ijar
ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

COMPARISON OF BICLUSTERING ALGORITHMS FOR DETECTION OF NOISY AND OVERLAPPING BICLUSTERS USING SIMULATED GENE EXPRESSION DATA

Hamid Alavi Majd[1], Ahmad Reza Baghestani[1], Seyyed Mohammad Tabatabaei[2], Soodeh Shahsavari[1]*, Mostafa Rezaei Tavirani[3], Mohsen Hamidpour[4]

1. Biostatistics Department, Faculty of Paramedical Sciences, ShahidBeheshti University of Medical Sciences, Darband Avenue, Qods Square, Tehran, Iran.
2. Medical Informatics Department, Faculty of Paramedical Sciences, ShahidBeheshti University of Medical Sciences, Darband Avenue, Qods Square, Tehran, Iran.
3. Proteomics Research Center, Faculty of Paramedical Sciences, ShahidBeheshti University of Medical Sciences, Darband Avenue, Qods Square, Tehran, Iran.
4. HematologyDeparment and Blood Banking, Faculty of Paramedical Sciences, ShahidBeheshti University of Medical Sciences, Darband Avenue, Qods Square, Tehran, Iran.

Manuscript Info

Manuscript History:

Received: 14 November 2015
Final Accepted: 19 December 2015
Published Online: January 2016

Key words:

Gene expression data; Biclustering algorithm; Simulation

***Corresponding Author**

Dr. Sameer P Darawade

Abstract

Biclustering is an important technique for pattern recognition in gene expression data to find groups with similar expression patterns. Issues exist with biclustering algorithms in general and it is not clear which algorithms are best suited for this task. The present study evaluated four biclustering algorithms using simulated data for efficacy of detection of overlaps. Scenarios were constructed by changing the size of the data matrix and the level of noise and overlap. Results showed that the Cheng and Church and Spectral algorithms were not sufficient for these scenarios. The BiMax algorithm was robust to noise but its efficacy decreased in the presence of overlap between biclusters. The Plaid algorithm was mostly robust for overlap, but its efficacy decreased as the noise level increased. These results are designed to aid researchers when selecting the most appropriate algorithm for a dataset.

INTRODUCTION

The cell is the basic structure of any organism. All cells of an organism carry the same genes, which have different expression levels under different conditions Crick (1970). Scientists have concluded that specific conditions influence whether a particular gene is expressed and how it is expressed. The health of an organism can be compromised by some types of expression; therefore, it is crucial to evaluate the levels of genome when exposed to tense factors Jae (2001).

DNA microarray technology can be used to monitor thousands of gene expression data simultaneously for cells under different conditions and processes. This technology accelerates and increases the efficiency of gene expression studies Tanay (2008). The development of this technique has led to the availability of a gene expression matrix with rows containing thousands of genes and columns containing hundreds of conditions Liu (2007). Clustering is an important technique for pattern recognition Tanay (2008), but traditional clustering methods can face difficulties in detecting pattern similarities in gene expression data Yang (1993).

Biclustering with a computational framework is a more flexible method of overcoming these constraints to find the relevant gene expression patterns CheaGan (2008). A bicluster is a subset of genes with similar expression patterns over a subset of conditions; thus, biclustering is used to determine homogeneous submatrices Tanay (2004). The first biclustering algorithm, called block clustering, was developed by Hartigan(1972). Cheng and Church proposed the first biclustering algorithm for analysis of high-dimensional gene expression data Cheng (2000).

An important aspect of gene expression data is the high level of noise. DNA chips provide only rough approximations of expression levels and are subject to errors of up to 2-fold the measured value The chipping forecast II(2002). Any analysis method, and biclustering algorithms in particular, should be robust enough to cope with significant levels of noise. Noise/error in data is the primary factor limiting discovery of biclusters. A secondary factor is the ability of the algorithm to find overlapping biclusters.

It is important to determine whether an algorithm can cope with these issues and find valid biclusters. Most algorithms ignore noise and discover biclusters using all gene expression data, but some are unable to find overlapping biclusters. Fair comparison of clustering and biclustering approaches is difficult because each method uses a different problem formulation; an algorithm may work well for certain types of data and fail for others.

The methods of comparison of different algorithms nearly all ignore different types of noise and overlap in the data and do not simulate different scenarios Santamar (2007). The present study evaluated biclustering models and algorithms that were developed for gene expression analysis using simulation.

