Correlation Quality Measure for Microarray Tricluster

C.Sentamil Selvan
Assistant Professor, Department of Computer Science
Morappur Kongu College of Arts & Science, Morappur

OPEN ACCESS

Volume : 6

Special Issue : 1

Month : September

Year: 2018

ISSN: 2321-788X

Impact Factor: 3.025

Citation:
Selvan, C. S. (2018). Correlation Quality Measure for Microarray Tricluster. *Shanlax International Journal of Arts, Science and Humanities*, 6(S1), pp.29–40.

DOI:
https://doi.org/10.5281/zenodo.1410955

Abstract

Triclustering technique is suitable for the analysis of 3D microarray data that is genes gathered with several samples are taken at multiple time points. This paper dealt with an effective correlation measure is introduced to find the highly coherent tricluster which is used to identify the various gene groups under a set of conditions with several time points for microarray data. Mean Correlation Value (MCV) of a tricluster is defined as the mean of the correlation coefficient between the same subset of gene and same subset of sample/conditions across all time points in the 3D dataset. The performance of MCV is studied using synthetic dataset. From the result, the correlation measure mines the highly correlated tricluster for the microarray data.

Keywords: Triclustering, Correlation, Coherent Tricluster, Gene expression, Mean Correlation Value, Microarray data, Tricluster

Introduction

Data mining technique is used for analyses of microarray data very frequently. Clustering and biclustering techniques have mostly used the approach in analyses of microarray data. For 3D microarray data, clustering or biclustering techniques are fails to extract most biologically significant pattern because it considers any two dimensions like gene and sample or sample and time point. Owing to the rapid growth of modern technologies, 3D Microarray data are available for research easily. The extension of a 2D data mining technique is the triclustering technique which includes three dimensions data for analysis. 3D microarray datasets consist of groups the genes and samples with several time points. Triclustering algorithms are used to analyse the 3D microarray.The existing triclustering algorithm is to find the coherent cluster for the gene*sample matrix. It gives maximal cliques in this multigraph to yield the set of biclusters for this time slice. A new tricluster algorithm for finding the 3D clusters over GST microarray data [9]. A technique determines a set of first modules in each unordered pair of Gene-sample planes which then extended to the last triclusters. δ-TRIMAX algorithm groups of co-expressed genes from time series gene of expression data, or from any 3D gene expression dataset using a planar similarity measure (PMRS) [2]. TriGen algorithm extracts groups of genes with similar patterns in subsets of conditions and time points, and these groups have shown to related regarding their functional annotations extracted from the Gene Ontology[8]. A novel concept is to propose the fast filter method to identify the significant features, and redundancy among those features are considered without pairwise correlation analysis [5]. Most of the triclustering algorithms used three dimensions Mean Squared Residue (MSR) as a quality measure to evaluate or extract triclusters. But, 3D MSR is not efficient measure to extract coherent triclusters.

A new triclustering algorithm is devised in this paper using this correlation measure to extract highly correlated triclusters [6].

Contributions of this paper are as follows

- A new measure based on correlation for tricluster is proposed called Mean Correlation Value (MCV) measure.
- To develop a new triclustering algorithm over GST microarray data using MCV.

Finally, an empirical study is conducted on a synthetic dataset to validate the effectiveness of the proposed MCV measure. The organisation of this paper is as follows: Section II describes the Literature Review needed for the research work. Section III discusses the Methods and Materials required for the study. The proposed Triclustering Algorithm described in section IV. Section V concludes this research work with possible future enhancement.

