Journal of Theoretical and Applied Information Technology
15th October 2011. Vol. 32 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

IMPROVED BICLUSTERING ALGORITHM FOR GENE EXPRESSION DATA

[1]SADIQ HUSSAIN, [2] PROF. G.C. HAZARIKA

[1]System Administrator, Examination Branch, Dibrugarh University, Dibrugarh, Assam, India

[2]Director i/c, Centre for Computer Studies, Dibrugarh University, Dibrugarh, Assam, India

E-mail: [1]sadiqdu@rediffmail.com , [2]gchazarika@gmail.com

ABSTRACT

Biclustering algorithms simultaneously cluster both rows and columns. These types of algorithms are applied to gene expression data analysis to find a subset of genes that exhibit similar expression pattern under a subset of conditions. Cheng and Church introduced the mean squared residue measure to capture the coherence of a subset of genes over a subset of conditions. They provided a set of heuristic algorithms based primarily on node deletion to find one bicluster or a set of biclusters after masking discovered biclusters with random values. The mean squared residue is a popular measure of bicluster quality. One drawback however is that it is biased toward flat biclusters with low row variance. In this paper, we introduce an improved bicluster score that removes this bias and promotes the discovery the most significant biclusters in the dataset. We employ this score within a new biclustering approach based on the bottom up search strategy. We believe that the bottom-up search approach better models the underlying functional modules of the gene expression dataset.

Keywords: *Bicluster, Gene expression Data, Clustering, Pattern Recognition*

1. INTRODUCTION

Advances in gene expression microarray technologies over the last decade or so have made it possible to measure the expression levels of thousands of genes over many experimental conditions (e.g., different patients, tissue types and growth environments). The data produced in these experiments are usually arranged in a data matrix of genes (rows) and conditions (columns). Results from multiple microarray experiments may be combined and the data matrix may easily exceed thousands of genes and hundreds of condition in size.

As datasets increase in size, however, it becomes increasingly unlikely that genes will retain correlation across the full set of conditions making clustering problematic. The gene expression context further exacerbates this problem as it is not uncommon for the expression of related genes to be highly similar under one set of conditions and yet independent under another set [4]. Given these issues it is perhaps more prudent to cluster genes over a significant subset of experimental conditions. This two-way clustering technique has been termed biclustering and was first introduced to gene expression analysis by Cheng and Church [5]. They developed a two-way correlation metric called the mean squared residue score to measure the bicluster quality of selected rows and columns from the gene expression data matrix. They employed this metric within a top-down greedy node deletion algorithm aimed at discovering all the significant biclusters within a gene expression data matrix. Following this seminal work, other metrics and biclustering frameworks were developed [6], [7], [8]. However, approaches based on Cheng and Church's mean squared residue score remain most prevalent in the literature [9], [10], [11], [12].

One notable drawback, however, of the mean squared residue score is that it is also affected by variance, favouring correlations with low variances. Furthermore, because variance changes by the square of the change in scale, the score tends to discover correlations over lower scales. These effects culminate in a bias toward 'flat' biclusters containing genes with relatively unfluctuating expression levels within the lower scales (fold changes) of the gene expression dataset. This issue has been articulated previously in [13]. In this paper, we introduce an improved bicluster scoring metric which compensates

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

for this bias and enables the discovery of biclusters throughout expression data, including those potentially more interesting correlations over the higher scales (fold changes).

Correlation Coefficient between two random variables may be used for studying the linear dependency between two genes. In this paper, this fact has motivated the use of measures based on proposed correlations among genes [21,22]. In [23] the correlation coefficient is used for forming biclusters with a greedy algorithm. In [24] an enumeration algorithm based on a tree structure for biclustering is presented and it uses an evaluation function based on the Spearman's Rank correlation.

