ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

Triclustering Algorithm for 3D Gene Expression Data using Correlation Measure

[1] Narmadha . N, [2] Rathipriya . R
[1] PhD Research Scholar ,[2] Assistant Professor
[1] Department of Computer Science, Periyar University, Salem, Tamilnadu, India.

***Abstract* - In gene expression data analysis there are various biological data mining methods have been proposed. Among them, Co-clustering or Biclustering is a common method to extract the gene groups that behave similarly/coherently under a subset of experimental conditions. Due to increasing the dimensionality of gene expression data, the three-way clustering method called triclustering is developed recently for three-dimensional gene expression analysis. It is used to mine the coherent cluster named tricluster in three-dimensional gene expression datasets. This paper introduces a novel correlation measure to find the correlation between the genes, condition and time point in 3D gene expression dataset. The proposed method has applied to time series data obtained from yeast cell cycle analysis. From the result, it is evident that proposed algorithm mines the highly correlated 3D cluster or tricluster from gene expression data.**

Keywords: Triclustering, Correlation, Gene expression, Yeast cell cycle, Bioinformatics

I. INTRODUCTION

Analysis of Gene expression data is a vital tool for identifying useful and hidden knowledge from raw data. This analysis is helpful in following: identification of diseases causing factors, examining for proper treatments and discovering hidden similarities among different species[4]. Gene expression data denoted as matrices of expression levels of genes under specific experimental conditions (environments, tissues, etc.). Data mining techniques can be used to analyse the gene expression data. In which, Clustering is the fundamental technique to analyse gene expression data. It can be used to group only genes or conditions or time point of three-dimensional (3D) Gene expression data. However, clustering algorithms fail to discover the relationship among the gene, samples and time point. The extension of clustering is called Biclustering which performs clustering in both dimensions (genes vs. conditions or genes vs. time point or sample vs. time point) simultaneously. Therefore, Biclustering is defined as a "simultaneous clustering of both rows and columns of two- dimensional data. Biclustering aims to find the set submatrices (biclusters) either from genes vs. conditions or genes vs. time point.

Due to the rapid growth of technologies, 3D gene expression data is available easily. For 3D data, clustering or biclustering techniques are fails to extract most biologically significant information. Therefore, clustering of 3D data is called Triclustering is introduced and explained well in the literature. The objective of this research work is to propose a novel correlation measure to identify the correlated genes, samples and time point from 3D gene expression data. A new triclustering algorithm is devised using this correlation measure to extract highly correlated triclusters [4].

Major contributions of this paper are as follows:

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

Finally, empirical study is conducted on a Yeast cell cycle dataset to validate the effectiveness of the proposed algorithm. This paper is organized 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 is described in section IV. Section V concludes the work with possible future enhancement.

II. LITERATURE REVIEW

Most of the work focused on MSR homogeneity for extracting and evaluating tricluster. But it fails to extract the highly correlated tricluster of genes, samples and time points. Because of high magnitude homogeneity criterion value there for in this a novel mean correlation value

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

tricluster is introduced to extract highly coherent from the tricluster from the 3D gene expression data.

Table I : List of Related Works for Triclustering Algorithm

S. No	Author	Technique	Description
1	T. R. Golub,D. K. Slonim P. Tamayo[11]	Clustering	The feasibility of cancer classification based solely on gene expression monitoring and suggests a general strategy for discovering and predicting cancer classes for other types of cancer, independent of previous biological knowledge.
2	K. Y. Yeung1, C. Fraley A. Murua,A. E. Raftery W. L. Ruzzo[7]	Clustering	On real expression data, the model-based approach produced clusters of quality comparable to a leading heuristic clustering algorithm, but with the key advantage of suggesting the number of clusters and an appropriate model
3	Xin Xu, Ying Lu, Anthony K. H. Tung, WeiWang [12]	Biclustering	A novel pattern based biclustering algorithm for identifying shifting and scaling co-regulation patterns, satisfying both coherence constraint and regulation constraint
4	T. George ,S. Merugu [10]	Biclustering	Current collaborative filtering (CF) techniques that involves simultaneous clustering of users and items to build an efficient real-time CF framework
5	Lizhuang Zhao, Mohammed J. Zaki [9]	Triclustering	Proposed tricluster algorithm coherent cluster for gene*sample matrix. It gives maximal cliques in this multigraph to yield the set of biclusters for this time slice.
6	Haoliang Jiang, Shuigeng Zhou1, Jihong Guan, and Ying Zheng [6]	Triclustering	A new tricluster algorithm for finding the 3D clusters over GST microarray data.
7	H. A. Ahmed, P. Mahanta, D. K. Bhattacharyya, J. K. Kalita, A. Ghosh[5]	Triclustering	A technique finds a set of initial modules in each unordered pair of Gene-sample planes which then extended to the final triclusters. To propose a planar similarity measure called PMRS to extend the initial modules to the final triclusters.
8	Anirban Bhar, Martin Haubrock, Anirban Mukhopadhya, Edgar Wingender[1]	Triclustering	δ-TRIMAX algorithm groups of co-expressed genes from time series gene expression data, or from any 3D gene expression dataset.
9	D. Gutiérrez-Avilés , C.Rubio-Escudero ,n, F.Martínez-Álvarez, J.C.Riquelme[2]	Triclustering	TriGen algorithm extracting groups of genes with similar patterns in subsets of conditions and times, and these groups have shown to related regarding their functional annotations extracted from the Gene Ontology.
10	Lei Yu, Huan Liu[8]	Triclustering	A novel concept, predominant correlation, and propose a fast filter method which can identify relevant features as well as redundancy among relevant features without pairwise correlation analysis

