Triclustering in Gene Expression Data Analysis: A Selected Survey

P. Mahanta, H. A. Ahmed
Dept of Comp Sc and Engg
Tezpur University
Napaam -784028, India
Email: priyakshi@tezu.ernet.in, hasin@tezu.ernet.in

D. K. Bhattacharyya
Dept of Comp Sc and Engg
Tezpur University
Napaam -784028, India
Email: dkb@tezu.ernet.in

Jugal K. Kalita
Dept. of Computer Science
University of Colorado
Colorado Springs, USA
Email: kalita@eas.uccs.edu

Abstract—Mining microarray data sets is important in bioin- formatics research and biomedical applications. Recently, mining triclusters or 3D clusters in a Gene Sample Time or 3D microarray data is an emerging area of research. Each tricluster contains a subset of genes and a subset of samples such that the genes are coherent on the samples along the time series. There is a scarcity of triclustering algorithms in the literature of microarray data analysis. We review some existing triclustering algorithms and discuss their merits and demerits. Finally we are trying to provide the researcher who are new to this field a base platform by exposing the issues which are still challenging in triclustering through our analysis of these algorithms.

Index Terms—triclustering, TRICLUSTER, gTRICLUSTER, GST data

I. Introduction

With advances in DNA microarray technology, expression levels of thousands of genes can be simultaneously measured efficiently during important biological processes while the traditional approach to genomic research focuse on the local examination and collection of data on single genes. The two major types of microarrays are the cDNA microarray and oligonucleotide arrays. Though the two types differ in the details of their experiment protocols, both involve three common basic procedures [1]: (i) Chip manufacture (ii) Target preparation, labeling and hybridization (iii) Scanning

Analyzing microarray data to identify localized co-expressed gene patterns is a new focus of researchers. Clustering techniques have proven to be useful in understanding gene function, gene regulation, subtypes of cells and cellular processes in gene regulation networks. Co-expressed genes can be clustered together with similar cellular functions. Cluster analysis of gene expression data can be done in three ways: (i) Gene based clustering (ii) Sample based clustering and (iii) Subspace clustering. Mining three-dimensional (3D) clusters in gene sample time (GST) microarray data is emerging as an emerging research topic. A tricluster consists of a subset of genes that are coherent on a subset of samples along a segment of time series. This kind of coherent clusters may contain information that may help users to identify useful phenotypes, potential genes related to these phenotypes and their expression rules.

A. Existing clustering approaches for gene expression data

Clustering is the process of grouping data objects into a set of disjoint classes, called clusters, so that objects within a class have high similarity to each other, while objects in separate classes are highly dissimilar. Cluster analysis of gene expression data can be done in three ways:

1) Gene based clustering
   In gene based clustering, the genes are treated as the objects, while the samples are the features.
2) Sample based clustering
   In sample based clustering, the samples are treated as the objects, while the genes are the features. Here, the samples are partitioned into homogeneous groups. Each group may correspond to some particular macroscopic phenotype, such as clinical syndromes or cancer types. Both gene-based and sample-based clustering approaches search for exclusive and exhaustive partitions of objects that share the same feature space.
3) Subspace clustering
   The current thinking in molecular biology holds that only a small subset of genes participate in any cellular process of interest and that a cellular process takes place only in a subset of the samples. This belief calls for subspace clustering to capture clusters formed by a subset of genes across a subset of samples. For subspace clustering algorithms, genes and samples are treated symmetrically, so that either genes or samples can be regarded as objects or features. Furthermore, clusters generated through such algorithms may have different feature spaces.
   One faces several challenges in mining microarray data using a subspace clustering technique (a) Subspace clustering is known to be an NP-hard problem and therefore many proposed algorithms for mining subspace clusters use heuristic methods or probabilistic approximations. This decrease the accuracy of the clustering results. (b) Due to varying experimental conditions, microarray data is inherently susceptible to noise. Thus, it is essential that subspace clustering methods be robust to noise. (c) Subspace clustering methods allow overlapping clusters that share subsets of genes, samples or time-

978-1-4244-9581-8/11/$26.00 © 2011 IEEE

courses/spatial-regions. (d)Subspace clustering methods should be flexible enough to mine several (interesting) types of clusters. (e) The methods should not be very sensitive to input parameters.