Literature Review

Table 1: List of Related Works for Correlation Measure

S. No	Author	Technique	Description
1	Rudi Cilibrasi and Paul M.B. Vit´anyi[13]	Clustering	To determine a parameter-free, universal, similarity distance, the normalised compression distance or NCD, computed from the lengths of compressed data files.
2	Masfin sileshi, Bjorn gam back[12]	Clustering	To presents an equalisation of four clustering algorithms: k-means, average linkage and ward's method. This measures regarding cluster cohesiveness and semantic cohesiveness.
3	Abraham B.Korol[1]	Cluster Analysis	In cluster analysis, it performs monitoring the expression of ten thousand gene simultaneously.
4	Gengxin chen, Saied A.Jaradat, Nila Banerjiee, Tetsuya S.Tanaka, Minoru S.H.Ko, Michael Q.Zhang[6]	Cluster	Most of the clustering algorithms used to analyse microarray gene expression data. Given embryonic stem cell of gene expression data to evaluate the performance of clustering algorithm.
5	Tulika kakati, Hasin A.Ahmed, Dhruba K.Bhattacharyya[15]	Bicluster	Biclustering or simultaneous clustering aims to cluster of the gene*sample dataset into groups of genes co-ex-pressed across a subset of conditions.

6	D.Gutierrez-aviles, C.Rubio Escudero, F.Martinez Alvarez, J.C Riruelme[4]	Bicluster	To present the trigen algorithm, a genetic algorithm that finds triclusters of gene expression that take the conditions and the time points simultaneously.
7	J.Bagyamani, K.Thangavel, R.Rathipriya[9]	Biclustering	The proposed query based algorithm SIMBIC+ first identifies a functionality rich query gene in microarray data.
8	Stefan Gremalschi, Gulsah Altun, Irina Astrovkaya, Alexander Zelikovsky[14]	Biclustering	A greedy deletion addition algorithm to find a given number of K-biclusters, whose mean squared residues(MSR) are below certain thresholds and missing values in matrix replaced with random numbers.
9	Boris.G.Mirkin[3]	Biclustering and Triclustering	A disjunctive model of box bicluster and tricluster analysis is considered. The least squares locally optimal one cluster method is proposed oriented towards analysis of binary data.
10	P.Mahanta, H.A Ahmed[11]	Triclustering	Mining microarray datasets is important in bioinformatics research and the biomedical application recently is used to tricluster or three dimensions of clusters in a gene

Table 1 show that the various data mining techniques to be performed the correlation-based clustering and biclustering techniques. Clustering and biclustering are the most widely used methods for mining microarray data even for the 3D dataset. Recently a very few triclustering techniques are used to mine the microarray data. Most of the triclustering algorithms extract the clustering of genes with related sample (gene*sample), and the various genes gathered at different time points (gene*time). Triclustering result is efficient when all three dimensions to mine tricluster are considered for analysis.

Methods and Materials

3D Gene Expression Data (GST)

Microarray technology can generate data called gene-sample-time microarray data (GST) or three-dimensional microarray data. It contains the expression levels of a group of genes under a set of samples/conditions during a series of time points. Three-dimensional (3D) Microarray dataset is a dataset contains three types of variables (gene, sample, and time point). In general, each cell mijk in a 3D dataset represents the value of ith row under jth column at kth time space. It can also be viewed as a two-dimensional matrix, such that each cell mi,j contains the time series on ith row under jth column[4].

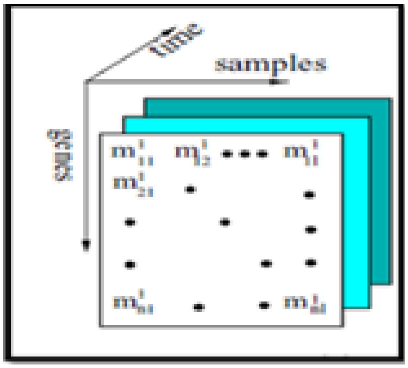


Figure 1: Representation of 3D Gene Expression Data (GST)

Triclustering

Triclustering is simultaneously clustering of rows and columns at the different time point of the three-dimensional dataset. A Tricluster TC is a submatrix of 3d dataset TC = r x c x t = {TCijk}, where r R, c C, and t T provided certain homogeneity condition is satisfied. It mines the maximal 3D clusters (or triclusters) satisfying the following correlation criterion as defined in equation 1: any 2 × 2 submatrices of a tricluster must obey a constant multiplicative or additive or coherent pattern/relationship.