Material and Methods

Generated Data

Gene expression data is often subject to noise inherent in the system under measurement and errors in the measuring process. The preferred algorithms are those that are robust with respect to noise. The biclustering algorithms examined in the present study were compared for their ability to resist random noise and discover overlaps in the data. For this purpose, two matrices with different degrees of overlap and noise were generated and the intrinsic structure of the data was compared for detection of biclusters.

Two biclusters were embedded in the matrices that overlapped 0%, 10%, and 25%. The levels of noise generated was 0%, 1%, 3%, 5% and 10%. The simulated data formed matrices 50×20 and 500×50 in size with a distribution of N(0,100). The two embedded biclusters had normal distributions with means of 6 and 15, respectively, and a variance of 0.1.The noise built into the data had a binary distribution and then values were generated using a normal distribution with a mean of 20 and variance of 4. The performance of the algorithms was evaluated based on the number of rows and columns and the degree of overlap of the biclusters. The algorithm was run for 1000 iterations and the averages of the criteria were reported. R software and the Biclust package were used to compare the strengths and weaknesses of the biclustering methods.

Selected Bicluster Algorithm

Cheng and Church Cheng (2000), Spectral Kluger (2003), Plaid Lazzeroni (2002), and BiMax Prelic (2006) biclustering algorithms were chosen for comparison.These algorithms demonstrate approaches developed for the identification of bicluster patterns in large matrices and for gene expression matrices, in particular. They can be roughly classified by their model and scoring schemes. The selected algorithms are briefly described below.

Cheng and Church's Algorithm (CC): The CC algorithm define a bicluster as a submatrix for which the mean-squared residue score is below user-defined threshold δ. They proposed a two-phase strategy to identify the largest δ-bicluster in the data. First, the rows and columns are removed from the original expression matrix until the constraint is fulfilled. Next, if the bicluster score does not exceed δ, the previously-deleted rows and columns are added to the resulting submatrix. This procedure is iterated several times after the previously-found biclusters are masked by random values.

Spectral biclustering: This approach uses techniques from linear algebra to identify bicluster structures in the input data. It is assumed that the expression matrix has a hidden checkerboard-like structure that can be identified using eigenvector computations.

Plaid model: The Plaid model is a statistically inspired modeling approach developed by Lazzeroni and Owen for analysis of gene expression data. The basic idea is to represent the genes-conditions matrix as a superposition of layers, corresponding to biclusters, where each layer is a subset of rows and columns on which a particular set of values takes place.

BiMax: The BiMax algorithm algorithm uses a binary representation of the gene expression matrix. BiMax formulates biclustering as the search for all maximal bicliques in a bipartite graph. The nodes are either genes or experiments and a connection between a gene and an experiment exists if the gene was significantly expressed in that experiment.

Result and Discussion

The algorithms were implemented for two generated datasets and then evaluated. The comparison focused on the identification of locally co-expressed genes.

Results for scenarios generated by the 50×20 and 500×50 matrices for biclusters with different values of overlap and increasing noise. This table shows the performance of the algorithms in identification of overlap. The BiMax algorithm was unable to discover overlap between biclusters and the spectral algorithm discovered only 1% to 5% of cells that overlapped. For the 10% overlap scenario and 0% noise, the Plaid algorithm discovered all overlapped cells in the small dataset and 88.32% in the large dataset. For the 25% overlap scenario and no noise, the Plaid algorithm discovered all overlapped cells in the small dataset and 75.25% in the large dataset. When the noise level increased, however, this algorithm could no long discover overlapping cells.

It can be seen that the BiMax algorithm is robust for noise, but its efficacy decreased in the presence of overlap between biclusters. This algorithm better identified biclusters in small datasets.

The Plaid algorithm was nearly robust for overlap, but efficacy decreased when the noise level increased. This algorithm performed similarly for both small and large datasets. The efficacy of the Plaid algorithm was very low in for noise levels >5% and it was unable to find biclusters. The CC algorithm was extremely sensitive to both noise and overlap. In the presence of any level of noise or overlap, the efficacy of this algorithm was <5%. The Spectral algorithm was unable to identify biclusters under all conditions and showed low efficiency overall.

Variation and drawbacks in gene expression data suggest that the best way to analyze biclustering performance is to generate a known dataset. The present study used simulated datasets as the main tool to determine efficacy of identifying biclusters.

The selected algorithms were evaluated for identification of biclustering in gene expression data by generating two datasets with different levels of noise and overlap to simulate a real dataset.

