Enhanced Biclustering for Gene Expression Data

R. Parimala

Abstract- *Microarray technology is a powerful method for monitoring the expression level of thousands of genes in parallel. Using this technology, the expression levels of genes are measured. Microarray data is represented in N × M matrix. Each row indicates genes and each column indicates condition. In Gene Expression data, standard clustering algorithms are called as global clustering. In global clustering, genes are analyzed under all experimental conditions based on their expression. Biclustering is a very popular method to identify hidden co-regulation patterns among genes and to identify the local structures of genes and conditions. In existing system, Cheng and Church biclustering algorithm is presented as an alternative approach to standard clustering techniques to identify local structures and also identify subsets of genes that shows similar expression patterns across specific subsets of experimental conditions and vice versa. Clustering the microarray data is based on user defined threshold value, this affects the quality of biclusters formed. In proposed scheme, threshold value ∂ is calculated rather than user defined threshold. Biclusters are formed based on the low mean squared residues and ∂, which would improve the quality of the biclusters.*

Keywords:- *Microarray technology, Clustering, Biclustering, gene expression data*

I. INTRODUCTION

The basic structural and functional unit of all known living organism is cells. The nucleus is the most important organelle found in a center part of a cell. Two different kinds of genetic material exist in a nucleus, deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). Most organisms are made of DNA, but a few viruses have RNA as their genetic material. The biological information contained in an organism is encoded in its DNA or RNA sequence. Millions of genes are present in our human body. A gene is a unit of heredity in a living organism. It normally resides on a stretch of DNA that codes for a type of protein or for an RNA chain that has a function in the organism.

All living things depend on genes, as they specify all proteins and functional RNA chains. Genes hold the information to build and maintain an organism's cells and pass genetic traits to offspring. The vast majority of living organisms encode their genes in long strands of DNA. The most common form of DNA in a cell is in a double helix structure, in which two individual DNA strands twist around each other in a right-handed spiral. DNA consists of a chain made from four types of nucleotide subunits such as adenine, cytosine, guanine, and thymine. In this structure, the base pairing rules specify that guanine pairs with cytosine and adenine pairs with thymine. Gene expression occurs in two steps:

- Transcription
- Translation

The process of genetic transcription produces a single-stranded RNA molecule known as messenger RNA from DNA. The process of translation produces defined sequence of amino acids from mRNA.

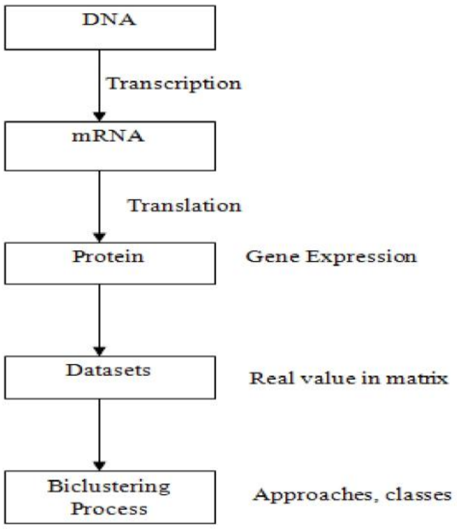


Fig.1 Steps involved in Gene Expression

To measure the gene expression levels, number of techniques has developed. Some of the traditional methods are Southern Blot, Northern blot**.** Microarray Technology measures the thousands of gene expression levels simultaneously. Some of the methods are DNA microarrays, Expressed cDNA Sequence Tag(EST), Serial Analysis of Gene Expression (SAGE), etc. Clustering is a data mining technique used to place data elements into related groups without advance knowledge of the group definitions. Popular clustering techniques include k-means clustering and expectation maximization (EM) clustering, etc. The traditional approach to research in Gene Expression has been an inherently local one, examining and collecting data on a single gene, a single protein or a single reaction at a time. A dataset (e.g., from microarray experiments) is normally given as a rectangular m × n matrix A, where each column represents a condition and each row represents a gene: where value of a_{ij} is the expression of i-th gene in j-th condition. Traditional clustering algorithms partition an expression matrix into submatrices that extend over the whole set of conditions, giving all conditions equal weight. To account for this, biclustering approaches carry out the grouping in both dimensions simultaneously, genes and conditions. This allows finding subgroups of genes that show the same response under a subset of conditions, example, if a cellular process is only active under these conditions. Furthermore, if a gene participates in multiple pathways that are differentially regulated, one would expect this gene to be included in more than one cluster this cannot be achieved by traditional clustering.

