Comparative Analysis and Evaluation of Biclustering Algorithms for Microarray Data

Ankush Maind and Shital Raut

Abstract From the last decade, the concept of biclustering becomes very popular for the analysis of gene expression data. This is because of the advantages of biclustering algorithms over the drawbacks of clustering algorithms on gene expression data. Many biclustering algorithms have been published in recent years. Some of them performed well on gene expression data and other have some issues. In this paper, analysis of some popular biclustering algorithms have been done with the help of experimental study. Along with this, survey of all the bicluster quality measures which have been used for extracting biologically significant biclusters in various biclustering algorithms is also given. For the experimental study, synthetic dataset has been used. Based on the experimental study, some comparative analyses have been done, and some important issues related to the biclustering algorithms have been pointed out. From this analytical as well as experimental study, newcomers who are interested to do the research in the area of biclustering will get proper direction for the better research.

Keywords Biclustering · Gene expression data · Biologically significant etc.

1 Introduction

Humans are very fast to implement new and advanced technology in their day-to-day activities. From last few decades, advancement in medical sciences is going on but to overcome some common diseases such as cancer, HIV is still a big challenge. Many times due to incorrect diagnosis and improper drugs, people have to lose their lives. There is a need of proper diagnosis of diseases for the proper treatment with the help of proper drugs for the recovery of the disease. To diagnose

A. Maind (✉) · S. Raut
Computer Science & Engineering Department, VNIT, Nagpur, Maharashtra, India
e-mail: ankushmaind@gmail.com

S. Raut
e-mail: saraut@cse.vnit.ac.in

G.M. Perez et al. (eds.), *Networking Communication and Data Knowledge Engineering*, Lecture Notes on Data Engineering and Communications Technologies 4, https://doi.org/10.1007/978-981-10-4600-1_15

the disease and to discover the proper drug for the complex diseases are again a very challenging task. For finding the better solution to this challenging task, the proper analysis on the biological data is required. Many researchers have worked in this area on various kinds of biological dataset for solving the different biological problems.

Various types of biological data have used for the analysis; Microarray gene expression data is one of them. Gene expression data play very important role in the field of the medical for the drugs discovery [1], disease diagnosis [2], gene identification [3], pathway analysis, and other. The functions of the genes and the mechanisms underlying diseases can be identified using gene expression data. So for that, one has to find the pattern from the microarray data, i.e., co-expressed genes. Gene expression data can be generated from the microarray chip. A single microarray chip can take the large amount of gene samples from the multiple tissues at the different conditions or situations. Then after some pre-processing, this chip will generate the gene expression data in the form of matrix in which row indicates the genes and column indicates the samples or conditions. The value in cell of the gene expression matrix represents the expression level of the particular gene at the particular conditions. Presently, gene expression data are used in very wide range for the research in the field of bioinformatics. Because by doing the proper analysis on gene expression data, one can find the solution to many biological issues.

For the research on gene expression data, various techniques have been used. Among these techniques, clustering is one of the important techniques used to extract the significant pattern from the data. Though it is very famous and favorite solution in machine learning, it has some disadvantages like: it works only on the one dimensional not simultaneously on two dimensional. Another drawback is that an element in gene expression matrix can be present either in one cluster or not in any cluster, but same element cannot be present in more than one cluster. But in biology, same gene can be participated in more than one biological process. So, to overcome these disadvantages of clustering, new techniques for finding the biologically significant pattern have been discovered known as biclustering technique. In biclustering, simultaneous clustering will be done on both the directions, i.e., on gene side and on conditions side. It will find the correlated genes across subset of conditions and also identify genes that are not behaved similar in all conditions. Therefore, biclustering technique is more efficient to find biologically significant patterns as compared to clustering techniques.

The concept of biclustering was introduced by the J. Hartigan [4] in 1972, but he has not applied it on gene expression data. Actual working of biclustering on gene expression data has been started by Y. Cheng and G. church in 2000 [5]. After that, plaid model [6], spectral biclustering algorithm [7], FLOC [8], SAMBA [9], ICS [10], CoBi [11], BICLIC [12], and so many algorithms on the biclustering have been published. Within a decade, biclustering became one of the popular techniques for finding the biologically significant patterns from gene expression dataset.

