Randomized Algorithmic Approach for Biclustering of Gene Expression Data

Sradhanjali Nayak[1], Debahuti Mishra[2], Satyabrata Das[3] and Amiya Kumar Rath[4]
[1,3,4] Department of Computer Science and Engineering, College of Engineering Bhubaneswar, Odisha, INDIA
[2] Institute of Technical Education and Research, Siksha O Anusandhan University, Bhubaneswar, Odisha, INDIA
sradha.mtech09@gmail.com, debahuti@iter.ac.in, satya.das73@gmail.com and amiyamaiya@rediffmail.com

***Abstract*—Microarray data processing revolves around the pivotal issue of locating genes altering their expression in response to pathogens, other organisms or other multiple environmental conditions resulted out of a comparison between infected and uninfected cells or tissues. To have a comprehensive analysis of the corollaries of certain treatments, deseases and developmental stages embodied as a data matrix on gene expression data is possible through simultaneous observation and monitoring of the expression levels of multiple genes. Clustering is the mechanism of grouping genes into clusters based on different parameters. Clustering is the process of grouping genes into clusters either considering row at a time(row clustering) or considering column at a time(column clustering). The application of clustering approach is crippled by conditions which are unrelated to genes. To get better of these problems a unique form of clustering technique has evolved which offers simultaneous clustering (both rows and columns) which is known as biclustering. A bicluster is deemed to be a sub matrix consisting data values. A bicluster is resulted out of the removal of some of the rows as well as some of the columns of given data matrix in such a fashion that each row of what is left reads the same string. A fast, simple and efficient randomized algorithm is explored in this paper, which discovers the largest bicluster by random projections**.

Keywords: Bicluster; microarray data; gene expression; randomized algorithm

I. INTRODUCTION

Gene expression data is typically arranged in the form of a matrix with rows corresponding to genes, and columns corresponding to patients, tissues, time points, etc. Gene expression data are being generated by DNA chip and other microarray technology and they are presented as matrices where each entry in the matrix represents the expression levels of genes under various conditions including environments, individuals and tissues. Each of the *N* rows represents a gene (or a clone, ORF, etc.) and each of the *M* columns represents a condition (a sample, a time point, etc.) [8]. It can either be an absolute value (e.g. Affymetrix GeneChip) or a relative expression ratio (e.g. cDNA microarrays). A row/column is sometimes referred to as the "expression profile" of the gene/condition [4]. Due to complex procedure of microarray experiment, gene expression data contains a huge amount of data. Clustering is applied to extract useful information from the gene expression data matrix. The process of grouping data objects into a set of disjoint class clusters, so that objects within a class have high similarity to each other, while objects in separate classes are more dissimilar [1]. Clustering can be applied either conditions (column clustering). Table 1 show the row clustering where, all the columns for the rows G2, G3 and G4 is selected and table 2 shows column clustering, where C3, C4 and C5 column is clustered with all the rows/genes.

TABLE 1: Row Clustering

	C_1	**C_2**	**C_3**	**C_4**	**C_5**	**C_6**	**C_7**
G_1	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	a_{17}
G_2	a_{21}	a_{21}	a_{23}	a_{24}	a_{25}	a_{26}	a_{27}
G_3	a_{31}	a_{32}	a_{33}	a_{34}	a_{35}	a_{36}	a_{37}
G_4	a_{41}	a_{42}	a_{43}	a_{44}	a_{45}	a_{46}	a_{47}
G_5	a_{51}	a_{52}	a_{53}	a_{54}	a_{55}	a_{56}	a_{57}
G_6	a_{61}	a_{62}	a_{63}	a_{64}	a_{65}	a_{66}	a_{67}

The classical approach to analyze microarray data is clustering. The process of clustering partitions genes into mutually exclusive clusters under the assumption that genes that are involved in the same genetic pathway behave similarly across all the testing conditions. The assumption might be true when the testing conditions are associated with time points. However, when the testing conditions are heterogeneous, such as patients or tissues, the clustering can be proven as the method of extraction information [6].

