Biclustering Gene Expression Data using MSR Difference Threshold

Shyama Das[#1], Sumam Mary Idicula[*2]

[#1,2]*Department of Computer Science, Cochin University of Science and Technology*
Kochi, Kerala, India

[1]shyamadas777@gmail.com

[2]sumam@cusat.ac.in

Abstract— **Biclustering is simultaneous clustering of both rows and columns of a data matrix. A measure called Mean Squared Residue (MSR) is used to simultaneously evaluate the coherence of rows and columns within a submatrix. In this paper a novel algorithm is developed for biclustering gene expression data using the newly introduced concept of MSR difference threshold. In the first step high quality bicluster seeds are generated using K-Means clustering algorithm. Then more genes and conditions (node) are added to the bicluster. Before adding a node the MSR X of the bicluster is calculated. After adding the node again the MSR Y is calculated. The added node is deleted if Y minus X is greater than MSR difference threshold or if Y is greater than MSR threshold which depends on the dataset. The MSR difference threshold is different for gene list and condition list and it depends on the dataset also. Proper values should be identified through experimentation in order to obtain biclusters of high quality. The results obtained on bench mark dataset clearly indicate that this algorithm is better than many of the existing biclustering algorithms.**

Keywords— **biclustering, gene expression data, data mining, K-Means clustering.**

I. Introduction

Gene expression study is revolutionized by microarray technologies which simultaneously measures the expression levels of thousands of genes in a single experiment. Gene expression data is organized in the form of a matrix where rows represent genes and columns represent experimental conditions. Each element in the matrix refers to the expression level of a particular gene under a specific condition

Clustering as the most popular data mining technique identifies genes with same functions or the same regulatory mechanisms. However clustering has its limitations. Clustering is based on the assumption that related genes behave similarly across all measured conditions. Based on a general understanding of the cellular process, subset of genes are co-regulated and co-expressed under certain experimental conditions but behaves almost independently under other conditions. Moreover clustering partitions genes into disjoint sets i.e. each gene is associated with a single biological function which is in contradiction to the biological system.

Biclustering is clustering applied in two dimensions simultaneously. This approach identifies group of genes that show similar expression level under a specific subset of experimental conditions. Hartigan introduced biclustering[1]. Cheng and Church were the first to apply biclustering for the analysis of gene expression data [2]. They introduced a measure known as mean squared residue score to assess the coherence of the elements of a bicluster.

In this work a novel algorithm is introduced based on the innovative idea of MSR difference threshold. Only genes or conditions having MSR difference value less than the MSR difference threshold is added to the bicluster. This results in the generation of biclusters with high coherence.

II. MATERIALS AND METHODS

A. *Problem Definition*

The gene expression dataset can be viewed as an NxM matrix A of real numbers. Let $X=\{G_1,G_2,....G_N\}$ be the set of genes and $Y=\{C_1,...C_M\}$ be the set of conditions in the gene expression dataset. A bicluster of a gene expression dataset is a subset of genes which exhibit similar expression patterns along a subset of conditions. That means a bicluster is a submatrix B of A and if the size of B is IxJ, then I is a subset of rows X of A, and J is a subset of the columns Y of A. The rows and columns of the bicluster B need not be contiguous as in the expression matrix A.

Cheng and Church defined a bicluster to be a subset of rows and subset of columns with high similarity score. Similarity is measured by the coherence of genes and conditions in the subset. Biclusters of this kind are called biclusters with coherent values. They are biologically more relevant than biclusters with constant values. In this work biclusters with coherent values are identified. The degree of coherence is measured by mean squared residue score or Hscore. It is the sum of the squared residue score. The residue of an element reveals its degree of coherence with the other elements of the bicluster it belongs to. The residue score of an element *bij* in a submatrix *B* is defined as *RS(bij)=bij-bIj-biJ+bIJ*

The residue score of an element *bij* provides the difference between the actual value and its expected value predicted from its row mean, column mean and bicluster mean. Hence from the value of residue, the quality of the bicluster can be evaluated by computing the mean squared residue. That is Hscore or mean squared residue score of bicluster *B* is
MSR(*B*) $= \frac{1}{|I||J|}\sum i \in I, j \in J\, (RS(bij))^2$ where

978-1-4244-4859-3/09/$25.00 ©2009

$$b_{iJ} = \frac{1}{|J|}\sum_{j \in J}(b_{ij})$$

$$b_{Ij} = \frac{1}{|I|}\sum_{i \in I}(b_{ij})$$

$$b_{IJ} = \frac{1}{|I||J|}\sum_{i \in I, j \in J}(b_{ij})$$

Here *I* denotes the row set, *J* denotes the column set, *bij* denotes the element in a submatrix, *biJ* denotes the ith row mean, *bIj* denotes the *j*th column mean, and *bIJ* denotes the mean of the whole bicluster.