Three Dimensions MSR

3D of MSR that measures the homogeneity of triclusters which contain subgroups of genes, conditions, and time points. This measure is said to be MSR3D. The formal definition can seen in [4].

$$MSR_{3D}(TC) = \frac{\sum_{g\in G, c\in C, t\in T} r^2 gct}{\#G * \#C * \#T}$$

Where r_{gct} can defined as

$$r_{gct} = TC_V(g,c,t) + M_{CT}(g) + M_{GT}(c) + M_{GC}(T) - M_G(c,t) - M_C(g,t) - M_T(g,c) - M_{GCT} \qquad (2)$$

Each of the members of (1) and (2) defined as follows:

TC: tricluster,

G: TC’s gene subset

C: TC’s condition subset,

T: TC’s condition subset,

#G: number of genes,

#C: number of conditions/samples,

#T: number of time points,

TCV(g,c,t): expression level of gene g under condition c at time t in TC,

MCT(g): mean of all conditions at all times for a gene g in TC

MGT(C): mean of all genes at all times for a condition c in TC

MGC(t): mean of all genes under all conditions at time t in TC,

MG(c,t): mean of the values of a condition c and a time t under all genes in TC,
MC(g,t): mean of the values of a gene g and a time t under all conditions in TC,
MT(g,c): mean of the values of a gene g and a condition c under all times in TC,
MGCT: mean value of all values in TC.

Most of the work in the literature used MSR3D as a homogeneity measure to evaluate the quality of tricluster.

Mean Correlation Value for Tricluster

$$\sum_{m}\sum_{n}(A_{mn}-\bar{A})(B_{mn}-\bar{B}) \;/\; \sqrt{\left(\sum_{m}\sum_{n}(A_{mn}-\bar{A})^2\right)\left(\sum_{m}\sum_{n}(B_{mn}-\bar{B})^2\right)} \qquad (3)$$

Where $\bar{A}=\frac{\sum_m\sum_n(A_{mn})}{m*n}$, $\bar{B}=\frac{\sum_m\sum_n(B_{mn})}{m*n}$

The range of MCV is [0 ,1]. A value close to ‘1’ signifies high correlated/coherent tricluster otherwise low or null correlated tricluster.

Various Kinds of Tricluster

Triclusters have different patterns. They are:

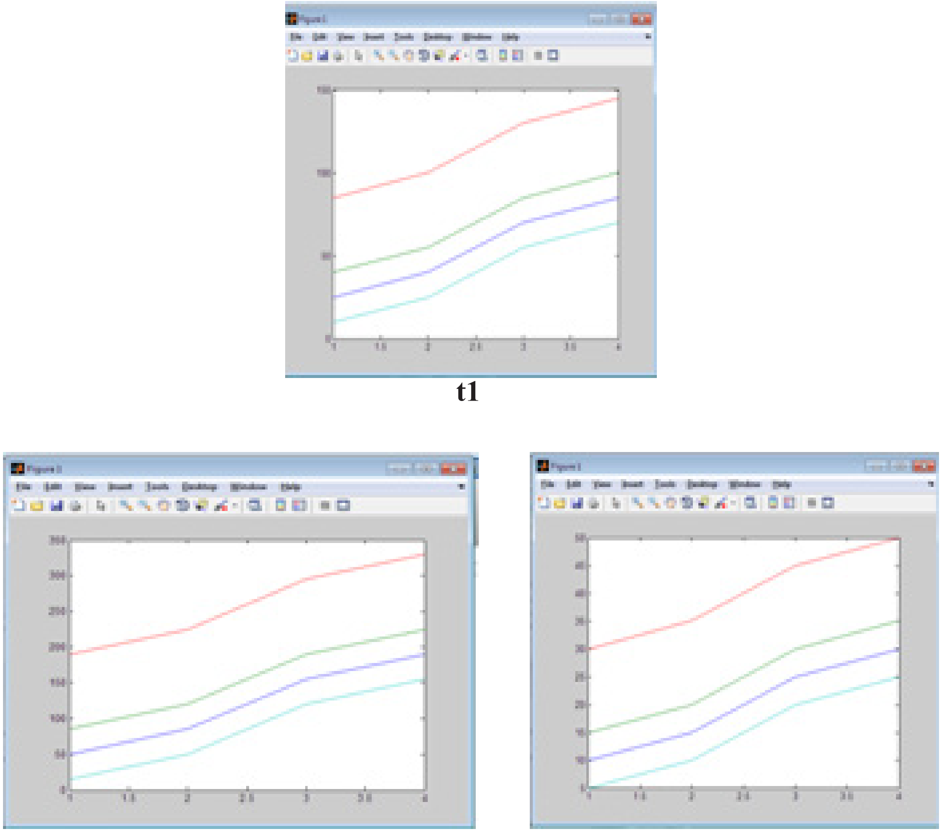
- Additive Tricluster: Tricluster with Additive pattern
- Multiplicative Tricluster : Tricluster with Multiplicative Pattern
- Coherent Tricluster: Tricluster with Coherent Pattern
- Coherent Evolution Tricluster: Tricluster with Coherent Evolution Pattern

