MICROARRAY GENE EXPRESSION DATA ANALYSIS USING ENHANCED K-MEANS CLUSTERING METHOD

Muhammad Rukunuddin Ghalib, Rittwika Ghosh, Priti Sasmal, Udisha Pande
School of Computer Science and Engineering, Vellore Institute of Technology
Vellore-632014, Tamil Nadu, India

ABSTRACT

This Clustering analysis method is one of the important methods which can influence clustering results directly. Among all the clustering methods, k-means clustering is one of the most popular schemes owing to its simplicity and practicality. In this paper we've discussed the standard clustering algorithm and analyzed the outcomes of an enhanced k-means algorithm such that it is possible to calculate the distance between all data objects and all cluster centre in each iteration which makes the efficiency of clustering high. This paper basically reviews the method utilized in processing and analysis of gene expression data generated using microarrays. This type of experiment allows determining relative levels of mRNA abundance in a set of tissues or cell populations for thousands of genes simultaneously. We have proposed and implemented an enhanced k-means algorithm which stabilizes and thereby increases the performance in terms of cluster output.

KEYWORDS: *Microarray Analysis, Clustering Algorithm, k-means algorithm, Euclidean Distance similarity Measure.*

I. INTRODUCTION

In order to make proteins, the gene from the DNA is coped by each of the chemical bases into messenger RNA (ribonucleic acid) or mRNA. The mRNA moves out of the nucleus and uses cell organelles in the cytoplasm called ribosomes to form the polypeptide or amino acid that finally folds and configures to form the protein.
All the DNA in the cell makes up the human genome. There are about 20,000 important genes located on one of the 23 chromosome pairs found in the nucleus or on long strands of DNA located in the mitochondria.

Gene Expression Analysis
The proper and harmonious expression of a large number of genes is a critical component of normal growth and development and the maintenance of proper health. Disruptions or changes in gene expression are responsible for many diseases. It is a term used to describe the transcription of the information contained within the DNA, the repository of genetic information, into messenger RNA (mRNA) molecules that are then translated into the proteins that perform most of the critical functions of cells. Scientists study the kinds and amounts of mRNA produced by a cell to learn which genes are expressed, which in turn provides insights into how the cell responds to its changing needs. Gene expression is a highly complex and tightly regulated process that allows a cell to respond dynamically both to environmental stimuli and to its own changing needs [1][4][11]. This mechanism acts as both an "on/off" switch to control which genes are expressed in a cell expression of particular genes as necessary. Gene expression is the process by which a gene's coded information is converted into the structures present and operating in a cell. Gene expression occurs in two major stages: transcription and translation. During transcription, a gene is copied to produce an RNA molecule (a primary transcript) with essentially the same sequence as the gene; and during translation, proteins are synthesized based on the RNA molecule [2][3][7][9].

Microarray Data Analysis

Microarray technology is arguably the most important recent breakthrough in molecular biology. It enables researchers to obtain snapshots of gene expression for all the genes in a genome in a single experiment. Microarray experiments generate massive amounts of data that can be analysed to extract new knowledge about the underlying biological processes.DNA Microarrays are small, solid supports onto which the sequences from thousands of different genes are immobilized, or attached, at fixed locations. Each spot on an array is associated with a particular gene. Each color in an array represents either healthy (control) or diseased (sample) tissue .Functional genomics involves the analysis of large datasets of information derived from various biological experiments. One such type of large-scale experiment involves monitoring the expression levels of thousands of genes simultaneously under a particular condition, called gene expression analysis [6][11]. Microarray technology makes this possible and the quantity of data generated from each experiment is enormous, dwarfing the amount of data generated by genome sequencing projects.

Analysis of Gene Expression Data

One of the reasons to carry out a microarray experiment is to monitor the expression level of genes at a genome scale. Patterns could be derived from analysing the change in expression of the genes, and new insights could be gained into the underlying biology.