TABLE 2: Column Clustering

	C_1	**C_2**	**C_3**	**C_4**	**C_5**	**C_6**	**C_7**
G_1	a_{11}	a_{12}	**a_{13}**	**a_{14}**	**a_{15}**	a_{16}	a_{17}
G_2	a_{21}	a_{21}	**a_{23}**	**a_{24}**	**a_{25}**	**a_{26}**	a_{27}
G_3	a_{31}	a_{32}	**a_{33}**	**a_{34}**	**a_{35}**	a_{36}	a_{37}
G_4	a_{41}	a_{42}	**a_{43}**	**a_{44}**	**a_{45}**	a_{46}	a_{47}

G_5	a_{51}	a_{52}	a_{53}	a_{54}	a_{55}	a_{56}	a_{57}
G_6	a_{61}	a_{62}	a_{63}	a_{64}	a_{65}	$\mathbf{a_{66}}$	a_{67}

However clustering has got its own limitations. Clustering is based on the assumption that all the related genes behave similarly across all the measured conditions. It may reveal the genes which are very closely co-regulated along the entire column. Based on a general understanding of the cellular process, the subsets of genes are co-regulated and co-expressed under certain experimental conditions. But they behave almost independently under other conditions. Moreover, clustering partitions the genes into disjoint sets i.e. each gene is associated with a single biological function, which is in contradiction to the biological system [8]. In order to make the clustering model more flexible and to overcome the difficulties associated with clustering the concept of biclustering was introduced (see table 3). Biclustering is clustering applied in two dimensions, i.e. along the row and column, simultaneously. This approach identifies the genes which show similar expression levels under a specific subset of experimental conditions. The objective is to discover maximal subgroups of genes and subgroups of conditions. Such genes express highly correlated [18] activities over a range of conditions.

One would expect that a group of genes would exhibit similar expression patterns only in a subset of conditions, such as the subset of patients suffering from the same type of disease. Under this circumstance, biclustering becomes the alternative to the traditional clustering paradigm. Biclustering is a process which performs clustering in two dimensions simultaneously. Clustering method derives a global model while biclustering produces a local model. Biclustering enables one to discover hidden structures in gene expression data in which many genetic pathways might be embedded [2]. It might also allow one to uncover unknown genetic pathways, or to assign functions to unknown genes in already known genetic pathways, while clustering technique is applied a given gene cluster is defined using all the conditions ,similarly each condition cluster is defined for all genes. But each gene in a bicluster is selected using only a subset of the conditions and each condition in a bicluster is selected using only a subset of genes [2]. The goal of biclustering is to identify subgroups of genes and subgroups of conditions by performing simultaneous clustering of both the rows and columns instead of in two dimensions separately as in clustering [2].

Randomized algorithm approach is based on the idea of randomly selecting a set of columns and rows [6]. It is a very simple, effective method to find bicluster on both the aspect of time complexity and space complexity. The sub matrix produced by the biclustering has the property that each row reads the same string, so such a sub matrix would therefore correspond to a group of genes that exhibit a coherent pattern of states over a subset of conditions. [3].

A. *Proposed Model*

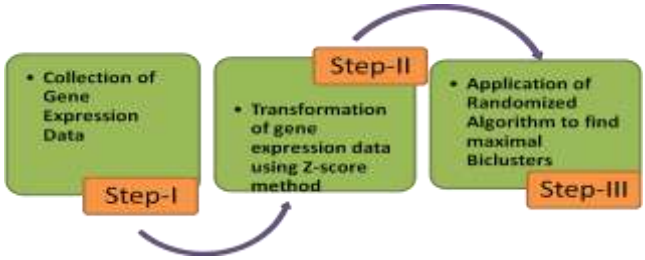

Figure 1: Our Proposed Model

Our proposed work is to find the biclusters from gene expression data using randomized algorithm. First, we have used a synthetic data set, and then we have validated our work with Yeast data set [20]. Second, we have pre-processed our data set using Z-score method to put the attribute values in a standard range of values. Finally, we validate our randomized model by comparing our model with existing biclustering models by considering various parameters. Our model (See figure 1) outperforms the existing model of Cheng and Church [15]on the basis of run time for finding number of patterns and also the scalability issues have been found to be improved significantly considering both the attributes and objects as they increases.

