ISSN: 2230-9926

Available online at http://www.journalijdr.com

IJDR

International Journal of Development Research
Vol. 06, Issue, 01, pp. 6401-6404, January, 2016

International Journal of
DEVELOPMENT RESEARCH

Full Length Research Article

COMPARATIVE STUDY OF GENE EXPRESSION DATA SET USING BICLUSTERING METHODS

***Kusum Rajput, Veda, N. and Pamela Vinitha**

Department of Information Science, Rajiv Gandhi Institute of Technology, Bangalore, India

ARTICLE INFO

Article History:

Received 22nd October, 2015
Received in revised form
14th November, 2015
Accepted 30th December, 2015
Published online 31st January, 2016

Key Words:

Gene Expression Dataset,
Clustering, Biclustering,
Biclustering Methods.

ABSTRACT

Expression technology, such as high density DNA microarray, allows us to monitor gene expression patterns at the genomic level. Advent of this technology leads to the new challenges of extracting biologically relevant knowledge from such large gene expression data sets. As a result, data mining of gene expression data has become an important area of research for biologists. Suitable mining techniques will contribute to get into the insight of the gene-gene relationships and that may further lead to discover hidden facts related to any species or microbes. To explore the gene-gene relationships through suitable data mining techniques in order to understand how genes relate and how they regulate one another. Moreover, above gene-gene relationship will be used further to extract intrinsic or embedded gene clusters which are prevalent in most of the expression of genes into clusters such that genes in the same cluster have similar gene expression patterns than genes in other clusters.

INTRODUCTION

From last few decades an extensive research works in the field of biological data mining contributes a lot. However, all of the techniques proposed are not sufficient up to the mark to solve the issues and challenges. The present trends of research in gene expression data analysis is to group genes sharing common functional characteristics. Mere grouping of genes based on some similarity measures are unable to provide the relationships between the genes that may gives rise to biologically relevant outcomes. DNA microarray technologies have enabled researchers to measure the expression levels of many genes at same time while collecting and processing of the sample. Clustering is one popular gene expression data analysis method that partitions a set of genes into clusters such that genes in the same cluster have similar gene expression patterns than genes in other clusters. The discovery of functional relationships between gene gene based on gene expression profiles may even identifies the relationship between the two genes changes in different diseases or conditions. In clustering based methods that have focused on differentially expressed genes, pair wise functional may be indicative of the functional relationships related to a disease mechanism. Biclustering is a methodology allowing for feature set and test set clustering (supervised or unsupervised) simultaneously.

Corresponding author: Kusum Rajput, Veda, N.,
Department of Information Science, Rajiv Gandhi Institute of Technology, Bangalore, India.

It finds clusters of samples possessing similar characteristics together with features creating these similarities. The required consistency of sample and feature classification gives biclustering an advantage over other methodologies treating samples and features of a dataset separately of each other.

Literature Survey

A gene expression used to detect differences in populations of cells on a genome of any organism. Every cell of an organism contain sets of chromosomes and identical of genes. Only a fraction of genes are turned on and the gene that is turned on is the subset that is ―expressed‖ that shows unique feature to each cell type. Gene expression is the information contained within DNA (www.ncbi.nlm.nih.gov/ About/primer/microarrys.html.), and it also has gene's ability to make a gene product. Gene Expression matrix has been extensively studied into two dimensions that is on the gene dimension and the condition dimension. Clustering can be considered the most important *unsupervised learning* problem; it finds a *structure* in a collection of unknown data. Clustering can be defined as could be ―the process of organizing objects into groups whose members are similar in some way (Roy and Bhattacharyya, 2008). A group of objects which are ―similar‖ between them and are ―dissimilar‖ other different group is known as clustering. Biclustering is a methodology allowing for feature sets and testing and clustering simultaneously. Its method can be applied in 2D simultaneously where as clustering can be applied either row or columns of the data matrix.

Clustering produce global model where as Biclustering local model. Clustering goals is to identify subgroups of conditions, by performing simultaneously clustering of both rows and columns of the gene expressions matrix. Biclustering approaches are the key techniques to use when one or more of the following situation applies (Sara *et al*., 2004):-

- Only a set of gene participates in a cellular process of interest.
- An interesting cellular process is active only in a subset of conditions
- A single gene can participate in multiple path ways that may or may not be coactive under all the conditions.