All biclustering algorithms cannot work properly on all types of gene expression dataset. A particular algorithm is bounded to the specific types of dataset. But the purpose of all biclustering algorithms is to find the biologically significant patterns.

Still some issues are present in existing algorithms which have been pointed out in this paper with the help of experiment.

The remaining paper is divided into following sections: Details about the microarray data, biclustering definition, biclustering types, and quality measures are described in background section. Section 2.5 describes and compares the most popular biclustering algorithms which have been used for the experimental study. Section 3 describes the experimental setup and results of experiment. Section 4 discusses the results of all biclustering algorithms and issues which have been pointed out from the experiment.

2 Background

In this section, some details about the microarray data, bicluster definition, types of bicluster, and quality measure for bicluster have been explained.

2.1 Microarray Data

Microarray is a key technology in genomics. DNA microarray [13] data have been used successfully in various research areas such as gene discovery, toxicological research, disease diagnosis, and drug discovery. DNA Microarray data can be used to measure the expression of thousands of genes at the same time. The functions of the genes and the mechanisms underlying diseases can be identified using microarray data. Generally, microarray data are called as gene expression data. The process for getting the gene expression data includes, first selection of cell, after that RNA/DNA preparation have to do, then hybridization process on it after that will get the array image. This resulted image has to analyze then finally will get the gene expression data in matrix form.

Main advantage of microarray is its intrinsic robustness and also it is cheap in cost. Microarrays are in the market from last several years, and today, the microarrays are extremely sensitive and reliable. Microarray data are easily customizable, and reproducible and can be adapted to many situations.

Gene expression data which are available in matrix form are called as gene expression matrix, in which rows represent the genes and column represents the conditions or samples under which gene expressed. The value in the cell represents the amount of mRNA expressed by the particular gene under particular condition. Figure 1 shows the gene expression matrix. This matrix is of m x n dimension, i.e., ‘m’ genes and ‘n’ conditions, in which from G1, G2 to Gm are the genes and from C1, C2, to Cn are the conditions; V11, V12, to Vmn are the amount of mRNA expressed by genes under respective conditions.

2.2 Bicluster Definition

Many people have defined the bicluster in their literature. Bicluster is the submatrix of subset of co-expressed genes across the subset of conditions under which these genes co-expressed. Process of searching bicluster is called biclustering.

Let 'X' be the gene expression matrix of dimension m x n as shown in Fig. 1. Where rows (1, 2,… m) are genes in the matrix and columns (1, 2,…. n) are the conditions under which respective genes expressed. Bicluster 'B' is defined as subset of matrix 'X' containing set I of |I| number of genes and set J of |J| number of conditions, in which b_{ij} indicates the expression levels of gene 'i' under condition 'j'. Following Fig. 2 shows the bicluster 'B'.

2.3 Types of Biclusters

All biclustering algorithms could not produce same results because of many special constraints defined by that specific biclustering algorithm. Many of the algorithms have their own modified dataset to produce the best results, but they are not able to produce the same results on real dataset. Therefore, the result of many algorithms shows different types of biclusters on same real dataset.

Many researchers have defined the types of biclusters in various ways. Table 1 shows the various types of biclusters [3] with their equations, where, 'b_{ij}' indicates the expression levels of gene 'i' under condition 'j', 'π' is constant values for 'B', 'α_i' is adjustment for rows i ϵ I, and 'β_j' is adjustment for column j ϵ J.

Fig. 1 Gene expression matrix

Condition → / Gene ↓	C1	C2	.....	Cn
G1	V11	V12	.....	V1n
G2	V21	V22	.....	V2n
G3	:	:	:	:
:	:	:	:	:
Gm	Vm1	Vm2	...	Vmn

(1)

Fig. 2 Representation of bicluster 'B'

$$B = \begin{bmatrix} b_{11} & b_{12} & \cdots\cdots & b_{1|J|} \\ b_{21} & b_{22} & \cdots\cdots & b_{2|J|} \\ : & : & \vdots & : \\ : & : & \vdots & : \\ b_{|I|1} & b_{|I|2} & \cdots\cdots & b_{|I||J|} \end{bmatrix}$$

(2)