2. RELATED WORK

Cheng and Church were one of the first who introduced the term "biclustering" in the context of expression data analysis [2]. They also introduced the mean squared residue as a homogeneity measurement and proposed heuristic algorithms exploiting mathematical properties of the mean squared residue. Since then several different biclustering approaches have been proposed, for example random walk strategies [14], evolutionary algorithms [12], [13] and parameter distribution identification [15], [16]. The biclustering problem is taken to be NP-complete [2], so strategies solving this task should aim only at approximating the optimal solution in order to save time and space resources. Theoretical aspects of the mean squared residue have only slightly or not even at all been analyzed. Many approaches just incorporate or extend the heuristic algorithm of Cheng and Church with slightly or no changes. Cho et al figured out theoretical aspects of the mean squared residue for the columns and for the rows explicitly, and applied k-means clustering on both dimensions separately to find non-overlapping k row and l column cluster minimizing the homogeneity score [17]. Kung et al analyzed the impact of using more than one metric model and proposed a classifier system based on fuzzy support vector machines to gather subsets of genes and subsets of conditions [18]. Empirical studies of the mean squared residue have shown that by using this measure one is able to find shifting but no scaling patterns [19]. This observation emphasizes the fact that the mean squared residue may not be the only appropriate homogeneity measure for a comprehensive analysis of biological or other related data sets. In order to increase the performance in the context of biologically meaningful outcome of biclustering algorithms, the incorporation of ontological database knowledge might be a key feature of a gene expression analysis, as suggested by [16]. Thus there exist several drawbacks and limitations in determining gene ontological relationships by recent available tools yet [20], and so with focusing on solving the general biclustering problem, this work concentrates on the establishment of enrichment algorithms based on numerical information provided by the data matrix only.

In the context of gene expression analysis, the input data is usually given as a two-dimensional data matrix A maintaining the expression value ai,j of object i under condition j by a floating-point number. The rows are then in general associated with the genes and the columns refer to the conditions of a biological experiment. In the following we will consider the general case of a finite two-dimensional matrix A $\subset$ IR2 given by a set of rows and a set of columns.

Definition 1: bicluster

Given a n × m data matrix A with the set of rows N and the set of columns M, we denote the element of row i and column j of the matrix as ai,j . The row mean, or the mean value of row i is then given as

$$a_{i,J} = \frac{1}{|J|} \sum_{j \in J} a_{i,j},$$

and equally, the mean value of column j is denoted as

$$a_{I,j} = \frac{1}{|I|} \sum_{i \in I} a_{i,j}.$$

aI,J refers to the mean value of the whole bicluster given as

$$a_{I,J} = \frac{1}{|I||J|} \sum_{i \in I} \sum_{j \in J} a_{i,j}.$$

Finally, a bicluster (I, J) is defined as a subset of rows I $\subseteq$ N and a subset of columns J $\subseteq$ M. The row mean, the column mean and the mean of the whole bicluster are also known as the base of a row, the base of a column and the base of a bicluster respectively [5].

Definition 2: mean row and mean column of a bicluster Given a bicluster (I, J), the mean row mr of (I, J) is defined as

The mean column mc is defined analogously:

$$m_c(I, J) = \{a_{0,J}, a_{1,J}, \ldots, a_{|I|-1,J}\}. \quad (2)$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

mr stands for the |J|-dimensional vector maintaining the mean values of all columns, and is therefore clearly defined for each bicluster. Note that the mean row mr differs from the row mean, the mean value ai,J of row i. Same considerations hold for the mean column of a bicluster. In order to quantify the homogeneity within the elements of a bicluster, the mean squared residue H has been introduced by [3] as:

Definition 3: mean squared residue

Given a bicluster (I, J), the mean squared residue H(I, J) is defined as

$$H(I,J) = \frac{1}{|I||J|}\sum_{i\in I}\sum_{j\in J} r(a_{i,j})^2 \quad (5)$$

with

$$r(a_{i,j}) = a_{i,j} + a_{I,J} - a_{I,j} - a_{i,J}$$

representing the residue of element ai,j . Additionally the row residue dr(i) of a row i $\in$ I of a bicluster (I, J) is defined as:

$$d_r(i) = \frac{1}{|J|}\sum_{j\in J} r(a_{i,j})^2.$$

The definition of the column residue dc(j) of a column j $\in$ J is analogously given as

$$d_c(j) = \frac{1}{|I|}\sum_{i\in I} r(a_{i,j})^2.$$

The mean squared residue measures the coherence of all elements within a bicluster, with low values indicating a high correlation. A bicluster (I, J) is called a perfect bicluster if H(I, J) = 0, and a bicluster with H(I, J) $\leq \delta$ is called δ-bicluster. δ describes the tolerated level of deviance within a bicluster motivated by technological boundaries of precision and by the occurrence of systematic and non-systematic bias in experimental measurements in general. For this reason, most approaches search for δ-bicluster rather than for perfect ones. Besides the degree of homogeneity, a bicluster (I, J) can also be characterized by measuring its information content. One often used function is known as the mean row variance Var(I, J) [2], defined as:

Definition 4: mean row variance

Given a bicluster (I, J), the mean row variance

Varr(I, J) is defined as

$$Var_r(I,J) = \frac{1}{|I|}\sum_{i\in I} v_r(i) \quad (7)$$

with

$$v_r(i) = \frac{1}{|J|}\sum_{j\in J}(a_{i,j} - a_{i,J})^2 \quad (8)$$

as the variance of row i. The mean column variance is defined analogously.

Definition 5: residual distance

Given the set of rows N and the set of columns M of an n ×m input matrix A, the residual distance dr(i, k) between any pair of rows (i, k) $\in$ I × I is defined as:

$$d_r(i,k) = \frac{1}{|J|}\sum_{j\in J} r(a_{i,j}, a_{k,j})^2 \quad (9)$$

with

$$r(a_{i,j}, a_{k,j}) = r(a_{k,j}) - r(a_{i,j}) = a_{k,j} - a_{k,J} + a_{i,J} - a_{i,j} \quad (10)$$

and r(ai,j), r(ak,j) known as the residue of element ai,j and ak,j respectively (see equation (4)). The residual distance between any pair of columns dc can be defined adequately as

$$r(a_{i,j}, a_{i,l}) = r(a_{i,l}) - r(a_{i,j}) = a_{i,l} - a_{I,l} + a_{I,j} - a_{i,j}. \quad (11)$$

The residual distance function dr measures the distance between two rows or two columns, whereby the outcome is related to the homogeneity function mean squared residue. The question of what kind of relation exists and how it can be exploited to establish biclustering algorithms will be answered in detail in the following sections.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3. THE ALGORITHM

Inputs : Data Matrix, $A=a_{ij}$, i = 1…N, j=1….M, bicluster (I , J), homogeneity bounds δ , δ_1 and δ_2 with $\delta_1 \leq \delta < \delta_2$, positive integer values $K_1 \in IN$

Output : New bicluster (I' , J) with I' $\supseteq$ I.

Procedure Row_Extension

1. Compute m_r (I' , J), $a_{I',J}$ and d_r (I, m_r (I' , J)) for all rows i $\in$ N, and let I" be an empty set of rows.
2. Add all rows i' $\in$ N, i $\notin$ I' to the bicluster (I' , J), if d_r (i' , m_r (I', J)) $\leq \delta_1$ holds.
3. Repeat Step 1 and Step 2 until no improvements observed or maximum number K_1 iterations holds.
4. Compute m_r (I' , J),), $a_{I',J}$ and d_r (I, m_r (I' , J)) for all rows i $\in$ N.
5. Sort the set of rows

   I" = { i" | i" $\notin$ I' ^ d_r (i" , m_r (I' , J)) $\leq \delta_2$ }

   in ascending order of d_r (i" , m_r (I' , J)) .