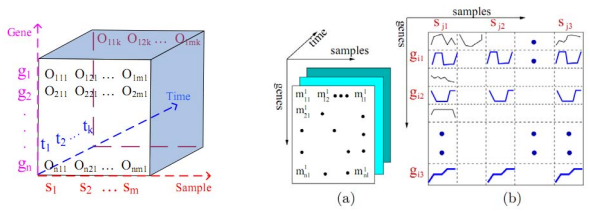
B. Problem Formulation

A microarray experiment typically measures a large number of DNA sequences(genes, cDNA clones, or expressed sequence tags[ESTs] under multiple conditions(including environments, individuals, and tissues) and are often presented as matrices of expression levels of genes under different conditions. Extracting the patterns hidden in gene expression data offers a tremendous opportunity for an enhanced understanding of functional genomics. However, the large number of genes and the complexity of biological networks greatly increase the challenges of comprehending and interpreting the data, which often consists of millions of measurements. A first step toward addressing this challenge is the use of clustering techniques to reveal natural structures and identify interesting patterns in the data. Coherent gene expression patterns may characterize important cellular processes and also be involved in the regulating mechanisms in the cells.

Some previous work on DNA microarray data clustering have aimed to find those genes that are coherent on a subset of the samples during the whole time series. A basic biological observation is that the genes that are biologically associated may behave in similar expression patterns only over a segment of the time series and beyond this time range their expression patterns could be completely irrelevant. To address this issue, triclustering algorithms find those genes coherent on a subset of the samples within a segment of the time series.

C. Triclustering in Gene Coherent Pattern Identification

Microarray datasets are mostly of two types: gene time datasets and genesample datasets.The genetime datasets record the expression levels of various genes over a series of time points.The gene sample datasets account the expression levels of various across related sample. With the latest advances in microarray technology, the expression levels of a set of genes under a set of samples can be monitored synchronically during a series of time points. Different from the previous gene time or gene sample microarray datasets, the new datasets have three variables: genes (G), samples (S) and time (T). We call such data gene sample time microarray data, or GST data for short. Each cell $m^k_{i,j}$ in a particular GST dataset represents the expression level of a particular gene g_i under sample s_j at time point t_k. Fig 1(b) illustrates an example of a 3D GST data where the expression levels of n genes are measured simultaneously under m tissue samples over a series of k time points. In gene sample microarray data, coherent gene expression patterns may characterize important cellular processes and also be involved in the regulating mechanisms in the cells. Therefore, it is interesting to identify a subset of genes g, a subset of samples s and a subset of time point t in a GST microarray dataset such that each gene g ∈ G has similar patterns over the subset of the samples in



(a) GST microarray data[2]

(b) Structure of GST microarray data[3]

Fig. 1. Gene sample time microarray data

s across subset of time spans t. Such a three dimensional cluster is referred as triclusters. Triclusters provide valuable information for the biologist. The sample sets may correspond to some phenotypes, while the corresponding set of genes may suggest the candidate genes correlated to the phenotypes and corresponding set of time points may refer to the time points when the phenotype is expressed.

D. Our contribution

In this paper the following contributions are made

- Formal definitions of some preliminary concepts on triclustering is provided.
- Basic challenges of triclustering is given through analysis of three triclustering algorithms namely Mining Coherent gene clusters from GST microarray data[3], TRICLUSTER[4] and gTRICLUSTER[5].
- Inability of TRICLUSTER to detect shifting patterns is established theoretically.
- Different parameters and metrics to evaluate triclustering algorithms are given.

E. Paper organization

The rest of the paper is organized as follows. Section II presents the existing triclustering algorithms and a discussion is given in section III. Research issues are discussed in section IV and finally Section V presents the conclusion..

II. TRICLUSTERING TECHNIQUES

Some preliminary concepts of relevant to triclustering based on [4] are given below

Definition 1: GST microarray data

Let $G = \{g_1, g_2, g_3, ...g_n\}$ be a set of n genes, let $S = \{s_1, s_2, s_3, ..., s_m\}$ be a set of m biological samples (e.g., different tissues or experiments), and let $T = \{t_1, t_2, t_3, ..., t_l\}$ be a set of l experimental time points. A three dimensional microarray dataset is a real-valued $n \times m \times l$ matrix $D = G \times S \times T$, whose three dimensions correspond to genes, samples and times respectively.

Definition 2: Tricluster

A tricluster C is a submatrix of the dataset D, where $C = X \times Y \times Z$ with $X \subseteq G$, $Y \subseteq S$ and $Z \subseteq T$, provided certain

conditions of homogeneity are satisfied.