III.METHODS AND MATERIALS

A. 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 3 types of variables (gene, sample, and time point). In general, each cell mijk in a 3D

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

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 with respect to ith row under jth column[4].

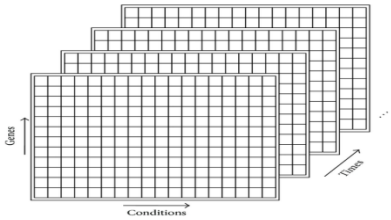

Fig 1 :Representation of 3D Gene Expression Data (GST)

B. Triclustering

Triclustering is simultaneously clustering of rows and columns with several time point of 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 homogeneity criterion as defined in equation 1: any 2 × 2 submatrix of a tricluster must obey a constant multiplicative or additive or coherent pattern / relationship.

C. Three dimensions MSR

Three dimensions 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 be seen in [4].

$$MSR_{3D}(TC) = \frac{\sum_{g\in G, c\in C, t\in T} r^2 gct}{\#G*\#C*\#T} \quad (1)$$

Where rgct can be defined as

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

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

TC: tricluster being evaluated,
G: subset of genes from TC,
C: subset of conditions from TC,
T: subset of times from TC,
#G: number of genes in TC,
#C: number of conditions in TC,
#T: number of times in TC,
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 the MSR3D as a homogeneity measure to evaluate the quality of tricluster.

D. Mean Correlation Value for Tricluster

Mean Correlation Value of a Tricluster (MCV) is defined as:

$$\sum_m \sum_n (A_{mn}-\bar{A})(B_{mn}-\bar{B}) \Big/ \sqrt{\left(\sum_m \sum_n (A_{mn}-\bar{A})^2\right)\left(\sum_m \sum_n (B_{mn}-\bar{B})^2\right)}$$

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

The range of MCV is [0 ,1]. Value close to '1' signifies high correlated tricluster otherwise low or null correlated tricluster.

E. Types of Tricluster

Triclusters have different patterns. They are:
They are:

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

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

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

1	2	5	0
2	3	6	1
4	5	8	3
5	6	9	4

101	102	105	100
102	103	106	101
104	105	108	103
105	106	109	104

51	52	55	50
52	53	56	51
54	55	58	53
55	56	59	54

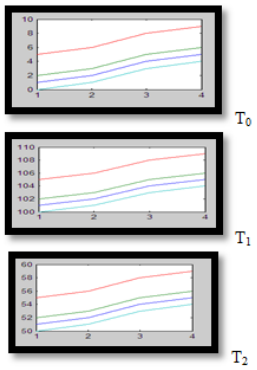

Fig. 2. Graphical Representation of Additive Tricluster

5	10	25	0
10	15	30	5
20	25	40	15
25	30	45	20

15	30	75	0
30	45	90	15
60	75	120	45
75	90	135	60

35	70	175	0
70	105	210	35
140	175	280	105
175	210	315	140

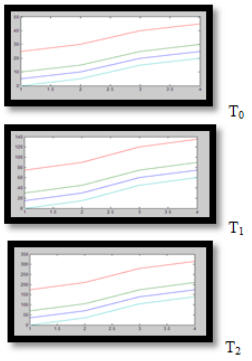

Fig. 3. Graphical Representation of Multiplicaivet Tricluster

Coherent Tricluster:

1.00	2.00	0.25	0.75
2.00	4.00	0.50	1.50
3.00	6.00	0.75	2.25
4.00	8.00	1.00	3.00
5.00	10.00	1.25	3.75

0.50	1.00	0.13	0.38
1.00	2.00	0.25	0.75
1.50	3.00	0.38	1.13
2.00	4.00	0.50	1.50
2.50	5.00	0.63	1.88

5.00	10.0	1.25	3.75
10.0	20.0	2.50	7.50
15.0	30.0	3.75	11.25
20.0	40.0	5.00	15.00
25.0	50.0	6.25	18.75