Clustering

Clustering is a way that classify the raw data reasonably and searchs the hidden patterns that may exist in datasets [3]. One of the goals of microarray data analysis is to cluster genes or samples with similar expression profiles together, to make meaningful biological inference about the set of genes or samples. Clustering is one of the unsupervised approaches to classify data into groups of genes or samples with similar patterns that are characteristic to the group. Clustering methods can be hierarchical (grouping objects into clusters and specifying relationships among objects in a cluster, resembling a phylogenetic tree) or non-hierarchical (grouping into clusters without specifying relationships between objects in a cluster) . An object may refer to a gene or a sample, and a cluster refers to a set of objects that behave in a similar manner. The widespread use of DNA microarray technology[2][6] to perform experiments on thousands of gene fragments in parallel has led to an explosion of expression data. To handle such huge amounts of data on entities whose interrelationships are poorly understood, exploratory analysis and visualization techniques are essential. Clustering (the assignment of each of a large number of items to one of a much smaller number of classes) is one widely used technique [1][4].

K-means clustering:

The k-means clustering algorithm is one of the popular data clustering approaches. The k-means clustering algorithm receives as input a set of points and the number k of desired centers or cluster representatives. With this input, the algorithm then gives as output a set of point sets such that each set of points have a defined center that they "belong to" that minimizes the distance to a center for all the possible choices of each set [5][7][13]. A drawback of the k-means algorithm is that the number of clusters k is an input parameter [8].

The manuscript is organized as introduction in chapter I, then we talk about our methodology for achieving the concept implemented in chapter II, followed by our proposed idea and its implementation in chapter III, then results and discussion on our findings are put up on chapter IV. The last chapter V is the conclusion and the future direction of our research.

II. RELATED WORK

The related work of clustering analysis can be divided into three categories. The first category is on similarity measurements that may affect the final clustering results directly. Euclidean distance and correlation coefficient are most popular similarity measures. The second category is on the clustering methods, which are the core of clustering analysis and have received extensive attentions. The third category is on validation techniques that are applied to evaluate the validity of the clustering results, the suitability of parameters, or the reliability of clustering algorithms.[6][9][13][14]

III. METHODOLOGY

At first need to define K centers, one for each cluster. These centers should be placed in such a way that it should not be very close to each other in the same location because different location causes different result. Better to place the cluster as much as possible far away from each other. The next step is to take each point belonging to a given data set (cluster) and associate it to the nearest center. If no point is pending, then the first step is completed and grouping is almost done. Assume X = $\{x_1,x_2,x_3,\ldots\ldots\ldots,x_n\}$ be the set of data points and V = $\{v_1,v_2,\ldots\ldots\ldots,v_c\}$ be the set of centers. After these points need to re-calculate k new centroids which we have obtained from the previous step, this is shown in equation (1) below,

$$v_i = (1/c_i)\sum_{j=1}^{c_i} x_i \qquad \ldots..(1)$$

Where, 'c_i' represents the number of data points in i^{th} cluster. Now we have k new centroids, a new binding has to be done between the same data set points and the nearest new center. So, a loop has been generated. As a result of this loop the k centers change their location step by step until centers do not move further.Measuring of the minimized distance is done by finding the square of the distance between each point to minimize the distance by applying the formula given below in equation (2):

$$J(V)=\sum_{i=1}^{c}\sum_{j=1}^{c_i}(\|x_i - v_j\|)^2 \qquad \ldots...(2)$$

'$||x_i - v_j||$' Euclidean distance between x_i and v_j.
'c_i' is the number of data points in i^{th} cluster.
'c' is the number of cluster centers.
Calculates stop condition. Stops if the point farthest from its centroid is within the average distance value.

IV. PROPOSED SYSTEM

In the beginning we process and transform the given expression data. Next, we take the clustering parameters to be used in the clustering algorithm. Then we calculate the Euclidian distance using the equations (1) and (2).Calculate the stop condition.We stop the calculation if the point farthest from its centroid is within the average distance value. The ovearall system design is given in fig 1.

