ISSN: 2348 - 2117

International Journal of Engineering Technology and Computer Research (IJETCR)
Available Online at www.ijetcr.org
Volume 3; Issue 6; November-December-2015; Page No. 50-63

Study and Analysis of Gene Expression Profile Patterns of Myocardial Infraction Dataset Using Biclustering Approach

Dr. R.Porkodi[1], R.Tamilarasi[2]

[1]Bharathiar University, Department of computer science, school of computer science and Engineering, Bharathiar University, Coimbatore-46

porikodi_r7@yahoo.com

[2]Bharathiar University, Department of computer science, school of computer science and Engineering, Bharathiar University, Coimbatore-46

tamiljesus.amala16@gmail.com

Abstract

The proposed research work is to study and analyse the four biclustering algorithms BCPlaid BCXmotifs, BCSpectral and BCBimax using myocardial infraction microarray dataset. The methodology of the proposed research work consists of 3 Phases Pre-processing phase is used to convert the expression data profile into Binarnized form and discretized form. The need for converting the dataset into these forms helps certain algorithms to give effective result. The BCBimax algorithm takes Binarnized form of dataset as input whereas the BCXmotifs algorithm takes input as discretized form of dataset. The second phase is to study four biclustering algorithms and to analyse the experimental dataset. The biclustering results are to be validated using coherence measure value and Jaccard index. The visualization of biclustering results to be implemented using the heat map and parallel coordinates.

Key Words: Data Mining, Biclustering, BCPlaid, BCXmotif, BCSpectral, BCBimax

1. Introduction

Data mining is the process of discovering insightful, interesting, and novel patterns, as well as descriptive, understandable, and predictive models from large-scale data. Data Mining is defined as extracting information from huge sets of data. In other words, we can say that data mining is the procedure of mining knowledge from data. Data collection and storage technology has made it possible for organizations to accumulate huge amounts of data at lower cost. Exploiting this stored data, in order to extract useful and actionable information, is the overall goal of the generic activity termed as data mining. Bioinformatics is the computational analysis of biological data, consisting of the information stored in the form of DNA and protein sequence in various biological databases. The bioinformatics is the field of science in which biology, computer science and information technologies merge into a single discipline. There are three important sub-disciplines within bioinformatics the development of new algorithms and statistic which assess relationships among members of large dataset Microarray is the collection of microscopic DNA spots attached to a solid surface. Scientists use DNA microarray to measure the expression levels of large number of gene concurrently or to genotype multiple regions of genome. These can be a short section of a gene or other DNA elements that are used to hybridize a cDNA or cRNA sample called target.

Biclustering is an unsupervised learning technique which over the last few years has been widely used in microarray analysis, outperforming traditional clustering. While clustering techniques group genes similarly expressed under all conditions or vice versa (clusters), biclustering techniques group them under a certain subgroup of conditions (groups of both genes and Conditions are called bicluster). Biclustering may be applied to group genes with constant expression value, samples with constant expression values, gene based on additive model and genes with multiplicative model. A gene or condition can be in more than one bicluster at the same time (overlapping), while in clustering a gene or condition

Corresponding author: Dr.R.Porkodi

is usually assigned to a unique cluster. The biclustering algorithms can be found in Madeira and Oliveira (2004). Biclusters are more flexible and fit biological behaviour better than clusters, but their special characteristics (overlapping and grouping of genes and conditions) make it difficult to apply cluster visualizations to bicluster. The set of values a_{ij} represent the relation between its rows i and its columns j and the goal is to identify subset of rows with certain coherence properties in a subset of the columns. Most biological applications of biclustering are performed using gene expression data obtained using microarray technologies that allow the measurement of the expression level of thousands of genes in target experimental conditions.

This paper is organized as follows: section1 describes the introduction on biclustering microarray gene expression data to identify in myocardial infraction data set using biclustering algorithm, Section 2 presents review of the literature for the clustering and a biclustering method over microarray dataset. Section 3 describes the methodology of the proposed work to identify the biclustering patterns for experimental dataset. Section 4 discusses the experimental result of the proposed work. Section 5 given the conclusion.

2. Literature Review

Osama Abu Abbas et al [3] proposed the study and comparison of various clustering algorithms. Algorithms have been compared based on factors such as Dataset size, number of clusters, dataset and used software. The R package includes an iterative algorithm for solving this objective. After computing an initial clustering with standard k-Means, the algorithm alternates between computing the optimal w for fixed L and C (with a closed form soft-thresholding type operation), and computing L and C for fixed using K–means on a reweighted dataset Z with Z(j) for each j. Since one of the factors we investigate in this work is the effect of initialization, we modify this code slightly to allow initialization with an arbitrary clustering.

Grothaus et al[4] represents overlapping biclusters in a single heat map and allow row and column duplications if biclusters cannot be represented contiguously. While being optimal with respect to the number of duplications, such an automatic layout algorithm does not allow for interactivity. In our work, we allow overlapping biclusters and the user may decide which biclusters to show contiguously in order to minimize row and column duplications.

Cheng and Church et al [6] presented the bicluster as a subset of rows and a subset of columns with a high similarity score. The similarity score introduced and called *mean squared residue*, was used as ameasure of the coherence of the rows and columns in the bicluster. Given the data matrix a bicluster was defined as a uniform sub-matrix having a low mean squared residue score. A sub matrix is considered bicluster for some In particular; they aim at finding large and maximal bicluster with scores below a certain threshold.

Kluger et al [15] refers, that addressed the problem of identifying bicluster with coherent values and looked for checkerboard structures in the data matrix by integrating biclustering of rows and columns with normalization of the data matrix. They assumed that after a particular normalization, which was designed to accentuate bicluster if they exist, the contribution of a bicluster is given by a multiplicative model as defined in (13). Moreover, they use gene expression data and see each value the data matrix as the product of the background expression level of gene the tendency of gene to be expressed in all conditions and the tendency of all genes to be expressed in condition.

Murali and Kasif et al [30] proposed the conserved gene expression motifs (xMOTIFs) they defined an xMOTIFs as a subset of genes (rows) that is simultaneously conserved across a subset of the conditions (columns). The expression level of a gene is conserved across a subset of conditions .The gene is in the same state in each of the conditions in this subset. The data may contain several xMOTIFs (bicluster) and aimed at finding the largest xMOTIFs.

Mir kin et al [10] presented the biclustering has been used to describe simultaneous clustering of both row and column sets in a data matrix. Other terms that have been associated to the same idea include direct clustering and box clustering.

Ben-Dor et al [11] defined a bicluster as an order-preserving sub matrix (OPSM). Specifically, a sub matrix is order-preserving if there is a permutation of its columns under which the sequence of values in every row is strictly increasing. They define a complete bicluster model as the pair (Y,π) where $\pi=$

© 2015 IJETCR. All Rights Reserved.

(*y*1… *ys*) is a linear ordering of the columns in *Y*. A row supports (*Y*,π) if the *s* corresponding values, ordered according to the permutationπ, are monotonically increasing. Since an exhaustive algorithm that tries all possible models is not feasible, the algorithm grows partial models iteratively until they become complete models.

Juan A Nepomucenol et.al [39] described that aim of finding bicluster from gene expression data. In this algorithm the proposed fitness function is based on the linear correlation among genes to detect shifting and scaling patterns from genes and an improvement method is included in order to select right now positively correlated genes. The proposed algorithm has been tested with dataset the performance of the proposed method and fitness function are compared to that of Bimax, xMOTIFs, plaid and spectral.

Amela Preli et al [1] focuses on common setting that reflects the general basic of the majority of the biclustering studies available and in particular of those techniques are considered. The comparison focused on the identification of co-expressed genes as in other biclustering algorithms.