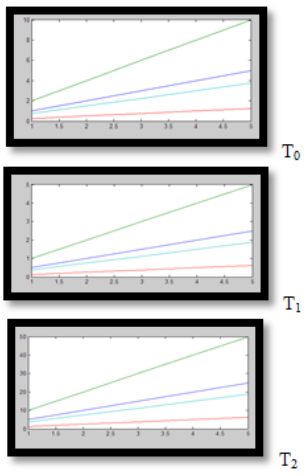

Fig. 4. Graphical Representation of Coherent Tricluster

Coherent Evolution

350	65	95	50
245	200	245	175
200	100	135	75
450	75	100	60

23.33	04.33	06.33	03.33
16.33	13.33	16.33	11.66
13.33	06.66	09.00	05.00
30.00	05.00	06.67	04.00

70	13	19	10
49	40	49	35
40	20	27	15
90	15	20	12

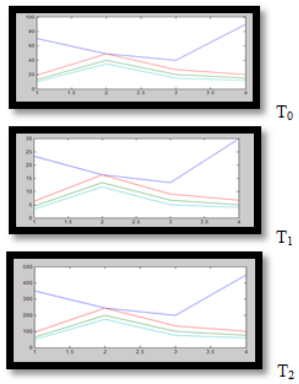

Fig. 5. Graphical Representation of Coherent Evolution Tricluster

TABLE II. PERFORMANCE OF MSR3D VS MCV

Tricluster	Additive	Multiplicative	Coherent	Coherent Evolution
MSR_{3D}	5.0759 e^{+04}	6.1162 e^{+07}	9.3726 e^{+03}	9.8709 e^{+07}
MCV	1	1	1	1

The correlation based measure Mean Correlation Value (MCV) evaluates well all types of triclusters since it can tolerate transformations like translation and scaling. The MCV of these triclusters is nearly one. Therefore MSR is not a perfect measure to discover patterns in data when the variance of hit count values is high for all the patterns. Hence in this work, correlation based merits function has been used in the entire triclustering model.

IV. TRICLUSTER ALGORITHM USING MCV

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

The pseudocode for identifying the bicluster with maximum Average Correlation Value [12] for each time point is shown in Figure 6(b). Tricor_Clust is the correlation based Triclustering for 3D gene expression data. It mines the tricluster with specified correlation

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

threshold from the given 3D dataset which shown in figure 6(c). It is efficient in identifying the triclusters with additive pattern, multiplicative pattern, coherent pattern and coherent evolution pattern well.

```
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
```

Fig : 6 (a): Pseudocode for Tricluster MCV

```
Function [pop,acvarr,R,C]=MaxACVpop(aa,popsize,flag)

// aa – 3D gene expression data of 'ng' genes, 'ns' samples and 'nt' time point
//popsize- number of individuals (ie bicluster)
// flag -0 or 1
//R- Row set for each individuals(ie bicluster)
//C- Column set for each individuals(ie bicluster)

[r,c]=size(aa);

if flag==1 % Random population creation
pop=randi([0 1],popsize,r+c);

else % population created using kmeans

rk=popsize; // number of gene cluster
ck=2; // number of column cluster

pop=zeros(rk*ck, r+c);

rid=kmeans(aa,rk, 'replicates', 10); // row clustering using kmeans
cid=kmeans(aa',ck, 'replicates', 10); // column clustering using kmeans

for i=1:rk
Rbic{i}=find(rid==i);
end

for i=1:ck
Cbic{i}=find(cid==i);
end
it=1;

for i =1:rk
 for j=1:ck

cp=[Rbic{i};r+Cbic{j}];
pop(it,cp)=1;
end
it=it+1;
end

[acvarr,R,C,vol]=gaacv1(pop,aa);

end
```

Fig 6. (b): Pseudocode for Initial Population Generation Tricorr_Clust(data,popsize)

```
Tricorr_Clust(data,popsize)

// Data – 3D Gene expression Data
//popsize- Number of individuals

[r,c,d]=size(data);// r- no. of genes, c- no. of samples, d- no. of time points
popsize=10;

flag=2; // using maximum ACV bicluster as a population

newpop=zeros(popsize, r+c); // initial population

for i=1:d // identifying the maximum ACV Bicluster

[pop,acvarr,R,C]=popcreation(data(:,:,i),popsize,flag);

[maxacv, maxacvid]=max(acvarr);
newpop(i,:)=pop(maxacvid,:);
// Assignment of ACV, Gene id , Sample id of Max. ACV bicluster into a variable respectively
newpopacv(i)=maxacv;
geneset{i}=R{maxacvid};
sampleset{i}=C{maxacvid};
end

// Extraction of Tricluster using initialized population
Count =10 // No. of Tricluster

cor-threshold=0.95 // correlation threshold

it=1 // iterator

while (it<=10) do
tset=randi([0 1],d,d); // random generation time point set

MCV_tri(it)=correlation-tri(A(geneset{t},Sample set{t},tset)

If MCV_tri(it)>0.95
Set_tricluster{it}=A(gene{t},Sample set{t},tset{t}); // Tricluster set
Set_triMCV{it}=MCV_tri// Tricluster's MCV
it=it+1
end (if)

end (while)
// Return Set_tricluster and Set_triMCV
```