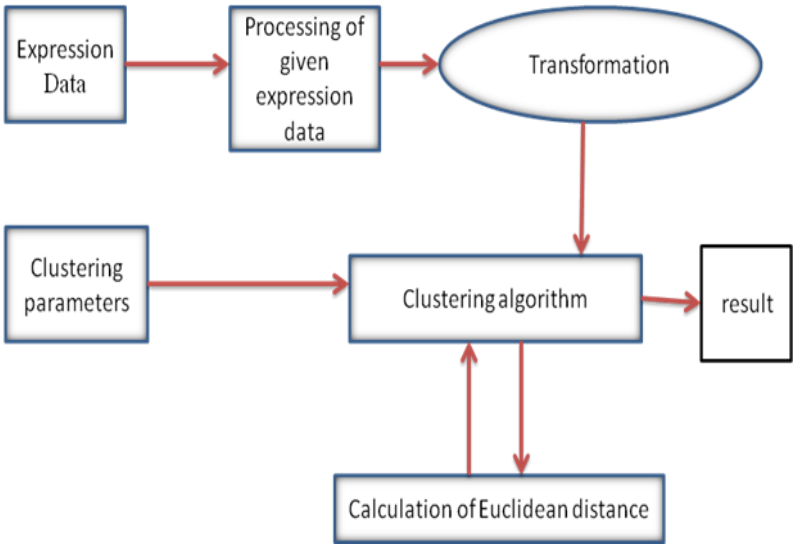

Fig.1: Overall design of the whole system

BASIC ALGORITHM

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Now again assign the object to each group that has the closest centroid.
5. After finishing the assignment need to recalculate the new distance of the centroid.
6. Stops if the point farthest from its centroid is within the average distance value.
7. If not then need to continue the above mentioned 3 to 5 no process.

The above algorithm is also depicted in fig 2 in a flowchart manner.

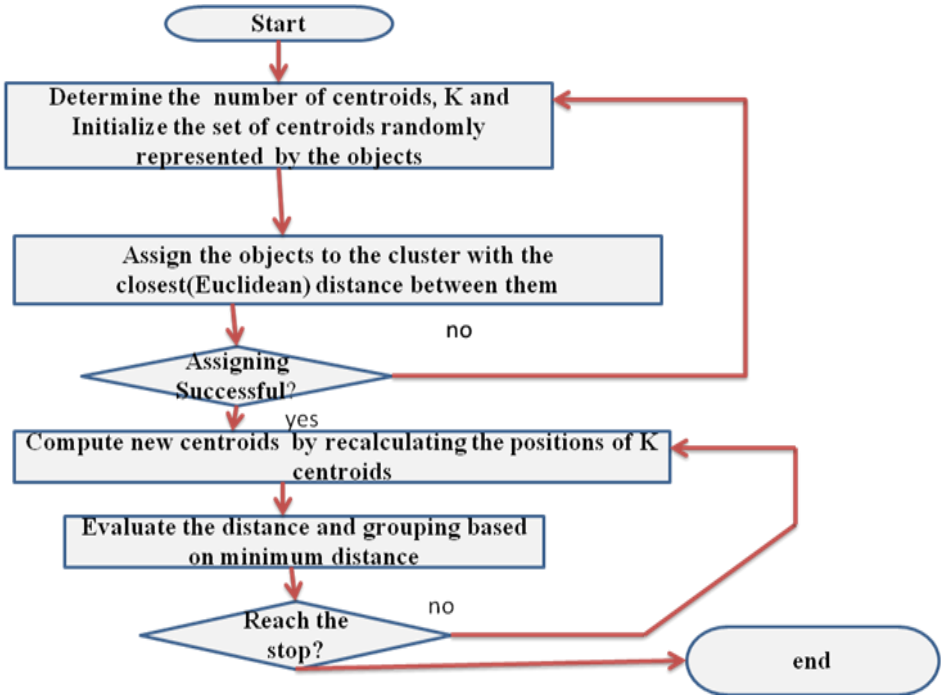

Fig.2: Flowchart of enhanced k-means clustering

V. RESULT ANALYSIS AND DISCUSSION

Suppose, we define five random clusters of five different colours and different attributes.

Table1: Different gene clusters with their attributes

No. of Gene clusters	*Attribute1*	*Attribute2*
A	***1***	***1***
B	***2***	***1***
C	***4***	***3***
D	***5***	***4***