The sample triclusters with graphical representation is shown in the following figure:
Additive Tricluster:

g/s	s1	s2	s3	s4
g1	6	7	10	5
g2	7	8	11	6
g3	9	10	13	8
g4	10	11	14	9

g/s	s1	s2	s3	s4
g1	61	62	65	60
g2	62	63	66	61
g3	64	65	68	63
g4	65	66	69	64

g/s	s1	s2	s3	s4
g1	112	112	115	110
g2	112	113	116	111
g3	114	115	118	113
g4	115	116	119	114



t1

t2 **t3**

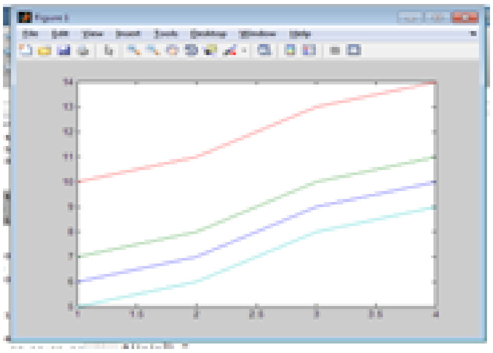
Figure 2 Graphical Representation of Additive Tricluster

Multiplicative Tricluster

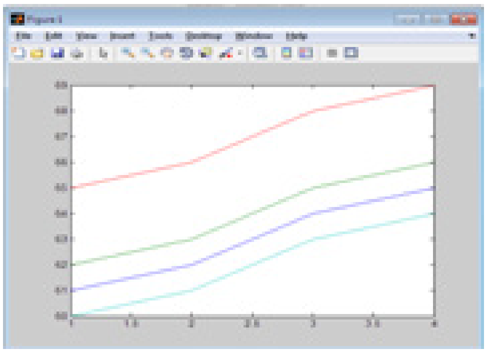
g/s	s1	s2	s3	s4
g1	10	15	30	5
g2	15	20	35	10
g3	25	30	45	20
g4	30	35	50	25

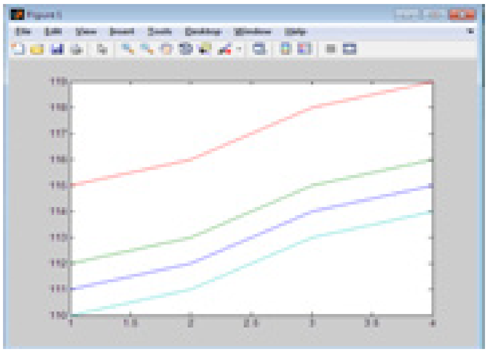
g/s	s1	s2	s3	s4
g1	25	40	85	10
g2	40	55	100	25
g3	70	85	130	55
g4	85	100	145	70

