STUDY OF CLUSTERING ALGORITHMS FOR GENE EXPRESSION ANALYSIS

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Technology

in

Computer Science and Engineering

By

Sunil Babu Guntur

Department of Computer Science and Engineering
National Institute of Technology, Rourkela
Rourkela, Orissa 769008, India

May 2007

STUDY OF CLUSTERING ALGORITHMS FOR GENE EXPRESSION ANALYSIS

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Technology

in

Computer Science and Engineering

Submitted By

Sunil Babu Guntur

Under guidence of

Prof. S. K. Rath

Department of Computer Science and Engineering
National Institute of Technology, Rourkela
Rourkela, Orissa 769008, India

May 2007

National Institute of Technology, Rourkela

Orissa, 769008, India.

CERTIFICATE

This is to certify that the Thesis entitled "**Study of Clustering Algorithms for Gene Expression Analysis** " submitted by Sri **Sunil Babu Guntur** in partial fulfillment of the requirements for the award of MASTER of Technology Degree in Computer Science and Engineering with specialization in "Computer Science and Engineering" at the National Institute of Technology, Rourkela, is an authentic work carried out by him under my supervision and guidance.

To the best of my knowledge, the matter embodied in the thesis has not been submitted to any other University/ Institute for the award of any degree or diploma.

Date: May 2007

Prof. S. K. Rath

Department of Computer Science and Engg

National Institute of Technology

Rourkela -769008

ACKNOWLEDGMENTS

No thesis is created entirely by an individual, many people have helped to create this thesis and each of their contribution has been valuable. I express my sincere gratitude to my thesis supervisor, Dr. S . K. Rath, Professor, CSE, for his kind and valuable guidance for the completion of the thesis work. His consistent support and intellectual guidance made me energize and innovate new ideas. I am grateful to Dr. S .K . Jena, Professor and Head, CSE for his excellent support during my work. I am also thankful to Prof B. Majhi, Dr. D. P. Mahopatra, Dr A. K. Turuk, Dr. R. Baliarsingh,Mr. B. D. Sahoo of CSE Department, for providing me support and advice in preparing my thesis work. Thanks to all my classmates for their love and support. Last, but not least I would like to thank my parents for supporting me to do complete my masters degree in all ways.

Sunil Babu Guntur
Roll No: 20506001
M. Tech, Computer Science and Engineering
N.I.T. Rourkela-769008.

ABSTRACT

Data Mining refers to as "the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable pattern in data". Based on the type of knowledge that is mined, data mining can be classified in to different models such as Clustering, Decision trees, Association rules, and Sequential pattern and time series. In this thesis work, an attempt has been made to study theoretical background and applications of Clustering techniques in data mining with a special emphasis on analysis of Gene Expression under Bioinformatics.

Bioinformatics is the study of genetic and other biological information using computer and statistical techniques. DNA microarray technology has now made it possible to simultaneously monitor the expression levels of thousands of genes during important biological processes and across collections of related samples. A flood of data means that many of the challenges in biology are now challenges in computing. A first step toward addressing this challenging is the use of clustering technique, which is essential in the data mining process to reveal natural structures and identifying interesting patterns in the underlying data.

In this thesis work, effort has been made to compare between few Clustering algorithms such as: K means, Hierarchical, Self Organization Map(SOM), and Cluster Affinity Search Technique(CAST) with proposed algorithm called CAST+. Strengths and Weaknesses of the above Clustering algorithms are identified and drawbacks like knowing number of clusters before clustering, and taking affinity threshold as input from the users are rectified by the proposed algorithm. Results show that Proposed Algorithm is efficient in comparison with other Clustering algorithms mentioned above.

The Clustering algorithms are compared on the basis of few Evaluation Indices such as Homogeneity Vs separation, and Silhouette width.

Abbreviations

SOM	Self Organizing Map
CAST	Cluster Affiniy Search Technique
ECAST	Enhanced Cluster Affinity Search Technique
CAST+	Proposed Algorithm
DNA	Deoxiribo Nucliec Acid
cDNA	colored Deoxiribo Nucleic Acid
KDD	Knowledge Discovery in Databases
CPU	Central Processing Unit
$\vec{O}$	Vector of object O
RNA	Ribo Nucleic Acid
mRNA	messenger Ribo Nucleic Acid
GB	Gene Bank

Contents

List of Figures

Chapter 1

Introduction

1.1 Data Mining and Knowledge Discovery

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making.

Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to as "the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable pattern in data". While data mining and knowledge discovery in databases (KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. Figure 1.1 shows data mining as a step in an iterative knowledge discovery process.

The task of the knowledge discovery and data mining process is to extract knowledge from data such that the resulting knowledge is useful in a given application. The Knowledge Discovery process in Databases comprises of a few steps leading from raw data collections to some form of retrieving new knowledge. The iterative process consists of the following steps:

1. Data cleaning: Also known as data cleansing, it is a phase in which noisy data and irrelevant data are removed from the collection.

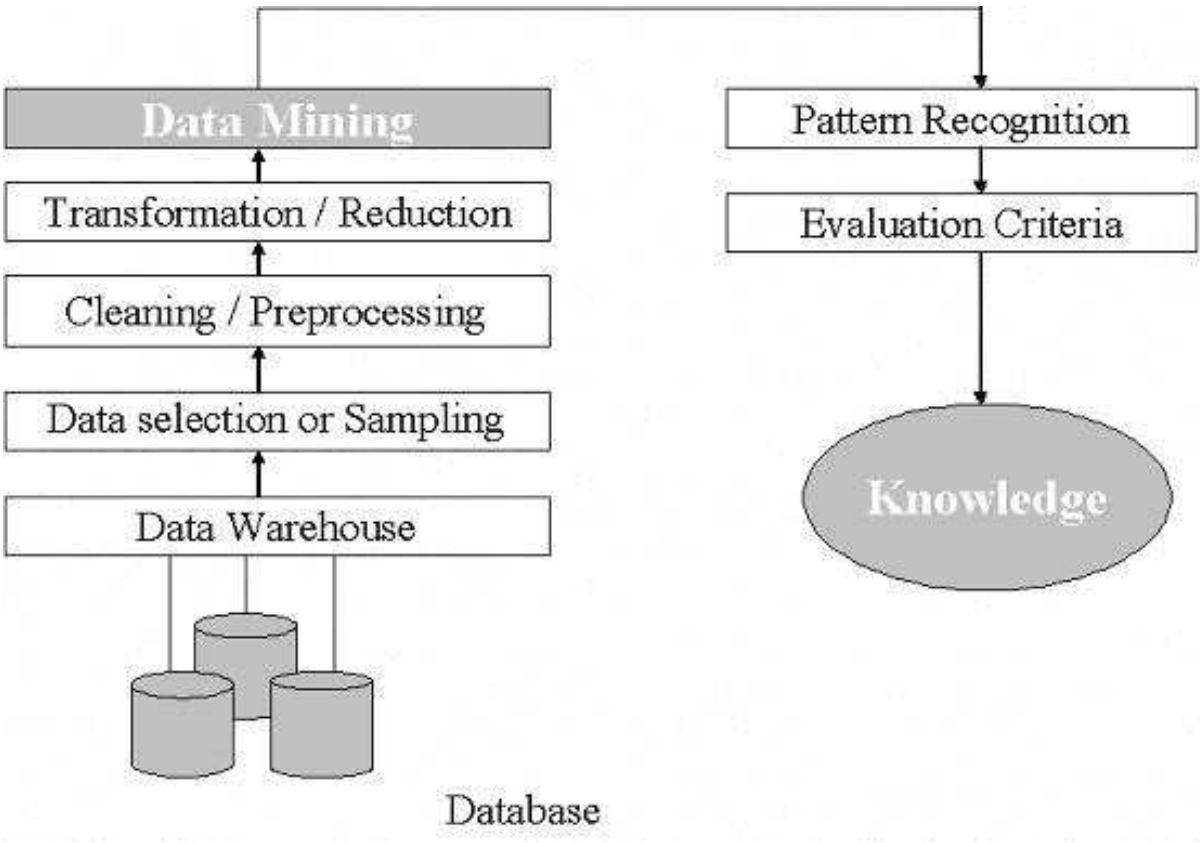

Figure 1.1: An Overview of the Steps Comprising the KDD Process

2. Data integration: At this stage, multiple data sources, often heterogeneous, may be combined in a common source.

3. Data selection: At this step, the data relevant to the analysis is decided on and retrieved from the data collection.

4. Data mining: It is the crucial step in which clever techniques are applied to extract data patterns potentially useful.