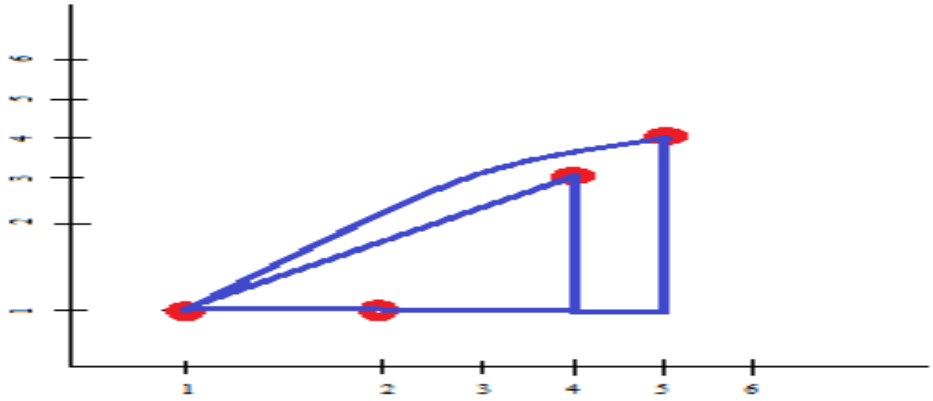

Fig. 3: Graph showing attribute1 Vs attribute2

Each gene cluster represents one point with 2 features- (x,y)
Initial value of centroid:
Assuming gene clusters A and B are the 1st two centroids (c1,c2)
Object centroid distance:
We need to calculate the distance between each cluster's centroid. Now use Euclidean distance matrix

$$\text{at iteration} = \begin{bmatrix} 0 & 1 & 3.6 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix}$$

In the above matrix, we are comparing the minimum values of gene clusters- A, B, C, D to specify the initial centroid for iteration1.
Object clustering: Assign each gene cluster based on the minimum distance. We can call it as group matrix, which shows that in which group we have assigned the gene clusters.

$$G_0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix}$$

First row forms group1 which includes only gene cluster A but the second row forms group2 which includes only gene cluster B, C and D.
Iteration1:
Here we determine centroids. Since we know the members of each gene cluster, so, it is possible to compute new centroids of each group based on this new membership.

$$\text{C1= (1, 1)} \qquad \text{C2}=\left(\frac{2+3+5}{3},\frac{1+3+5}{3}\right)=\left(\frac{11}{3},\frac{8}{3}\right)$$

$$\text{Distance matrix, } D_1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.49 & 1.89 \end{bmatrix}$$

Iteration1 Object Clustering:

$$G_1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

First row forms group1 which includes gene clusters A and B but the second row forms group2 includes gene cluster C and D.
Iteration2 (Determining second new centroids):

$$\text{C1}= \left(\frac{1+2}{2},\frac{1+1}{2}\right)=\left(\frac{3}{2},1\right)$$

$$\text{C2}= \left(\frac{4+5}{2},\frac{3+4}{2}\right)=\left(\frac{9}{2},\frac{7}{2}\right)$$

$$\text{Distance matrix, } D_2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.3 & 3.54 & 0.71 & 0.71 \end{bmatrix}$$

Iteration2 Object Clustering:

$$G_2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

Hence we can see, $G_1 = G_2$.So, we can conclude that it is not possible to group the gene cluster further. Therefore enhanced k-means has reached its stability and no more iteration is needed.
The fig 4 and fig 5 shows the simulated scatter plots of clusters formed in five colors and two colors format respectively.