Cheng & Church's algorithm: constructs one bicluster at a time using a statistical criterion – a low mean squared residue .Once a bicluster is created, its entries are replaced by random numbers, and the procedure is repeated iteratively.Mean squared residue (H) was used as measure of the coherence of the rows and columns in the bicluster. Cheng and Church defined the mean squared residue, H, of a bicluster (I, J) as the sum of the squared residues. The mean squared residue score is given by (Yizong Chengzx and George, 2000):

$$H(I, J) = \frac{1}{\text{[illegible]}} \sum_{\in \square, \square} (a_{\square} - \quad - \quad + a_{\square\square} \in \square \quad \square \quad a_{\square\square} \quad a_{\square\square})^2$$

Here it consists of five algorithms

- Brute Force Deletion and Addition.
- Single Node Deletion.
- Multiple Node Deletion.
- Node addition.
- Finding the given number of biclusters.

Bimax Algorithm follow divide-and conquer strategy. It uses a simple data model reflecting the fundamental idea of biclustering, while aiming to determine all optimal biclusters in reasonable time. This method has the benefit of providing a basis to investigate (Alexe *et al*., 2002) the usefulness of the biclustering concept in general, independently of interfering effects caused by approximate algorithms, (Preli'c *et al*., 2006) and the effectiveness of more complex scoring schemes and biclustering methods in comparison to a plain approach.

Samba presented a graph-theoretic approach to biclustering which use statistical data model. In this framework, the expression matrix is modelled as a bipartite graph, a bicluster is defined as a subgraph, and a likelihood score is used in order to assess the significance of observed subgraphs (Amons Tanay *et al*., 2002). A corresponding heuristic algorithm called Samba aims at finding highly significant and distinct biclusters. This approach has been extended to integrate multiple types of experimental data. Order Preserving Submatrix Algorithm (OPSM) (Amela Preli *et al*., 2006) bicluster is defined as a sub matrix that preserves the order of the selected columns and
is proposed that is run on different random seeds, similarly to ISA (Amela Preli *et al*., 2006). rows. An identical linear ordering is given for the data value that is given in a bicluster. Based on a stochastical model, a deterministic algorithm is developed to find large and statistically significant biclusters (Amela Preli *et al*., 2006).

Iterative Signature Algorithm (ISA) (Amela Preli *et al*., 2006) considers a bicluster to be a transcription module, i.e., a set of co-regulated genes together with the associated set of regulating conditions. Starting with an initial set of genes, all samples are scored with respect to this gene set and those samples are chosen for which the score exceeds a predefined threshold. In the same way, all genes are scored regarding the selected samples and a new set of genes is selected based on another user-defined threshold. The entire procedure is repeated until the set of genes and the set of samples converge, i.e., do not change anymore. Multiple biclusters can be identified by running the iterative signature algorithm on several initial gene sets (Amela Preli *et al*., 2006).

xMotif (Amela Preli *et al*., 2006) biclusters are sought for which the included genes are nearly constantly expressed—across the selection of samples. In a first step, the input matrix is preprocessed by assigning each gene a set of statistically significant states. These states define the set of valid biclusters: a bicluster is a submatrix where each gene is exactly in the same state for all selected samples. To identify the largest valid biclusters, an iterative search method Here it consists of five algorithms:

MSBE define a similarity score for a sub- matrix.Using the similarity score, a polynomial time algorithm is design to find an optimal bi- cluster. the algorithm handles various kinds of other cases. The algorithms have the following advantages: (1) no discretization procedure is required, (2) performs well for overlapping bi- clusters (Xiaowen Liu and Lusheng Wang, 2006).

Data Sets Analysis

Data set Information

The data set that are use here are taken from GEO(Gene Expression Omnibus),GEO is a data base where high throughput gene express data, hybridization array,chips and microarray data set can be found. It is under NCBI. The data set present here are in soft format, the gene expression datasets deals here are one yeast and three mammals. The five gene expression datasets used is given in the Table 1. Null rows/columns and rows/columns with all zeros are deleted from the datasets before applying biclustering.

