Volume 11, Special Issue I, May 2020
International Journal of Advanced Research in Computer Science
CONFERENCE PAPER
Available Online at www.ijarcs.info
ISSN No. 0976-5697

IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES USING BI-CLUSTERING

B.SATHWIKA
Computer science and information technology,
REVA UNIVERSITY,
Bengaluru, India
sathwikasindhura024@gmail.com

CH.NARESH BABU
Computer science and information technology,
REVA UNIVERSITY,
Bengaluru, India
nareshchintha888@gmail.com

D.VINUSHA
Computer science and information technology,
REVA UNIVERSITY,
Bengaluru, India
vinushad10@gmail.com

G.C.MANVITHA
Computer science and information technology,
REVA UNIVERSITY,
Bengaluru, India
manvithagc@gmail.com

PAVITHRA.S
Computer science and information technology,
REVA UNIVERSITY,
Bengaluru, India
pavithra.s@reva.edu.in

***Abstract*:** In order to identify the biologically relevant gene module and also which are the genes responsible for causing a disease, we use Biclustering technique which is also a useful co-clustering technique.In this paper, we present a exceptional method to state specific gene modules and also functionally related gene modules which are the reason for disease causing ,by applying a genetic algorithm to genetic data which is in the form of microarray data. To detect these differentially expressed gene modules, the anticipated method finds biclusters in which genes are overexpressed or under expressed, and also which are differentially-expressed in the samples of genetic data. Inorder to get the differentially expressed we perform three steps in which we use K means alogorithm for clustering and Cheng and Chruch algorithm for biclustering. As to overcome the drawbacks in clustering we use Biclustering technique which reduce redundancy in the data.The ensuing gene modules uncover preferable exhibitions over near techniques in the GO (Gene Ontology) term enhancement test and an analyzed association between gene modules and infection.

Keywords: Bioinformatics, Raw data, analysing, Preprocessing, Clustering, Bi-clustering

INTRODUCTION

Basically Bioinformatics is a field in which biology, computer sciences,information technology, statistics, mathematics etc…merged into a single thing and performs operations to analyse genetic or biological information via computers and statistical approaches. For better understanding of complex condtions in biological system there is a need of advancements in genetic factor related to that feature and that advancement can be done by improving our information about gene expression .The procedure of gene expression of data which is a genetic factor used in the combination of gene product of functional group that may be called as proteins.[12].

There have been many research studies are working for finding out the new biological or genetic treasured information because there is a huge amount of genetical information available these days. There is a technique called clustering that identifies the genetic factor modules that displays the same expression across the set of samples. The significant study of expression of genes is explorative to statistical information. It is finding and grouping of clustering genes shows the same over the conditions of expression forms and also group of circumstances profile to across that set of gene expression. Clustering is a high successful technique with two measures. 1.Highest similarity among the identical clusters and elements 2.Lowest similarity between the elements from dissimilar clusters .It comprises the grouping of similar substances into a set known as cluster and the substances in one cluster are likely to be diverse when compared to objects grouped under another cluster. In this paper we are using K means clustering algorithm[5] to find and group gene modules. Though clustering is an effective approach that displays the same expression among the given group of samples for recognizing related information, it fails in finding infection related genetic module as it does not involve simultaneous or parallel clustering of rows and columns which shows the vital expression on definite samples. To overcome this limitation

bi-clustering technique is used to cluster the data. In this paper we are using cheng and church bi-clustering algorithm[2] to cluster the data. Cheng and Church is known as the first bi-clustering execution and the idea was presented in 1972 by Hartigan[3]. It defines a bi-cluster of maximum similarity between the group of columns and rows. The projected score is also known as mean square residue. This is used to calculate its own coherence set of columns and rows

Bi-clustering technique is a beneficial co-clustering approach that is used for grouping the genes as well as

conditions in both the dimensions instantaneously. It permits simultaneous clustering of rows and columns to find out the gene module over the set of samples. Bi-clustering methods were used by many researchers for analysing gene expression data. some researchers were used order preserving submatrix method to find the calculates of mean squared residue score[7-11]. OPSM model was introduced to invent genes among the subgroup of conditions[5] and few people used maximum similarity bi-cluster algorithm. MSBE is for finding an optimal bi-clusters that have maximum similarity. It is used for finding submatrix alike expressions in the mico array data sets[10]

On the basis of clustering algorithms that have been used to analyse gene appearance data, genes are showing similar expression assumed to be co-regulatory path way.