Segal et al [22] described the probabilistic model, which is based on the probabilistic relational models (PRMs). These models extend Bayesian networks to a relational setting with multiple independent objects such as genes and conditions. By using this approach manage to discover a set of biclusters with constant values on their columns.

Barkow et al and Cheng et al [17] discussed the Biclusters are more flexible and fit biological behavior better than clusters, but their special characteristics (overlapping and grouping of genes and conditions) make it difficult to apply cluster visualizations to bicluster. While some cluster visualization techniques can be adapted to the representation of single bicluster.

Hibbs et al [20] present that linked-views approach, so two visualizations, heat maps and cluster projections, are displayed simultaneously, boosting the visual analysis. The projection used is similar to that of gCLUTO but now in a 3D space. It improves heat map comparison of transcription levels and a "karyoscope" visualization that represents the transcription levels of the genes under one condition, ordered as they are located in the chromosomes.

Pavan and Pelillo et al [34] approaches the algorithm is based on the dominant set approach of a bicluster, genes and conditions are iteratively sorted using weight vectors, which are also iteratively refined using the sorting vector of the previous iteration. In each iteration the matrix is transposed and the process is repeated over the other dimension, thus alternating from genes to conditions.

Wang et al [36] presented the algorithm that performs exhaustive bicluster enumeration, subject to a restriction that they should posses a minimum number of rows and a minimum number of columns. To speed up the process and avoid the repetition of computations, they use a suffix tree to efficiently enumerate the possible combinations of row and column sets that represent valid bicluster.

Kaiser et al [32] refers the extension for the R environment package proposes a variety of computation algorithms and also many visualization techniques to represent the biclustered data. In addition to the traditional heat map representation and parallel-coordinate plots also bubble charts are provided. Our tool allows using BiClust or any other algorithm provided in R for the computation of bicluster. BiCat is another application that can be used to analyze biological data such as gene expression data. In contrast to R it is not based on command line arguments but provides a graphical user interface to manipulate and navigate in the data. Expression View is another R package that allows heat map-based browsing of bicluster obtained from gene expression experiments. The tool uses an ordering that maximizes the areas of the largest contiguous parts of bicluster.

Cheng at el [41] developed BiVisu an open-source software tool for detecting and visualizing biclusters embedded in a gene expression matrix and display the co -regulated genes. BiVisu clustering methods in which partition data based on whole set of genes or conditions, biclustering groups a subset of genes (rows) over a subset of conditions (columns).

3. Methodology

The methodology of the proposed research work is to study and analysis the four biclustering algorithms namely BCPlaid, BCSpectral, BCXmotifs and BCBimax using myocardial infraction microarray data set. Then, the biclustering results are validated and

© 2015 IJETCR. All Rights Reserved.

verified using coherence value and Jaccard index. Finally, visualize the biclustering results using the heat map and parallel coordinates. Fig.1 shows the methodology of proposed research work.

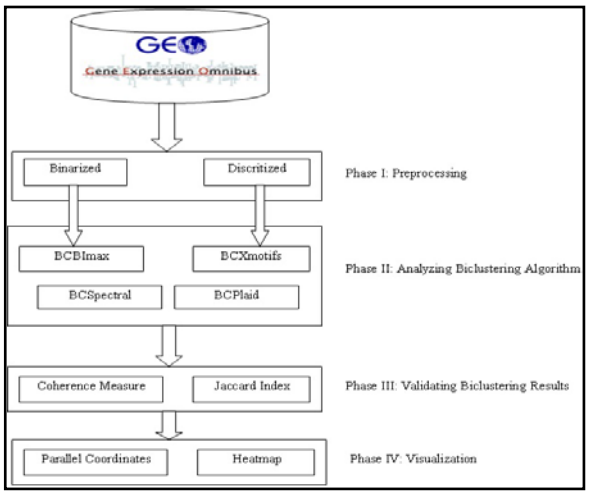

Figure 1: Framework for Proposed Methodology

3.1. Preprocessing

The preprocessing phase involves two different functions. These two functions are useful for certain biclustering method and these functions are used to make the dataset into acceptable form
Which is required for certain algorithms such as BCBimax and BCXmotifs, etc.As the proposed research work use four biclustering methods, BCBimax takes input as Binarnized from and BCXmotifs takes input as discritized from.

Binarnized method: - Binarnized which the converts real dataset values into 0's and 1's.The bimax algorithm uses input as this binarized data. The divide and conquer method used in bimax biclustering algorithm only works with binary matrices the non – binary data converted in to binary data number of ways more sophisticated binarizing method be used.

Descritized method: - Descritized method needs a discrete matrix function delivers a discrete matrix with either a given number of levels of equally spaced intervals from minimum to maximum or levels of same size using the quantiles. The xMOTIFs biclustering algorithm searches for rows with constant values over a set of columns.

3.2. Analyzing Biclustering Algorithms

This section describes the basic idea behind in each bicluster method and how these algorithms implemented in R using bicluster package.

A.BCPlaid:- The Plaid biclustering algorithm is a statistically inspired modeling approach developed by Lazzeroni and Owen for the analysis of gene expression data. The basic idea is to represent the genes-conditions matrix as a superposition of layers, corresponding to biclusters in our terminology, where each layer is a subset of rows and columns on which a particular set of values takes place. Different values in the expression matrix are thought of as different colors, as (false colored) heat maps of chips.

B.BCSpectral:- The Spectral Methods are a class of spatial discretizations for differential equations: they provide a way of translating an equation expressed in continuous space and time into a discrete equation which can be solved numerically. There are three primary benefits of Spectral Methods over alternative approaches, such as Finite Elements or Finite Differences. The purpose of changing the representation need not be solely to minimize the error of discretizations. Analysis of the spectrum of a signal is the most natural and powerful approach for solving many signal processing problems.

C.BCXmotifs:- The XMOTIFs biclustering algorithm for rows with constant values over a set of columns. For gene expression data, they call the biclusters conserved genes expression motifs, short XMOTIFs. Again, finding a good preprocessing method is crucial, because the main aspect of their algorithm is to define a gene state where a gene (row) is called conserved, if it has the same state in all samples (columns). One way to deal with gene states is to simply descritized the data. This algorithm finds sub matrices where all rows have the same value structure over the columns. So here it is possible to find groups with a large variance in their values in the row direction.

D.BCBimax:-The Bimax clustering algorithm presented by Prelic et al. (2006) finds subgroups in a binary matrix where all entries are one. The algorithm iterates rearrange the rows and columns to concentrate ones in the upper right of the matrix. The divide the matrix into two sub matrices. Whenever in one of the sub matrices only ones are found, this sub matrix is returned. In order to get satisfying results the method has to be restarted several times with different starting points. Although the algorithm was originally designed to deliver ideas for bicluster validation, it can also be used as a bicluster method itself.

© 2015 IJETCR. All Rights Reserved.

3.3. Validating Bicluster Results

The proposed approach is validated using validation metric. The verification of the algorithm is done using coherence metric and Jaccard index. The proposed bicluster based on maximum threshold is compared with biclustering R packages. The proposed work generates constant and row wise clusters. Proposed work does not provide any additive and multiplicative clusters for the dataset. The myocardial infraction microarray dataset is analyzed and the biclusters are generated using four algorithms. The biclustering extracted results are verified and validate using coherence measure and Jaccard index.