Definition 3: Shifting patterns

A group of genes has a shifting pattern when the values G_i vary in the multiplication of a multiplicative constant α. For two genes $g_i=(a_1,a_2,...,a_n)$ and $g_j=(b_1,b_2,...,b_n)$ if ($b_1=a_1+\alpha$, $b_2=a_2+\alpha$,..., $b_n=a_n+\alpha$), then the genes have a shifting patterns.

Definition 4: Scaling patterns

A group of genes has a scaling pattern when the values G_i vary in the addition of a additive constant α. For two genes $g_i=(a_1,a_2,...,a_n)$ and $g_j=(b_1,b_2,...,b_n$ if $b_1=a_1*\alpha$, $b_2=a_2*\alpha$,..., $b_n=a_n*\alpha$, then the genes have a scaling patterns.

The existing triclustering algorithms for mining coherent clusters in three-dimensional (3D) gene expression datasets that this paper deals with are:

- Mining Coherent gene clusters from GST microarray data[3]
- TRICLUSTER[4]
- gTRICLUSTER[5]

A. Mining Coherent gene clusters from GST microarray data

This algorithm tries to find coherent genes under a subset of samples across entire time range of the GST dataset. It accepts three input parameters i.e. coherence threshold (δ), minimum number of genes min_g and minimum number of samples min_s and outputs the complete set of coherent gene clusters. The basic steps of the algorithm are given below

(i) For each gene find the maximal coherent subsets of samples. Coherence between two samples is measured using Pearson correlation coefficient across the whole time range.

(ii) To find the complete coherent triclusters from the maximal coherent samples of genes as computed in the previous step the authors proposed two algorithms.
   (a) Sample Gene Search
   (b) Gene Sample Search

The steps of the both algorithms can be summarized as follows

(i) Enumerate all the possible subsets of samples/genes. The authors use set enumeration tree along with prunning operation to find these subsets efficiently.

(ii) For each of the subsets of samples S/genes G find the maximal subsets of genes G/samples S such that G×S is a coherent gene clusters. This is done using inverted list of maximal coherent sample set of genes that was prepared in the earlier step of the algorithm.

The features of this algorithm are stated below

(1) It uses set enumeration tree and prunning operations on it to find all the possible combinations of samples or genes.

(2) The algorithm takes care of inter temporal coherence in the extracted triclusters.

(3) Due to use of Pearson correlation coefficient it can detect both shifted and scaled form of inter temporal coherence.

(4) The authors provide two ways to find all the possible triclusters.

(5) It is a deterministic algorithm and it allows overlapping of triclusters.

Some limitations of this algorithm are stated below

(1) The algorithm detects triclusters that include the entire time span.

(2) The second approach proposed by the author to find the complete set of coherent gene clusters i.e. GeneSample Search is computationally very costly.

(3) Though the algorithm considers inter temporal coherence, it does not take care of coherence among genes in triclusters.

(4) None of the proposed approaches is capable enough to find all the possible triclusters and it is practically infeasible to use the both at the same time.

B. TRICLUSTER

It is a efficient and deterministic triclustering algorithm that accepts four input parameters i.e. maximum ratio threshold (ε), minimum gene threshold (mx), minimum sample threshold (my) and minimum time threshold (mz) and outputs coherent clusters along gene-sample-time dimension. The basic steps of the algorithm are given below

(a) For each G×S time slice matrix, find the valid ratio-ranges for all pair of samples, and construct a range multigraph.

(b) Mine the maximal biclusters by performing depth first search on the range multigraph.

(c) Construct a graph based on the mined biclusters and get the maximal triclusters.

(d) Delete or merge clusters if certain overlapping criteria are met.

Some facts about the TRICLUSTER algorithm based on[4] are given below

- It can mine only the maximal triclusters satisfying certain homogeneity criteria.
- The clusters can be arbitrarily positioned anywhere in the input data matrix and they can have arbitrary overlapping regions.
- It can mine several types of triclusters.
- It is a deterministic and complete algorithm, which utilizes the inherent unbalanced property.

Some limitations of TRICLUSTER algorithm are stated below

1) The algorithm depends on four input parameters.