B. *Paper Layout*

This paper is arranged in the following manner, section I gives the introduction as well as our proposed model is also outlined, section II deals with related work on biclustering models. In section III the preliminary information about gene expression data, bicluster, randomized approach, problem statement and algorithms are described. Section VI describes our proposed algorithm. Section V gives the analysis of our work and shows its significance over the Cheng and Church[15] algorithm. Finally, section VI gives the conclusion and future directions of our work.

II. RELATED WORK

Shyama Das et al [13] proposed a greedy randomized adaptive search procedure to find the biclusters. The bicluster seeds are generated using k-means algorithm, and then these seed are enlarged using GRASP. GRASP happens in two phases i.e construction and local search. In the construction phase a feasible solution is developed iteratively by adding one element each time which will generate a feasible solution whose neighborhood will be searched until a local minimum is identified during the local search phase. The best solution is stored as the result.

In this study GRASP is applied for the first time to identify biclusters from Human Lymphoma dataset. In this paper the GRASP meta heuristics is used for finding biclusters in gene expression data. In the first step K-Means algorithm is used to group rows and columns of the data matrix separately. Then they are combined to produce small biclusters.

Bing Liu et al [7] proposed an efficient semi-unsupervised gene Selection method via spectral biclustering. From biological and clinical point of view finding smaller number of important genes help the doctor to concentrate on these genes and investigating the mechanism for cancer causes and its remedies.

Haider Banka et al. [8] give an evolutionary biclustering of gene expression data. They have proposed to uncover genetic pathways (or chains of genetic interactions) which is equivalent to generating clusters of genes with expression levels that evolve coherently under subsets of conditions, *i.e.*, discovering biclusters where a subset of genes are co-expressed under a subset of conditions. Such pathways can provide clues genes that contribute towards a disease. This emphasizes the possibilities and challenges posed by biclustering. The objective here is to find sub matrices or maximal subgroups of conditions where the genes exhibit highly co-related activities over a range of conditions.

Stefano Lonardi et al. [6] find biclusters by random projection. From a given matrix *X* composed of symbols, a bicluster is a sub matrix of *X* obtained by removing some of the rows and columns, so that each row left will read the same string. An efficient randomized approach is used to find largest bicluster which is probabilistic that is each entry of the matrix is associated with the probability.

Daxin Jiang et al. [11] proposed an interactive exploration of gene expression patterns from a gene expression data set. Analyzing coherent gene expression patterns is an important task in bioinformatics research and biomedical applications. The development of microarray technology provides a great opportunity for functional genomics. Identifying co-expressed genes and coherent expression patterns in gene expression data can help biologists understand the molecular functions of the genes and the regulatory network between the genes. However, due to the distinct characteristics of gene expression data and the special requirements from the biology domain, mining coherent patterns from gene expression data presents several challenges, which cannot be solved by traditional clustering algorithms.

III. PRELIMINARIES

A. *Microarray or Gene Expression Data*

Microarrays is a small chip made of chemically coated glass , nylon membrane or silicon onto which thousands of DNA molecules are attached in fixed grids[19]. Microarray is used in the medical domain to produce molecular profiles of diseased and normal tissues of patients. Microarray captures the expression level of thousands of genes under one experiment. Microarray operations are done under different condition to have parallel comparison between the experimental levels of gene. The relative abundance of mRNA of a gene is called the expression level of a gene [9].This is measured using DNA microarray technology which revolutionized the gene expression study by simultaneously measuring the expression levels of thousands of genes in a single experiment [8][13].

The data generated by these experiments high dimensional matrix contain thousands of rows (genes) and hundreds of conditions. The experimental conditions can be patients, tissue types, different time points etc. Each entry in this matrix is a real number which denotes the expression level of a gene. Genes participating in the same biological process will have similar expression patterns. Clustering is the suitable mining method for identifying these patterns [1][13]. The ability of arrays to monitor thousands of separate but unrelated events simultaneously has captured the thoughts of scientists practicing in both basic and applied research [8][16][17].