5. Pattern evaluation: In this step, strictly interesting patterns representing Knowledge is identified based on given measures.

6. Knowledge representation: Is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

It is common to combine some of these steps together. For instance, data cleaning and data integration can be performed together as a pre-processing phase to generate a data warehouse. Data selection and data transformation can also be combined where the consolidation of the data is the result of the selection, or, as for the case of data warehouses, the selection is done on transformed data.

The KDD is an iterative process. Once the discovered knowledge is presented to the user, the evaluation measures can be enhanced, the mining can be further refined, new data can be selected or further transformed, or new data sources can be integrated, in order to get different, more appropriate results.

1.2 Data Mining Models

There are several data mining models, some of these are narrated below which are conceived to be important in the area of "Data Mining".

- Clustering: It segments a large set of data into subsets or clusters. Each cluster is a collection of data objects that are similar to one another with the same cluster but dissimilar to objects in other clusters.

- Classification: Decision trees, also known as classification trees, are a statistical tool that partitions a set of records into disjunctive classes. The records are given as tuples with several numerics and categorical attributes with one additional attribute being the class to predict. Decision trees algorithm differs in selection of variables to split and how they pick the splitting point.

- Association Mining: It uncovers interesting correlation patterns among a large set of data items by showing attribute value conditions that occur together frequently.

- Sequential Pattern and Time series: Sequential pattern and time -series mining looks for patterns where one event (or value) leads to another later event (or value). One example is that after the inflation rate increases, the stock market is likely to go down.

1.3 Bioinformatics

Bioinformatics is the study of genetic and other biological information using computer and statistical techniques. The explosive growth in the amount of biological data demands the use of computers for the organization, the maintenance and the analysis of these data. The aims of Bioinformatics are:

1. The organization of data in such a way that allows researchers to access existing information and to submit new entries as they are produced.

2. The development of tools that help in the analysis of data.

3. The use of these tools to analyze the individual systems in detail, in order to gain new biological insights.

Application of Data Mining techniques for Bioinformatics is vast area of study. It includes

- **Gene Expression in Datamining** : Gene expression analysis is the use of quantitative mRNA-level measurements of gene expression (the process by which a gene's coded information is converted into the structural and functional units of a cell) in order to characterize biological processes and elucidate the mechanisms of gene transcription.

- **Data mining in genomics**: Genomics is the study of an organism's genome and deals with the systematic use of genome information to provide new biological knowledge.

- **Data Mining in Proteomics**: Proteomics is the large-scale study of proteins, particularly their structures and functions.

1.4 Introduction to Microarray Technology

Compared with the traditional approach to genomic research, which is focused on the local examination and collection of data on single genes, microarray technologies have now made it possible to monitor the expression levels for tens of thousands of genes in parallel. The two

major types of microarray experiments are the cDNA microarray and oligonucleotide arrays (abbreviated oligo chip). Despite differences in the details of their experiment protocols, both types of experiments involve three common basic procedures :

1. **Chip manufacture**: A microarray is a small chip (made of chemically coated glass, nylon membrane or silicon), onto which tens of thousands of DNA molecules (probes) are attached in fixed grids. Each grid cell relates to a DNA sequence.

2. **Target preparation, labeling and hybridization**: Typically, two mRNA samples (a test sample and a control sample) are reverse-transcribed into cDNA (targets), labeled using either fluorescent dyes or radioactive isotopics, and then hybridized with the probes on the surface of the chip.

3. **The scanning process**: Chips are scanned to read the signal intensity that is emitted from the labeled and hybridized targets.