In 2000, bi-clustering technique was first applied to gene expression data in all the conditions for the identification of co-expressed genes for a matrix data set. Bi-clustering algorithms have developed to improve the ability for a large data sets generated to high-throughput omic technologies.

The richness of gene expression data sets offers an opening to find out genes with parallel expression patterns across numerous circumstances that is co-expression gene modules(CEMS)[6]. Micro array platforms are used in generating genetic factor expression data because of its easy accessibility and low cost.

METHODOLOGY

Block diagram:-

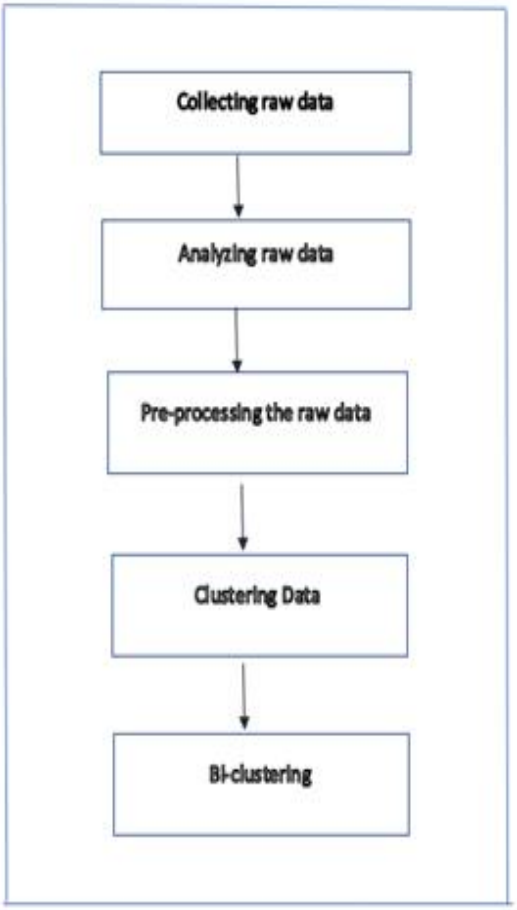


Fig 1.1

Fig 1.1 shows the step by step procedure for doing bi-clustering.

1.Collecting raw data:-The primary step is to collect raw data. After collecting raw data, the raw data need to be analysed before giving it as a input. Here, we have collected the raw data sets from[13]**www.ncbi.nlm.nih.gov.**

2.Analyzing and Preprocessing the data:-

We need to create a file for representing the information to load the data into R. The created file has to be opened in text editor. There will be a single column representing the set of list files. The last file needs three columns that contains name, file name, and target name. In this case the name column and file name column are identical. Target column contains the information about the samples that we have taken and we have to label the samples appropriately and when comparision of cell line and tissues of the sample is done, it defines duplicate groups for further analysis.

3.Normalizing data: A normalization phase will helps in computing and modifying the biases to precise the data. It is the process of elimination of any systematic biases in the data set. It is used to remove the non-biological dissimilarity as much as possible.

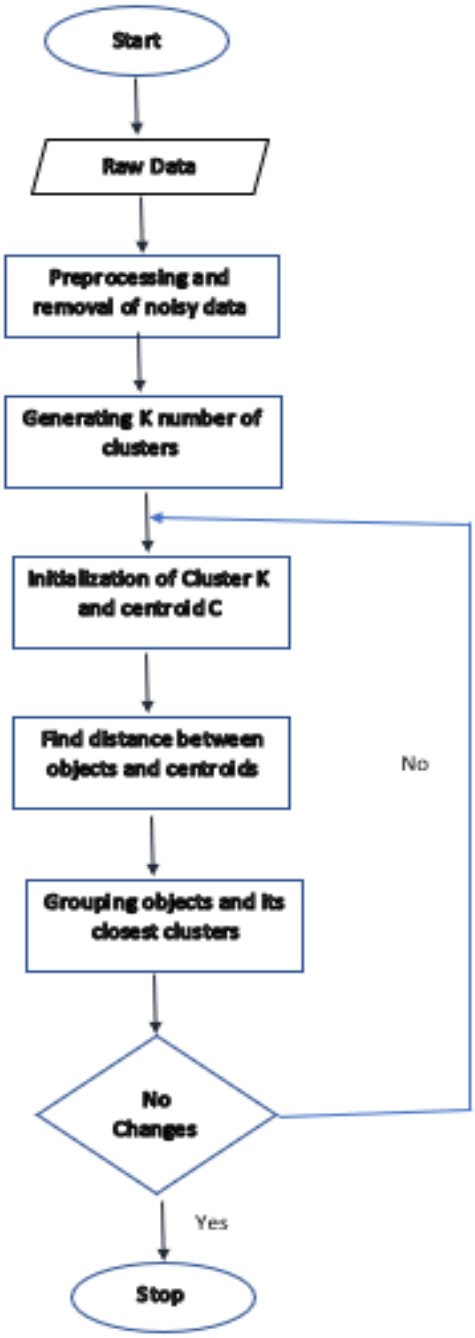


Fig 1.2

Quality controller check:- Before analyzation of data we have to do some quality control checks to make definite that there are not any more problems with the data set.

