ISSN: 2277-9655
Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Identification of Biclustering Algorithms for Gene Extraction

K.Sathishkumar[*1], Dr.V.Thiagarasu[2], M.Ramalingam[3]
[*1,2,3]Assistant Professor, Dept. of Information Technology, Gobi Arts & Science College (Autonomous), Gobichettipalayam, India
sathishmsc.vlp@gmail.com

Abstract

Many clustering approaches have been proposed for the analysis of gene expression data obtained from microarray experiments. However, the results from the application of standard clustering methods to genes are limited. This limitation is imposed by the existence of a number of experimental conditions where the activity of genes is uncorrelated. For this reason, a number of algorithms that perform simultaneous clustering on the row and column dimensions of the data matrix have been proposed. This work explores the use of sub matrices, sub group of genes and sub groups of conditions to exhibit the genes highly correlated activities and identifies class of algorithms called biclustering suitable for gene extraction. Biclustering has also been widely used in fields such as information retrieval and data mining. In this comprehensive analysis a large number of existing approaches to biclustering has been examined and are classified in accordance with the type of biclusters that are identified, the patterns of biclusters that are discovered, the methods used to perform the search, the approaches used to evaluate the solution, and the target applications.Biclustering approach facilitates an efficient output by considering only a subset of conditions when looking for similarity between genes.The subset of genes exhibits significant homogeneity within the subset of homogeneity criteria.Moreover,it is observed that biclustering techniques are also used for revealing sub matrices showing unique patterns.

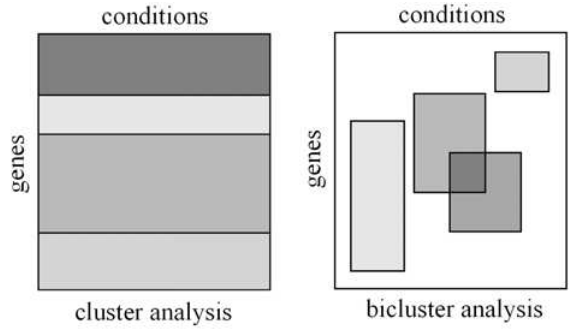
Keywords: Biclustering.Classification, Possibilistic approach.

Introduction

DNA chips and other techniques measure the expression level of a large number of genes, perhaps all genes of an organism, within a number of different experimental samples [1]. The samples may correspond to different time points or different environmental conditions. The samples may have come from different organs, from cancerous or healthy tissues, or even from different individuals. Simply visualizing this kind of data, which is widely called gene expression data or, simply, expression data, is challenging and extracting biologically relevant knowledge is harder still [11].Usually, gene expression data is arranged in a data matrix, where each gene corresponds to one row and each condition to one column. Each element of this matrix represents the expression level of a gene under a specific condition, and is represented by a real number, which is usually the logarithm of the relative abundance of the mRNA of the gene under the specific condition. Gene expression matrices have been extensively analyzed in two dimensions: the gene dimension and the condition dimension. These analysis correspond, respectively, to analyze the expression patterns of genes by comparing the rows in the matrix, and to analyze the expression patterns of samples by comparing the columns in the matrix.

Clustering techniques can be used to group either genes or conditions and, therefore, to pursue the required objective. However, applying clustering algorithms to gene expression data runs into a significant difficulty. Many activation patterns are common to a group of genes only under specific experimental conditions. In fact, general understanding of cellular processes leads to expect subsets of genes to be coregulated and coexpressed only under certain experimental conditions, but to behave almost independently under other conditions. Discovering such local expression patterns may be the key to uncover many genetic pathways that are not apparent otherwise. It is therefore desirable to move beyond the clustering paradigm, and to develop approaches capable of discovering local patterns in microarray data [2].