Table 1 Types of bicluster with equation

Types of bicluster	Equation	Eq. No.
Constant	$b_{ij} = \pi$	1
Constant rows	$b_{ij} = \pi + \alpha_i$, $b_{ij} = \pi * \alpha_i$	2
Constant column	$b_{ij} = \pi + \beta_j$, $b_{ij} = \pi * \beta_j$	3
Shifting (coherent value)	$b_{ij} = \pi + \alpha_i + \beta_j$	4
Scaling (coherent value)	$b_{ij} = \pi * \alpha_i * \beta_j$	5
Coherent evolution	No equation	

Biclusters are of various types as mentioned in Table 1. Details about the mentioned bicluster types are as follows,

Constant bicluster: In constant types of patterns, expression levels of all genes under all conditions are same, i.e., constant is available in all cells of bicluster. Equation (1) of Table 1 represents constant bicluster. Constant biclusters also have two categories, first is row constant in which row-wise constant expression levels are present. Equation (2) of Table 1 represents row constant biclusters. Another category is column constant in which column-wise constant expression levels are present. Column constant bicluster is represented with the help of Eq. (3) of Table 1.

Coherent values: In this type of biclusters, expression levels of all genes are in the form of additive or multiplicative. Biclusters with additive expression levels are called additive biclusters. Additive type of bicluster is also called shifting bicluster. Bicluster with multiplicative expression levels is called multiplicative bicluster. Multiplicative type of bicluster is also called scaling bicluster.

Shifting biclusters: In shifting biclusters, expression levels of bicluster are added with constants after every condition, so that the expression levels of genes will shifted with the same difference into next expression level. Shifting bicluster is represented with the help of Eq. (4) in Table 1, in which α_i and β_j are added to the π.

Scaling biclusters: In scaling biclusters, expression levels of bicluster are multiplied with constants after every condition, so that the expression levels of genes will be scaled with some multiplicative difference into next expression level. Scaling bicluster is represented with the help of Eq. (5) in Table 1, in which α_i and β_j are multiplied with constant 'π'.

Coherent evolutions: This type of bicluster is increasing or decreasing type of bicluster which cannot have particular types of patterns. It may be up-regulated or down-regulated. There is no mathematical equation for these types of biclusters.

2.4 Bicluster Quality Measure

Bicluster quality measure plays very important role for searching biologically significant biclusters from the gene expression data. It is useful to decide the quality of biclusters. Many biclustering algorithms have been used various types of quality measures. Each quality measure is used for extracting specific types of biclusters. Till date, no single quality measure was discovered, which is useful for finding all types of biclusters from the gene expression data. Very few quality measures are

Table 2 Quality measures for extracting various types of biclusters with details

Sr. no.	Name of quality measure	Types of bicluster extract	Advantages	Who has applied first time
1	Variance (Var)	Constant	Minimizes the sum of bicluster variance	J. Hartigan [4]
2	Mean square residue (MSR)	Shifting	Efficiently used for extracting additive types of patterns	Cheng and Church [5]
3	Scaling mean square residue (SMSR)	Scaling	Efficiently used for extracting multiplicative types of patterns	Mukhopadhyay [17]
4	Relevance index (RI)	Constant Row, Constant Column	Efficiently used for extracting constant rows or constant column types of patterns	Yip K, Cheung D, Ng M [18]
5	Pearson's correlation coefficient (PCC)	Shifting	Efficiently used for accessing shifting types of patterns. Access the linear relationship among the genes	L. Teng and L. Chan [19]
6	Average Spearman's Rho (ASR)	Shifting, Scaling	Use to measures the statistical dependency between two variables, assessing how well their relationship can be described using a monotonic function, even if their relationship is not linear	Ayadi et al. [20]
7	Average correlation value (ACV)	Shifting, scaling	Use to evaluate the homogeneity of a bicluster or a data matrix	Teng and Chan [19]
8	Virtual error (VE)	Shifting, Scaling	Best method for extracting shifting or scaling patterns	F. Divina and B. Pontes [21]
9	Transposed virtual error (VE^t)	Shifting, Scaling, shifting + Scaling,	It is efficient to recognize shifting, scaling and shifting + scaling patterns in biclusters either simultaneously or independently	B. Pontes, R. Giráldez, and Jesús S. [22]

able to extract the biologically significant biclusters from the microarray data perfectly. Table 2 shows the bicluster quality measures which are used in various biclustering algorithms along with its supported bicluster types, advantages, and researchers who have firstly used this quality measures in biclustering algorithm.