The efficacy of the BiMax algorithm was better than the others, but it was still not able to identify overlapping cells. In real gene expression data, there is usually overlap between genes for which discovery is important. The Plaid algorithm performed better than the others in scenarios having moderate noise and overlap. This algorithm was able to discover biclusters with different degrees of overlap. The present study was designed to aid researchers in the choice of a biclustering algorithm to select a practical method for analysis.

Table1 Statistics for rows and columns number of generated biclusters

size of Matrices	Bicluster	Overlap Degree					
		0%		10%		20%	
		Rows	Cols	Rows	Cols	Rows	Cols
500*50	1	120:220	3:22	80:200	3:22	80:275	3:22
	2	280:400	30:49	166:370	30:49	200:383	30:49
50*20	1	12:32	1:6	18:37	1:6	18:37	1:6
	2	36:48	11:15	32:43	11:15	27:48	11:15

Table 1 lists the statistics for the generated biclusters. Two biclusters were embedded in the 2 matrices that overlapped 0%, 10%, and 25%. The simulated data formed matrices 50×20 and 500×50 in size with a distribution of N(0,100). The two embedded biclusters had normal distributions with means of 6 and 15, respectively, and a variance of 0.1. Elements of each biclusters in the generated matrices denoted in this table.

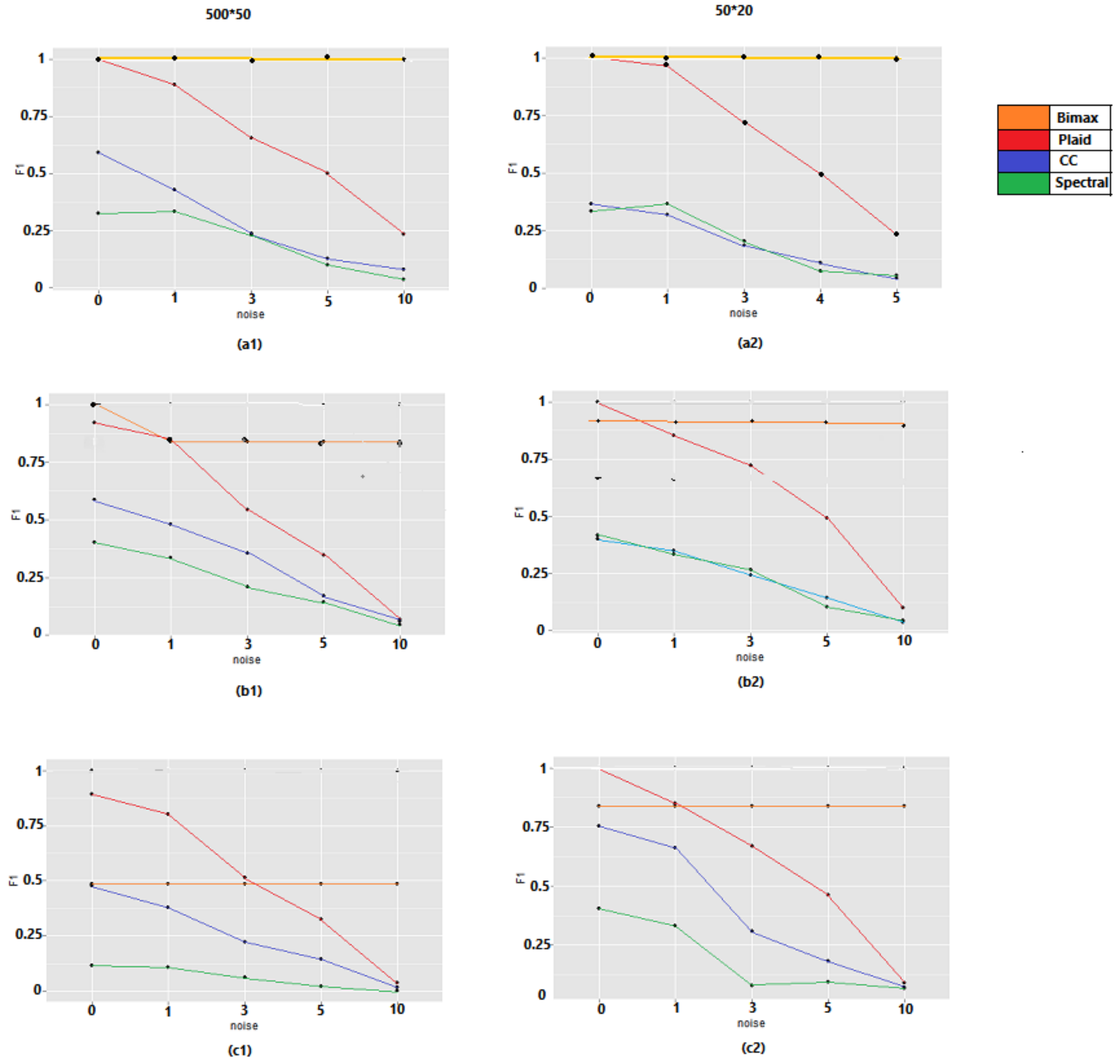
Table 2. Percentage of correctly-discovered overlap in matrices