Variation with respect to threshold

Correlation threshold value should be between −1 and +1 but most likely to be positive, and depends on data. The Biclustered data set that is loaded. For the purpose of the functional enrichment analysis an algorithm TANGO

is used. TANGO (Tool for AN alysis of GO enrichments) which is use to find the maximum p-value in the give biclusters, it use empirical distribution to deter

Table 1. The datasets used in analysis

Name (organism)	Name of genes	Name of samples
GDS2938(H.sapiens)	22283	12
GDS958(Mouse)	22690	12
GDS1038(Rat)	2879	4
GDS1551(sporulation)	6178	8
GDS3717(H.sapiens)	23826	12

Thus, for determination of correlation threshold θ, one can vary correlation threshold between 0 and 1, and then for each biclustering result, the average number of functionally enriched attributes is determined. From a plot of average number of functionally enriched attributes (computed using *P*- values) versus correlation threshold value, the correlation threshold value associated with the highest average number of functionally enriched attributes can be selected. Presence of common transcription factors in the promoter regions of a set of the genes is good evidence toward co-regulation. We have considered the biclusters containing less than or equal to 50 genes.

MATEIALS AND METHODS

BicAT_plus is a software which provides a step- by-step workflow for the purpose of biclustering in the datasets. In these software many biclustering methods are assemble together. In BicAT_pluse there is a option where the data set file is loaded, these data set file should be in text format. When any biclustering algorithm is run in BicAT_plus , preprocessing of data is need, so discretization of the data set to the binary value is done. After the discretization of the data set now the biclustering algorithm can be run in the data set . In the window that appear there is a Run dialog , in this dialog algorithm like BiMax, OPSM, MSBE, CC,ISA, xMotifs appears and any of this algorithm can be run .These data set are in ORF format .

Data format of clustered

Gene ID	Gene Name	Bicluster No
(Tab delimited)		

This format will help us to run the biclusted data in EXPANDER .EXPANDER (EX Pression AN alyzer and Display ER) is a software that is use to find the transcription factor. Fun Associated is also another type of software which helps us in finds the TF but its result is not accurate as the EXPANDER give us and it take less time than FunASSociated (Ron Shamir *et al*., 2005). . The comparative study of biclustering result using Expander Tests is:

- Functional Analysis using Tango.

RESULTS

Comparison Study of the Biclustering data set of each organism is done in which the graph is plot. Give below is the graph of each algorithm for given data set of an organism. By studying all the Biclustering data set, Sporulation(Yeast) is the organism that give great response on any biclustering algorithm as it functional enrichment is any in any other organism for any biclustering data and Cheng & Church's algorithm is the algorithm that give highest value of functional enrichment.

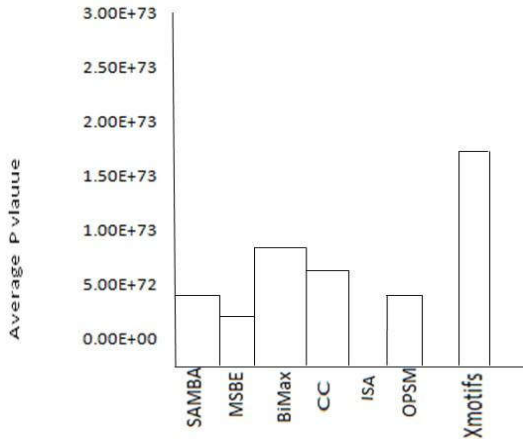
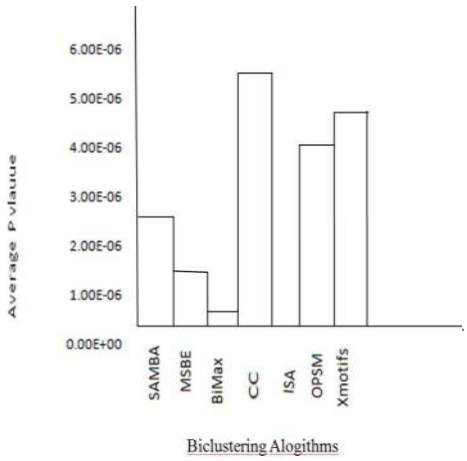
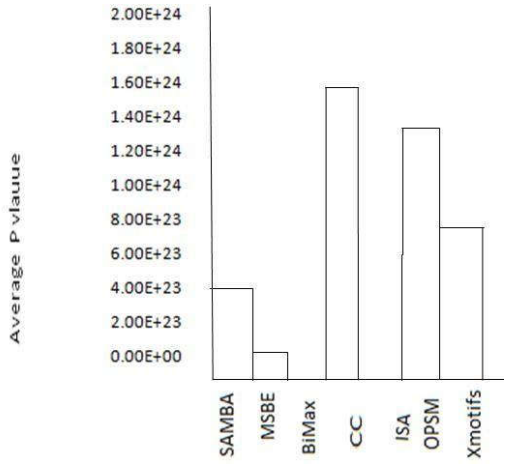


Figure 5.1. A Comparison of the Biclustering Algorithms for



Human (GDS2938).