ISSN: 2277-9655
Impact Factor: 1.852

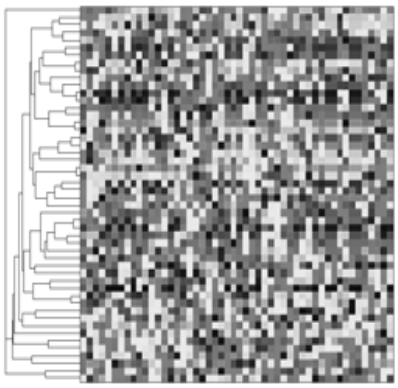


(a) **(b)**

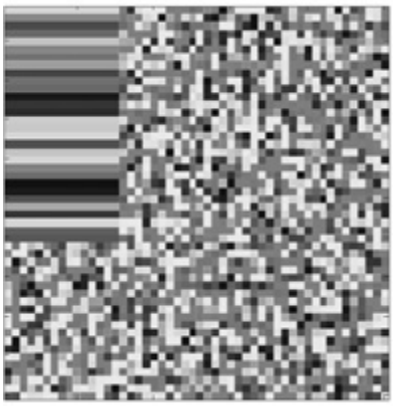
Figure 1: Conceptual difference between (a) cluster analysis and (b) bicluster analysis

Figure 1 shows the conceptual difference between traditional clustering and biclustering. Traditional clustering considers the entire set of conditions when clustering similar genes, whereas biclustering considers subset of genes and subset of conditions simultaneously.

In Figure 2, an example is presented where conventional hierarchical clustering fails but biclustering works. Figure 2a shows a data matrix, which appears random visually even after hierarchical clustering. However, if the rows and columns are appropriately permuted as in bicluster analysis, a hidden pattern embedded in the data would be uncovered as shown in Figure 2b.



(a) data matrix, which appears random visually even after hierarchical clustering



(b) A hidden pattern reviewed by appropriate permutation of subset of rows and columns

Figure 2: An example where conventional clustering fails but biclustering works

Survey on Gene extraction Using Biclustering

The biclustering method is very useful analysis tool when some genes have multiple functions and experimental conditions are diverse in gene expression measurement. This is because the biclustering approach, in contrast to the conventional clustering techniques, focuses on finding a subset of the genes and a subset of the experimental conditions that together exhibit coherent behavior. However, the biclustering problem is inherently intractable, and it is often computationally costly to find biclusters with high levels of coherence. Sungroh and Nardini [21] proposed a novel biclustering algorithm that exploits the zero-suppressed binary decision diagrams (ZBDDs) data structure to cope with the computational challenges. This method can find all biclusters that satisfy specific input conditions, and it is scalable to practical gene expression data.

Microarrays have become a standard tool for investigating gene function and more complex microarray experiments are increasingly being conducted. For example, an experiment may involve samples from several groups or may investigate changes in gene expression over time for several subjects, leading to large three-way data sets. In response to this increase in data complexity, Turner et al., [23] proposed the extensions to the plaid model, a biclustering method developed for the analysis of gene expression data. This model-based method lends itself to the incorporation of any additional structure such as external grouping or repeated measures.

Bing Liu et al., [3] presented an efficient method for selecting relevant genes. First, spectral biclustering to obtain the best two eigenvectors for class partition has used. Then gene combinations are selected based on the similarity between the genes and the best eigenvectors. Bing Liu et al., [3] demonstrated semi-unsupervised gene selection method using two

microarray cancer data sets, i.e., the lymphoma and the liver cancer data sets, where method is able to identify a single gene or two-gene combinations which can lead to predictions with very high accuracy.

Cheng and Church described that an NP-complex problem, biclustering algorithms are more complex than the classical one dimensional clustering technique, particularly requiring multiple computing platforms for large and distributed datasets. Tchagang and Tewfik [22] proposed an extension of the robust biclustering algorithm (RoBA) that is capable of performing biclustering on extremely large or geographically distributed set of gene expression data. The distributed version will divide the cluster tasks among A' processors with negligible communication costs thus making it scalable over large number of computing nodes. The proposed algorithm has been implemented using Matlab MPI and the performance results are reported based on executions on a 1, 2, 3, 4, and 5 nodes Windows PC cluster connected over 100 Mbits links. The experimental results show increased performance with the increased number of nodes on the same set of data.