Generally, both cDNA microarray and oligo chip experiments measure the expression level for each DNA sequence by the ratio of signal intensity between the test sample and the control sample, therefore, data sets resulting from both methods share the same biological semantics. In this thesis work, unless explicitly stated, we will refer to both the cDNA microarray and the oligo chip as microarray technology and term the measurements collected via both methods as gene expression data.

1.4.1 Gene expression data

A microarray experiment typically assesses a large number of DNA sequences (genes, cDNA clones, or expressed sequence tags) under multiple conditions. These conditions may be a time series during a biological process or a collection of different tissue samples (e.g., normal versus cancerous tissues). In this thesis work, emphasis is given on the cluster analysis of gene expression data without making a distinction among DNA sequences, which will uniformly be called "genes". Similarly, it is referred to all kinds of experimental conditions as "samples", if no confusion will be caused. A gene expression data set from a microarray experiment can be represented by a real-valued expression matrix $M = \{W_{ij} \mid 1 \leq i \leq n, 1 \leq j \leq m\}$ as shown in Figure 1.2, where the rows ($G = \{g_1 ... g_n\}$) form the

expression patterns of genes, the columns ($S = \{S_1 ... S_m\}$) represent the expression profiles of samples, and each cell is the measured expression level of gene i in sample j. Table 1.1 includes some notation that is used in thesis work.

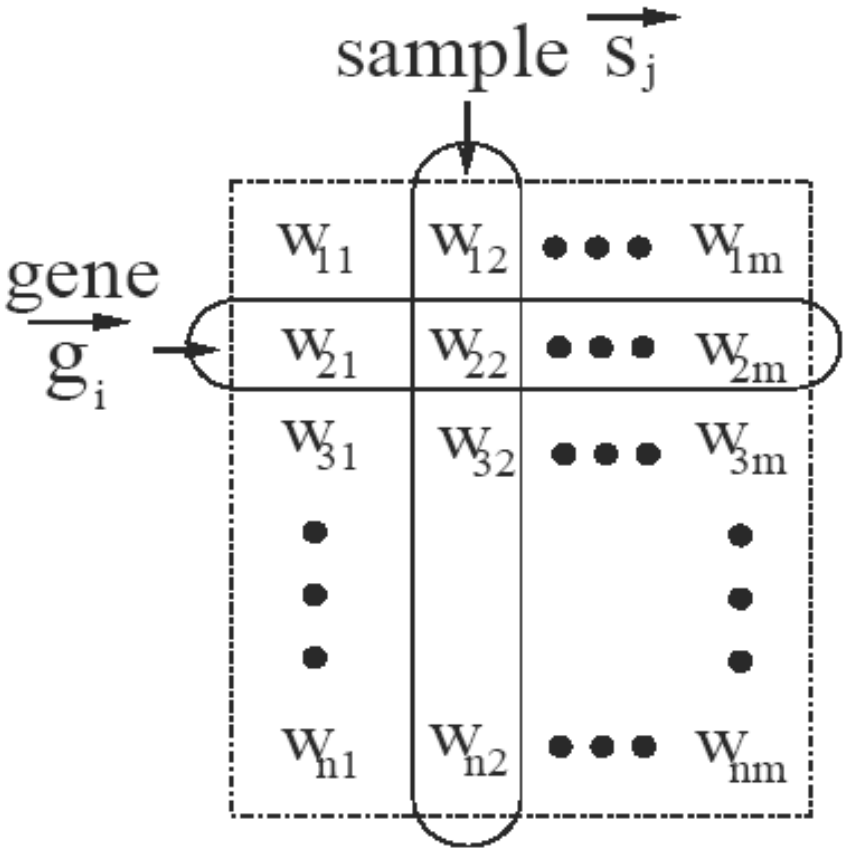

Figure 1.2: A Gene Expression Matrix

1.5 Motivation

Analysis of microarrays presents a number of unique challenges for data mining. Typical data mining applications in domains like banking or web, have a large number of records (thousands and sometimes millions), while the number of fields is much smaller (at most

n	number of genes
m	number of samples
M	a gene expression matrix
w_{ij}	each cell in gene expression matrix
g_i	a gene
S_j	a sample
$G, G_0, ...$	a set of genes
$S, S_0, ...$	a set of samples

Table 1.1: Notation in this Thesis

several hundred). In contrast, a typical microarray data analysis study may have only a small number of records (less than a hundred), while the number of fields, corresponding to the number of genes, is typically in thousands. Given the difficulty of collecting microarray samples, the number of samples is likely to remain small in many interesting cases. It need especially robust methods to validate the models and assess their likelihood.