A. Coherence measure: - Most common method in biclustering is to measure the degree of coherence or similarity in behavior among objects in a bicluster. A bicluster can be presented in a plot based on the degree of coherence. Ideally the element at row i and column j of the bicluster has value given in following Eq (1).

$$\alpha ij = F(\alpha i, \beta j, \mu) \quad (1)$$

Where αi is a constant specific to row i, βj is a constant specific to column j, μ is a constant specific to the bicluster, and F is some function. Coherent values on both rows and columns. This kind of biclusters identifies more complex relations between genes and conditions, either in an additive or multiplicative way given in Eq (2) and (3) respectively

Additive Coherence

$$F(\alpha i, \beta j, \mu) = \mu + \alpha i + \beta j \quad (2)$$

Multiplicative Coherence

$$F(\alpha i, \beta j, \mu) = \mu * \alpha i, * \beta j \quad (3)$$

Where F represents any constant value for B, βi ,(1 ≤ i ≤ |I|)and βi ,(1 ≤ i ≤ |I|) refer to constant values used in additive models for each gene i or condition j; and αi ,(1 ≤ i ≤ |I|) and α_j ,(1 ≤ j ≤ |J|) correspond to constant values used in multiplicative models for each experimental gene i or condition j.

This section presents with the detailed of biclustering proposed work using constant Measure value. The addition methods for contains coherence measures and Constance for single bicluster. This allows simultaneous clustering of the rows and columns of a matrix [33]. The work for classifying bicluster, the function constant variance returns the corresponding variance rows as the average the sum of Euclidean distances d_{xy} between all rows of the bicluster given in following.Eq(4).

$$d_{xy} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (4)$$

B.Jaccard Index: -The Jaccard index is a measurement of similarity among two cluster results, zero means no concordance, and one means that the results are identical. This can be partly explained by different pre-processing steps which were necessary such that the data conform to the respective assumptions of the algorithms. A first validation can be done by simply comparing the Jaccard index of the results. The matrix jac result contains the result of comparing each algorithm to all other algorithm calling jaccardind () given in following Eq(5).

$$jac(BCiBCj) = jacij = \frac{BCi \cap BCj}{BCi \cup Bcj} \quad (5)$$

3.4 Visualization of Bicluster Results

Visualization of biclustering results are stored in consistent classes, it is easy to implement visualization techniques which work for results of different algorithms. Parallel coordinates (function parallel Coordinates ()) can be used to visualize similarity of rows over columns within a bicluster. Heat maps (draw Heat map ()) highlight the difference between the bicluster and the surrounding rows and columns.

A. Heat map: - The data as an overview in a single static view. Many implementations of the heat map use multi-hue color maps to indicate up- and down regulated genes separately. In this work maps data values to grayscale values using linear interpolation between the smallest and largest value of the data matrix, as changes in single-hue color maps are perceived more accurately than red to green color scales for contiguous regions in the matrix using row and column.

B.Parallel-Coordinate Plots: - Parallel-coordinate plots to display biclusters. In proposed work, visualization tool exploits linking and brushing techniques to link heat map and parallel-coordinate plots. This allows a user to explore the data from different points of view. The parallel-coordinates plot, each poly line represents the expression of a gene over all conditions (which are represented by vertical axes). Genes belonging to a bicluster are

© 2015 IJETCR. All Rights Reserved.

rendered using the same color as the corresponding bicluster in the heat map.

4. Analysis of Biclustering Patterns

The proposed research work consisting of following four phases and the results obtained are discussed in next coming sections. Pre-processing, Proposed Biclustering algorithm, Validation, Visualization.

4.1 Result of Pre-processing

There are two pre-processing functions are used in the proposed research work to convert the gene expression data profiles in to some acceptable form which is required for some biclustering functions. The Fig. 2 show the binarized form of data set for threshold value 8 is required to run Bimax algorithm. The expression value below the threshold is replaced with 0 and above threshold is replaced with 1 in binary matrix.

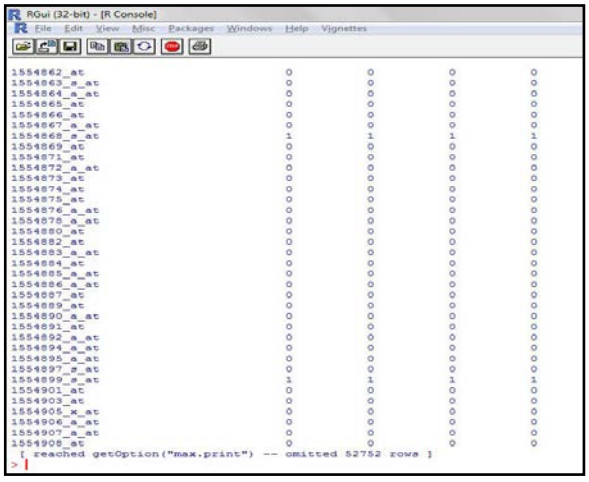

Figure 2: Binarized form of dataset

The Figure3.shows the discretized form of dataset which is required as an input for Xmotifs algorithm

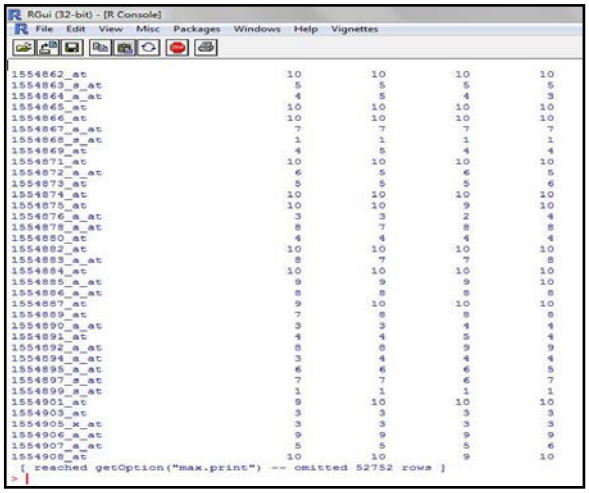

Figure 3: Discretized form of dataset

4.2. Result of Biclustering Algorithms

A.BCPlaid Algorithm

The BCPlaid algorithm produces bicluster with constant rows and constant columns. This algorithm was implemented using ordinary least squares (OLS) in which mean, row and column are considered to compute bicluster by calculating the sum of squares of the layer (LSS). Finally return bicluster if LSS is higher. The result of BCPlaid algorithm is tabulated in Table 2. The required variables and the values initialized are given below. back.fit = 2, shuffle = 3, fit.model = ~m+a+b, iter.startup = 5, iter.layer = 30.

B.BCXmotifs Algorithm

The BCXmotifs algorithm considers rows with constant values over a set of columns and produces clusters of sub matrices where all row have same value structure over the columns. The biclustering result of above algorithm is represented in Table 3. The required variables and the values initialized are given below.Ns = 200, nd = 200, sd = 4, alpha = 0.05, number = 5.

Table 1: Bcplaid algorithm

Number of Bicluster	BC1	BC2	BC3	BC4
Rows	4244	443	55	22
Columns	12	12	10	9

Table 2: Bcxmotifs algorithm

Number of Bicluster	BC1	BC2	BC3	BC4	BC5
Rows	26077	10484	4397	2810	1974
Columns	5	5	5	5	5

C.BCSpectral Algorithm

The BCSpectral algorithm takes the log2 transformation of the experimental dataset and produces the number of clusters based on checker board structure. This algorithm uses singular value Decomposition (SVD) and the resulting eigen value and eigen vector to retrieve bicluster. The bicluster identified by BCSpectral algorithm is tabulated in 4. The required variable and the value initialized are given below. WithinVar = 4.

© 2015 IJETCR. All Rights Reserved.