By using the results of biclustering on discrete data as a starting point for a local search function on continuous data, the algorithm avoids the problem of heuristic initialization. Similar to OPSM, the algorithm aims to detect biclusters whose rows and columns can be ordered such that row values are growing across the bicluster's columns and vice-versa. Results were generated on the yeast genome (Saccharomyces cerevisiae), a human cancer dataset and random data. Results on the yeast genome showed that 89% of the one hundred biggest non-overlapping biclusters were enriched with Gene Ontology annotations. A comparison with OPSM and ISA demonstrated a better efficiency when using gene and condition orders. Christinat et al., [8] presented results on random and real datasets that show the ability of the algorithm to capture statistically significant and biologically relevant biclusters.

It is an important task for biologists to analyze gene expression data with microarray technology development. Biclustering of gene expression data is the process of grouping a subset of genes over a subset of conditions into a class, in which each gene behaviors similarly over the selected conditions and each condition is related to a certain classification. Juan Liu and Feng Liu [10] presented a random projection method to find the largest biclusters from gene expression data. To avoid sampling the column uniformly, Authors adopted the bucketing technology to estimate the probability to sample each column. Experiments show that this method can find the largest biclusters in simulation data and real data.

Noureen et al., [16] the study showed that among the chosen five biclustering algorithms SAMBA and ISA showed the best performance on the basis of functional enrichment. Biclusters were also obtained through remaining three algorithms also but were not functionally enriched.

Among biclustering ability, binary inclusion maximal algorithm (BiMax) forms biclusters when applied on a gene expression data through divide and conquer approach. The worst-case running-time complexity of BiMax for matrices containing disjoint biclusters is O(nmb) and for arbitrary matrices is of order O(nmb min{n, m}) where b is the number of all inclusion-maximal biclusters in matrix. Noureen and Qadir [17] presented an improved algorithm, BiSim, for biclustering which is easy and avoids complex computations as in BiMax. The complexity of approach is O(n*m) for n genes and m conditions, i.e, a matrix of size n*m. Also it avoids extra computations within the same complexity class and avoids missing of any biclusters.

Although most biclustering formulations are NP-hard, in time series expression data analysis, it is reasonable to restrict the problem to the identification of maximal biclusters with contiguous columns, which correspond to coherent expression patterns shared by a group of genes in consecutive time points. This restriction leads to a tractable problem. Madeira et al., [13] proposed an algorithm that finds and reports all maximal contiguous column coherent biclusters in time linear in the size of the expression matrix. The linear time complexity of CCC-Biclustering relies on the use of a discretized matrix and efficient string processing techniques based on suffix trees. Authors also propose a method for ranking biclusters based on their statistical significance and a methodology for filtering highly overlapping and, therefore, redundant biclusters. Authors reported results in synthetic and real data showing the effectiveness of the approach and its relevance in the discovery of regulatory modules.

Especially biclustering has also been proved to be very useful to analyze data matrix other than gene expression data. Compared with the traditional clustering methods, bicluster detection is very different since the elements of one bicluster may be greatly distributed among the original data matrix. A novel one-way bicluster detection method is proposed. It makes use of the existing traditional clustering algorithms such as K-means as an intermediate tool to do data clustering. Based on the clustering results and a characteristic of bicluster, the biclusters are detected one by one. Furthermore an efficient submatrices and tables creation method is proposed to save the memory storage and accelerate the processing speed. At the end of the paper

ISSN: 2277-9655
Impact Factor: 1.852

an experiment with the simulated data are presented by Zhang et al., [24].