2.5 Biclustering Algorithms

From the last decade, biclustering algorithms became very popular. Many research papers have published it. Everyone tried to improve the performance of their own biclustering algorithm over the existing algorithm. Some of the popular algorithms have chosen for experiments are as follows.

Cheng and Church's (CC) Approach: Cheng and Church [5] invented first biclustering approach which is applied to the gene expression data in 2000. Dataset are used by them are yeast dataset and human (gene dataset). This approach has adopted the strategy of iterative greedy search for finding the bicluster. CC's approach is divided into four steps:

1. Single node deletion
2. Multiple node deletion
3. Node addition
4. Finding a Given Number of Biclusters

They have used the mean square residue (MSR) in their approach as a quality measure of the bicluster. Advantage of this method is, it is very simple. Drawbacks of this approach are, first is, MSR only able to capture shifting tendencies within the data not the scaling, therefore it cannot find the scaling bicluster. Second is, there are always need of masking and need to calculate the threshold every time. Third drawback is, there is a problem of random interference caused by masked biclusters. Fourth is, it cannot find the overlapped biclusters.

Plaid Model (PM): Plaid model [6] is an algorithm for exploratory analysis of multivariate data. It was introduced by L. Lazzeroni, A. Owen, in 2000. In plaid model, gene-condition matrix is represented as a superposition of layers related to biclusters. They have used yeast DNA data, nutrition data, and foreign exchange data for their experiment. They have also used the hierarchical clustering algorithm for ordering. The advantages of this model are, it can find the interpretable structure and allows cluster to overlap. Another drawback of this method is all bicluster membership functions are re-estimated at each step of the iteration.

xMotif: xMotif [2] algorithm was discovered by T.M. Murali, S. Kasif in 2003 with the aim of extracting conserved gene expression motifs from gene expression data. This algorithm is based on the probabilistic model and can identify genes which are

conserved in more than one class but are in different states in different classes. This algorithm is also helpful for extracting the coherent evolution types of patterns.

Spectral Biclustering (**SB**): Spectral biclustering [7] is a linear algebra-based technique proposed by Y. Kluger, R. Basri, J. Chang, M. Gerstein in 2003. This method is designed to cluster populations of different tumors assuming that each tumor type has a subset of marker genes that exhibit overexpression and that typically are not over expressed in other tumors. They have used lymphoma microarray dataset and lymphoma Affymetrix dataset for the analysis. This is the first method of biclustering in which cancer dataset used. This method identifies the distinctive checkerboard like structure from the microarray data.

Iterative Signature Algorithm (**ISA**): ISA [14] was linear algebra-based approach invented by S. Bergmann, J. Ihmels, N. Barkai in 2003. For noisy expression data, this approach leads to better classification due to the implementation of the threshold. Drawback of this method is that there is no evaluation of the statistical significance and additionally two threshold parameters have to define. ISA is very much suitable to any dataset that consists of multicomponent measurements. Applications of the ISA could include the analysis of biological data on protein–protein interactions or cell growth assays, as well as other large-scale data, where a meaningful reduction of complexity is needed.

Flexible Overlapped Biclustering (**FLOC**): FLOC [8] is move-based probabilistic algorithm introduced by J. Yang, H. Wang, W. Wang, and P.S. Yu in 2005. This approach is the extension of CC's approach. Most of the issues of CC's approach have been overcome by this method. FLOC approach proved that 'random interference phenomenon' plays a very important role for the discovery of the high-quality bicluster. FLOC can discover a set of k possibly overlapping biclusters simultaneously. They have used the MSR as a quality measure.

Binary inclusion-maximal biclustering algorithm (**BIMAX**): Bimax [15] is one of the simple and fast method for the bicluster searching which was invented by Prelic, Bleuler, Zimmermann, Wille, Buhlmann, Gruissem, Hennig, Thiele, and Zitzler in 2006 with the aim of comparative study of existing algorithms such as CC [5], OPSM [16], ISA [14], SAMBA [9], and xMotif [2] with Bimax [15]. This approach applied to the synthetic and real dataset, i.e., gene expression data. An advantage of Bimax algorithms is that it is capable of generating all optimal biclusters, given the underlying binary data model. Bimax also requires less memory resources. In their approach, they have proved that Bimax perform well as compared to mentioned algorithms.