Data filtering:-In this stage we are going to filter un wanted and useless data such as investigation sets and eliminating genes that have low variance and that could be doubtful for passing the statistical tests for different expression, or expressed homogeneously.

3.Clustering:- clustering is the next step to be followed after analyzation and preprocessing. Clustering aims in finding the genes that shows the alike expression among the given group of samples and also aims in grouping the genes that shows similar expression pattern. We have to cluster the data that we got in analysis and preprocessing steps. Though clustering is an efficient technique it fails in removing the duplicates or duplicates are possible in clustering. To overcome this limitations bi-clustering method has to be followed.

4.Bi-clustering:- Duplicates are not possible in bi-clustering as this technique permits simultaneous clustering of rows and columns at a time. Bi-Clustering is a practice that allows parallel clustering of genetic factor and samples to identify the genetic module that displays dangerous expression between the samples.

ALGORITHM 1:-

Clustering is done using k-means algorithm.

K-MEANS is a clustering algorithm which categorize or to collect the objects accordingly to the attributes which are into k number of clusters. K-means computational correspondence will be started by allocating k points into the space with will be defined by the objects which are actually clustered. There points specifies initial centroids. We can take any random objects as an initial centroids. Then the k-means algorithm will execute few steps.

1) Decide the centroid coordinates.
2) Decide the distance of each object from the centroids
3) Assemble the objects which have the minimum distance.
4) Iterate the third steps untill no object is changed from its assigned group.

Each iteration of the k means changes the present partition by checking all possible ways for the adjustment of the solution, also one of the element is changed to another cluster and this is carried out by reducing the sum of distances between the objects and also the centers of the clusters. This method is changed till no further changes are carried out and all objects will be grouped into the final number of needed clusters.

K means flowchart:

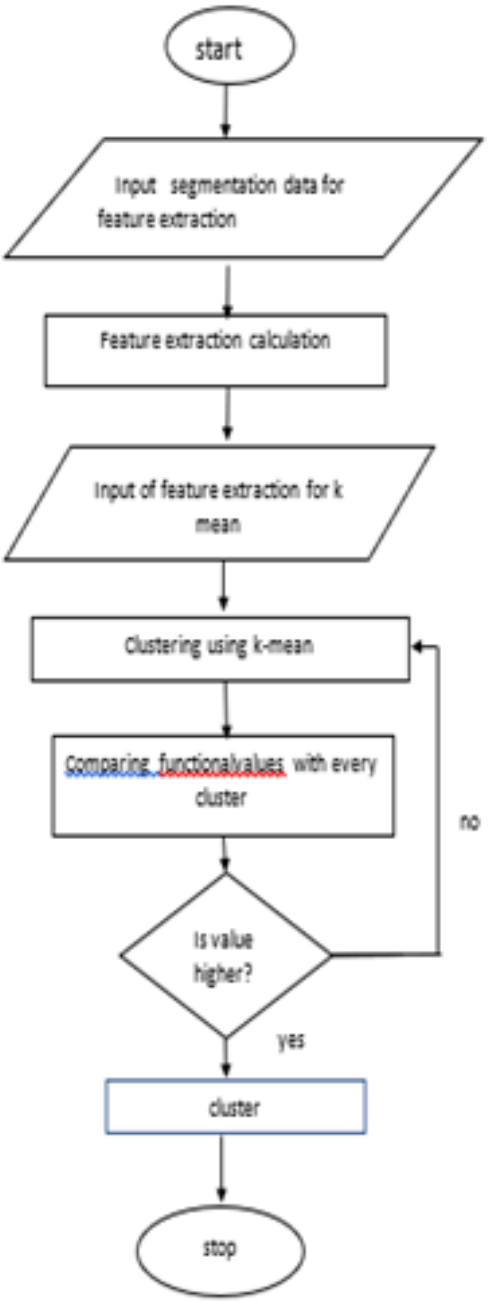


Fig 1.3

K means algorithm:

This most familiar algorithm will use the iterative technique. Because of this ubiquitousness, this is also called "the k-means algorithm". In some of the situations it is also refered as "n k-means", due to this there also exist much faster options.