Fig 6 (c): Pseudocode for Correlation Based Triclustering Algorithm

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

V. RESULT DISCUSSION

Dataset Description:

The yeast cell cycle analysis project's goal is to identify all genes whose mRNA levels regulated by the cell cycle. The resources used are public and available in http://genome-www.stanford.edu/cellcycle/. Here to find information about gene expression values obtained from different experiments using microarrays. In particular, it created a dataset elu from the elutriation experiment with 7744 genes, 13 experimental conditions, and 14-time points. The three dimensions of elu data are used to find the correlation of taking measures from 0 to 390 minutes.

Characteristics of Triclusters

Tricluster ID	No. of Genes	No. of Samples	No. of Time Points	MCV
Tricluster 1	12	9	6	0.9364
Tricluster 2	9	9	8	0.9562
Tricluster 3	11	9	6	0.9604
Tricluster 4	10	9	6	0.955
Tricluster 5	11	9	6	0.9445
Tricluster 6	12	9	4	0.9364
Tricluster 7	12	9	5	0.9364
Tricluster 8	13	9	6	0.9363
Tricluster 9	12	9	8	0.9364
Tricluster 10	12	9	8	0.9364

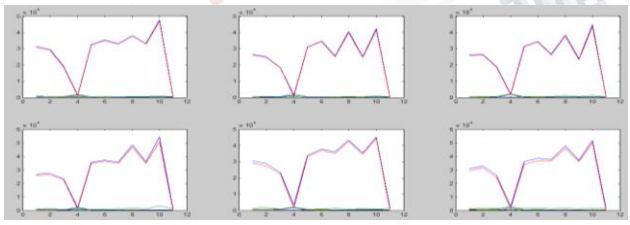

Fig 7. Graphical Representation of Tricluster found in Yeast Cell Cycle Data.

Table 3 shows the correlation analysis for the 10 triclusters obtained. The correlation coefficients are very high, in most cases the values is close to one. This indicates almost perfect homogeneity between the genes, samples and times points of the tricluster.

Figure 6 shows the graphical representation of Tricluster 3 with genes 9 samples and 6 time points.

CONCLUSION

A novel correlation measure for tricluster is introduced in this paper. This work dealt with the correlation based triclustering algorithm method to identify the highly correlated 3D cluster of genes,conditions and time points from 3D microarray expression data. Empirical study was conducted to test the performance of proposed triclustering algorithm using yeast cell cycle analysis dataset. It has been observed from the result, the proposed algorithm shows better performance in extracting highly correlated tricluster. In future this algorithm will be tuned based on biological significance.

REFERENCE

[1] 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)

[2] 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

[3] Daxin Jiangy Jian Peiyz Murali Ramanathany Chun Tangy Aidong Zhangy "Mining Coherent Gene Clusters from Gene Sample Time Microarray Data" This research is partly supported by NSF grants DBI-0234895, IIS-0308001 and NIH grant 1 P20 GM067650-01A1.

[4] Duygu Dede, Hasan Ogul, "A three-way clustering approach to Cross-Species gene Regulation Analysis", IEEE 2003.

[5] 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.

[6] 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.

[7] K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery ,W. L. Ruzzo, "Model-based clustering and data transformations for gene expression data" , Oxford Journals 2001, Volume 17, Issue 10

[8] 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

[9] Lizhuang Zhao, Mohammed J. Zaki, "TriCluster: An Effective Algorithm for Mining Coherent Clusters in 3D Microarray Data", SIGMOD 2005 June 14-16, 2005, Baltimore, Maryland, USA.

[10] T. George ; S. Merugu , "A scalable collaborative filtering framework based on co-clustering, Data Mining", Fifth IEEE International Conference

[11] T. R. Golub, D. K. Slonim, P. Tamayo, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring", Science Vol 286, Issue 5439 15 October 1999.

[12] Teng L, Chan L: Discovering biclusters by iteratively sorting with weighted correlation coefficient in gene expression data. J Signal Process Syst 2008, 50(3):267-280.

[13] Xin Xu, Ying Lu, Anthony K. H. Tung, WeiWang, "Mining Shifting-and-Scaling Co-Regulation Patterns on Gene", Data Engineering, 2006. ICDE '06, Proceedings of the 22nd International Conference.