g/s	s1	s2	s3	s4
g1	50	85	190	15
g2	85	120	225	50
g3	155	190	295	120
g4	190	225	330	155



t1





t2 **t3**

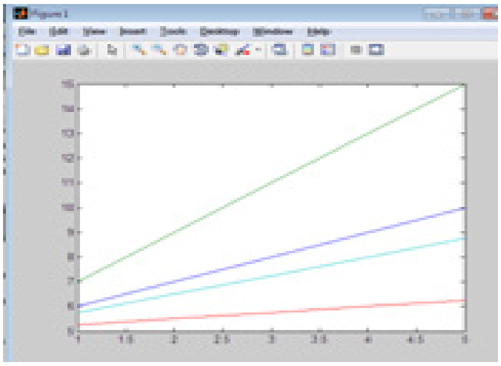
Figure 3 Graphical Representation of Multiplicative Tricluster

Coherent Tricluster

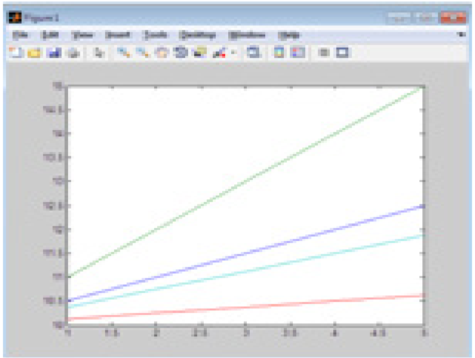
g/s	s1	s2	s3	s4
g1	365	80	110	65
g2	260	215	260	190
g3	215	115	150	90
g4	465	90	115	75

g/s	s1	s2	s3	s4
g1	33	14	16	13
g2	26	23	26	22
g3	23	17	19	15
g4	40	15	17	14

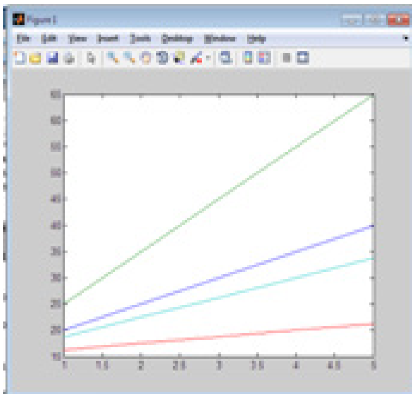
g/s	s1	s2	s3	s4
g1	75	18	24	15
g2	54	45	54	40
g3	45	25	32	20
g4	95	20	25	17



t1



t2



t3

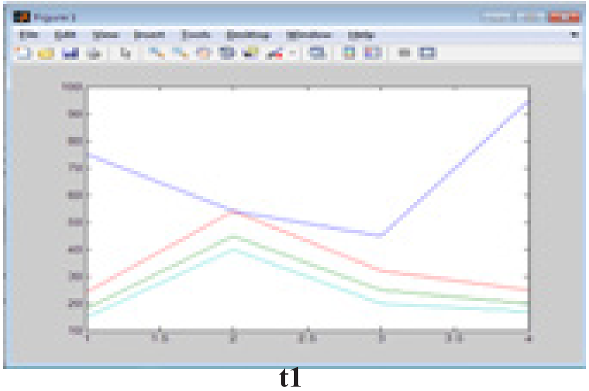
Figure 4 Graphical Representation of Coherent Tricluster

Coherent Evolution Tricluster

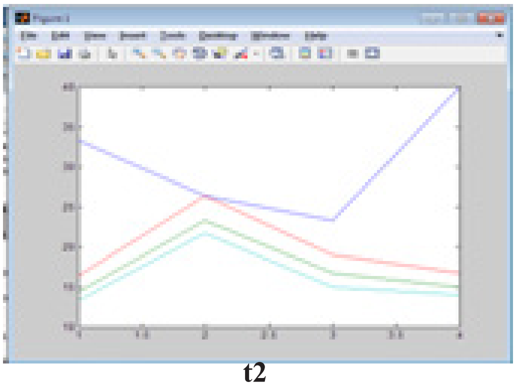
g/s	s1	s2	s3	s4
g1	365	80	110	65
g2	260	215	260	190
g3	215	115	150	90
g4	465	90	115	75