2) Computation to find the genes to be included in terms of two samples is costly. The condition can be made simpler by replacing $\frac{\max(|ru|,|rl|)}{\min(|ru|,|rl|)}-1$ by $\max(|ru|,|rl|)-\min(|ru|,|rl|)$

3) Due to the aggregate nature of comparison, a single value of ε is not sufficient to detect relevant biclusters that may need different thresholding for their recognition. Fig.2 shows that due to the aggregate nature of comparison, gene g3 and g4 may get included in a cluster because these two genes produce a value

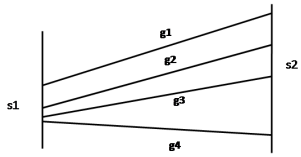


Fig. 2. Expression levels of genes g1, g2, g3 and g4 at samples s1 and s2

of the aggregate measure that is almost same if we take g1, g2 and g3. So a mutual measure may be more effective from this aspect.

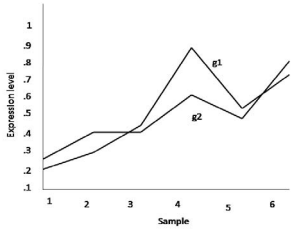


Fig. 3. Expression levels of two genes g1, g2

4) It is very difficult to control the required level of coherence using parameter ε. Practically recommended value of ε is less than 0.01. But such a value can't detect biologically relevant patterns such as the one in Fig .3.
5) The algorithm can detect clusters with scaling patterns, but can't detect the ones with shifting patterns.

To establish (5), we provide the following proof

Proof:

The value of the parameter $\frac{\max(|ru|,|rl|)}{\min(|ru|,|rl|)}$-1 should be less than ε for a subset of genes against a pair of samples to get the set included in a tricluster. Let us consider expression levels of two samples S_k and S_{k+1} of 3 genes g_1, g_2 and g_3 to be $\{e_1,e_2\},\{\beta e_1,\beta e_2\};\{\alpha e_1,\alpha e_2\}$ where $\alpha \geq \beta$

Now, the ratio of the expression values of gene g_1, g_2 and g_3 in columns s_k and s_{k+1} will be

$r_{g1}^{S_k S_{k+1}}=e_1/e_2$

$r_{g3}^{S_k S_{k+1}}=\alpha e_1/\alpha e_2=e_1/e_2$

$r_{g2}^{S_k S_{k+1}}=\beta e_1/\beta e_2=e_1/e_2$

and maximum ratio, $\max(ru,rl)=e_1/e_2$ and minimum ratio, $\min(ru,rl)=e_1/e_2$

$$\text{i.e} \quad \frac{\max(|ru|,|rl|)}{\min(|ru|,|rl|)}=1$$

$$\text{or} \quad \frac{\max(|ru|,|rl|)}{\min(|ru|,|rl|)}-1=0$$

So the value 0 would be always less than any positive value(recommended) of ε. Hence it can detect clusters with scaling patterns.

Now considering the expression levels of two samples S_k and S_{k+1} of 3 genes g_1, g_2 and g_3 are $\{e_1,e_2\};\{\beta+e_1,\beta+e_2\}$ and $\{\alpha+e_1,\alpha+e_2\}$ where $\alpha \geq \beta$

$r_{g1}^{S_k S_{k+1}}=e_1/e_2$

$r_{g3}^{S_k S_{k+1}}=\alpha+e_1/\alpha+e_2$

$r_{g2}^{S_k S_{k+1}}=\beta+e_1/\beta+e_2$

Now, $\max(ru,rl)=\alpha+e_1/\alpha+e_2$ and $\min(ru,rl)=e_1/e_2$

$$\text{i.e} \quad \frac{\max(|ru|,|rl|)}{\min(|ru|,|rl|)}-1=0$$

Hence it can't always detect clusters with shifting patterns.

C. gTRICLUSTER

gTRICLUTER is a triclustering algorithm that accepts four input parameter minimum similarity threshold (δ), minimum sample threshold (m_s), minimum gene threshold (m_g) and minimum time threshold (m_t) and outputs coherent clusters along gene-sample-time dimension. The basic steps of the algorithm are given below

(a) Identify the maximal coherent sample subset for each gene during the time series segment.
(b) Perform a depth first search in the sample space to enumerate all possible maximal cliques
(c) Find the maximal coherent gene set for a subset of samples by generating an inverted list.
(d) Find the intersection of the inverted list, check whether result is a maximal coherent gene cluster and combine with time segments information

The features of gTRICLUSTER is that it considers inter temporal coherence while generating the triclusters.

