.250	848.664
Medulloblastoma	**Positive**	13	280	9	3.775	16781.595
	Negative	4	59	9	-1.429	206.022
PNET	**Positive**	10	236	5	3.817	7570.136
	Negative	3	23	5	-2.221	133.515

Centring and standardization are necessary step in this algorithm because assume distribution of null model is standard normal. This algorithm iteratively search submatrices and sequentially discovered bicluster with large score and process had been stoped when the score had been negative. In addition for positive and negative values of gene expression data separately obtained submartices with significant averages. In this gene expression data range of column average was (5.30,5.49) and range of column std deviation was (0.95,1.42). In addition value for average column kurtosis was 14.67 then normal transformation was performed. After normalization, LAS algorithm was performed and biclusters was extracted.

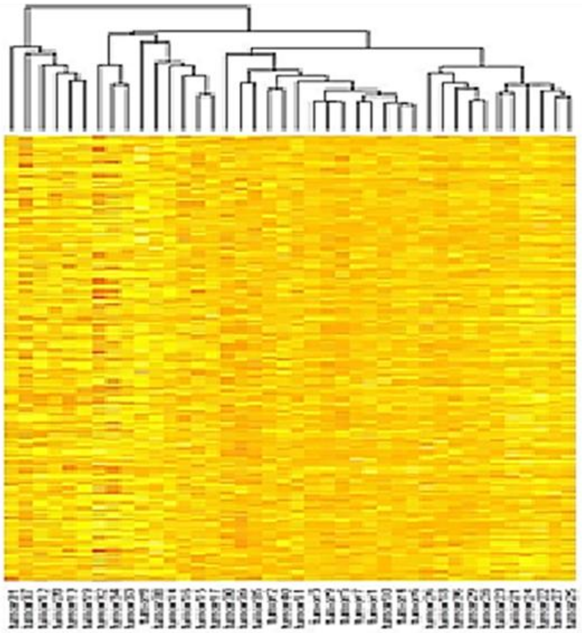

Figure1: a heatmap of all gene expression data

In table 1 result was shown and contained summary information about biclusters. Each rows information about single bicluster. In this study 31 bicluster was discovered that 26 of them was for positive gene expression values and other was for negative.. In this study, dimensions of input matrix 7129×40 had reduced to 729×39 with LAS algorithms. Information about biclusters was given in table 1. Also this procedure had provided 26 biclusters for positive expression and 5 bicluster for negative expression. In addition, biclusters had discovered for tumor type. Number of biclusters, size of representative bicluster, average and score had calculated. Representative bicluster was a bicluster with maximum genes and conditions. Maximum and minimum number of genes had discovered in brain and PNET tumor respectively. Finally, confirmation and evaluation of results had considered by GO. Biological validity for LAS procedure in biological process, in molecular function and in cellular component were 77.96%, 62.28% and 74.39% respictively. This result shown that LAS algorithm could be finding biclusters include genes and conditions well and had discovered rows and columns relations with different size.

DISCUSSION

Biclustering algorithms are convenient methods for finding sample-variable data and have advantages[11]. LAS has been applied to high-dimentional genomic dataset that discovered for large average submatrices. One of the important advantages of using this approach is ability to account noise in searching biclusters and ability to search smaller novel biclusters than much of traditional algorithms. Furthermore LAS can be applied to binary matrix of gene expression data that entries in matrix showed that gene r in condition d expressed or not[12]. In this study LAS biclustering algorithm was applied to tumor gene expression data. The result of biological validation of LAS algorithm in this study that was at least 62.28% and in Sacchromyces cerevisiae [15], breast cancer and lung cancer[12] had shown LAS algorithm effectively convenient in discovering good biclusters.

Most of biclustering algorithms generally provide a large number of biclusters that often may bias the evaluation. Therefore we usually use a selection methodology and accept bicluster by known number of genes and conditions[11]. For example in this study because number of genes for negative gene expression data in Brain and Renal was obtained 1 then rejected them as biclusters.

REFERENCES

1.Ashburner M. Ball C. A. Blake J. A. Botstein D. et.al., Gene Ontology:tool for the unification of biology. Nat Genet 2000;,25(1):25-29.

2.Tanay A. Sharan R. Shamir R. Discovering statistically significant biclusters in gene expression data. Bioinformatics 2002;18:136-144.

3.Ayadi W. Elloumi E. Hao J-K. Pattern-driven neighborhood search for biclustering of microarray data. Bioinformatics 2012;13:11.

4.Rodriguez-Baena D. S. Peterz-Pulido A. J. Aguilar-Ruiz J. S. A biclustering algorithm for extracting bit-patterns from binary datasets. Bioinformatics 2011; 27:2738-2745.

5.Agilar-Ruiz J. S. Shifting and scaling patterns from gene expression data. Bioinformatics 2005; 21(3840-3845).

6.Cheng Y. Church G. M. Biclustering of gene expression data. Proc lnt conf Intell Syst Mol Biol 2000, 8:93-103.

7.Prelic A. Bleuler S. Zimmermann P. Wille A. .et.al., A Systematic Comparison and Evaluation of Biclustering Methods for Gene Expression Data. Bioinformatics2006.

8.Dharan S. Nair A. S. Biclustering of gene expression data using reactive greedy randomized adaptive search procedure. BMC Bioinformatics 2009; 10(27).

9.Maderia SC. Oliveria AL. Biclustering algorithms for biological data analysis: a survey. IEE/ACM Trans Comput Biol Bioinform 2004; 1:24-45.

10.Liu X. Wang L. Computing the maximum similarity biclusters of gene expression data. Bioinformatics 2007; 23:50-56.

11.Gupta R. Rao N. Kumar V. Discovery of error-tolerant biclusters from noisy gene expression data, BMC Bioinformatics 2011;12(1).

12.Shardin A. A. Weigman V. J. Perou C. M. Nobel A. B. Finding large average submatrices in high dimensional data. The Annals of Applied Statistics 2009; 3(3): 985-1012.

13.Chen C. H. Generalized Association Plots:Information Visualization via Iteratively Generated Correlation Matrices. Statistica Sinica2002; 12:7-29.

14.Weinstein J. A postgenomic visual icon. Science 2008; 319: 1772-1773.

15.Alavi majd H, Younespoor S, Zayeri F, Rezaei Tavirani M. Biclustering of DNA microarray gene expression data by Large Average Submatrices Method. daneshvar Medicine 2011; 93