Manuscript received on April, 2013.

R.Parimala, Assistant Professor, CSE dept, Info Institute of Engineering. Coimbatore, Tamilnadu, India.

Condition Gene	Condition 1	Condition 2	… … .	Condition m
Gene 1	a11	a12	… ...	a1m
Gene 2	a21	a22	… …	a2m
……	….	…..	… ..	……
Gene n	an1	an2	… ...	anm

Fig.2 Gene Expression Matrix

II. LITERATURE REVIEW

Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Clustering is being considered as the main approach for analysis of gene expression data.

- Gene based clustering
- Sample based clustering
- Subspace clustering

One of the characteristics of gene expression data is to cluster both genes and samples. In gene based clustering, the genes are treated as the objects, while the samples are the features. Some of the gene based clustering algorithms are K-means, Hierarchical clustering, Model-based clustering. In sample based clustering regards the samples as objects and genes as features. One of the sample based clustering algorithm is CLIFF. The major drawbacks of commonly used clustering algorithms are that they assign each gene to a single cluster. Both gene based and sample based clustering algorithms are examples of global clustering. To identify the local clustering in microarray data, subspace clustering is introduced. Subspace clustering capture clusters formed y subset of genes across a subset of samples. A single gene may participate in multiple pathways that may or may not be coactive under all conditions, so that a gene can participate in multiple clusters. Some of the subspace clustering algorithms is Coupled Two-Way clustering, Plaid model, Biclustering.

A. Subspace Clustering

Subspace clustering is the task of detecting all clusters in all subspaces. Subspace clustering is an extension of traditional clustering that seeks to find clusters in different subspaces within a dataset. Standard clustering methods for the analysis of gene expression data only identifies the global models while missing the local expression patterns. One of the types of subspace clustering is Biclustering. Biclustering is an important approach in microarray data analysis.

Biclustering, co-clustering, or two-mode clustering is a data mining technique which allows simultaneous clustering of the rows and columns of a matrix. Biclustering of the gene expressing data is an important task in bioinformatics. Using biclustering algorithms, one can obtain sets of genes that are co-regulated under subsets of conditions. Biclustering is one of the types of Subspace Clustering. In gene expression data, a biclusters is a subset of genes exhibiting a consistent pattern over a subset of conditions.Different from traditional clustering methods, such as hierarchical clustering and k-means clustering, Cheng and Church used a biclustering method for the analysis of gene expression data.

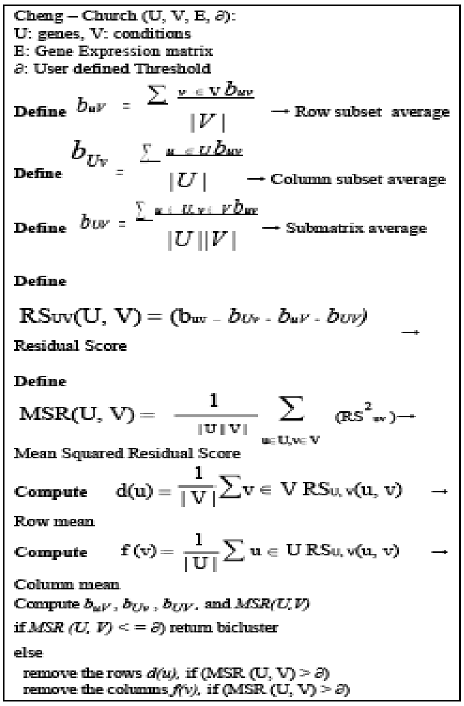
B. Biclustering Types

Bicluster can be categorized into the following types based on the patterns exhibited by the underlying sub matrices. They are

1. Bicluster with constant values
2. Bicluster with constant values on rows or columns
3. Bicluster with coherent values (additive and multiplicative)