Senario		Corrected Overlap (%)				Corrected Overlap (%)			
		50×20				500×50			
O%	N%	CC	Sptrl	Bimax	Plaid	CC	Sptrl	Bimax	Plaid
10	**0**	28.32	4.26	0	100	0	3.26	0	88.32
10	**1**	24.03	8.30	0	75.33	0	0.36	0	72.62
10	**3**	12.66	7.80	0	27.23	5.63	0.26	0	46.78
10	**5**	3.05	3.13	0	8.59	0	0	0	18.94
10	**10**	1.66	1.70	0	0.86	0.54	0	0	0.91

25	**0**	17.82	2.13	0	100	4.54	1.16	0	75.25
25	**1**	13.91	4.06	0	69.49	0.78	1.50	0	68.47
25	**3**	12.72	4.53	0	18.12	0	0	0	34.22
25	**5**	5.63	4.73	0	6.66	1.09	1.50	0	20.88
25	**10**	1.27	2.96	0	0.38	2.87	0.42	0	1.09

The results of the evaluation of algorithms in identification of overlap is shown in Table 2.This table shows the results for scenarios generated by the 50×20 and 500×50 matrices for biclusters with different values of overlap and increasing noise. This table shows the performance of the algorithms in identification of overlap. The BiMax algorithm was unable to discover overlap between biclusters and the spectral algorithm discovered only 1% to 5% of cells that overlapped. For the 10% overlap scenario and 0% noise, the Plaid algorithm discovered all overlapped cells in the small dataset and 88.32% in the large dataset. For the 25% overlap scenario and no noise, the Plaid algorithm discovered all overlapped cells in the small dataset and 75.25% in the large dataset.

Figure 1. Efficacy of algorithms in identification of biclusters



(a1) (a2)

(b1) (b2)

(c1) (c2)

The efficacy of the algorithms for identification of biclusters (when there is different overlp and noise level) in two generated data matrix with different dims, is presented in Figure 1.Figure 1 shows the ability of the algorithms to identify co-expressed biclusters (a1 and a2), non-overlapping modules with increasing noise levels (b1 and b2), 10% overlap with increasing noise (c1 and c2), and 25% overlap with increasing noise.

References

CheaGan X., Wee A., Liew C., Yan H (2008), Discovering Biclusters in Gene Expression Data Based on High-dimentional Linear Geometrics. BMC Bioinformatics, 9, 209.

Cheng Y., Church G.M. (2000), Biclustering of Gene Expression Data. Intelligent Systems in Molecular Biology, 93-103.

Crick F. (1970) Central Dogma of Molecular Biology. Nature, 227, 561-563.

Hartigan J.A. (1972), Direct Clustering of a Data Matrix. American statistical association (JASA), 67, 123-129.

Jae K. L. (2001), Analysis Issues for Gene Expression Array Data. Clinical Chemistry, 47, 1350-1352.

Kluger Y., Basri R., Chang J.T., and Gerstein M (2003), Spectral biclustering of microarray data: Coclustering genes and conditions. Genome Research. Genome Research, 13, 703-716.

Lazzeroni L., Owen A (2002), Plaid Models for Gene Expression Data. Citeseer, 61-86.

Prelic A., Bleuler S., Zimmermann P., Wille A., Buhlmann P., Gruissem W., et al. (2006), A Systematic Comparison and Evaluation of Biclustering Methods for Gene Expression Data. Bioinformatics, 22, 1122-1129.

19.

Tanay A., Sharan R., Shamir R (2004), Biclustering Algorithms: A Survey. Science, 9, 1-20.

Tanay A., Sharan R., Shamir R (2008), Discovering Statistically Significant Biclusters in Gene Expression Data. Bioinformatics, 18, 136-144.

The chipping forecast II. Special supplement to Nature Genetics Vol 32, 2002.

Santamar´ıa R, Quintales L, Ther´on R. Methods to bicluster validation and comparison in microarray data, In Proc. 8th Int'l Conf. Intelligent Data Engineering and Automated Learning, pages 780–789, 2007.