Given aset of k means , the algorithm will be done by swapping or exchanging between two steps

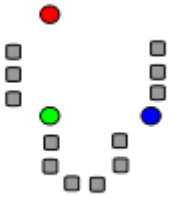
1) step 1: We should assign each considered cluster to the closest mean.

2) step 2: Re-calculate the means of the observations whichare given for each and every cluster.

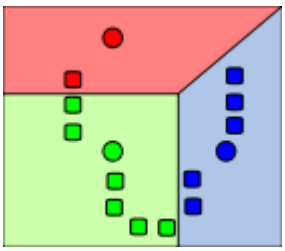
The algorithm will be existed until there is assignments no longer will be converted. The algorithm will not promise to find the optimum.

The algorithm is used for assigning the objects to the closest cluster by distance. many adjustments of k-means such as spherical k-means and also the k-medoids will be requsted to use the other distance measures.

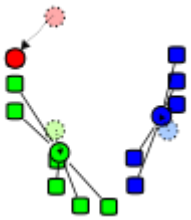
- **Demonstration of standard alg**



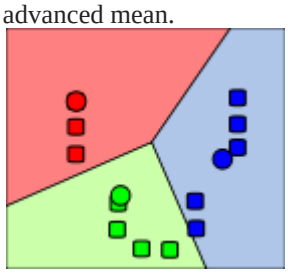
k initial are randomly generated within its data domain



k clusters will be created by connecting every observation to its nearest mean.



The centroid of each of the *k* clusters will be turned into the advanced mean.



Steps 2 and 3 will be repated till merging to be reached.

ALGORITHM 2

Chung and Church algorithm:-

Biclustering is done using cheng and church algorithm.This is the first Biclustering algorithm applied by CC as it cluster the both rows and columns simultaneously.

CC algorithm uses Greedy iterative method whose complexity stands as 0((n+m)nm) and prediction ability is coherent values.The most significant invention of CC algorithm is that they set a significance called residue score. Attribute residue, object residue, and δ-cluster residue are the three parts of an expression model divided by this algorithm

$$e_{IJ} = \frac{\sum_{i\in I, j\in J} e_{ij}}{|I||J|}$$

Here, I and J are the row and column vector sets of sub matrix, $|I|$ and $|J|$ are the number of rows and columns, respectively. e_{ij} is the component of sub matrix. e_{Ij}, e_{iJ}, , e_{IJ} are the attributes residue, object residue, and δ-clustering residue, respectively.

The meaning of the residue score is as mentioned below:

$$RSIJ(i,j) = eij - eIj - eiJ + eIJ$$

Let X be the collection of the genes and Y *be* the collection of the conditions. Let e_{ij} be the component of the genetic factor-condition expression matrix on behalf of the logarithm of the relative profusion of the mRNA of the i^{th} gene under the j^{th} condition. Let $I \subset X$ and $J \subset Y$ be the subsets of the genetic factors and conditions. The pair (I,J) specifies a sub matrix A_{IJ} with the below *mean squared residue score*:

$$H(I,J) = \sum_{i\in I, j\in J} \frac{RS_{ij}^2}{|I||J|}$$

The lowest score $H(I,J) = 0$ specifies that the gene expression levels adjust in unity. This contains the small or constant bi clusters where there is no chance of fluctuation. These insignificant biclusters might not be very fascinating but need to be revealed and disguised so that more remarkable ones can be found. The row variance may be an complementary score to discard insignificant bi clusters.

$$V(I,J) = \frac{1}{|J|}\sum_{j\in J}(e_{ij} - e_{Ij})^2$$

The greater the value of H is, the greater the disordered data . In CC algorithm, a greedy method is used to choice the sub matrix with a small H score which is distributed into two stages. Initially, the method is to eliminate the row or column to reach the largest decrease of the score. For the present submatrix, they calculate the average residue score of each and every row using

$$d(i) = \frac{1}{|J|}\sum_{j\in J} RS_{IJ}(i,j$$

And the average residue score of each column using

$$e(j) = \frac{1}{|I|}\sum_{i\in I} RS_{IJ}(i,j$$

Then select the row or column with the utmost score and eliminate that from the existing sub matrix, till $H(I,J) < \delta$. They use a restriction α, so that they can delete a set of nodes to each time previously the score is re calculated. Without bringing up-to-date the score after the removal of each node, the matrix may contract too much and henceforth, might miss some large δ-clusters.