Table 3: Bcspectral algorithm

Number of Bicluster	BC1	BC2	BC3	BC4	BC5
Rows	6	6	6	6	6
Columns	6	8	17	12	9

D.BCBimax Algorithm

This algorithm finds subgroups in a binarized form of data set where all entries are one. The table 5 presents the biclustering result of Bimax Algorithm. The required variables and the values initialized are given below. Minr = 5, minc = 5, number = 50

Table 4: Bcbimax algorithm

Number of Biclustering	BC1	BC2	BC3	BC4	BC5
Rows	10307	10557	10560	10469	10302
Column	5	5	5	5	5

4.3. Result of Bicluster Validation

The biclustering results of the four algorithms are compared with each other using Jaccard index. The Jaccard index is useful to compare the calculated results between each other. To compare to biclustering it calculates the fraction of row, column combination in both bicluster from all row-column in at least one bicluster [27]. Two bicluster with one hundred percent similarity have got a Jaccard index of 1 and two bicluster with no equal elements have a Jaccard index of 0 by four biclustering algorithm.

A. Coherence measures: - In addition to Jaccard index for comparing bicluster result of two algorithms; there is important to calculate the coherence measure for single bicluster. This measure is calculated based on additive, constant and multiplicative variance of a bicluster. The table 7 show the coherence of first bicluster found by BCPlaid, BCXmotifs, BCSpectral and BCBimax using constant, additive and Multiplicative variances. The constant variance has been calculated for the dimension 'column' and the additive and multiplicative variances are calculated based on the dimidiation 'both' which consider both rows and column.

Table 5: Coherence measure

Biclustering Algorithm	Additive variance	Multiplicative variance	Constant variance
BCPlaid	0.48448774	0.06476704	7.775096
BCXmotifs	0.326719	0.234157	17.09955
BCSpectral	1.271691	0.2364157	4.941058
BCBimax	0.075812	1.357417	22.29504

4.4.Visualization of Bicluster

The visualizing bicluster results based on gene expression data. The proposed research work uses heatmap representations and parallel-coordinate plots so that more than one bicluster can be visualized simultaneously. The contiguous representation of selected biclusters allowing rows and columns.

A.Heat Map: - Heat maps are a good choice to represent a large portion of the data as An Over view in a single static view. The data values to grayscale values using linear interpolation between the smallest and largest value of the data matrix, as changes in single-hue color maps are perceived more accurately than red to green color scales for continuous data values. In general, it is not possible to represent more than two biclusters in a way that all of them are located in contiguous regions in the matrix using row and column permutations only. The visualization of bicluster obtained by BCPlaid, BCXmotifs BCSpectral and BCBimax algorithm are given in below paragraphs.

BCPlaid:-The Figure. 8 shows the heat map of the first bicluster out of 4 bicluster found by BCPlaid biclustering algorithms.

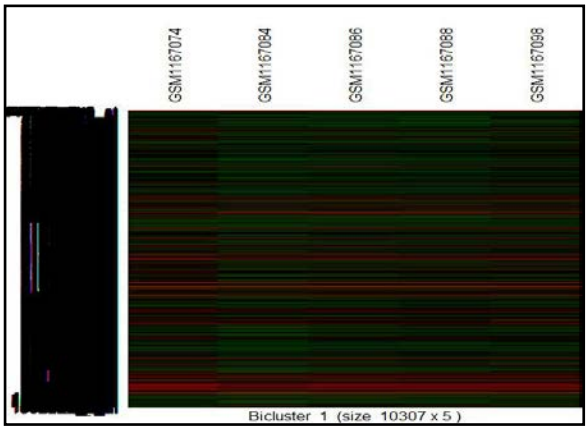

Figure 4: Bcplaid heat map

© 2015 IJETCR. All Rights Reserved.

BCXMOTIFs:-The Figure. 9 shows the heat map of the first bicluster out of 5 bicluster found by BCXmotifs biclustering algorithms.

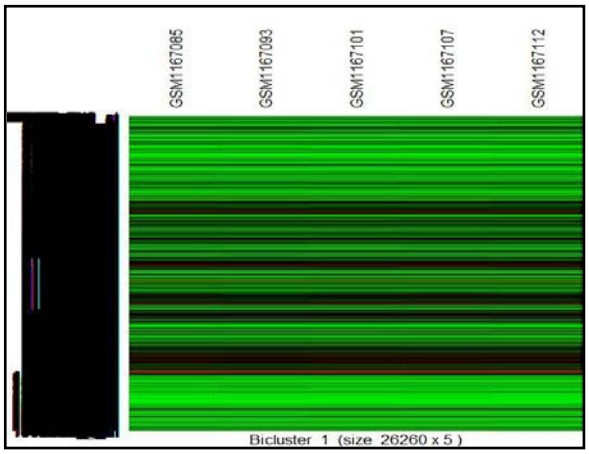

Figure 5: Bcxmotifs heat map

BCSpectral:-The Figure 6 shows the heat map of the first bicluster out of 5 bicluster found by BCXmotifs biclustering algorithm.

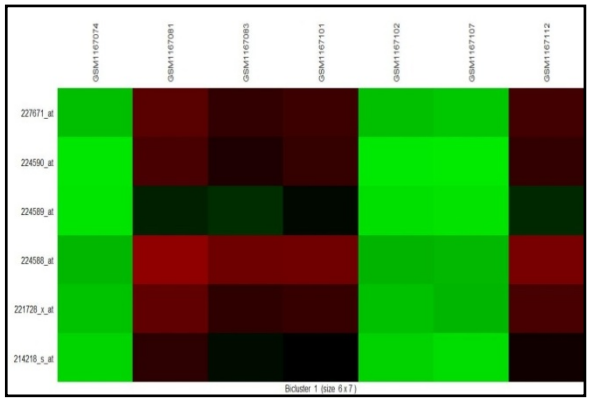

Figure 6: Bcspectral heat map

BCBimax:-The Figure .7 shows the heat map of the first bicluster out of 5 bicluster found by BCPlaid biclustering algorithms.

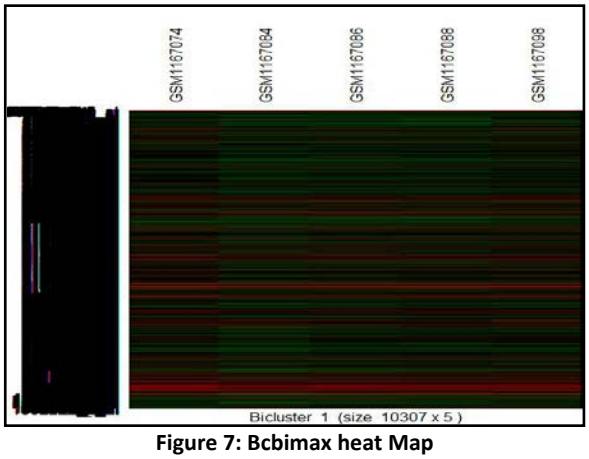

Figure 7: Bcbimax heat Map

B.Parallel-Coordinates Plots:-Parallel-coordinate plots are useful technique visualize to bicluster result. In the parallel-coordinates plot, each poly line represents the expression of a gene over all conditions (which are represented by vertical axes). Genes belonging to a bicluster are rendered using the same color as the corresponding bicluster in the heatmap.

BCPlaid

BC1:-The constant biclustering BC 1is found with 4244 rows which represents the gene of patients who undergone in the clinical tests and 12columns which represents the samples of patients from myocardial infraction dataset as shown in Fiureg 8.