III. EXISTING SYSTEM

The phenomenon of biclustering used to analyze gene expression data was firstly introduced by Cheng and Church, as an optimization problem in the algorithmic framework. The algorithm aims to extract biclusters followed by solving the restricted optimization problem defined by the respective scoring function. This algorithm works on greedy iterative search method, based on the idea of maximizing the local gain by adding or removing rows or columns from the biclusters. The algorithm considers a gene expression data matrix. The row subset average, column subset average and submatrix average are computed. By using this subset averages, residual score and mean square residual score are calculated. After this calculation is done, row mean and column mean is calculated. Cheng and Church Biclustering algorithm consists of two phases. In the first phase, the rows mean or column mean having the higher value are deleted. In the second phase, the rows and columns are being added by looking for the lowest mean squared residues without exceeding the user defined threshold value.



```
Cheng – Church (U, V, E, ∂):
U: genes, V: conditions
E: Gene Expression matrix
∂: User defined Threshold
Define $b_{uV} = \frac{\sum_{v \in V} b_{uv}}{|V|}$ → Row subset average
Define $b_{Uv} = \frac{\sum_{u \in U} b_{uv}}{|U|}$ → Column subset average
Define $b_{UV} = \frac{\sum_{u \in U, v \in V} b_{uv}}{|U||V|}$ → Submatrix average
Define
$RS_{UV}(U, V) = (b_{uv} - b_{Uv} - b_{uV} - b_{UV})$ →
Residual Score
Define
$MSR(U, V) = \frac{1}{|U||V|} \sum_{u \in U, v \in V} (RS^2_{uv})$ →
Mean Squared Residual Score
Compute $d(u) = \frac{1}{|V|} \sum v \in V\ RS_{U,V}(u, v)$ →
Row mean
Compute $f(v) = \frac{1}{|U|} \sum u \in U\ RS_{U,V}(u, v)$ →
Column mean
Compute $b_{uV}$, $b_{Uv}$, $b_{UV}$, and $MSR(U,V)$
if $MSR(U, V) <= ∂)$ return bicluster
else
  remove the rows $d(u)$, if (MSR (U, V) > ∂)
  remove the columns $f(v)$, if (MSR (U, V) > ∂)
```

Fig.3 Cheng and Church algorithm

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

IV. DRAWBACKS IN EXISTING SYSTEM

Clustering the microarray data is based on user defined threshold value, this affects the quality of biclusters formed. This value will affect the quality of biclusters by missing some of the genes or by including some unwanted genes. Once a bicluster is created, its entries are replaced by random numbers, preventing the identification of overlapping biclusters. Problem of finding the minimum set of biclusters to cover all the elements in a data matrix is very hard.

V. MOTIVATION

Biclustering is an important approach in microarray data analysis. The underlying bases for using bi-clustering in the analysis of gene expression data are (1) similar genes

may exhibit similar behaviors only under a subset of conditions, not all conditions, (2) 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. Cheng and Church biclustering algorithm is based on the user defined threshold value, this affects the quality of biclusters formed. This value will affect the quality of biclusters by missing some of the genes or by including some unwanted genes. To overcome these problems, develop an algorithm to find an optimal bicluster with threshold value ∂ rather than user defined threshold.

VI. PROPOSED SYSTEM

In proposed system, the threshold value is calculated by using volume of the matrix and number of rows and columns in the matrix. This model defined a bicluster as a submatrix that exhibits some coherent tendency, each bicluster can be uniquely identified by its set of genes and conditions. The proposed method was somewhat similar to the model proposed by Cheng and Church

(2000), but the parameter ∂ is calculated by using volume of the matrix and the number of rows and columns in the matrix. Gene expression data as an n x m matrix A_{IJ} where I represent the set of genes and J represents the set of conditions. An entry a_{ij} of this matrix corresponds to the expression level of gene i under a condition j. The concept of bicluster was introduced by Cheng and Church (2000) to capture the coherence of a subset of genes and a subset of conditions. While the mean squared residue represents the variance of the selected genes and conditions with respect to the coherence, the goal of biclustering is to find biclusters with low mean squared residue.

A bicluster is defined as a submatrix A_{IJ} that exhibits some coherency. The volume V_{IJ} of a bicluster is defined as the number of non-missing entries present in the bicluster. Since the biclusters identified may not be perfect, introduce the notion of residue to evaluate the quality of biclusters and its elements. In proposed model consists of two phases.