A bicluster *B* is called a δ bicluster if MSR (*B*)< δ for some δ >0. If the MSR value is high it means that the data is uncorrelated. If the MSR value is low then there is correlation in the matrix. The value of δ(MSR threshold) depends on the dataset and it should be calculated in advance. For Yeast dataset the value of δ is 300. The volume of a bicluster or bicluster size is the product of number of rows and the number of columns in the bicluster. Larger the volume and smaller the MSR or Hscore of the bicluster better is the quality of the bicluster. In this algorithm the identified biclusters neither need to be disjoint nor be covering the entire matrix.

B. *Encoding of Bicluster*

Bicluster is represented by a binary string of fixed size n+m where *n* and *m* are the number of genes and conditions of the microarray dataset, respectively. The first n bits are associated to n genes and the following m bits to m conditions. If a bit is set to 1, it means that the corresponding gene or condition belongs to the bicluster; otherwise it does not. This encoding presents the advantage of having fixed size [4].

C. Seed Finding

The gene expression dataset is divided into n gene clusters and m sample clusters using the K-Means algorithm. For getting a maximum of 10 genes per gene cluster, it is further divided according to the cosine angle distance from the cluster centre. In the same manner each sample cluster is further divided into sets of 5 samples according to cosine angle distance from the cluster centre. The number of gene clusters, having a maximum of 10 close genes is p and the number of sample clusters having maximum 5 conditions is q. The initial gene expression data matrix is thus partitioned into p*q submatrices and bicluster seeds having Hscore value below a certain limit is selected as seeds.

D. Biclustering using MSR difference Threshold.

Each seed obtained from K-Means clustering algorithm is enlarged separately by adding more conditions and genes from the condition list and gene list respectively. Condition list and gene list is formed by the set of genes and conditions not included in the bicluster. MSR difference of a gene or condition is the incremental increase in MSR after adding the same to the bicluster. In this method a gene or condition (node) is added to the bicluster. If the incremental increase in MSR after adding the node is greater than MSR difference threshold or if the MSR of the resulting bicluster is greater than δ, the added node is removed from the bicluster. MSR difference threshold is different for genelists as well as condition list. And it varies depending on the dataset. The identification of suitable value needs experimentation. It is observed that if MSR difference threshold for condition list is set to 30 it is possible to get biclusters with all 17 conditions for the Yeast dataset. For gene list the MSR difference threshold is set to 10. By properly adjusting the MSR difference threshold biclusters of high quality can be obtained.

In the case of novel greedy search algorithm [6], the added node is removed only when the MSR of the bicluster exceeds δ i.e. MSR threshold. In this algorithm there is one more constraint called MSR difference threshold. Hence this method can produce better biclusters compared with novel greedy search algorithm.

```
Algorithm biclusteringMSRdiffernce(seed, δ, msrdiffgenethresh, msrdiffcondthresh )
bicluster := seed
previous=MSR(seed)
j:= 1;
While (j <= total _no_conditions)
If  condition[ j]  is not included in bicluster
  Changed=1;
  Add all elements of condition[ j]  corresponding to genes already included to bicluster
 present= MSR(bicluster)
      if (present> δ) or (present-previous)>msrdiffcondthresh
       remove elements of  condition[ j]  from bicluster
changed=0;
endif
if changed==1
previous=present
 endif
endif
j:= j+1
end(while)
 i := 1;
prev=MSR(bicluster)
While (i <= total _no_ genes)
If  gene[i]  is not included in bicluster
  Changed=1;
 Add all elements of gene[i] corresponding to conditions
already included to bicluster
 present= MSR(bicluster)
        if (present> $\delta$) or (present-previous)>msrdiffgenethresh
             remove elements of gene[i] from bicluster
            changed=0
        endif
  if changed==1
      previous=present
 endif
endif
i:= i+1
end(while)
return bicluster
end(biclusteringMSRdifference)
```

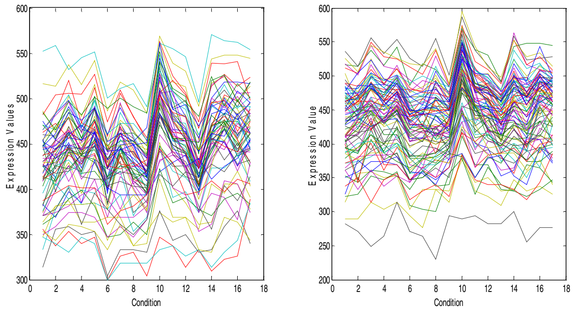
III. EXPERIMENTAL RESULTS

A. Dataset used

Experiments are conducted on the Yeast Saccharomyces Cerevisiae cell cycle expression dataset to evaluate the quality of the proposed method. The dataset is based on Tavazoie et al [7]. Dataset consists of 2884 genes and 17 conditions. The values in the expression dataset are integers in the range 0 to 600.Missing values are represented by -1. The datasetis obtained from http://arep.med.harvard.edu/biclustering

B. *Bicluster Plots*

Figure 1 shows eight biclusters obtained by this algorithm on Yeast dataset. Expression values of the genes in the biclusters show similar up regulation and down regulation under a set of experimental conditions.