g/s	s1	s2	s3	s4
g1	33	14	16	13
g2	26	23	26	22
g3	23	17	19	15
g4	40	15	17	14

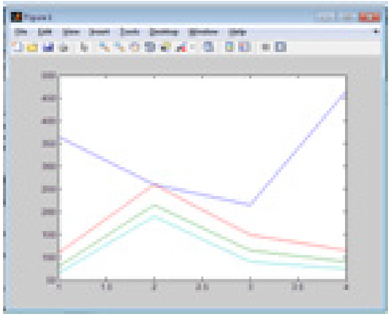
g/s	s1	s2	s3	s4
g1	75	18	24	15
g2	54	45	54	40
g3	45	25	32	20
g4	95	20	25	17



t1



t2



t3

Figure 5 Graphical Representation of Coherent Evolution Tricluster

Table 2 tabulates the quality of tricluster using MSR3D and MCV. It clearly is seen that MSR3D has very high value than MCV. Table 2: Performance of MSR3D Vs MCV

Tricluster	MSR3D	MCV
Additive	1.5097e+07	1
Multiplicative	8.9106e+07	1
Coherent	1.1300e+05	0.9807
Coherent Evolution	1.2995e+08	0.9996

Tricluster algorithm using MCV

The proposed algorithm is developed using Mean Correlation Value for Tricluster (as given in equation 3). This algorithm developed in Matlab Toolbox R2013. In figure 6, the pseudocode for identifying the MCV for tricluster given.

Correaltion_tri(Data)

```
[m,n,z]=size(data) // Data is 3D gene expression dataset
 for k=1:z
     for j=1:z
         corr_mat(k,j)=corr2(data(:,:,k),data(:,:,j))
     end
 end
 MCV=abs(sum(sum(corr_mat))-z)/(z^2-z);// Mean Correlation Value of a Tricluster
```

Figure 6 Pseudocode for Tricluster MCV

Result Discussion

Table 3 Characteristics of Synthetic Triclusters

Tricluster ID	No. of Genes	No. of Samples	No. of ime Point	MCV
Tricluster 1	7744	9	5	0.9122
Tricluster 2	7744	9	5	0.9033
Tricluster 3	7744	9	5	0.8996
Tricluster 4	7744	9	5	0.9028
Tricluster 5	7744	9	5	0.9085

Table 3 shows the correlation analysis for the 5 synthetic triclusters obtained. The correlation coefficients are very high, in most cases, the values are close to one. This indicates almost perfect homogeneity between the genes, samples and times points of the tricluster.

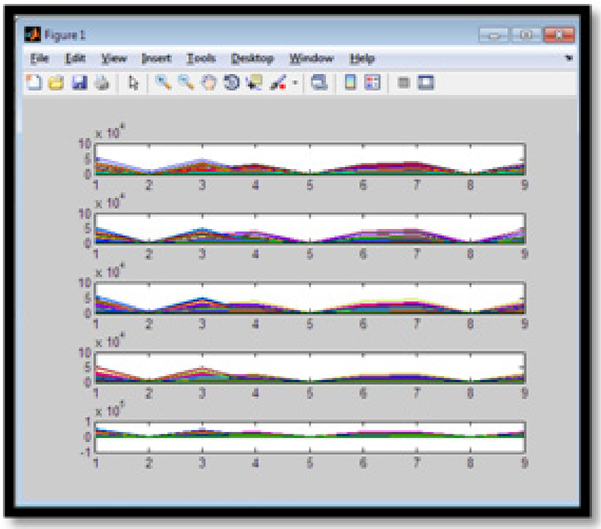


Figure 7 Graphical Representation of Synthetic Tricluster

Figure 7 shows the graphical representation of 5 Tricluster with 7744 genes 9 samples and 5 time points.