Clustering is a potential area of Data mining that can be dealt with the large data simultaneous. Due to special characteristics of gene expression data and particular requirements from biological domain. Gene based clustering presents several challenges.

- The purpose of clustering gene expression data is to reveal the natural data structures and gain some initial insights regarding data distribution. Therefore, a good clustering algorithm should depend as little as possible on prior knowledge, which is usually not available before cluster analysis.

- Due to complex procedures of microarray experiments, gene expression often contain a huge amount of noise. Therefore clustering algorithms for gene expression data should be capable of extracting useful information from a high level of background noise.

- Users of microarray data may not only be interested in the clusters of genes, but also be interested in the relationship between the clusters, and the relation between the genes within the same cluster.

1.6 Organization of thesis work

This thesis work is divided into five chapters. This Chapter gives introduction to what is data mining, bioinformatics and our motivation towards clustering the gene expression data generated by microarray. **Chapter 2** presents the literature survey that is carried out. **Chapter 3** presents clustering concepts and different types of methods used in clustering gene expression data. **Chapter 4** presents the Cluster Affinity Search Technique(CAST) and rectified its drawbacks using proposed algorithm, and also showed the results proving the proposed algorithm is better than existing. And the **last chapter** presents the conclusion and proposals for possible extension of this thesis work.

1.7 Conclusion

Biological data analysis has been increasing at explosive rates due to improvements of existing technologies and the introduction of new ones such as the microarrays. The explosive growth in the amount of biological data demands the use of computers for the organization, the maintenance and the analysis of these data. Consequently, data mining in bioinformatics has become a research area with increasing importance. Data mining is the confluence of many fields such as Database, Statistics, and Artificial Intelligence etc.

Chapter 2

Literature Review

According to *Reichhardt T(1999)* , Biological data are being produced at a phenomenal rate [31]. For example as of April 2001, the GenBank repository of nucleic acid sequences contained 1,15,46,000 entries and the SWISSPROT database of protein sequences contained 95,320 entries. On an average, these databases are doubling in size every 15 months. In addition, since the publication of the H. influenza genome [14], complete sequences for nearly 300 organisms have been released, ranging from 450 genes to over 100,000. At the same time, there have been major advances in the technologies that supply the initial data. Anthony Kervalage of Celera recently cited that an experimental laboratory can produce over 100 gigabytes of data a day with ease [20]. Figure 2.1 shows the growth of DNA sequence in Gen Bank during a period from 1982 to 2003. This incredible processing power has been matched by developments in computer technology; the most important areas of improvements have been in the CPU speed, disk storage and Internet, allowing faster computations, better data storage and revolutionalised the methods for accessing and exchanging data.

According to *Dr. Diego Kuonen*, Data mining and bioinformatics are fast expanding research frontiers. It is important to examine what are the important research issues in bioinformatics and develop new data mining methods for scalable and effective analysis. The effective interactions and collaborations between these two fields have just started and lots of exciting results will appear in the feature. Bioinformatics and Data mining will inevitably grow toward each other because bioinformatics will not become knowledge