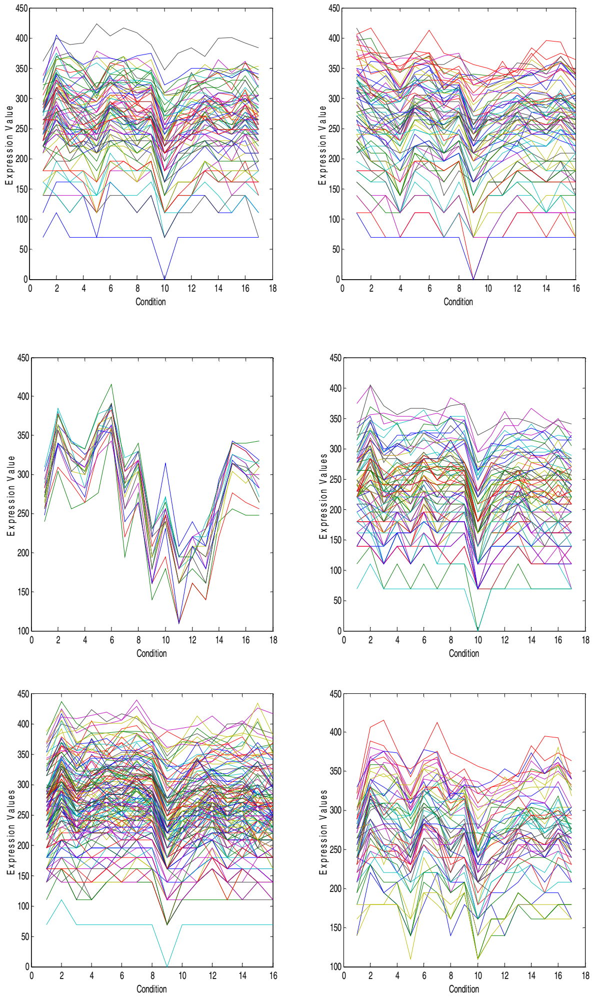


Fig.1.Eight biclusters found for the Yeast dataset. Bicluster labels are (a), (b), (c), (d), (e), (f), (g) and (h) respectively. In the bicluster plots X axis contains conditions and Y axis contains expression values. The details about biclusters can be obtained from Table I using bicluster label. All the means squared residues are lower than 200.Only biclusters with different shapes are selected.

TABLE I
INFORMATION ABOUT BICLUSTERS OF FIG. 1.

Label	Rows	Columns	volume	MSR
(a)	65	17	1105	198.8756
(b)	86	17	1462	198.3953
(c)	74	17	1258	199.6859
(d)	77	16	1232	199.5544
(e)	16	17	272	199.3662
(f)	81	17	1377	199.9548
(g)	140	16	2240	199.6735
(h)	55	17	935	199.4912

In the above table the first column contains the label of each bicluster. The second and third columns report the number of rows (genes) and of columns (conditions) of the bicluster respectively. The fourth column reports the volume of the bicluster and the last column contains the mean squared residue of the bicluster. For yeast dataset all conditions are obtained when the MSR difference threshold for conditionlists is set to 30. For genelist MSR difference threshold is set to 10.

IV. COMPARISON

The Table II given below provides a comparative summarization of results obtained by various biclustering algorithms for the Yeast dataset. The MSR value of biclusters obtained by all the algorithms listed in Table II are more or less equal to 200, even though the maximum limit of δ is 300. Hence the value of δ is set to 200 in this study. The performance of biclustering using MSR difference threshold is compared with novel greedy algorithm [6], SEBI [8], Cheng and Church's algorithm (CC) [2], FLOC [9] and DBF [10]. In novel greedy node addition follows node deletion. The added node is deleted only if the MSR value of the bicluster exceeds δ. SEBI (Sequential Evolutionary Biclustering) is based on evolutionary Computation. In the Cheng and Church approach, rows/columns were deleted from the gene expression data matrix to find a bicluster. Yang et al (2003) generalized the model of bicluster proposed by Cheng and Church for incorporating null values and for removing random interference. They developed a probabilistic algorithm FLOC that discovers a set of possibly overlapping biclusters simultaneously. Zhang et al. presented DBF (Deterministic Biclustering with frequent pattern mining).