Biclustering algorithms have been proven to be able to group the genes with similar expression patterns under a number of experimental conditions. Qinghua et al., [19] proposed a new biclustering algorithm based on evolutionary learning. By converting the biclustering problem into a common clustering problem, the algorithm can be applied in a search space constructed by the conditions. To further reduce the size of the search space; randomly separate the full conditions into a number of condition subsets (subspaces), each of which has a smaller number of conditions. The algorithm is applied to each subspace and is able to discover bicluster seeds within a limited computing time. Finally, an expanding and merging procedure is employed to combine the bicluster seeds into larger biclusters according to a homogeneity criterion. Test the performance of the proposed algorithm using synthetic and real microarray data sets. Compared with several previously developed biclustering algorithms, algorithm demonstrates a significant improvement in discovering additive biclusters.

Chandran and Iswaryalakshmi [4] presented a new algorithm, Enhanced Bimax algorithm which was based on the Bimax algorithm. The normalization technique was included which was used to display a coregulated biclusters from gene expression data and grouping the genes in the particular order. In this work, Synthetic dataset was used to display the coregulated genes.

In this paper, the Biclustering analysis of coregulated biclusters from gene expression data is carried out. Gene expression is the process, which produces functional product from the gene information. Data mining is used to find relevant and useful information from databases. Clustering groups the genes according to the given conditions. Biclustering algorithms belong to a distinct class of clustering algorithms that perform simultaneous clustering of both rows and columns of the gene expression matrix. In this paper a new algorithm, Enhanced Bimax algorithm is proposed. The normalization technique is included which is used to display a coregulated biclusters from gene expression data and grouping the genes in the particular order 11 I. The Synthetic Gene Expression dataset is used to display the coregulated genes, developed by Prelic et al., [18] It contains constant values and coherent values over the conditions and non-overlapping and overlapping clusters. The data matrix contains 10 overlapping cluster and each cluster extends over 5 genes and 15 conditions

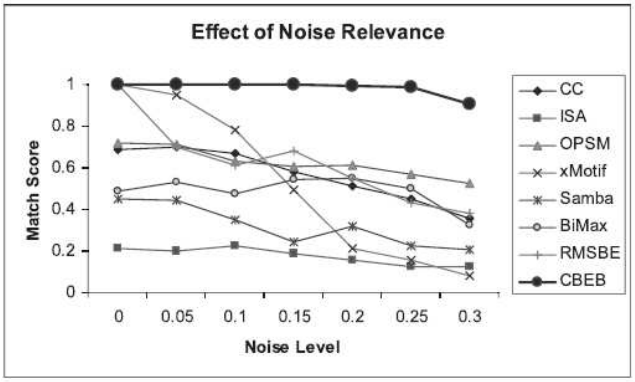
Experiment Evolution

The procedure for generating a synthetic bicluster is shown in Figure 3. Qinghua Huang et al., [19] compared with several previously reported algorithms including CC, Samba, ISA, OPSM, xMotif, Bimax, and RMSBE. They used the software BicAT developed by Prelic et al., [18] EXPANDER by Shamir et al., [20], and MSBE by Liu and Wang [12]. Table 1 shows the parameters used for each algorithm.

Table 1
The Parameters Set For the Biclustering Algorithm

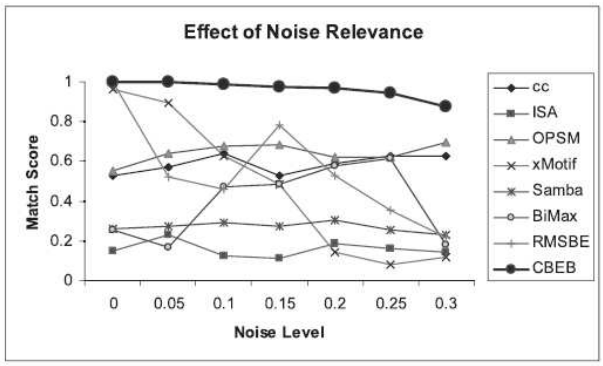
Algorithm	Parameters
CC	δ=0.5, α=1.2, seeds = 50
OPSM	l=100
ISA	t_g = 2.0, t_c = 2.0, seeds = 500
Bimax	Minimum number of genes and chips = 4
xMotif	n_s=10, n_d=1000, s_d=7, α=0.1, P value=10^{-10}, *max_length*=0.7m
Samba	D=40, N1 = 4, N2 = 6, k = 20, L=10
RMSBE	α=0.4, β = 0.5, $\gamma = \gamma_c$=1.2
CBEB	δ=0.005 for synthetic data, δ=0.01 for real data, T=0.02, N_{sc} = 15