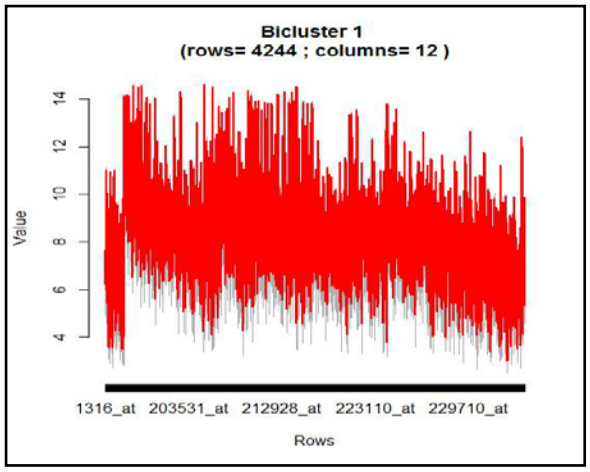

Figure 8: Bcplaid bc1 parallel-coordinates plots

BC2:-The biclustering BC 2 is found with 443 rows which represents the gene of patients who undergone in the clinical tests and 12 columns which represents the samples of patients from myocardial infraction dataset as shown in Figure9.

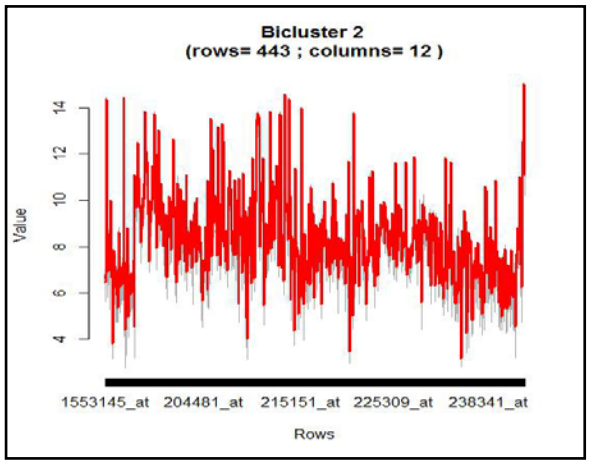

Figure 9: Bcplaid bc2 parallel-coordinates plots

© 2015 IJETCR. All Rights Reserved.

BC3:- The biclustering BC 3 is found with 55 rows which represents the gene of patients who undergone in the clinical tests and 10 columns which represents the samples of patients from myocardial infraction dataset as shown in Figure. 10.

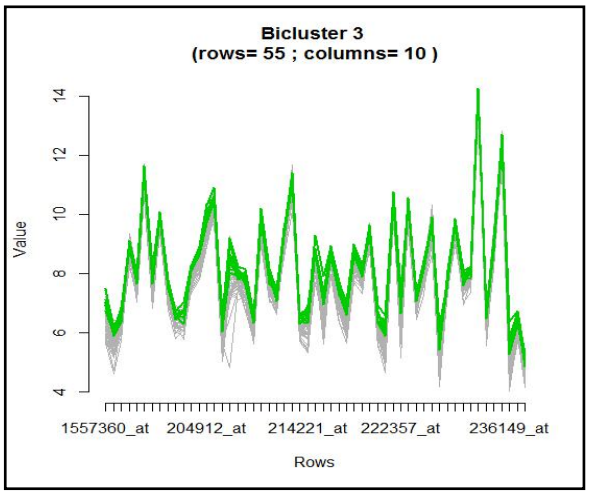

Figure 10: Bcplaid bc3 parallel- coordinates plots

BC4:- The biclustering BC 4 is found with 22 rows which represents the gene of patients who undergone in the clinical tests and 19 columns which represents the samples of patients from myocardial infraction dataset as shown in Figure 11.

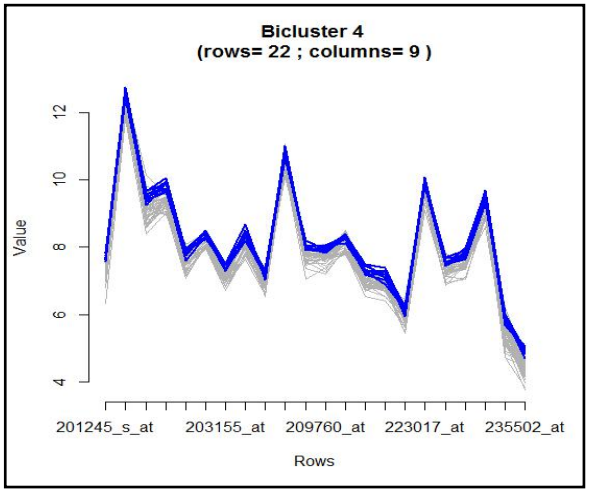

Figure 11: Bcplaid bc4 parallel-coordinates plots

BCXmotifs

BC1:-The biclustering BC 1is found with 26260 rows which represents the gene of patients who undergone in the clinical tests and 5 columns which represents the samples of patients from myocardial infraction dataset as shown in Figure12.

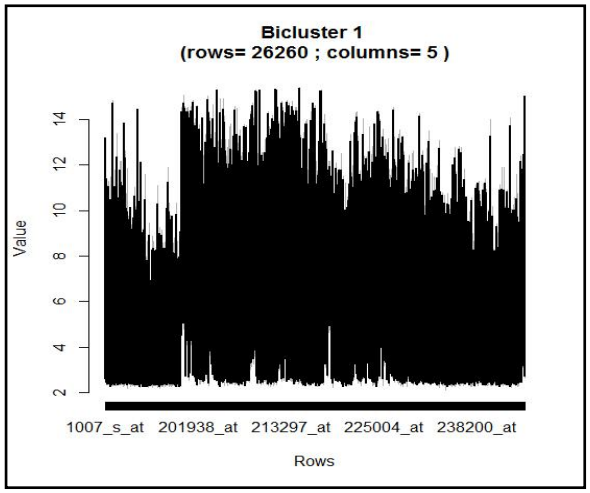

Figure 12: Bcplaid bc4 parallel-coordinates Plot

BC2:-The biclustering BC 2 is found with 10054 rows which represents the gene of patients who undergone in the clinical tests and 5 columns which represents the samples of patients from myocardial infraction dataset as shown in Figure. 13.

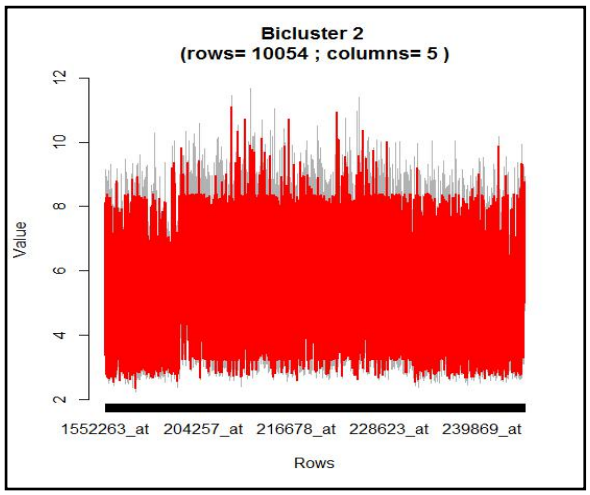

Figure 13: Bcxmotifs bc2 parallel-coordinates plots

BC3:-The analysis of constant biclustering BC3 is found with 4617 rows which represents the gene of patients who undergone in the clinical tests and 5 columns which represents the samples of patients from myocardial infraction dataset as shown in Figure 14.

© 2015 IJETCR. All Rights Reserved.

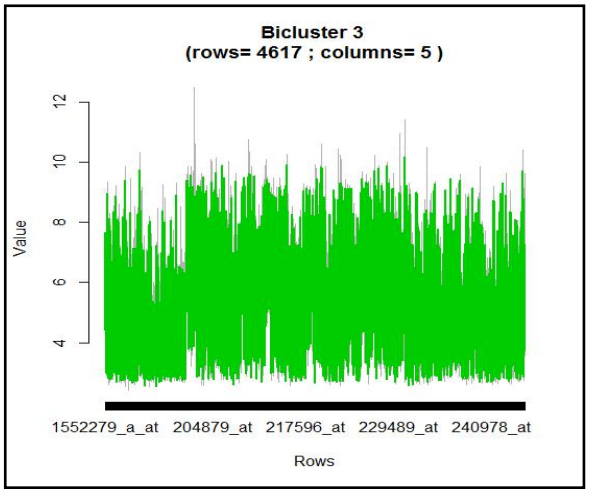

Figure 14: Bcxmotifs bc3 parallel-coordinates plots

BC4:-The biclustering BC 4 is found with 3293 rows which represents the gene of patients who undergone in the clinical tests and 5columns which represents the samples of patients from myocardial infraction dataset as shown in Figure 15.