The limitations of gTRICLUSTER are

(1) It depends on too many input parameters.
(2) Though the algorithm recovers from one of the limitations of TRICLUSTER by considering inter temporal coherence, it avoids considering inter gene coherence which is even more significant than inter temporal coherence.
(3) If there is a time latency between two similar patterns, gTRICLUSTER cant detect such patterns, since two similar patterns may appear in different time ranges.
(4) Though the Spearmen correlation coefficient is applied to capture the coherence across time dimension in gTRICLUSTER, the measure may not be effective always in serving the purpose. As shown in figure 4(a), two patterns which are additive produce a value 0 for the spearmen coefficient. But when we consider two patterns

which are not additive as in figure 4(b), it is supposed to produce a non zero value.However, in reality the measure is found to produce the same value i.e. 0. The measure is found to produce non zero value, when the patterns are overlapped as in figure 4(c). Therefore, we can raise question about the effectiveness of the measure in capturing the coherence properly.

III. Discussion

A brief comparison of triclustering algorithms is given in table I

Though the discussed algorithms seem to discover triclusters, they are quite different both in terms of their approach as well as the triclusters that they produce. The following are some of our observations in support of the fact:

1) Inter temporal coherence: The algorithms seem to operate differently in terms of the structure of the triclusters that they produce. Mining GST Microarray data does not take care of coherence between two genes in a single time point. Instead it captures coherence of genes across time points. This coherence is measured using Pearson corelation coefficient at the time of extracting the maximal coherent sample subset of a gene. TRICLUSTER captures coherence of different genes in a single time point in generating biclusters, but doesn't consider coherence of genes across time while forming triclusters. This is a consequence of using intersection operation to mine triclusters from the set of detected biclusters which are produced using any coherence measure. This fact can be visually perceived through figure 5. In figure 5 the solid line pattern represents the common pattern of selected genes across selected subset of samples in a gene sample space. TRICLUSTER may include all these time points corresponding to each of these dimensions along with the selected genes and samples in a tricluster though the solid line patterns are not coherent at all. The term "Time dominated tricluster" is used to refer to such a tricluster. TRICLUSTER can detect only time domain triclusters whereas gTRICLUSTER tries to overcome this limitation through the use of spearman correlation coefficient across time points. But the measure is not capable of serving the purpose as we have explained in section II-C. Moreover, the algorithm ignores the coherence of different genes in a single time point across samples while generating the final triclusters.

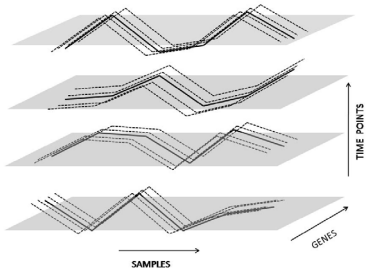
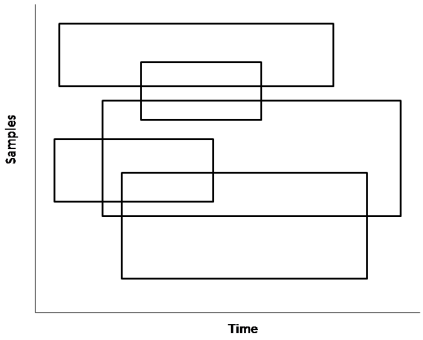
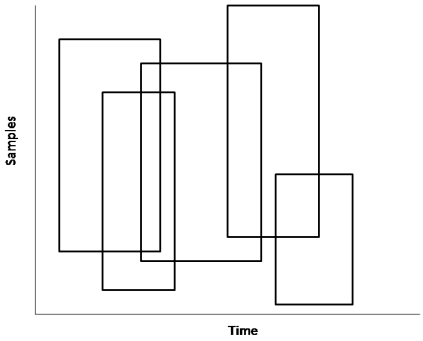


Fig. 5. Ignorance of coherence across time points by TRICLUSTER

2) Time dominated and Sample dominated approaches of triclustering: The triclusters procuded by Mining Coherent gene clusters from GST microarray data include all the time points of the GST dataset. We refer such triclusters as time dominated triclusters. This may miss some potential tricluster candidates with a subset of time points. TRICLUSTER initially generates the biclusters in each GS plane separately. Each GS plane correspond to different time points. While generating these biclusters, the algorithm tries to include as many samples as possible in each bicluster by finding the maximal subset of samples. Finally, intersection operations on these biclusters produce the set of triclusters. While including samples in the first phase time dimension isn't considered at all. As a result, we may miss out certain subsets of samples that may lead to include a large set of time points into the tricluster. We use the term "Time dominated tricluster" to refer to such a tricluster. TRICLUSTER can detect only time domain triclusters. On the other hand gTRICLUSTER tries