At each noise level, the proposed algorithm and the other seven algorithms and calculate the match score for the set of biclusters found by each biclustering algorithm for the purpose of comparison. Under each noise level, each algorithm is performed on each synthetic data set for 10 runs and the overall match scores are averaged. The averaged match score for a specific algorithm can demonstrate its performance in finding the embedded synthetic bicluster at different noise levels.

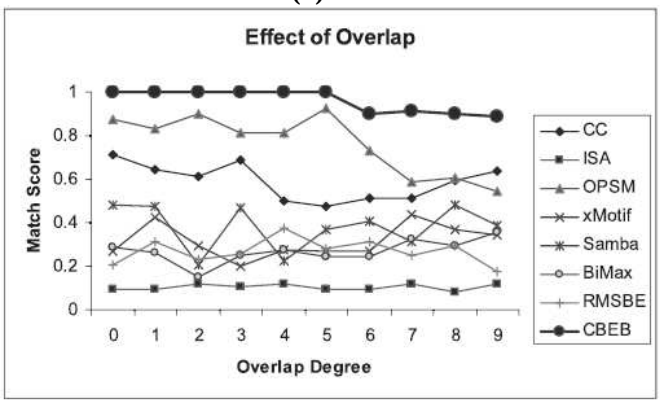


(a)

ISSN: 2277-9655
Impact Factor: 1.852



(b)



(c)

Figure 3: Results for the synthetic data sets. (a) The match scores for different algorithms using the first type of bicluster. (b) The match scores for different algorithms using the second type of bicluster. (c) The match scores for different algorithms using two overlapped biclusters.

Figures.3a and 3b illustrate the quantitative comparison results using the synthetic additive biclusters with different sizes. With the noise free data matrix, the xMotif and RMSBE algorithms can accurately find the embedded bicluster. Since the CC and OPSM algorithms rely on the behavior of the found elements to determine biclusters, they may add or delete some rows and columns of the embedded bicluster. However, they present relatively stable performance in finding biclusters with some noises. Therefore, the two algorithms are able to identify medium percentage (> 30 percent and < 80 percent) of the embedded bicluster at different noise levels. In contrast, the xMotif algorithm shows a significant sensitivity to noise.

As the noise level increases, the xMotif is hardly able to find part of the embedded bicluster. For the Bimax algorithm, the identified percentage of the bicluster varies in a relatively large range (from 20 to 60 percent). The RMSBE algorithm also shows a large variety of identifying percentage when the noise is added to the bicluster. This may be due to the random selection of reference rows and columns in the biclustering algorithm. The match scores obtained by the Samba and ISA algorithms are relatively lower. In contrast, our algorithm (CBEB) can obtain the best match scores with both of the two types of synthetic biclusters at the noise levels of less than 0.3. Fig. 3c illustrates the averaged match scores obtained by all the biclustering algorithms. It is shown that the CBEB algorithm obtains the best results in comparison with the others.

Problem and Directions

Ramalingam.M yet al., proposed a ad-hoc routing simulation technique in MANET using the biclustering method in large MANET to divide the small cluster based ad-hoc network.This cluster based ad-hoc used to improve the packet delivery ratio in MANET. [25].