Procedure Column_Extension

1. Compute m_c (I' , J), $a_{I',J}$ and d_c (I, m_c (I' , J)) for all rows i $\in$ N, and let I" be an empty set of rows.
2. Add all rows i' $\in$ N, i $\notin$ I' to the bicluster (I' , J), if d_c (i' , m_c (I', J)) $\leq \delta_1$ holds.
3. Repeat Step 1 and Step 2 until no improvements observed or maximum number K_1 iterations holds.
4. Compute m_c (I' , J),), $a_{I',J}$ and d_c (I, m_c (I' , J)) for all rows i $\in$ N.
5. Sort the set of rows

   I" = { i" | i" $\notin$ I' ^ d_c (i" , m_c (I' , J)) $\leq \delta_2$ }

   in ascending order of d_c (i" , m_c (I' , J)) .

Procedure Biclusters_finding ( )

Inputs : Data matrix $A=a_{ij}$, i=1…N, j=1…M, score limit δ , limit constant α , constant T.

Outputs : A set of T biclusters with scores $\leq \delta$

1. randomly select J $\subseteq$ {1,2,……..,M}, randomly select i $\in$ {1,2,…..N}, set I={i}, score =0
2. $\theta = \alpha\delta$
3. Row_Extension
4. If no extension is achieved go to step 2;
5. $\delta' = (\text{score} + \delta) / 2$
6. Column_Extension
7. $\delta' = \text{score}$
8. Column Extension based on δ'
9. Row Extension based on θ, δ'
10. print bicluster (I, J)
11. return;

Note that the value of θ is computed from the value of δ as $\theta = \alpha\delta$, where α is another constant. Lower value of α is another constant. Lower value of α (<1) leads to detection of more coherent biclusters with lower score, while higher value of α ($\geq$ 1) finds biclusters with higher score.

4. EXPERIMENTAL RESULTS

The Biclustering algorithm is tested on one set of expression data used in [2] and downloaded from http://arep.med.harvard.edu/biclustering . The dataset is the yeast data containing 2,884 genes and 17 conditions. Integer valued elements range between 0 and 600 with 34 missing values. We replaced the missing values with uniformly distributed random numbers within data range.

Our algorithm can detect biclusters with lower or higher score within the given limit of δ depending

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

on the selected value of α. This is demonstrated with the three biclusters shown in table 1 extracted from yeast dataset using three different values of α. All the biclusters extracted from the dataset may not be biologically interesting. We have not studied biological significance of the biclusters. We have extracted 100 biclusters from the dataset.

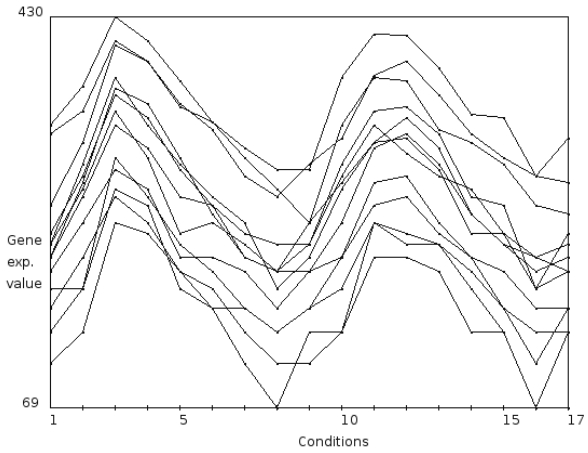

Figure I : Bicluster extracted from our algorithm

(genes: 216, 217, 526, 616, 1022, 1184, 1476, 1623,1795, 2086, 2278, 2375, 2538)

Table I. Sample biclusters in yeast dataset

α	δ	score	genes	conditions
1.0	300	144.15	8	17
1.2	300	198.93	13	17
2.0	300	295.03	31	17

Table II. Performance of proposed algorithm on yeast data set

Algorithm	Avg. score	Avg. gene	Avg. cond	Avg. Vol.
Cheng & Church	204.3	166.8	12.1	1577.0
Our	199.0	195.3	11.9	1773.2

5. CONCLUSION

In this paper we provided a bottom up algorithm that detects one biclusters at a time. Considering the impact of proposed algorithm , it is quite promising enrichment method with regard to mean square residue. An initial bicluster needs to be created or accepted as input and then it is extended by adding rows and columns. A set of biclusters are created with different initializations. Only a few passes (6, for example) over the data matrix is required to find a bicluster. The method may not be able to detect some very small biclusters as it adds rows incrementally. Because if there is no bicluster with 2 genes, it cannot detect bicluster with 3 genes.