The process of microarray formation experiment is associated with a collection of experimental factors describing the variables under study, e.g. "disease state","gender state". Each microarray in an experiment takes on a specific value for each of the experimental factors, e.g. "disease state = normal" and "gender = male" [9][19]. In the very first stage the mRNA (messenger RNA) of normal male cell and a cancer male cell is obtained through RNA isolation process. Then by the reverse transcriptage enzyme the cDNA is obtained from mRNA. The cDNA (complimentary DNA) of the diseased cell is labeled with red color and the normal cell cDNA is labeled with green color. Then by the hybridization process the diseased cell and normal cell is hybridized to a small chip made of chemically coated glass, nylon membrane or silicon called as microarray in a fixed form (grids). Gene expression data are being generated by DNA chip and other microarray technology and they are presented as matrices of expression levels of genes under various conditions including environments, individuals and tissues. Gene expressions provide a fundamental link between genotypes and phenotypes, and play a major role in biological processes [18][19] and systems including gene regulation, evolution, development and disease mechanism.

A gene expression data from microarray experiment is represented by a real valued matrix. $M = \{ A_{ij} | 1 \le i \le n , 1 \le j \le m\}$where rows ,$G = \{ g_1, g_2, g_3, \ldots\ldots\ldots g_r\}$represents the expression pattern of the genes and the column, $S = \{ s_1, s_2, s_3, \ldots\ldots s_c\}$ represents expression profiles for samples and each element w_{ij} is measured expression level of gene i in sample j_1 which is shown in the below table 3.

TABLE 3: Gene Expression Data

Gene	Condition 1	...	Condition j	...	Condition c
$Gene_1$	a_{11}	...	a_{1j}	...	a_{1c}
Gene...	...	...	...	...	...
$Gene_i$	a_{i1}	...	a_{ij}	...	a_{ic}
Gene...	...	...	...	...	...
$Gene_r$	a_{r1}	...	a_{rj}	...	a_{rc}

Where, r = no of genes, c = no of samples, M = gene expression data matrix, a_{ij} = element in the gene expression matrix, gene = different whose expression levels are taken in the row and condition = genes are studied under different conditions which are taken in the column.

Gene expression data set contains thousands of genes while the no. of tissue sample ranges from tens to hundreds, while analyzing expression profiles, a major issue is gene selection for target phenotype [5][19]. For example Cancer is a disease that begins in the cells of the body. Cancer is ultimately the result of cells that uncontrollably grow and don't die. Cancer occurs when cells become abnormal and keep dividing and forming more cells without order or control. From biological and clinical point of view finding the small number of important genes can help medical researchers to concentrate on these gens and investigating the mechanism for cancer development. Clustering is a reputed algorithmic technique that partitions a set of input data (vectors) into subsets such that data in the same subset are close to one another in some metric [3]. Recent developments require finding the largest bicluster satisfying some additional property with the largest area. For a given matrix of size $n \times m$ over a alphabet set $\sum$, a bicluster is a sub matrix composed of selected columns and rows satisfying a certain property [3]. Bicluster is a subset of genes that jointly respond across a subset of conditions, where a gene is termed responding under some condition if its expression level changes significantly under that condition with respect to its normal level. A bicluster of a gene expression data is a local pattern such that the gene in the bicluster exhibit similar expression patterns through a subset of conditions [6].