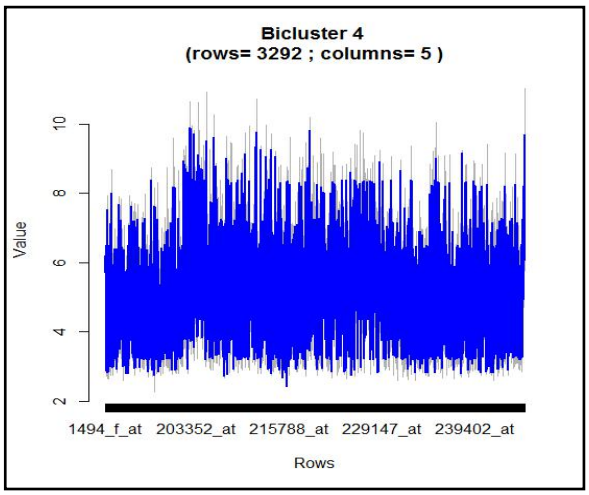

Figure 15: Bcxmotifs bc4 parallel-coordinates plots

BC5:-The biclustering BC 5 is found with 1912 rows which represents the gene of patients who undergone in the clinical tests and 5 columns which represents the samples of patients from myocardial infraction dataset as shown in Figure 16.

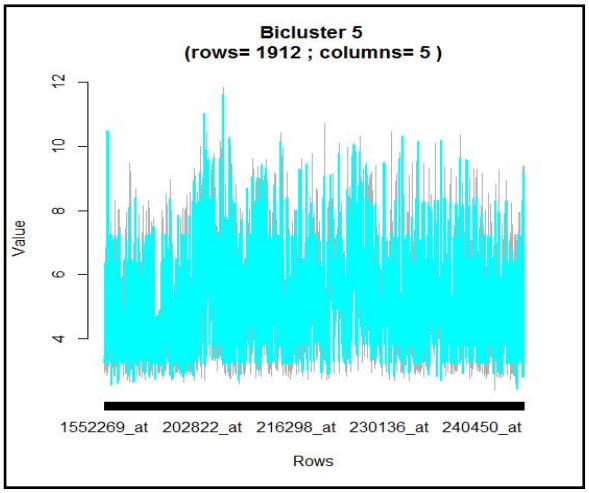

Figure 16: Bcxmotifs bc5 parallel-coordinates plots

BCSpectral

BC1:- The biclustering BC 1 is found with 6 rows which represents the gene of patients who undergone in the clinical tests and 7columns which represents the samples of patients from myocardial infraction dataset as shown in Figure 17.

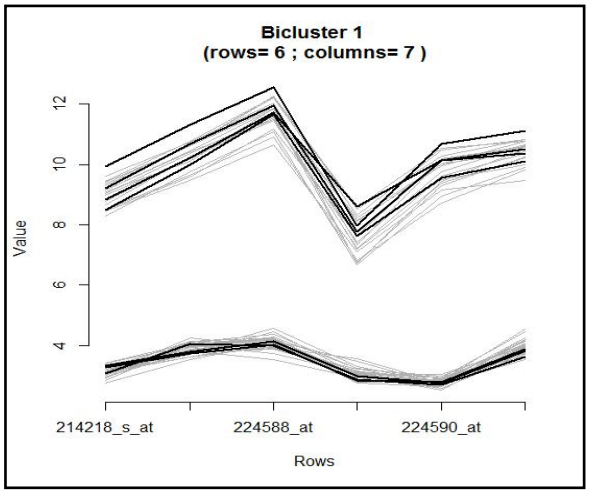

Figure 17: Bcspectral bc1 parallel-coordinates plots

BC2:- The biclustering BC 2 is found with 6 rows which represents the gene of patients who undergone in the clinical tests and 6 columns which represents the samples of patients from myocardial infraction dataset as shown in Figure 18.

© 2015 IJETCR. All Rights Reserved.

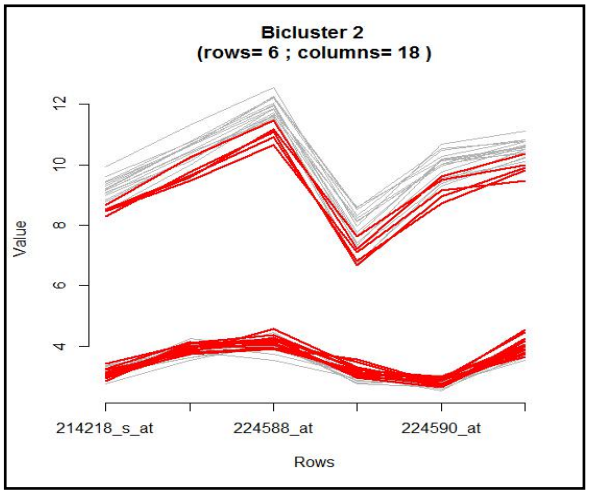

Figure 18: Bcspectral bc1 parallel-coordinates plots

BC3:-The biclustering BC 3 is found with 6 rows which represents the gene of patients who undergone in the clinical tests and 16 columns which represents the samples of patients from myocardial infraction dataset as shown in Figure 19.

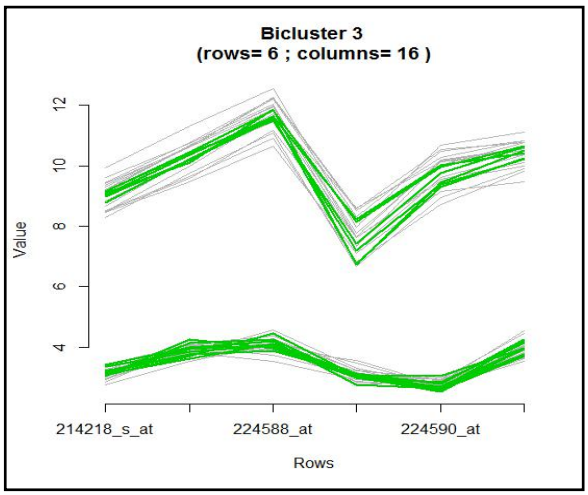

Figure 19: Bcspectral bc3 parallel-coordinates plots

BC4:-The biclustering BC4 is found with 6 rows which represents the gene of patients who undergone in the clinical tests and 6columns which represents the samples of patients from myocardial infraction dataset as shown in Figure 20.

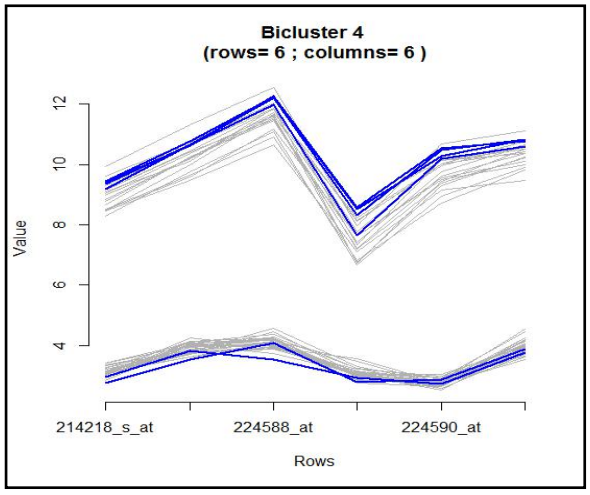

Figure 20: Bcspectral bc4 parallel-coordinates plots

BC5:-The biclustering BC5 is found with 6 rows which represents the gene of patients who undergone in the clinical tests and 5 columns which represents the samples of patients from myocardial infraction dataset as shown in Figure 21.