Factor Analysis for Bicluster Acquisition (**FABIA**): FABIA [1] is based on a multiplicative model and it was introduced by S. Hochreiter, U. Bodenhofer, M. Heusel, A. Mayr, A. Mitterecker in 2010. This algorithm helps to model heavy tailed data as observed in gene expression data. They have applied FABIA biclustering algorithm to 100 dataset. They also have applied the FABIA

successfully to drug design to find compounds with similar effects on gene expression data.

Biclustering by Correlated and Large Number of Individual Clustered seeds (BICLIC): BICLIC [12] is one of the popular biclustering algorithms for extracting biologically significant bicluster from the gene expression data. It was published by T. Yun, G.-S. Yi in 2013. In this algorithm, Pearsons correlation coefficient was used as a quality measure. BICLIC has solved the problem of changing output in multiple executions on same dataset. But, BICLIC does not have the overlapping control strategy.

3 Experimental Study

In experimental study, synthetic dataset have been taken. For this experiment, some of the popular biclustering algorithms which are already implemented in 'R' language have been used. 500 × 50 data matrices with random values which are normally distributed have been generated. Then, six types of biclusters with various sizes have been implanted. Types of biclusters such as constant bicluster, constant rows bicluster, constant column bicluster, shifting bicluster, scaling bicluster, and coherent evolution bicluster have been implanted into synthetic data matrix. After that, all biclustering algorithms as mentioned applied one by one. Table 3 shows the result of nine biclustering algorithms. First, column of Table 3 is name of algorithms. Number of bicluster found is a second column, in which maximum limit of biclusters are as ten, but some of these algorithms have extracted less than ten biclusters. Extracting more number of bicluster is good quality of biclustering algorithms, but these biclusters should be biologically significant otherwise it is of useless. Next column of Table 3 is maximum size of bicluster; it is a very important parameter for the biological application because generally the size of bicluster is more then it will give more biologically significance. One can predict more

Table 3 Biclustering algorithms with results

Algorithm	No. of bicluster found	Max. Size of bicluster	Avg. size of bicluster	Nature of output
CC	10	28 × 16	20 × 14	Changing
PM	7	17 × 11	10 × 6	Changing
xMotif	4	20 × 20	18 × 13	Fixed
SB	2	16 × 6	16 × 6	Changing
ISA	4	30 × 3	25 × 2	Changing
FLOC	10	23 × 20	11 × 11	Changing
Bimax	7	20 × 20	12 × 9	Fixed
FABIA	10	10 × 10	4 × 14	Changing
BICLIC	10	50 × 49	36 × 32	Fixed

accurately. Here, BICLIC algorithm has extracted the largest size of bicluster. Fourth column of Table 3 shows the average size of biclusters, from which, will get the idea about the average size of biclusters extracted from gene expression data using respective algorithms. Largest average size bicluster was extracted by the algorithm 'BICLIC'. Fifth column is 'nature of output' which indicates the output of biclustering algorithms is changing or fixed after every execution on same dataset. If the output of the algorithms is changing, then such result one cannot easily apply to the biological applications. Accuracy of the result will not maintain due to which one cannot predict easily about the disease and other thing also. Therefore, output of the algorithms should be fixed after the multiple executions on same dataset.

Based on experiment analysis, comparative analysis of all these algorithms have been done. Table 4 shows the comparative analysis of experimental study. First column of table is of name of algorithms. Second column represents the complexity of the algorithms which have been taken from their respective own authors publications as it is. Here, M is the number of genes, N indicates number of conditions, B indicates number of biclusters, K indicates number of seeds, N_{iter} indicates iterations, 'ns' indicates number of samples randomly selected, 'nd' indicates number of sets of genes for each sample, 'Ni' indicates input sets, M' is average number of genes, and N' is the average number of conditions. Next column of table is Extracted Bicluster type. From the experimental study, some observations have been pointed out, such as which algorithms extract which types of patterns, for example, constant, coherent values, scaling, shifting, or combinations as mentioned in Table 1. Last column of table represents the strategy of biclustering algorithms like one at a time, simultaneous, or set of bicluster at a time. It represents how the bicluster is extracted whether it is in set or one at a time or simultaneously all.