Each bicluster is represented as a tightly co-regulated sub matrix of the gene expression matrix. A (X, Y) is a matrix, I = subset of rows, J = subset of columns and (I, Y) = a subset of rows that exhibits similar behavior across the set of columns = cluster of rows. (X, J) = a subset of columns that exhibit similar behavior across set of all rows = cluster of columns. (I, J) = is a bicluster i.e. subset of genes and subsets of conditions, where the genes exhibit similar behavior across the conditions and vice versa. Cluster of columns (X, J) = (C_3,C_4,C_5), Cluster of Rows (I, Y) = (G_2, G_3,G_4), Bicluster (I, J) = { (G_2, G_3,G_4), (C_3,C_4,C_5)}. (See table 4)

TABLE 4: Bicluster

	C_1	C_2	C_3	C_4	C_5	C_6	C_7
G_1	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	a**17**
G_2	a_{21}	a_{21}	a_{23}	a_{24}	a_{25}	a_{26}	a_{27}
G_3	a_{31}	a_{32}	a_{33}	a_{34}	a_{35}	a_{36}	a_{37}
G_4	a_{41}	a_{42}	a_{43}	a_{44}	a_{45}	a_{46}	a_{47}
G_5	a_{51}	a_{52}	a_{53}	a_{54}	a_{55}	a_{56}	a_{57}
G_6	a_{61}	a_{62}	a_{63}	a_{64}	a_{65}	**a**$_{66}$	a_{67}

The basic goal of biclustering is to identify subgroups of genes and subgroups of conditions, where the genes exhibit highly correlated activities for every condition, Identify sub-matrices with interesting properties and to perform simultaneous clustering on the rows and column dimensions of the genes. The underlying bases for using bi-clustering in the analysis of gene expression data are similar genes may exhibit similar behaviors only under a subset of conditions, not all conditions, genes may participate in more than one function resulting in one regulation pattern in one context and a different pattern in another.

B. *Randomized Approach for finding Biclusters*

Biclustering algorithms may have two different objectives: to identify one or to identify a given number of biclusters. Randomized algorithm is an approach to find one bicluster at a time which is very easy to understand and implement [6].

Let's assume that, given a large set of a data matrix, $X \in \sum_{n \times m}$ from which a sub matrix, $x_{(r^*,c^*)}$ has to be discovered where the sub matrix $x_{(r^*,c^*)}$ is the largest one from the data matrix set. For the simplicity $r^* = |R^*|$ and $c^* = |C^*|$. The concept of the algorithm owes its origin to the following simple observation. It is analyzed that if we can know what is the value of R^* then we can easily determine C^* by selecting the clean columns with respect to R^* or if instead we know C^*, then, R^* could be obtained by taking the maximal set of rows which read the same string. Unfortunately, if neither R^* nor C^* is known then the approach is to "sample" the matrix by random algorithm, with the expectation that at least some of the projections will overlap with the solution (R^*, C^*), one can focus to either rows or columns, but here, in this algorithm, it is described how to retrieve the solution by sampling columns.

The steps for the algorithm are as described below:

1. Select a random subset of columns as S of size k uniformly from the set of columns {1, 2, . . . , m}.
2. Lets assume that for the instant that $S \cap C^* \neq \Phi$. If we know $S \cap C^*$, then (R^*,C^*) could be determined by the following three steps:

   a. select the string(s) w that appear exactly r^* *times* in the rows of X [1:n.$S \cap C^*$]
   b. set R^* to be the set of rows in which w appears and
   c. set C^* to be the set of clean columns corresponding to R^*.

Given a selection of rows R, we say that a column j, $1 \leq j \leq m$, is *clean* with respect to R if the symbols in the j^{th} column of X restricted to the rows R, are identical. In general, a solution of the largest bicluster can contain a column of zeros, as long as they appear in all rows of the sub matrix [6].

C. *Problem Statement*

The main problem behind this algorithm is to find the largest bicluster from the given data matrix.

Largest Biclsuter*(f)* problem
Instance: let $\sum$ denotes the set of nonempty symbols.
Let X be a gene expression data matrix asdefined over the alphabet $\sum^{n \times m}$ of symbol.

n = no of rows or genes
m = no of columns or conditions

The set $\sum$ denotes a non-empty *alphabet* of *symbols* and a *string* over $\sum$ *is* an ordered sequence of symbol Largest Biclsuter (*f*) problem.

Objective: To find a row selection *R* and a column selection *C* such that the rows of *X (R,C)* are identical strings and the objective function *f (X(R,C))* is maximized from the alphabet set.