- Addition Phase
- Deletion Phase

Threshold value calculation

The threshold value is calculated by using the formula

$$\partial = \frac{|D|}{N \times M} \qquad (2)$$

where |D| is the volume of original matrix which has N rows and M columns.

The row subset average, column subset average and submatrix average are computed. By using this subset averages, residual score and mean square residual score are calculated. After this calculation is done, row mean and column mean is calculated. Proposed algorithm consists of two phases. In the first phase, the rows mean or column mean having the higher value are deleted, if the MSR is greater than the ∂. In the second phase, the rows and columns are being added by looking for the lowest mean squared residues without exceeding the ∂.This iterative algorithm on convergence results into ∂-biclusters, having low mean squared residue and locally maximal size. The bicluster elements are masked with randomly generated uniform values in the original matrix, for usage in next iteration.

```
Proposed algorithm (U, V, E, $\partial$):
U: genes, V: conditions
E: Gene Expression matrix
$\partial$: Threshold
Compute $\partial = \frac{D}{N \times M}$ → Threshold value
Define $b_{uV} = \frac{\sum_{v \in V} b_{uv}}{|V|}$ → Row subset average
Define $b_{Uv} = \frac{\sum_{u \in U} b_{uv}}{|U|}$ → Column subset average
Define $b_{UV} = \frac{\sum_{u \in U, v \in V} b_{uv}}{|U||V|}$ → Submatrix average
Define
$RS_{UV}(U, V) = (b_{uv} - b_{Uv} - b_{uV} - b_{UV})$ → Residual Score
Define
$MSR(U, V) = \frac{1}{|U||V|} \sum_{u \in U, v \in V} (RS_{uv})^2$
→Mean Squared Residual Score
Compute $d(u) = \frac{1}{|V|} \sum v \in V\ RS_{U,V}(u, v)$ → Row mean
Compute $f(v) = \frac{1}{|U|} \sum u \in U\ RS_{U,V}(u, v)$ → Column mean
Compute $b_{uV}$, $b_{Uv}$, $b_{UV}$, and $MSR(U,V)$
```

```
if MSR (U, V) < = ∂) return bicluster
else
        remove the rows d(u), if (MSR (U, V) > ∂)
        remove the columns f(v), if (MSR (U, V) > ∂)
```

Fig.4 Proposed algorithm

threshold value and biclusters are formed based on the mean squared residual score. In proposed algorithm, the threshold value ∂ is calculated based on the volume of the matrix and the number of rows and columns in the matrix.

Table I shows the performance comparison of different algorithms for synthetic dataset. The user defined threshold value is set as 500 in CC algorithm.

Table II shows the quality of different biclustering algorithms by calculating the average (residue/volume) ratio of the biclusters obtained from them:

$$\text{Quality} = \frac{1}{n}\sum_{i=1}^{n}\frac{\text{Residue}_i}{\text{Volume}_i} \qquad (3)$$

The rationale for this metric is obvious, the smaller the residue and /or the larger the volume, the better is the quality of a bicluster.

Table I. Performance comparison of different algorithms for synthetic dataset

	Dataset size	Threshold value	Avg. residue	Avg. gene num	Avg. cond. num
CC algorithm	100 × 100	500	486.5	9.3	32.8
Proposed algorithm	100× 100	1.400842E31	417.5	10.6	12

Table II. Quality comparison of different algorithms

	Cheng and Church algorithm	Proposed algorithm
Dataset size - 100× 100	22.4%	36.03%

VII. EXPERIMENTAL RESULTS

Microarray synthetic dataset is considered. It comprises of 100 rows and 100 columns. Cheng and Church biclustering algorithm is implemented with user defined