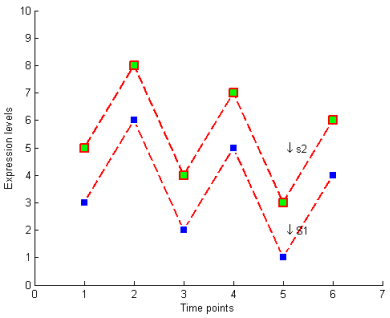


(a) Time dominated tricluster

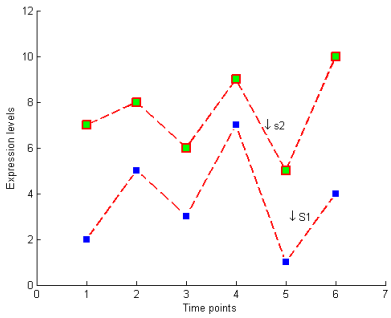


(b) Sample dominated tricluster

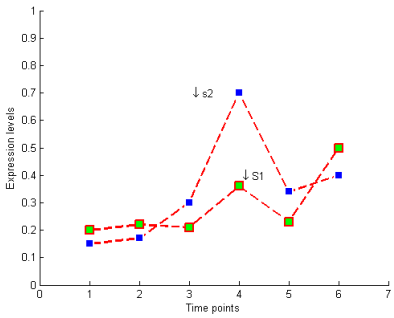
Fig. 6. Time dominated and Sample dominated triclusters



(a) Expression levels of two samples of a gene across different time points



(b) Expression levels of two samples of a gene across different time points



(c) Expression levels of two samples of a gene across different time points

Fig. 4. Effect of Spearman corelation coefficient

TABLE I
COMPARISON OF TRICLUSTERING ALGORITHMS

Parameter for comparison	Mining GST Microarray data	TRICLUSTER	gTRICLUSTER
No of input parameters	3	4	4
Cluster pattern	Doesn't take care of	Multiplicative	Doesn't take care of
Types of triclusters	Time dominant Yes	Sample dominated	Time dominated
Inter temporal coherence	No	No	Yes
Inter gene coherence	No	Yes	No
Hierarchical Representation of clusters	Pearson Correlation	No	No
Proximity measure		No	No

to find the longest time segment while mining the maximal coherent sample subset. While doing so, the algorithm doesn't consider other samples when it processes a sample pair. So the triclusters tend to give priority to the inclusion of samples leading to triclusters that we term "Sample dominated triclusters".

3) Time latent triclusters: We are using the term "Time latent triclusters" to refer to those triclusters that contain non consecutive time points. TRICLUSTER detects time latent triclusters, but gTRICLUSTER and Mining GST Microarray data can't detect such triclusters.

IV. Research Issues

The various reserach issues and challenges associated with triclustering algorithms are: (i) Development of a cost effective triclustering algorithm with minimum number of input parameters is an important research issue. (ii) Coverage of all possible triclusters by triclustering algorithms is another research issue. (iii) The triclustering algorithms should detect time latent triclusters and take care of both inter gene coherence in time points as well as inter temporal coherence. (iv) An appropriate validity measure to validate the triclustering results is another challenging issue.

V. Conclusion

Triclustering is a new research topic in the field of microarray gene expression data analysis. In this paper, we have presented a comprehensive survey of three triclustering techniques namely Mining Coherent gene clusters from GST microarray data, TRICLUSTER and gTRICLUSTER. Some interesting directions for future research have been uncovered by this work. We believe that this work can be used by the interested researcher. An algorithm that caters all the research issues seems to make an end to our quest for finding the optimal triclustering algorithm for GST microarray data.

References

[1] D. Jiang,C . Tang and A. Zhang, Cluster Analysis for Gene Expression Data: A Survey, IEEE transaction on knowledge and data engineering, 2004.

[2] J. Liping, Mining localised co-expressed gene patterns from microarray data, 2006.

[3] D. Jiang, J. Pei, M. Ramanathany, C. Tang and A. Zhang, Mining coherent gene clusters from gene-sample-time microarray data., In Proc of the 10 th ACM SIGKDD Conference(KDD04), 2004.

[4] L. Zhao and M. J. Zaki, TRICLUSTER: An effective algorithm for mining coherent clustersin 3D microarray Data. In Proc of SIGMOD05, 2005.

[5] H. Jiang, S. Zhou, J. Guan and Y. Zheng, gTRICLUSTER: A More General and Effective 3D Clustering Algorithm for Gene-Sample-Time Microarray Data. In Proc. BioDM, pp.48-59, 2006.