Assume that we are given a large matrix $X \in \sum^{n \times m}$ in which a sub matrix $X \in (R^*, C^*)$ is to be selected. Assume also that the sub matrix $X (R^*, C^*)$ is maximal. To simplify, let the notations are $r^* = |R^*|$ = set of rows and $C^* = |C^*|$ = set of columns. Let the examples of objective functions which are used as a basis to find the bicluster are as follows:

- $f1(X(R,C)) = |R| + |C|$;
- $f2(X(R,C)) = |R|$ provided that $|C| = |R|$; and
- $f3(X(R, C)) = |R||C|$.
- $f4(x_{(r^*,c^*)}) = |R^*| \cap |C^*|$

IV. Our Proposed Algorithm

Randomized search (step 1): Select a random subset *S* of size *k* uniformly from the set of columns {*1, 2, …, m*} ;

Example: Let us take an example of data matrix as follows:

TABLE 5: Example Data Matrix

2	1	0	1	1	2
0	0	1	1	0	2
0	1	2	0	1	1
2	0	0	1	0	2
1	2	1	2	1	2
0	0	2	1	0	1

Let us randomly select 3 columns as: $C_1^* = (1,2,3,4)$ Let us randomly select another 4 columns as: $C_2^* = (2,4,5,6)$

Randomized search (Step-2): From the selected columns take the common columns which are the subset of the given matrix *X*. As per our example the common column is : $C_1^* \cap C_2^* = (2,4)$. The common column is shown in green color in the below table 6.

Randomized search (Step-3): For all the subset of *S*, find the occurrences of string *w* that appears at least *r* times in each subset of *S*.

As per our example,
The String 11 appears = 1
The String 01 appears = 3
The string 12 appears = 1
The string 02 appears = 1

TABLE 6: Example

2	1	0	1	1	2
0	0	1	1	0	2
0	1	2	0	1	1
2	0	0	1	0	2
1	2	1	2	1	2
0	0	2	1	0	2

Randomized search (Step-4): Record the maximum no string which appears in the subset and record the corresponding rows. As per our example, the maximum string which appears is 01 and the corresponding rows are rows are (2,4,6)

Randomized search (step-5):

- Select the set of clean columns *C* with size at least ' *c*' corresponding to each *R*
- A column *j* is clean with respect to *R* if the symbols in the j^{th} column of *X* restricted to the rows *R*, are identical.

As per our example, the clean column with respect to the rows are (5,6).

Randomized search (step-6): Save the solutions and repeat step 1 to 4 for *t* iterations. As per our example, the largest bicluster is:

$$X'' = \begin{pmatrix} 0 & 1 & 0 & 2 \\ 0 & 1 & 0 & 2 \\ 0 & 1 & 0 & 2 \\ 0 & 1 & 0 & 2 \end{pmatrix}$$

Parameters used in the algorithms are as follows:

- Projection size *k* (k_{min})
- Column threshold *c*
- Row threshold *r*
- Number of iterations *t*
- In our example the clean column w.r.t the rows are (5,6)

V. Result Analysis

In this paper, we have simulated the randomized biclustering algorithm to find the maximal bicluster embedded in the data matrix using the synthetic data set as well as Yeast data set [20]. We have also implemented the the Minimum Square Residue (MSR) approach of Cheng and Church [15] to find the biclusers.

We have tested both the approaches in Intel Dual Core machine with 2GB HDD. The OS used is Microsoft XP and all programs are written in C. We have observed the similar trends on runtime versus number of biclusters found in both MSR based approach and our proposed randomized approach, the figure 2 shows the running time is significantly less as

compared to MSR approach. Table 7 shows the comparative study on both the approaches.