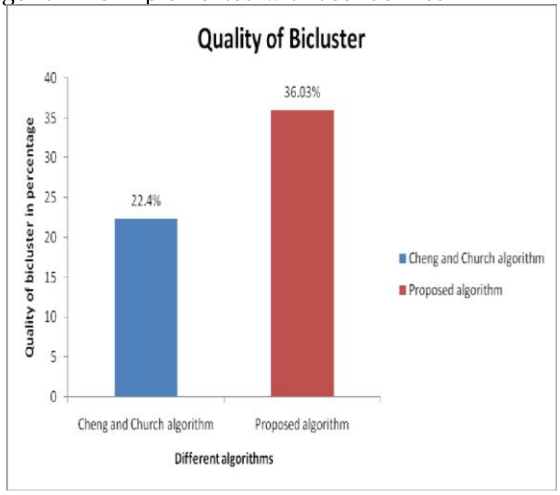


Fig.5 Quality of Bicluster

VIII. CONCLUSION

Biclustering is an important technique to identify submatrices in gene expression dataset. Biclustering allows finding subgroups of genes that show the same response under a subset of conditions. Cheng and Church algorithm is implemented and this algorithm is based on mean square residual score. For a given dataset, residual score is calculated for each row and column. Finally, bicluster are formed with low mean squared residue score and user defined threshold value.

Clustering the microarray data is based on user defined threshold value, this affects the quality of biclusters formed. This value will affect the quality of biclusters by missing some of the genes or by including some unwanted genes. Problem of finding the minimum set of biclusters to cover all the elements in a data matrix is very hard. To overcome these problems, develop an algorithm to find an optimal bicluster with threshold value ∂ rather than user defined threshold. The quality of bicluster 14% is improved in our proposed method. Therefore, biclusters are formed based on the low mean squared residues and ∂, which improved the quality of the biclusters.

REFERENCES

1. Amos Tanay, Roded Sharan and Ron Shamir (2004),"Biclustering Algorithms: A Survey", Handbook of Computational Molecular Biology.
2. Anupam Chakraborty and Hitashyam Maka (2005), "Biclustering of Gene Expression Data Using Genetic Algorithm", IEEE.
3. Cheng Y and Church G.M (2000), "Biclustering of expression data", Proceedings of International Conference on Intelligent System and Molecular Biology, Vol. 8, pp. 93–103.
4. Cheng K.O, Law N.F, Siu W.E and Liew (2007), "Biclusters Visualization and Detection Using Parallel Coordinate Plots", American Institute of Physics Conference Proceedings, Vol. 952, pp. 114-123.
5. Daxin Jiang, Chun Tang and Aidong Zhang (2004), "Cluster Analysis for Gene Expression Data: A Survey", IEEE Transactions on Knowledge and Data Engineering, Vol. 16, no. 11.
6. Jiong Yang, Haixun Wang, Wei Wang, Philip (2003), "Enhanced Biclustering on Expression Data", Proceedings of Third IEEE Conference on Bioinformatics and Bioengineering, pp.321-327.
7. Nighat Noureen and Muhammad Abdul Qadir (2009), "BiSim: A Simple and Efficient Biclustering algorithm", International Conference of Soft Computing and Pattern Recognition, IEEE.
8. Nishchal K. VermaL, Sheela Meena, Amarjot Singh, Shruti Bajpai, Van Cui, Aditya Nagrare (2010), " A Comparison of Biclustering Algorithms", Proceedings of International Conference on Systems in Medicine and Biology, IEEE.
9. Prelic A, Bleuler S, Zimmermann P, Wille A, Buhlmann P (2006), "A systematic comparison and evaluation of biclustering methods for gene expression data", Bioinformatics, Vol. 22,no. 9, pp. 1122-1129.
10. Sara C. Madeira and Arlindo L. Oliveira (2004), "Biclustering algorithms for biological data analysis: a survey", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 1, pp. 24-45.
11. Wen-Hui Yang, Dao-Qing Dai and Hong Yan (2011)," Finding Correlated Biclusters from Gene Expression Data", IEEE Transactions on Knowledge and Data Engineering, Vol. 23, no. 4.
12. Xiaowen Liu and Lusheng Wang (2007)," Computing the maximum similarity biclusters of gene expression data", Bioinformatics, Vol. 23, no. 1, pp. 50–56.

AUTHORS PROFILE

Assistant Professor R.Parimala, completed her ME CSE in Bannari Amman institute of Technology, Erode in 2012. And completed her BE CSE in Vivekananda College of engineering for women, tiruchengode in 2010. Her interested area is data mining.