The limitation of the algorithm is that although it generates biclusters; not all biclusters are found to be interesting.

Future works will focus on some improvements for the proposed algorithm with regard to the overlapping among genes and to the fitness function.

REFERENCES

[1] A. Ben-Dor, B. Chor, R. Karp, and Z. Yakini, “Discovering local structure in gene expression data : the order-preserving submatrix problem”, Journal of Computational Biology, vol. 10, No. 3-4, pp. 373-84, 2003.

[2] Y. Cheng and G.M. Church, “Biclustering of expression data”, in Proceedings of the Eight International Conference on Intelligent Systems for Molecular Biology (ISMB), 2000, pp. 93-103.

[3] A. Tanay, R. Sharan, and R. Shamir, “Discovering statistically significant biclusters in gene expression data”, Bioinformatics, vol. 18, pp. 36-44, 2002.

[4] L. Lazzeroni, and A. Owen, “Plaid models for gene expression data”, Statistica Sinica, vol. 12, pp. 61-86, 2002.

[5] Y. Kluger, R. Basri, J.T. Chang, and M. Gerstein, “Spectral biclustering of microarray data : Coclustering genes and conditions”, Genome Research, vol. 13, pp. 703-716, 2003.

[6] J. Yang, H. Wang, W. Wang, and P. Yu, “Enhanced biclustering on expression data”, in IEEE Third Symposium on Bioinformatics and Bioengineering, 2003.

[7] H. Cho, I.S. Dhillon, Y. Guan, and S. Sra, “Mimimum sum squared residue co-clustering

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

of gene expression data", SIAM international conference on datamining, 2004.

[8] K. Bryan, P. Cunningham, and N. Bolshakova, "Biclustering of expression data using simulated annealing", in Proceedings of the eighteenth IEEE Symposium on Computer Based Medical Systems, 2005.

[9] S. Bleuler, A. Prelic, and E. Zitzler, "An EA framework for biclustering of gene expression data", in Congress on Evolutionary Computation (CEC-2004), Piscataway, NJ : IEEE, 2004, pp.166-173.

[10] J. Aguilar-Ruiz, "Shifting and scaling patterns from gene expression data", Bioinformatics, vol. 21, No. 20., pp. 3849-3845, 2005.

[11] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: A survey," IEEE Transactions on computational Biology and Bioinformatics, Vol.1, No.1, January-March 2004, pp. 24 – 45, 2004.

[12] F. Divina and J. S. Aguilar-Ruiz, "Biclustering of expression data with evolutionary computation," IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 5, pp. 590–602, 2006.

[13] A. Chakraborty and H. Maka, "Biclustering of gene expression data using genetic algorithm," in CIBCB '05: Proceedings of the 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology,(CIBCB'05), 2005, pp. 1–8.

[14] F. Angiulli and C. Pizzuti, "Gene expression biclustering using random walk strategies." in Data Warehousing and Knowledge Discovery, 7th International Conference, DaWaK 2005, Copenhagen, Denmark, Aug 2005, pp. 509–519.

[15] A. Tanay, R. Sharan, and R. Shamir, "Discovering statistically significant biclusters in gene expression data," Bioinformatics, Vol.18, pp. 136–144, 2002.

[16] D. J. Reiss, N. S. Baliga, and R. Bonneau, "Integrated Biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks," BMC Bioinformatics, vol. 7, p. 280.