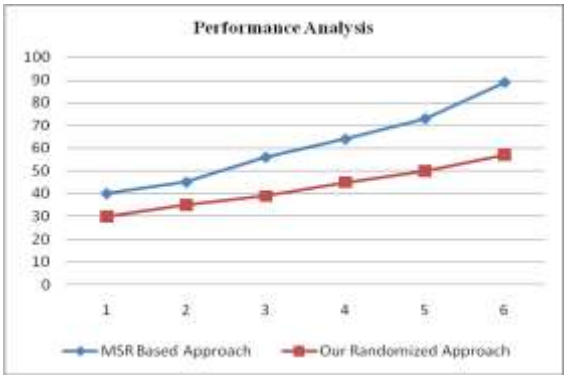

Figure 3: Performance Analysis

TABLE 7: Comparative Analysis

Model	Run Time (ms) For synthetic data set	Run Time (ms) for Yeast data set	No.of Biclusters found for synthetic data set	No.of Biclusters found for Yeast data set
MSR based Method	4000	24000	18	286
Our Randomized Approach	1500	20000	24	788

VI. CONCLUSION

The simultaneous clustering of the rows and columns of a matrix falls under diversified names such as biclustering, co-clustering or two mode clustering. The unique features of gene expression data and the specific demands from the domain of biology, propels different challenges on the front of coherent patterns from gene expression data which is a taxing task for traditional clustering algorithms. The underlying basis for using biclustering in the analysis of gene expression data are similar genes may exhibit similar behaviors only under a subset of conditions, not all conditions. Genes are regulated by multiple factors/processes concurrently. Genes may participate in more than one function resulting in one regulation pattern in one context and a different pattern in another. Using biclustering algorithms, one can obtain sets of genes that are co-regulated under subsets of conditions. Here, we have presented a rather simple algorithm based on random projections. We have presented a probabilistic analysis of the largest bicluster problem, which allows one to determine the statistical significance of a solution. In future we plan to extend our system to the following aspects randomized approach provides a flexible and consistent model to organize the expression patterns in individual data sets. In future this approach can be extended to the application of various soft computing techniques as genetic algorithm, pattern matching, unsupervised learning algorithm which can able to find more than one bicluster from the single data set. We are hopeful that this concept of biclustering will meet the future challenges and will prove itself as more effective and result oriented.

REFERENCES

[1] Daxin jiang, Chun Tang, and Aidong Zhang ,"Cluster analysis for gene expression data: a survey", *IEEE Transaction On Knowledge and Data Engineering*, Vol. 16, no.11. November 2004.

[2] S.C. Madeira and A.L. Oliveira, "Biclustering Algorithms for Biological Data Analysis: A Survey," IEEE/ACM Trans. Computational Biology and Bioinformatics, Vol. 1, No. 1, pp. 24-45, 2004.

[3] Gahyun Park and Wojciech Szpankowsk, "Analysis of biclusters with application to gene expression data". International Conference on Analysis of Algorithms, pp.267–274, 2005

[4] Mel N Kronick," Creation of the whole human genome microarray, *Technology Profile, Future Drugs Limited*, pp.19-28,www.fututе-drugs.com

[5] L. Lazzeroni, A. Owen, "Plaid models for gene expression data", *Statistica Sinica* 12 (1), pp.61–86, 2002

[6] Stefano Lonardia,,Wojciech Szpankowskib, QiaofengYanga," Finding biclusters by random projections", *Theoretical Computer Science* , 368, pp. 217 – 230, 2006

[7] Bing Liu, Chunru Wan, and Lipo Wang, "An Efficient Semi-Unsupervised Gene Selection Method via Spectral Biclustering". *IEEE Transactions on Nano Bioscience*, Vol. 5, No. 2, 2006.

[8] Alain B. Tchnag and Ahmed H.Tewfik," DNA Microarray Data Analysis: A Novel Biclustering Algorithm Approach". Volume 2006, pp. 1- 12 , DOI 10.1155/ASP/2006/59809,2006

[9] Jos´e Caldas, Nils Gehlenborg , Ali Faisal , Alvis Brazma and Samuel Kaski, "Probabilistic retrieval and visualization of biologically relevant microarray experiments", *BMC Bioinformatics.*, 10(Suppl 13), pp.1-9, 2009.

[10] Arifa Nisar, Waseem Ahmady Wei-keng Liao, Alok Choudhary,"High Performance Parallel/Distributed Biclustering Using Barycenter Heuristic". *SIAM*, pp. 1050-1061, 2009.