Conclusion

This paper introduced an efficient correlation quality based measure called Mean Correlation Value (MCV). The MCV evaluates the quality of all types of triclusters (i.e. triclusters with coherent pattern) well since it can tolerate transformations like translation and scaling. The MCV of highly correlated or coherent tricluster is nearly equal to one. Therefore, from the study, it is came to known that MSR is not a perfect measure to discover coherent patterns in data when the variance of expression values is high. Hence in this work, correlation based quality measure for tricluster is proposed and studied with the synthetic data. In future, triclustering algorithm using MCV will be developed to study its biological importance in microarray data analysis.

References

Abraham B.Korol “Microarray cluster analysis and applications”, Institute of Evolution, University of Haifa Jan 2003.

Anirban Bhar, Martin Haubrock, Anirban Mukhopadhya, Edgar Wingender, “Multiobjective triclustering of time-series transcriptome data reveals key genes of biological processes”, Bhar et al. BMC Bioinformatics (2015)

Boris.G.Mirkin and Andrey, “Approximate Bicluster and Tricluster Boxes in the Analysis of Binary Data”, Springer-Verlag Berlin Heidelberg 2011.

D. Gutiérrez-Avilés, C.Rubio-Escudero,n, F.Martínez-Álvarez, J.C.Riquelme, “TriGen: A genetic algorithm to mine triclusters in temporal gene expression data”, & 2013 ElsevierB.V.All rights reserved

D.Gutierrez Aviles, C.Rubio Escudero, F.Martinez Alvarez, J.C Riquelme “TriGen: A genetic algorithm to mine triclusters in temporal gene expression data” &2013Elsevier B.V.All rights reserved.

Duygu Dede, Hasan Ogul, “A three-way clustering approach to Cross-Species Gene Regulation Analysis”, IEEE 2003.

Gengxin chen, Saied A.Jaradet, Nila Banerjee, Tetsuya S.Tanaka, Minoru S.H.Ko, Michael Q.Zhang, “Evaluation and Comparison of Clustering Algorithms in Analyzing ES Cell Gene Expression Data” , Laboratories of Genetics, National Institute on Aging, National Institutes of Health, Baltimore, MD 21224, USA.

H. A. Ahmed, P. Mahanta, D. K. Bhattacharyya, J. K. Kalita, A. Ghosh “Intersected Coexpressed Subcube Miner: An effective triclustering algorithm” research project supported by DST, Govt. of India in collaboration with ISI, Kolkata.

Haoliang Jiang, Shuigeng Zhou, Jihong Guan, and Ying Zheng “gTRICLUSTER: A More General and Effective 3D Clustering Algorithm for Gene-Sample-Time Microarray Data” J. Li et al. (Eds.): BioDM 2006, LNBI 3916, pp. 48–59, 2006 c Springer-Verlag Berlin Heidelberg 2006.6

J.Bagyamani, K.Thangavelu, R.Rathipriya, “Biological Significance of Gene Expression Data using Similarity-based Biclustering Algorithm”, International Journal of Biometrics and Bioinformatics (IJBB), Volume (4): Issue (6).

Lei Yu, Huan Liu, “Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution”, Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.8

Mahanta, H.A.Ahmed, D.K Bhattacharyya, Jugal K.Kalita, “Triclustering in Gene Expression Data Analysis” 2011 IEEE.

Masfin Sileshi, Bjorn Gamback, “Evaluating Clustering Algorithms: Cluster Quality and Feature Selection in Content-Based Image Clustering”, Computer Science and Information Engineering, 2009 WRI World Congress

Rudi Cilibrasi and Paul M.B. Vitanyi ” Clustering by Compression” IEEE Transactions On Information Theory, Vol. 51, No 4, April 2005, 1523–1545

Stefan Gremalschia, Gulsah Altunb, Irina Astrovskayaa, and Alexander Zelikovskya, “Mean Square Residue Biclustering with Missing Data and Row Inversions” , Georgia State University, Atlanta, GA 30303.