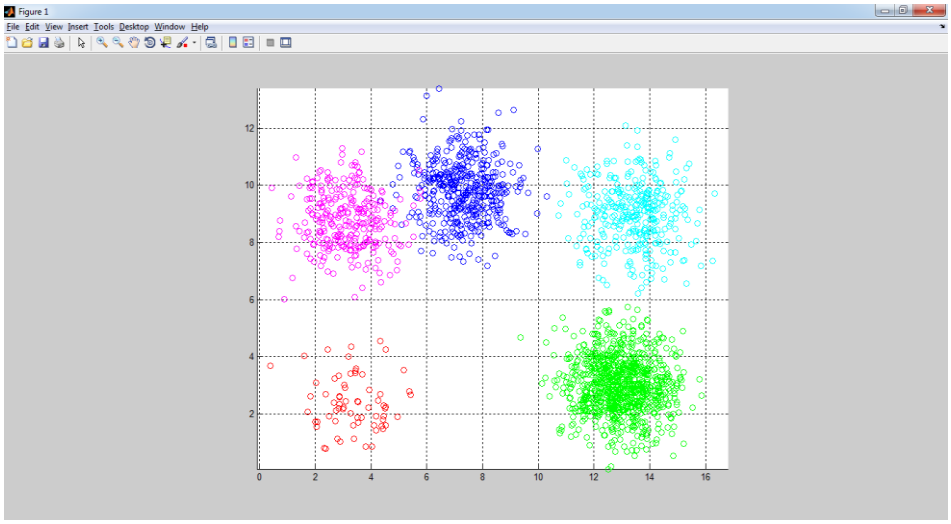

Fig. 4: Scatter Plot of clusters of five different colors

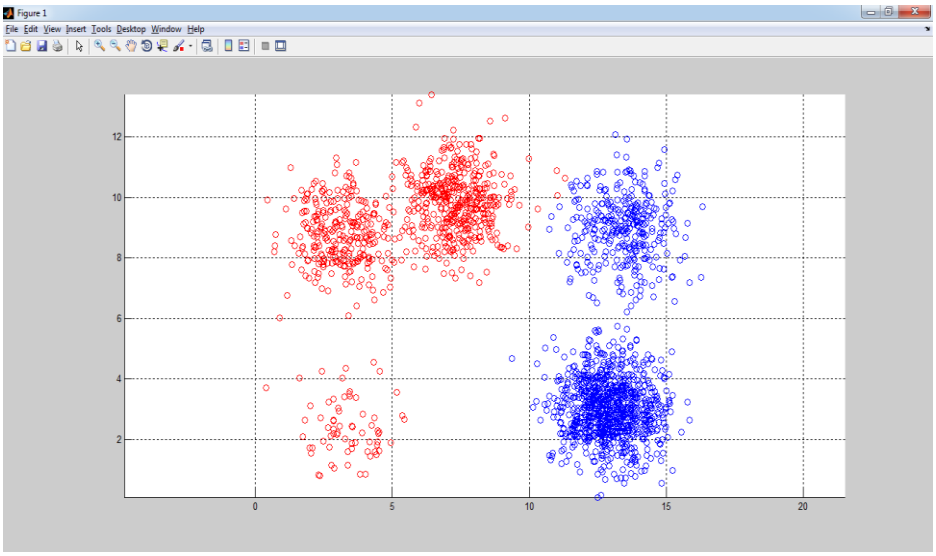

Fig 5: Scatter Plot of clusters of two different colors

The average centroids of the cluster points of the different clusters are shown in fig 6 below.

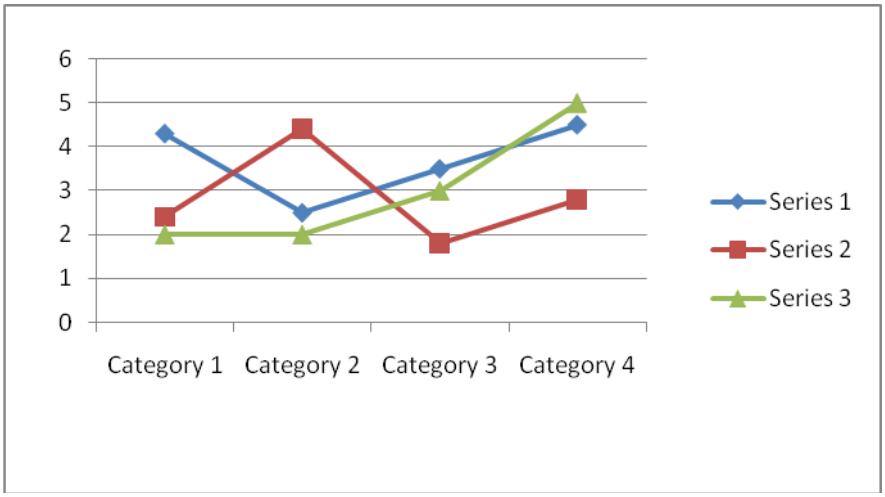

Fig. 6: Average centroid cluster points

The following table 2 shows the result of number of clusters formed with variation of k values in terms of different color representations.

Table 2: Results of clusters for different values of 'k'

Input	*Expected result*	*Remarks*
When k=2	***Shows 5 different clusters of 2 different colours***	***Refer Fig 4 Clustering successful***
When k=5	***Shows 5 different clusters of 5 different colours***	***Refer Fig 3 Clustering successful***

VI. CONCLUSION AND FUTURE WORK

As per the random datasets used, we could achieve 5 different clusters of two different colours when we made k=2 and followed by 5 different clusters of five different colours when we altered the value of k=2 to k= 5. In our research we could justify our finding by achieving enhance performance and increased cluster validity to simple k-means algorithm by reaching the stability without further iterations required. But it was difficult to assess the validity/significance of the results. Even "random" data with no structure can yield clusters or exhibit interesting looking patterns.
We could enhance this paper by taking on more real datasets in future and considerations of memory requirements in today's computing environment while dealing with huge dataset is a must and efficient validation technique be applied which can also be looked into in this paper.