[17] H. Cho, L. S. Dhillon, Y. Guan, and S. Sra, "Minimum sum-squared residue co-clustering of gene expression data," Fourth SIAM International Conference of Data Mining, 2004.

[18] S. Y. Kung, M.-W. Mak, and I. Tagkopoulos, "Multi-metric and multisubstructure biclustering analysis for gene expression data," in CSB '05: Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference (CSB'05). Washington, DC, USA: IEEE Computer Society, 2005, pp. 123–134.

[19] J. S. Aguilar-Ruiz, "Shifting and scaling patterns from gene expression data," Bioinformatics, vol. 21, no. 20, pp. 3840–3845, 2005.

[20] G. O. Consortium, "The gene ontology (go) project in 2006," 2006. [Online]. Available: www.geneontology.org/

[21] Nepomuceno JA, Troncoso A, Aguilar-Ruiz JS: Evolutionary metaheuristic for biclustering based on linear correlations among genes. SAC 2010 : Proceedings of the 2010 ACM Symposium on Applied Computing (SAC), Sierre, Switzerland, March 22-26, 2010, 1143-1147.

[22] Nepomuceno JA, Troncoso A, Aguilar-Ruiz JS: Correlation-Based Scatter Search for Discovering Biclusters from Gene Expression Data. EvoBIO 2010 : Proceedings of the 8th European Conference on Evolutionary Computation, Machine Learning and Data Mining, Istanbul, Turkey, April 7-9, 2010, 122-133.

[23] Bhattacharya A, De RK: Bi-correlation clustering algorithm for determining a set of co-regulated genes. Bioinformatics 2009, 25(21): 2795-2801.

[24] Ayadi W, Elloumi M, Hao JK: A biclustering algorithm based on a Bicluster Enumeration Tree : application to DNA microarray data. BioData Mining 2009, 2:9.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AUTHOR PROFILES:

Sadiq Hussain MCA from Tezpur University, Assam,India in the year 2000 with CGPA 7.85. Currently, he is working as System Administrator of Dibrugarh University. He is in this position since December, 2008. He is in the charge of Computerization of Examination System and MIS of Dibrugarh University.

Prof. G.C. Hazarika
Date of birth : 01-01-1954
Academic Qualification: M.Sc. (Math.), Ph.D. (Math).
Positions held : Director i/c, Centre for Computer Studies, Dibrugarh University, and Professor, Department of Mathematics, Dibrugarh University

Academic Positions held:

a. Computer Programmer: Joined as Computer Programmer, Dibrugarh University Computer Centre in Dec, 1977 and served till April, 1985.
b. Lecturer: Joined as Lecturer in the Department of Mathematics, Dibrugarh University in April, 1985.
c. Reader: Joined as Reader in a regular post in June, 1990.
d. Professor: Joined as Professor in a regular post in August, 1998.

Publications (a few)

1.Magnatic effect on flow through circular tube of non-uniform cross section with permeable walls
- Applied Science Periodical Vol. V. No.1, February, 2003
Jointly with B.C. Bhuyan.

2.Influence of Magnetic filed on Separation of a Binary Fluid Mixture in Free Convection flow Considering Soret Effect
- J. Nat. Acad. Math. Vol. 20 (2006), pp. 1-20
Jointly with B.R. Sharma and R.N. Singh

3. Effects of Variable viscosity and Thermal Conductivity on flow and heat transfer of a Stretching Surface of a rotating micropolar fluid with suction and blowing
- Bull. Pure and Appl. Sc. – Vol.-25 E No. 2, PP-361-370, 2006.
Jointly with P.J. Borthakur.

4. Effects of Variable viscosity and Thermal Conductivity on boundary Layer flow and heat transfer of micropolar fluid near an axisyusmetric Stagnation point on a moving cylinder- Proc. 51st .cong. of ISTAM, Dec-2006.

Research Guidance:

Have guided 11 Ph. D students and 9 M Phil students.