Table 4 Comparative details of biclustering algorithm

Algorithm	Complexity	Extracted bicluster type	Strategy
CC	$O(M \times N)$	Additive coherent values (Shifting)	One at a time
PM	……	Coherent values	One at a time
xMotif	$O(N \times ns \times nd)$	Coherent evolution	Simultaneous
SB	……	Coherent values	Simultaneous
ISA	$O(N_{iter} \times Ni \times (N \times M' + M \times N'))$	Coherent values	One at a time
FLOC	$O((N + M)2 \times K \times N_{iter})$	Additive coherent values (Shifting)	Simultaneous
Bimax	$O(M \times N \times \beta min)$	Up-regulated	One set at a time
FABIA	$O(M \times N \times B^2)$	Constant values	One at a time
BICLIC	……	Coherent values, negative correlation	One set at a time

4 Discussion

In this paper, both the analytical and experimental studies of biclustering algorithms have been done. In analytical study, analysis of the different types of biclusters, types of quality measures used in various biclustering algorithms, and popular biclustering algorithms with their advantages and disadvantages along with complexity of biclustering algorithms have been done. From this study, some issues have been pointed which are, the coherent evolution types of bicluster are very difficult for the extraction from the gene expression. Second difficult pattern for the extraction is scaling types of bicluster. For the extraction of various types of biclusters, quality measure plays very important role. In this paper, analyses of all the quality measures which have been used in various biclustering algorithms have been done. Transposed virtual error quality measure is useful for the extraction of three types of biclusters such as shifting, scaling, and shifting + scaling have been observed. VE^t is the only measure which is helpful for the extraction of these three types of biclusters.

For the experimental analysis, synthetic dataset have been used. In synthetic dataset, some biclusters of different types such as constant, scaling, shifting, coherent evolutions have been implanted. All the results of experimental study have mentioned in Sect. 3. From this experiment, one can say that some algorithms have performed well but some have not. So, by using the same dataset, one cannot compare all algorithms performance. Very few algorithms have extracted the patterns as it is like implanted patterns in the matrix. Most of them have extracted the biclusters but not accurately. So, to find the perfect bicluster from the gene expression data is a challenging task. Many researchers have claimed that their algorithms are perfect but as per the observation from the experimental study, not a single algorithm has extracted all these implanted biclusters accurately. Therefore, one can work in this area to improve the biclustering techniques. From this experimental study and theoretical analysis of other existing biclustering algorithms, some issues related to the biclustering algorithms have been pointed out which are as follows:

1. Most of the existing biclustering algorithms fail to efficiently find the sets of biologically significant biclusters from gene expression data perfectly.
2. Many biclustering method results shows lack of stability. Because these biclustering algorithms depend on random starting seeds, due to which contents of resulting biclusters are changing every time even though the same biclustering algorithm is applied to the same gene expression microarray dataset.
3. One cannot guess the truth of real biological gene expression dataset because it is unknown. It is challenging to verify the biological relevance. Existing algorithms also fail to completely extract the significant pattern from the gene expression dataset, i.e., accurate bicluster will not get.
4. Every existing algorithm is bound to the particular types of dataset. These algorithms will not perform well on all types of dataset.

5. Processing of large-scale gene expression data will take more computational cost. Many existing biclustering algorithms have this problem of high complexity.
6. Most of the gene expression dataset contain noise, due to which one cannot get the proper results. For removing noise, proper pre-processing on gene expression dataset is required, so that, one can find the biologically significant patterns.
7. Searching constant, shifting, scaling, shifting + scaling types of patterns is challenging task from the single biclustering algorithm. Many existing algorithms fail to search all these types of patterns at the same execution.
8. To find the set of larger size of biclusters efficiently is difficult task by using the simple workstation. There is a need of high configuration server. So more hardware cost also required.
9. After finding the pattern from the dataset, one needs to do the statistical validation and biological validation of the patterns to check the biological significance of the patterns. This is also a challenging task.
10. To find the biologically significant bicluster, there is a need of proper quality measures. So choosing the proper quality measure for the specific types of dataset is also challenging task.