In general, the bicluster problem is NP-hard as proven by Cheng and Church [5]. Thus, finding an exact solution could be time consuming. Cheng and Church [5] proposed a heuristic for discovering biclusters using MSR. This strategy succeeds in avoiding the overlapping, however it presents two main drawbacks:

(1) As biclusters are discovered, more and more elements of the original expression matrix are lost, since they are substituted with random values. It follows that the expression matrix the algorithm is working on contains more and more random values as biclusters are being discovered. As a consequence, the algorithm may return biclusters that are obtained using random values, whereas these random values will be later replaced by the original ones. Moreover, in this way some biclusters might not be found.

(2) During the execution of the algorithm, the MSR value of the biclusters considered has to be computed. If a bicluster contains random values its computed MSR is not real, since it is influenced by the presence of random values. This has a negative influence of the overall search process, since the algorithm cannot compute the real values of MSR for some biclusters [5].
Many issues in biclustering algorithm design also remain open and should be addressed by the scientific community. From these open issues, we select the analysis of the statistical significance of biclusters as one of the most important ones, since the extraction of a large number of biclusters in real data may lead to results that are difficult to interpret.

In order to overcome the drawbacks of the heuristic approach in Biclustering, efficient meta heuristic techniques such as Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), Artificial Bee Colony (ABC) etc can be used. Fuzzy related approaches would assist in better performance of the biclustering.

ISSN: 2277-9655
Impact Factor: 1.852

Conclusion

A comprehensive survey of the models, methods, and applications developed in the field of biclustering algorithms are studied and analyzed. The list of available algorithms is also very complex, and many combinations of ideas can be adapted to obtain new algorithms potentially more effective in particular applications. The tuning and validation of biclustering methods by comparison with known biological data is certainly one of the most important open issues. Another interesting area is the application of robust biclustering techniques to new and existing application domains.

References

[1] Baldi P. and Hatfield G.W., "DNA Microarrays and Gene Expression". From Experiments to Data Analysis and Modelling. Cambridge Univ.Press, 2002.

[2] Ben-Dor, Chor B., Karp R., and Yakhini Z., "Discovering Local Structure in Gene Expression Data: The Order-Preserving Submatrix Problem," Proc. Sixth Int'l Conf. Computational Biology (RECOMB '02), pp. 49-57, 2002.

[3] Bing Liu, Wan, C., Lipo Wang, "An efficient semi-unsupervised gene selection method via spectral biclustering" NanoBioscience, IEEE Transactions on 2006,Vol:5, No:2,pp.110 – 114.

[4] Chandran C.P. and IswaryaLakshmi K., "Biclustering Analysis of Coregulated Biclusters from Gene Expression Data", International Journal of Computational Intelligence and Informatics, Vol. 2: No.1, April – June 2012.

[5] Cheng Y. and Church G.M., "Biclustering of Expression Data," Proc. Eighth Int'l Conf. Intelligent Systems for Molecular Biology (ISMB '00), pp. 93-103, 2000.

[6] Cheng, Y. and Church, G. Biclustering of expression data, Proc. ISMB., 2000.

[7] Cheng, Y. and Church, G.M. "Biclustering of expression data", Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology, pp. 93-103, 2000.

[8] Christinat, Y, Wachmann, B., Lei Zhang, "Gene Expression Data Analysis Using a Novel Approach to Biclustering Combining Discrete and Continuous Data" Computational Biology and Bioinformatics, IEEE/ACM Transactions on 2008, pp. 583 – 593.

[9] Hartigan J.A., "Direct Clustering of a Data Matrix," J. Am. Statistical Assoc. (JASA), vol. 67, no. 337, pp. 123-129, 1972.

[10] Juan Liu and Feng Liu "Biclustering gene expression data by random projection based on bucketing" Information Technology and Applications in Biomedicine, ITAB International Conference on 2008, pp.229 – 232.