In the case of the biclustering using MSR difference threshold, MSR value is better than that of novel greedy, SEBI and CC. Largest bicluster size is better than that of all other algorithms except CC. Average volume is better than that of all other algorithms. It is better than novel greedy in all aspects. In multi-objective evolutionary biclustering[11] maximum number of conditions obtained is only 11. In this method almost all biclusters are with 17 conditions. Moreover this algorithm provides better performance in terms of speed compared to all the metahueristic and evolutionary algorithms.

TABLE 1I
PERFORMANCE COMPARISON BETWEEN BICLUSTERING USING MSR DIIFERENCE THRESHOLD AND OTHER ALGORITHMS

Algorithm	Avg. MSR	Avg. Volume	Avg. gene num.	Avg. cond. num	Largest Bicluster size
MSR diff. Threshold	199.63	2264.80	170.16	14.83	4400
Novel Greedy	199.78	1422.87	94.75	14.75	2112
SEBI	205.18	209.92	13.61	15.25	1394
CC	204.29	1576.98	166.71	12.09	4485
FLOC	187.54	1825.78	195.00	12.80	1408
DBF	114.70	1627.00	188.00	11.00	4000

V. CONCLUSION

One of the major applications of biclustering is for the gene expressions of cancerous data in the identification of coregulated genes, gene functional annotation and sample classification. In this paper a new algorithm is developed based on MSR difference threshold which is a newly introduced concept 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 submatrices. From these submatrices those with MSR value below a certain threshold are selected as seeds which are small tightly coregulated submatrices. Then suitable values for MSR difference threshold are identified for gene list and condition list. More genes and conditions are added to these seeds only when the MSR difference of the added node is less than MSR difference threshold or when the MSR value of bicluster generated after the addition of new node is less than MSR threshold. Otherwise the added node is removed from the bicluster. The algorithm is implemented on the Yeast dataset. Based on the algorithm implementation on the above mentioned benchmark dataset a comparative assessment of the results are given in order to demonstrate the effectiveness of the proposed method. In terms of the average gene number, average number of conditions, average MSR, average volume and largest bicluster size the biclusters obtained in this method is far better than many of the biclustering algorithms. Moreover from the bicluster plots it is clear that this method finds high quality biclusters. The expression values of genes in the bicluster show strikingly similar up-regulation and down-regulation under a set of experimental conditions

REFERENCES

[1] J. A. Hartigan, “Direct clustering of Data Matrix”, Journal of the American Statistical Association Vol.67, no.337, pp. 123-129, 1972.

[2] Yizong Cheng and George M. Church, “Biclustering of expression data”, Proc. 8th Int. Conf. Intelligent Systems for Molecular Biology, pp. 93-103, 2000.

[3] Madeira S. C. and Oliveira A. L., “Biclustering algorithms for Biological Data analysis: a survey” IEEE Transactions on computational biology and bioinformatics, pp. 24-45, 2004.

[4] Anupam Chakraborty and Hitashyam Maka “Biclustering of Gene Expression Data Using GeneticAlgorithm” Proceedings of Computation Intelligence in Bioinformatics and Computational Biology CIBCB, pp. 1-8, 2005.

[5] Chakraborty A. and Maka H., “Biclustering of gene expression data by simulated annealing”,HPCASIA ’05, pp. 627-632, 2005.

[6] Shyama Das and Sumam Mary Idicula “A Novel Approach in Greedy Search Algorithm for Biclustering Gene Expression Data” International Conference on Bioinformatics, Computational and Systems Biology (ICBCSB) which was held in Singapore during Aug 26-28, 2009.

[7] Tavazoie S., Hughes J. D., Campbell M. J., Cho R. J. and Church G. M., “Systematic determination of genetic network architecture”, Nat. Genet., vol.22, no.3 pp. 281-285, 1999.

[8] Federico Divina and Jesus S. Aguilar-Ruize, “Biclustering of Expression Data with Evolutionary computation”, IEEE Transactions on Knowledge and Data Engineering, Vol. 18, pp. 590-602, 2006.

[9] J. Yang, H. Wang, W. Wang and P. Yu, “Enhanced Biclustering on Expression Data”, Proc. Third IEEE Symp, BioInformatics and BioEng. (BIBE’03), pp. 321-327, 2003.

[10] Z. Zhang, A. Teo, B. C. Ooi, K. L. Tan, “Mining deterministic biclusters in gene expression data”, In: Proceedings of the fourth IEEE Symposium on Bioinformatics and Bioengineering(BIBE’04), 2004, pp. 283-292, 2004.

[11] Banka H. and Mitra S., “Multi-objective evolutionary biclustering of gene expression data”, Journal of Pattern Recognition, Vol.39 pp. 2464-2477, 2006.