[11] R. Sharan and R. Shamir. Click: A clustering algorithm with applications to gene expression analysis. In ISMB, pages 307-216, 2000.

[12] Q. Sheng, Y. Moreau, B.D. Moor, "Biclustering Microarray data by Gibbs sampling", *Proceeding of European Conf. on Computational Biology". (ECCB'03)*,pp. 196–205,2003.

[13] Shyama Das, Sumam and Mary Idicula," Application of Greedy Randomized Adaptive Search Procedure to the Biclustering of Gene Expression Data", International Journal of Computer Applications, Volume 2 – No.3,pp. 0975 – 8887, 2010.

[14] Haider Banka, Sushmita Mitra, "Evolutionary Biclustering of Gene Expressions", *ACM Ubiquity*, Volume 7, Issue 42 , 2006

[15] Yizong Cheng and George M. Church. "Biclustering of expression data". *In proceedings of the 8th International conference on intelligent systems for molecular Biology* (ISMB '00'), pages 93-103, 2000.

[16] Keisuke Lida, Ichiro Nishimura, "Gene expression profiling by DNA Microarray Technology", *Critical Reviews in Oral Biology and Medicine*, Vol. 13 no. 1,pp. 35-50, 2002.

[17] Daxin Jiang ,Jian Pei, Aidong Zhang ,"Towards Interactive Exploration of Gene Expression Patterns ", *SIGKDD Explorations*, 5(2), pp.79-90 ,2003.

[18] Haider Banka, Sushmita Mitra " Evolutionary biclustering of gene expression data", *Proceedings of the 2nd international conference on Rough sets and knowledge technology*, Pages: 284-291, 2007.

[19] Madan Babu,M., Luscombe, N., Aravind, L., Gerstein, M., Teichmann, S.A. , Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.* ,2004

[20] UCI Repository for Machine Learning Data bases retrieved from the *World Wide Web: http://www.ics.uci.edu*

[21] Uetz P., et al." A Comprehensive analysis of protein protein interaction in saccharomyces cerevisiae", *Nature*, 403(6770): 601-3, Feb-2000.

[22] Gavin A.C., et. al. "Functional organization of yeast proteome by systematic analysis of protein complexes". *Nature* 415(6868) :13-4, Jan-2002.

AUTHORS PROFILE

Sradhanjali Nayak is a scholar of M.Tech (CSE) at College of Engineering, Biju Pattanaik University, Bhubaneswar, Odisha, INDIA. Her research areas includes Data mining, Soft Computing Techniques etc.

Debahuti Mishra is an Assistant Professor and research scholar in the department of Computer Sc. & Engg, Institute of Technical Education & Research (ITER) under Siksha O Anusandhan University, Bhubaneswar, Odisha, INDIA. She received her Masters degree from KIIT University, Bhubaneswar. Her research areas include Datamining, Bio-informatics, Software Engineering, Soft computing. Many publications are there to her credit in many International and National level journal and proceedings. She is member of OITS, IAENG and AICSIT. She is an author of a book Aotumata Theory and Computation by Sun India Publication (2008).

Satyabrata Das is as Assistant Professor and Head in the department of Computer Sc. & Engineering, College of Engineering Bhubaneswar (CEB). He received his Masters degree from Siksha O Anusandhan University, Bhubaneswar. His research area includes Data Mining, Adho-network etc.

Dr.Amiya Kumar Rath obtained Ph.D in Computer Science in the year 2005 from Utkal University for the work in the field of Embedded system. Presently working with College of Engineering Bhubaneswar (CEB) as Professor of Computer Science & Engg. Cum Director (A&R) and is actively engaged in conducting Academic, Research and development programs in the field of Computer Science and IT Engg. Contributed more than 30 research level papers to many national and International journals. and conferences Besides this, published 4 books by reputed publishers. Having research interests include Embedded System, Adhoc Network, Sensor Network, Power Minimization, Biclustering, Evolutionary Computation and Data Mining.