RESULT

For this experiments which is performed, we make use of Windows 10 operating system of Intel Core i5-6200 and also with is up 2.8Hz, and we enforced our algorithm using R programming.

Differentiation of gene expression

The comparison steps are executed on the gene expression data. Intending to estimate the capability of the algorithms to group the utmost number of genes of which expression designs are same and the GO category also will be the same.Pre-processing of all the collected data is done. In a matrix of gene expression data in which the rows are indicated by genes and columns are indicated by the p-value of each is given as input to the clustering methods.

k-mean is a clustering algorithm which categorize or to collect the objects according tothe attributes which are into k number of clusters. K-means algorithms are used for clustering

Clustering methods are used to cluster the genes into homogeneous clusters they are differentially expressed and non-differentially obtained from various clustering techniques for different data sets. All these techniques show a small difference with respect to each other in selecting differentially expressed genes.

Bi-clustering is done intending to collect the genes and orders in both of the extensions accordingly. This accepts in finding subgroups of genes which express the similar feedback under a subtype of orders, but for not all the situations in the conditions. The gene may perform in multiple function which is more than one function, and which also results in one adjusted pattern in one set of conditions and a different pattern in another.

Cheng and church algorithm is evaluated as the first algorithm in bi-clustering implementation. This mainly performs the following two steps.

Takes out the rows and columns with a greater difference.

Add rows and columns with same difference.

We use k-means algorithm and Cheng and church algorithm to cluster the data. When this is done select the GO category and select the significant level and then comparison is done. The comparison results are displayed using statistical graphical charts.

CONCLUSION

A bi-clustering approach is needed here to integrate the methods and also the tools which are used biological field and also in the biomedical fields.

In this paper, we designed an algorithm method which not only finds the functionally-related gene modules, but it also fins the state of the specific gene modules by using an upgraded genetic algorithm. Through this genetic algorithm which improvise the process of natural selection, it may have problem like loss of diversity problem. So that, we additionally we add a selection pool for the genetic algorithm to solve this problem which came across. The experimental

results will prove that the method which is designed performance will be better than the existing algorithm for finding the disease-related gene modules. Hence, we proved that the method which is designed will acts well in finding the disease related gene modules and also state the specific gene modules which are over expressed and under expressed.

REFERENCES

1. Hartigan JA. Direct clustering of a data matrix. J AM stat Assoc. 67(337): pp.123-9,1972.
2. Y.Cheng and G. M. Church, Biclustering of expression data. Inproc.of the international conference on intelligent system for molecular biology. pp.93-103,2000.
3. Hua QU, Liu-Pu Wang and Chun-Guo Wu. An improved biclustering algorithm and its applications to gene expression spectrum analysis. Genomics, Proteomics and Bioinformatics, Elsevier. 3(3):pp.189-193,2016
4. Ben-Dor A, Chor B, Karp R, and Yakhini Z. Discovering local structure in gene expression data: The order-preserving sub A matrix problem, In Proc. International Conference on Computational Biology, pp.49-57, 2002.
5. Fadhl M. Al-Akwaa. Analysis of gene expression data using Bi-clustering algorithms.2012.
6. Wang Z, Gerstein M and Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 10(1):pp.57-63, 2009.
7. Bozdag D, Kumar A and Catalyurek UV, Comparative analysis of biclustering algorithms, In: Proceedings of 1st ACM, International Conference Bioinformatics and Computational Biology, pp.265274, 2010.
8. Cano C, Adarve L, Lopez L and Blanco A, Possibilistic approach for biclustering microarray data, Computers in Biology and Medicine, 37, pp.1426-1436, 2007.
9. Ahn, Youngmi Yoon, Jaegyoon Sanghyun Park, Noise-robust algorithm for identifying functionally associated biclusters from gene expression data, Information Sciences, 181 pp.435-449, 2011.
10. Shahreen Kasim, Safaai Deris, Razib M. and Othman, Multi-stage filtering for improving confidence level and determining dominant clusters in clustering algorithms of gene expression data, Computers in Biology and Medicine, 43(9), pp.1120-1133, 2013.
11. Tanay, A., Sharan, R., and Shamir, R, Biclustering algorithms: A survey, In Handbook of Computational Molecular Biology, S. Aluru, Ed, Chapman and Hall, 2006.
12. www.ncbi.nlm.nih.gov
13. G.F. Berriz, O.D. King, B. Bryant, C. Sander, F.P. Roth, Characterizing gene sets with Func Associate, Bioinformatics, 19, pp.2502-2504, 2003.