References

1. Hochreiter S, Bodenhofer U, Heusel M, Mayr A, Mitterecker A, FABIA: factor analysis for bicluster acquisition. Bioinformatics, Vol. 26. (2010) 1520–1527.
2. T.M. Murali, S. Kasif, Extracting conserved gene expression motifs from gene expression data, Pacific Symposium on Biocomputing, (2003) 77–88.
3. Madeira, S.C. and Oliveira, A.L. Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Trans. Comput. Biol. Bioinform. Vol. 1. (2004) 24–45.
4. J. Hartigan, Direct clustering of a data matrix, J. Am. Stat. Assoc. Vol. 67. (1972) 123–129.
5. Cheng, Y. and Church, G. Biclustering of expression data. Proc. Int. Conf. Intell. Syst. Mol. Biol. (2000) 93–103.
6. L. Lazzeroni, A. Owen, Plaid models for gene expression data, Stat. Sinica. Vol. 12. (2002) 61–86.
7. Y. Kluger, R. Basri, J. Chang, M. Gerstein, Spectral bicluster of microarray data: coclustering genes and conditions, Genome Res. Vol. 13. (2003) 703–716.
8. J. Yang, H. Wang, W. Wang, P.S. Yu., An improved biclustering method for analyzing gene expression profiles, Int. J. Artif. Intell. Tools. Vol. 14. (2005) 771–790.
9. A. Tanay, R. Sharan, R. Shamir, Discovering statistically significant biclusters in gene expression data, Bioinformatics, Vol. 18. (2002) 136–144.
10. H. Ahmed, P. Mahanta, D. Bhattacharyya, J. Kalita, Shifting-and-scaling correlation based biclustering algorithm, IEEE/ACM Trans. Comput. Biol. Bioinform. Vol. 11. (2014) 1239–1252.
11. S. Roy, D.K. Bhattacharyya, J.K. Kalita, CoBi: pattern based co-regulated biclustering of gene expression data, Pattern Recogn. Lett., Vol. 34. (2013) 1669–1678.
12. T. Yun, G.-S. Yi, Biclustering for the comprehensive search of correlated gene expression patterns using clustered seed expansion, BMC Genom., Vol. 14. (2013) 144.

13. P. Baldi and G.W. Hatfield, DNA Microarrays and Gene Expression. From Experiments to Data Analysis and Modelling. Cambridge Univ. Press, 2002.
14. S. Bergmann, J. Ihmels, N. Barkai, Iterative signature algorithm for the analysis of large-scale gene expression data, Phys. Rev., Vol. 67. (2003) 031902.
15. Prelic A, Bleuler S, Zimmermann P, Wille A, Buhlmann P, et al., A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics. Vol. 22. (2006) 1122–1129.
16. A. Ben-Dor, B. Chor, R.M. Karp, Z. Yakhini. 2003. Discovering local structure in gene expression data: the order-preserving submatrix problem. J. Comput. Biol. 10, 3–4 (2003), 373–384.
17. A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, A novel coherence measure for discovering scaling biclusters from gene expression data, J. Bioinform. Comput. Biol. Vol. 7. (2009) 853–868.
18. Yip K, Cheung D, Ng M, Harp: A practical projected clustering algorithm. IEEE Transactions on Knowledge and Data Engineering, Vol. 16. 1387–1397.
19. Li Teng and Laiwan Chan. Discovering biclusters by iteratively sorting with weighted correlation coefficient in gene expression data. Signal Processing Systems. Vol. 50. 267–280.
20. Ayadi W, Elloumi M, Hao J, A biclustering algorithm based on a bicluster enumeration tree: application to dna microarray data. BioData mining, Vol. 2. (2009) 1–16.
21. F. Divina, B. Pontes, R. Giráldez, J.S. Aguilar-Ruiz, An effective measure for assessing the quality of biclusters, Comput. Biol. Med., Vol. 42. (2012) 245–256.
22. Pontes B, Giráldez R, Aguilar-Ruiz J Measuring the quality of shifting and scaling patterns in biclusters. Pattern Recognition in Bioinformatics, (2010) 242–252.