ACKNOWLEDGEMENTS

We avail this opportunity to thank VIT University for helping us by providing the necessary sources and resources.

REFERENCES

[1]. Francis D. Gibbons and Frederick P. Roth (2002), "Judging the quality of gene expression-based clustering methods using gene annotation", Genome Research, Page 1575.
[2].Juntao Wang and Xiaolong Su (2011), ""An improved k-means clustering algorithm", IEEE research paper, page 45
[3].Shi Na, Liu Xumin, Guan Yong (2010), "Research on k-means clustering algorithm", IEEE research paper, pp – 64- 65
[4].Shuhua Ren and Alin Fan (2011), "k-Means clustering algorithm based on coefficient of variation", IEEE research paper, page – 2079
[5].Gregory A. Wilkin and Xiuzhen Huang (2007), "k-Means clustering algorithms: implementation and comparison", IEEE research paper, page – 133-134
[6].Ankur Mazumdar, Muhammad Rukunuddin Ghalib,(2011)"Qualitative and Quantitative metrics based analysis of Gene Expression Data Clustering Algorithms",Intl. J. of Computer Information Systems, Vol II, Issue IV, Pages 44-48
[7].Shou-Qiang Wang and Da-Ming Zhu (2008), "Research on selecting initial points for k-means clustering", IEEE research paper, page – 2675
[8]. Zhang Chen and Xia Shixiong (2009), "k-means clustering algorithm with improved initial center", IEEE research paper, page – 790-802.
[9] Vincent S. Tseng and Ching-Pin Kao, (2005)" Efficiently Mining Gene Expression Data via a Novel Parameterless Clustering Method", IEEE/ACM TCBB, Vol.2.No.4, PP355-365.
[10].Jieming Wu and Wenhu Yu (2009), "Optimizaation and improvement based on k-means clustering algorithm", IEEE research paper, page – 335- 336
[11].Yujun Lin, Ting Luo, Sheng Yao, Kaikai Mo, Tingting Xu and Caiming Zhong (2012), "An improved clustering method based on k-means", IEEE research paper, page – 734-736

[12]. Patrick C. H. Ma, Keith C. C. Chan, Xin Yao, Fellow, IEEE, and David K. Y. Chiu (2006)"An evolutionary clustering algorithm for gene expression microarray data analysis", ", IEEE research paper, page 296.
[13] Muhammad Rukunuddin Ghalib, D. K. Ghosh (2010) "CSTuEPM: An Efficient Clustering Algorithm for Microarray Gene Expression Data",Intl. J. of Advanced Research in Computer Science, Vol I, Issue IV, Pages 370-377.
[14] Muhammad Rukunuddin Ghalib, B.Sathiyabhama (2008) "Mining Gene Expression Data using CST Based Euclidean Proximity Measure" Proc. of Threads' 08 CSI Natl.Conf.

Author's Biography:

Muhammad Rukunuddin Ghalib: He is an Assistant Professor (Senior) in VIT University, Vellore, India and currently in his final stage of Ph.D degree in CSE from Anna University ,Chennai, India. A life member of Computer Society of India (CSI). His areas of interests are data mining, bioinformatics, algorithms and graph theory. He has published several research papers in refereed international journals and conferences.

Priti Sasmal: She is pursuing her final year B.Tech (CSE) in VIT University, Vellore. Her areas of interest in research are data mining in bioinformatics, soft computing and Microarray analysis. She has co-authored several papers in national and international conferences.

Ritwika Ghosh: She is pursuing her final year B.Tech (CSE) in VIT University, Vellore. Her areas of interest in research are Neural Networks, and Fuzzy systems and clustering techniques. She has also co-authored several papers in national and international conferences.

Udisha Pande: She is pursuing her III year B.Tech (CSE) in VIT University, Vellore. Her areas of interest in research are Gene Expression Microarray data analysis and Data mining. She has co-authored several papers in national and international conferences.