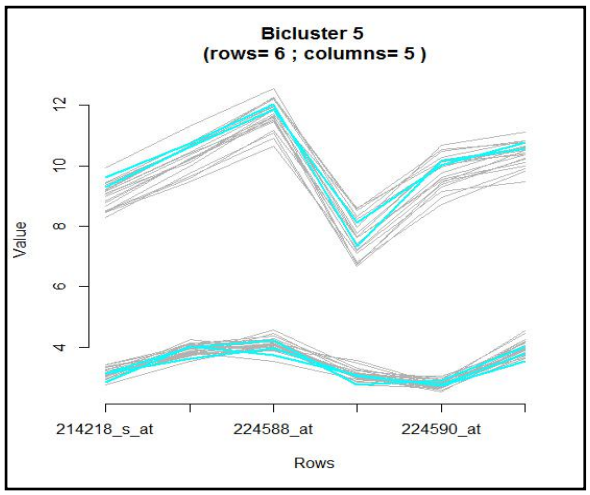

Figure 21: Bcspectral bc5 parallel-coordinates plots

BCBimax

BC1:-The biclustering BC 1 is found with 10307 rows which represents the gene of patients who undergone in the clinical tests and 5 columns which represents the samples of patients from myocardial infraction dataset as shown in Figure 22.

© 2015 IJETCR. All Rights Reserved.

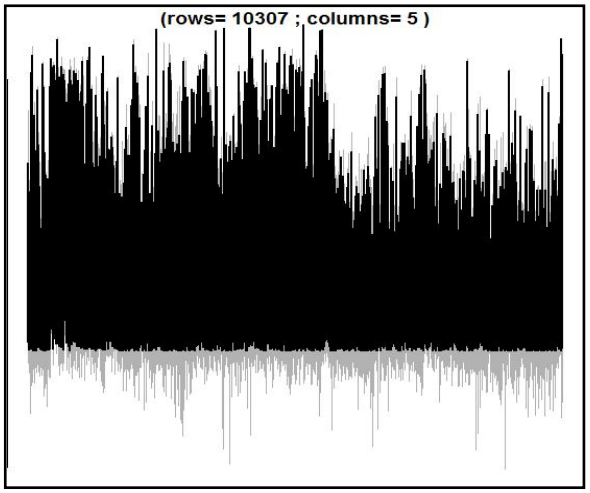

Figure 22: Bcbimax bc1 parallel-coordinates plots

BC2:-The biclustering BC2 is found with 10307 rows which represents the gene of patients who undergone in the clinical tests and 5 columns which represents the samples of patients from myocardial infraction dataset as shown in Figure 23.

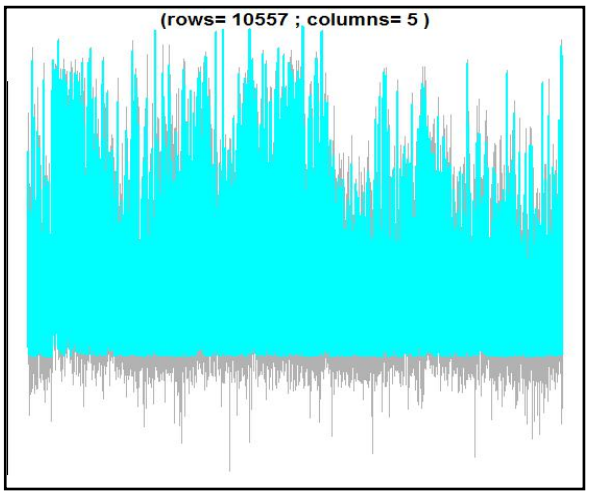

Figure 23: Bcbimax bc2 parallel-coordinates plots

BC3:-The biclustering BC3 is found with 10560 rows which represents the gene of patients who undergone in the clinical tests and 5 columns which represents the samples of patients from myocardial infraction dataset as shown in Figure 24.

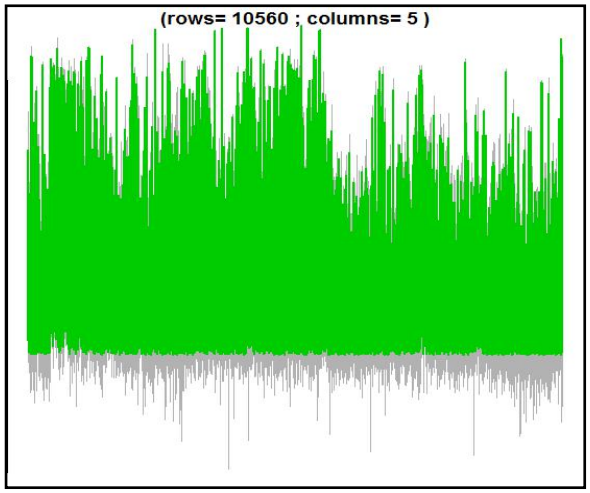

Figure 24: Bcbimax bc3 parallel-coordinates plots

BC4:-The biclustering BC 4 is found with 10469 rows which represents the gene of patients who undergone in the clinical tests and 5 columns which represents the samples of patients from myocardial infraction dataset as shown in Figure 25.

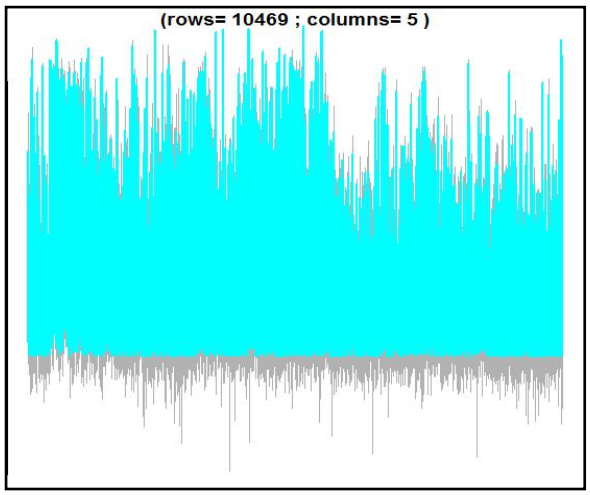

Figure 25: Bcbimax bc4 parallel-coordinates plots

BC5:-The biclustering BC5 is found with 10302 rows which represents the gene of patients who undergone in the clinical tests and 5 columns which represents the samples of patients from myocardial infraction dataset as shown in Figure 26.

© 2015 IJETCR. All Rights Reserved.

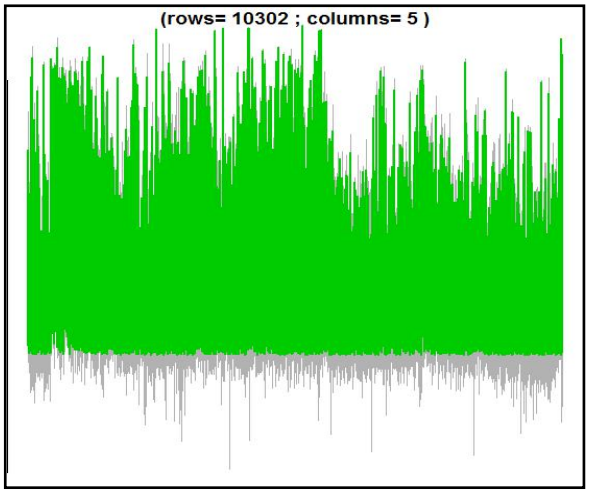

Figure 26: Bcbimax bc5 parallel-coordinates plots

5. CONCLUSION

The proposed research work has been successfully implemented in R to identify sub matrices or subgroups called bicluster using BCPlaid, BCXmotifs, BCSpectral and BCBimax biclustering algorithms from myocardial infarction dataset. The identified bicluster by all four algorithms are analyzed and discussed clearly. The bicluster obtained by four algorithms are visualized using heatmap and parallel-coordinate plots. The bicluster found by four algorithms are validated using Jaccard index measure. The intra cluster coherence is identified using coherence measure by constant, additive and multiplicative variance.