Figure 5.1. b Comparison of the Biclustering Algorithms for Mouse (GDS958)



Biclustering Alogithms

Figure 5.1. C Comparison of the Biclustering Algorithms for Rat (GDS1038)

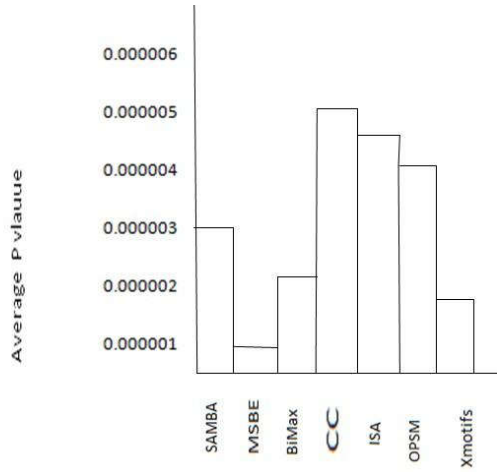


Figure 5.1.D Comparison of the Biclustering Algorithms for Sporulation(GDS1551)

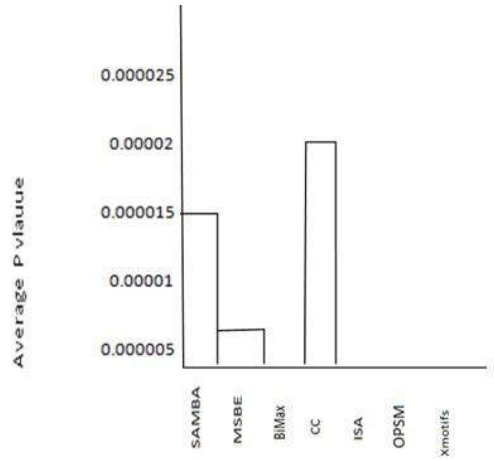


Figure 5.1.e Comparison of the Biclustering Algorithms for Human (GDS3717)

Conclusion

A suitable data mining based techniques has been studies for finding pair wise gene-gene relationship that can further be used to explore gene-gene interaction network from gene expression data.

The discovery of functional relationships between two genes are based on gene expression profiles may even identifies the relationship between the two genes. Biclustering techniques can identify co- expressed genes which can help us to study many disease and fine there cure.

REFERENCES

www.ncbi.nlm.nih.gov/About/primer/microarrys.html.

S Roy, D. K. Bhattacharyya, OPAM: An Efficient One Pass Association Mining Technique without Candidate Generation, September 2008.

Sara, C. Madeira, Arlindo, L. Oliveira, Biclustering Algorithms for Biological Data Analysis: A Survey, Jan 2004.

Yizong Chengzx_ and George M. Churchzy, Biclustering of Expression Data, 2000.

Alexe, G., Alexe, S., Crama, Y., Foldes, S., Hammer, P.L. and Simeone,B, Consensus algorithms for the generation of all maximal bicliques,Technical Report TF-DIMACS-2002-52.2002.

Preli'c A., Bleuler S., Zimmermann P., Wille A., Buhlmann 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 no. 9 2006.

Amons Tanay, Roded Sharan and Ron Shamir, Dicovering Statically significant biclusterings in gene expression data, Vol 18 Suppl.1 2002, pages S135-S144, March 31,2002.

Mela Preli, Stefan Bleuler, Philip Zimmermann, Anja Wille, Peter Buhlmann, Wilhelm Gruissem, Lars Henning, Lothar Thiele, Eckart Zitzler, A Systematic Comparison and Evaluation of Biclustering Methods for Gene Expression Data, 2006.

Xiaowen Liu and Lusheng Wang,Computing the maximum similarity bi-clusters of gene expression data, Vol. 23 no. 1 2007, pages 50–56 doi:10.1093/bioinformatics/ btl560, November 7, 2006.

Fadhl M. Al-Akwaa, Analysis of Gene Expression Data Using Biclustering Algorithms,2012.

Ron Shamir, Adi Maron-Katz, Amos Tanay, Chaim Linhart, Israel Steinfeld, Roded Sharan, Yosef Shiloh and Ran Elkon, EXPANDER – an integrative program suite for microarray data analysis, 21 September 2005.

****