[11] Lazzeroni L. and Owen A., "Plaid Models for Gene Expression Data," technical report, Stanford Univ., 2000.

[12] Liu X. and Wang L., "Computing the Maximum Similarity Bi- Clusters of Gene Expression Data," Bioinformatics, vol. 23, pp. 50- 56, 2007.

[13] Madeira, S.C., Teixeira, M.C., Sa-Correia, I., Oliveira, A.L., "Identification of Regulatory Modules in Time Series Gene Expression Data Using a Linear Time Biclustering Algorithm", Computational Biology and Bioinformatics, IEEE/ACM Transactions on 2010, pp.153 – 165.

[14] Mirkin B., "Nonconvex Optimization and its Applications," Math Classification and Clustering, Kluwer Academic Publishers, 1996.

[15] Mishra, D., Shaw, K., Mishra, S., Rath, A.K., Acharya, M., "Hash based biclustering for class discovery from gene expression data: A pattern similarity approach" Electronics Computer Technology (ICECT), 3rd International Conference on 2011, pp.137 – 141.

[16] Noureen, N., Kulsoom, N., de la Fuente, A., Fazal, S., Malik, S.I., "Functional and promoter enrichment based analysis of biclustering algorithms using gene expression data of yeast" Multitopic Conference INMIC IEEE 13th International 2009, pp.1 – 6.

[17] Noureen, N., Qadir, M.A., "BiSim: A Simple and Efficient Biclustering Algorithm", Soft Computing and Pattern Recognition, SOCPAR '09. International Conference of 2009, pp.1 – 6.

[18] Prelic A., Bleuler S., Zimmermann P., Wille A., Bu¨hlmann P., Gruissem W., Hennig L., Thiele L., and Zitzler E., "A Systematic Comparison and Evaluation of Biclustering Methods for Gene Expression Data," Bioinformatics, vol. 22, pp. 1122-1129, 2006.

[19] Qinghua Huang, Dacheng Tao, Xuelong Li, Liew, A.W.-C., "Parallelized Evolutionary Learning for Detection of Biclusters in Gene Expression Data", Computational Biology and Bioinformatics, IEEE/ACM Transactions on 2012,Vol:9 , No: 2, pp.560 – 570.

[20] Shamir R., Maron-Katz A., Tanay A., Linhart C., Steinfeld I., Sharan R., Shiloh Y., and Elkon R., "EXPANDER—An Integrative Program Suite for Microarray Data Analysis," BMC Bioinformatics, vol. 6, No:232, 2005.

[21] Sungroh Yoon, Nardini, C., Benini, L., De Micheli, G., "Discovering coherent biclusters from gene expression data using zero-suppressed binary decision diagrams", Computational Biology and

Bioinformatics, IEEE/ACM Transactions on 2005,Vol:2 , No:4, pp. 339 – 354.

[22] Tchagang, A.B. and Tewfik, A.H., "Distributed Robust Biclustering Algorithm for Gene Expression Analysis", Genomic Signal Processing and Statistics, GENSIPS IEEE International Workshop on 2007. Pp.1 – 4.

[23] Turner, H.L., Bailey, T.C., Krzanowski, W.J., Hemingway, C.A., "Biclustering models for structured microarray data" Computational Biology and Bioinformatics, IEEE/ACM Transactions on 2005 ,Vol:2 , No: 4, pp, 316 – 329.

[24] Zhang Yanjie, Wang Hong, Zhanyi Hu, "A novel one-way clustering based gene expression data biclustering method" Computer Engineering and Technology (ICCET), 2nd International Conference on 2010, pp.V4-364 - V4-368.

[25] Ramalingam.M,Dr.Thiagarasu.V,Narendran." Periodical and On Demand Topology Dissemination in routing protocols:A comprehensive Analysis based on Delay, Delivery Ratio and Throughput", International Journal of Advanced and Innovative Research on 2013,Vol:2,No:9, pp.123-127.