The proposed research work can be extended in further by considering other biclustering algorithms except the four biclustering algorithms used in this research work. It is also decided to go with other inter bicluster and intra bicluster metrics which are really used to validate the bicluster and to identify which one gives better results.

6. REFERENCES

1. Ben-Dor, B. Chor, R. Karp, and Z. Yakhini. Discovering local structure in gene expression data: the order-preserving sub matrix problem. In RECOMB 02: "*Proceedings of the sixth annual international conference on Computational biology"*, Page No 49–57, New York, NY, USA, 2002. ACM.
2. R.Tamilarasi, Dr R. Porkodi, "*A Study and Analysis of Disease Prediction Techniques in Data Mining for Healthcare "*International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 Vol-4, Issue-3 Page No 76-82 March 2015.
3. Y. Cheng and G. M. Church. Biclustering of expression data. In "*Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology"*, Page No 93–103. AAAI Press, 2000. 15. Gene Ontology Consortium. Gene ontology: vol-2
4. A.Dharan and A. S. Nair." *Biclustering of gene expression data using reactive Greedy randomized adaptive search procedure"*. BMC Bioinformatics, Page No 70-76 10(Suppl1):S27, 2009. Vol-9.
5. T. M. Murali and S. Kasif. "*Extracting conserved gene expression motifs from gene expression data"*. Pac. Symp. Biocomput, Page No 8:77–88, 2003. Vol-1
6. Q. Sheng, Y. Moreau, and B. De Moor. "*Biclustering microarray data by gibbs sampling.* Bioinformatics", 19: II Vol -7 Page No 196–II205, 2003.
7. K. O. Cheng, N. F. Law, W. C. Siu, and A. W. Lie, "Identification of coherent patterns in gene expression data using an efficient biclustering algorithm and parallel coordinate visualization". *BMC Bioinformatics*, Vol-1, Page No 9(210):1282–1283, 2008.
8. Kluger. et al. (2003) "*Spectral biclustering of microarray cancer data: co-clustering genes and conditions"*, Genome Res, Page No 13, 703–716.
9. B. Mirkin, "*Mathematical Classification and Clustering"*, Dordrecht: Kluwer, Page No 68-74, 1996.
10. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Buhrmann, W. Gruissem, L. Hennig, L. Thiele, E. Zitzler, Comparison of Biclustering Methods: "*A Systematic Comparison and Evaluation of Biclustering Methods for Gene Expression Data"*, Bioinformatics 22:9 (2006) Page No 1122-1129.
11. E. Segal, B. Taskar, A. Gasch, N. Friedman, and D. Koller," *Rich probabilistic models for gene expression", Bioinformatics*, 17(suppl 1): Page No 243–252, 2001.
12. H. Wang, W. Wang, J. Yang, and P. S. Yu. "*Clustering by pattern similarity in large data sets".* In SIGMOD, Page No 45- 50 2002.
13. J. Nepomucenol, A. Troncoso, J. Aguilar-Ruiz, "*Biclustering of gene expression data by*

© 2015 IJETCR. All Rights Reserved.

correlation-based scatter search", Bio Data Mining 4 (1) (2011) Page No 34-40

14. Amela, Bleuler, Stefan, Zimmermann, Philip, Wille, Anja, Buhrmann, Peter, Gruissem, Wilhelm, Hennig, Lars, Thiele, Lothar, and Zitzler, Eckart. "*A systematic comparison and evaluation of biclustering methods for gene expression data"*. Bioinformatics, 22(9): Page No 1122–1129, 2006.

15. Kaiser S, Leisch F (2008) "*A toolbox for bicluster analysis in r"*. Technical report, Department of Statistics University of Munich. Page No 68-73.

16. NeelamTyagi and Simple Sharma, "*Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Page*", International Journal of Computing and Engineering (IJSCE) July 2012 Page 50-57.

17. R. Davidson and D. Harel. Drawing graphs nicely using simulated annealing. "*ACM Transactions on Graphics* ", (TOG), 15(4): Page No 301- 331, 1996.

18. S. P. Borgatti, M. G. Everett, and L. C. Freeman. Ucinet for windows: "*Software for social network* analysis". 2002. Page No 56-66.

19. Barkow, S Bleuler, S Prelic, A, Zimmermann and Zitzler, E (2006):" Bichat: *a biclustering analysis toolbox"*. *Bioinformatics, 22, Page No 1282–1283*.

20. Kaiser, S., Santamaria, R., Theorn, R., Quintales, L., Leisch, F.: Bicluster algorithms. http://cran.r project.org/web/packages/biclust/biclust.pdf Page No 34-45 (2009)

21. Anirban Mukhoppadhyay and Ujjwal Maulik, "*An improved algorithm for clustering gene expression data* ", vol23 no.21, 2007, Page.No: 2859-2865doi:10.1093/bioinformatics/btm418, http://bioinformaticsOxford journals.org/Tamil Nadu veterinary and Animal Science University on June 18, 2005

22. Kardi Teknomo, "*K-Means Clustering Tutorials,"* July 2007,http://people revoledu.com /Kardi /tuto-rial/kmean/index.html Page No 90- 100.

23. Mohammad Taha Khan, Dr.Shamimul Qamar and Laurent F. Massin, "*A Prototype of Cancer/Heart Disease Prediction Model Using Data Mining"*, International Journal of Applied Engineering Research, 2012. Page No 89-100

24. Miss. Chaitrali S. Dangare, Dr. Mrs. Sulabha S. Apte, "*A data mining approach for prediction of heart disease using neural networks"*, international journal of computer engineering and technology, 2012. Page No 97-109.

25. Dietzsch, J., Heinrich, J., Nieselt, K., Bartz, and D.: SpRay: "*A visual analytics approach for gene expression data.*" In: IEEE Symposium on Visual Analytics Science and Technology. (2009) Page No 179-186.

26. Bergmann, S., J. Ihmels, and N. Barkai (2003). Iterative signature algorithm for "*the analysis of large-scale gene expression data"*. Physical Review Page No 6703-1902,

27. Grothaus, G., Mufti, A., Murali, T.: "*Automatic layout and visualization of biclusters"*. Algorithms for Molecular Biology 1 (2006) Page No 102-111

28. Lazzeroni, L. and A. Owen (2002)." *Plaid models for gene expression data"*. Statistical Sonica 12, Page No 61-86.

29. Jaccard, P. (1901). Distribution de la ore alpine dansle basin des dranses et dans quell ques regions voisines. Bulletin de la Societ Vaudoise des Sciences Naturelles 37, Page No 241-272.

30. Madeira, S. and Oliveira, A. (2004). "*Biclustering algorithms for biological data analysis: a survey*" IEEE Transactions on Computational Biology and Bioinformatics, 1(1): Page No 24-45.

31. Yang, J., Wang, H., Wang, W., and Yu, P. (2005). "*An improved biclustering method for analyzing gene expression*". International Journal on Artificial Tools, 14(5):771-789.

© 2015 IJETCR. All Rights Reserved.