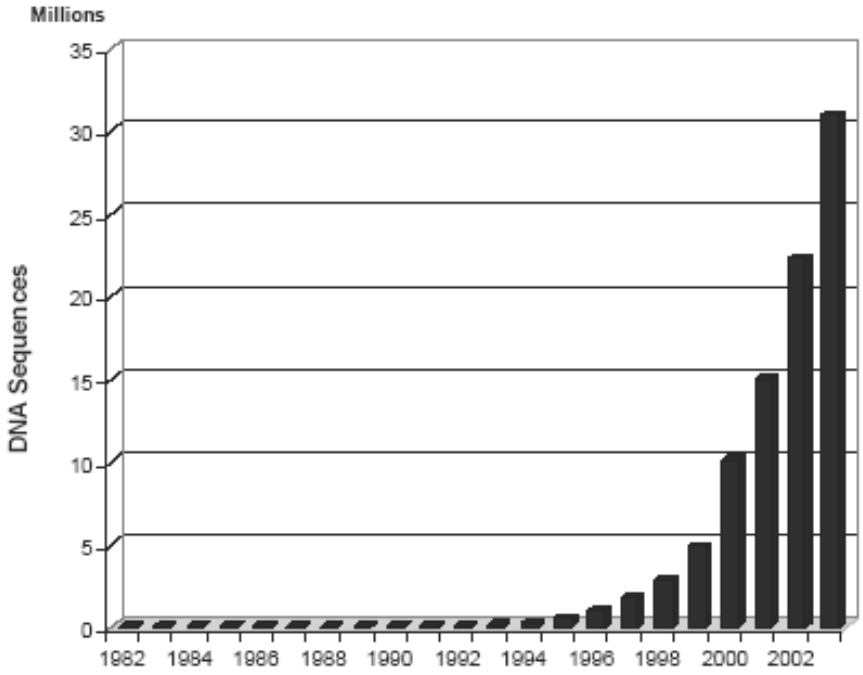

Figure 2.1: Figure showing the growth of gen bank

discovery without statistical datamining and thinking[8].

According to *P.Tamayo*, The main types of data analysis needed to for biomedical applications include:

- **Clustering**: finding new biological classes or refining existing ones [10].
- **Gene Selection**: In mining terms this is a process of attribute selection, which finds the genes most strongly related to a particular class.
- **Classification**: classifying diseases or predicting outcomes based on gene expression patterns, and perhaps even identifying the best treatment for given genetic signature.

One important clustering task is to identify groups of co expressed genes recognize coherent expression patterns. Microarrays are a revolutionary new technology with great potential to provide accurate medical diagnostics, help find the right treatment and cure for many diseases and provide a detailed genome-wide molecular portrait of cellular states.

Houle et al. (2000) refer to a classification of three successive levels for the analysis of biological data, that is identified on the basis of the central dogma of molecular biology: Application of Data Mining techniques for Bioinformatics is vast area to study. It includes [11]

- **Gene expression in Datamining**: Gene expression analysis is the use of quantitative mRNA-level measurements of gene expression (the process by which a gene's coded information is converted into the structural and functional units of a cell) in order to characterize biological processes and elucidate the mechanisms of gene transcription
- **Data mining in genomics**: Genomics is the study of an organism's genome and deals with the systematic use of genome information to provide new biological knowledge.
- **Data mining in proteomics**: Proteomics is the large-scale study of proteins, particularly their structures and functions.

Clustering techniques have proven to be helpful to understand gene function, gene regulation, cellular processes, and subtypes of cells[3]. Genes with similar expression patterns (co-expressed genes) can be clustered together with similar cellular functions. This approach may further understanding of the functions of many genes for which information has not been previously available [13]. Furthermore, co-expressed genes in the same cluster are likely to be involved in the same cellular processes, and a strong correlation of expression patterns between those genes indicates co-regulation. Searching for common DNA sequences at the promoter regions of genes within the same cluster allows regulatory motifs specific to each gene cluster to be identified and cis-regulatory elements to be proposed [15]. The inference of regulation through the clustering of gene expression data also gives rise to hypotheses regarding the mechanism of the transcriptional regulatory network. Finally, clustering different samples on the basis of corresponding expression profiles may reveal sub-cell types which are hard to identify by traditional morphology-based approaches.

According to *Kjersti Aas*[18], DNA microarray makes it possible to quickly, efficiently and accurately measure the relative representation of each mRNA species in the total cellular mRNA population. A DNA experiment consists of measurements of the relative representation of a large number of mRNA species (typically thousands or tens of thousands) in a set of related biological samples, e.g. time points taken during a biological process or clinical samples taken from different patients. Each experiment sample is compared to a common reference sample and the result for each gene is the ratio of the results of such experiments are represented in a table, with each row representing a gene, each column a sample, and each cell the log(base - 2) transformed expression ratio of the appropriate gene in the appropriate sample.

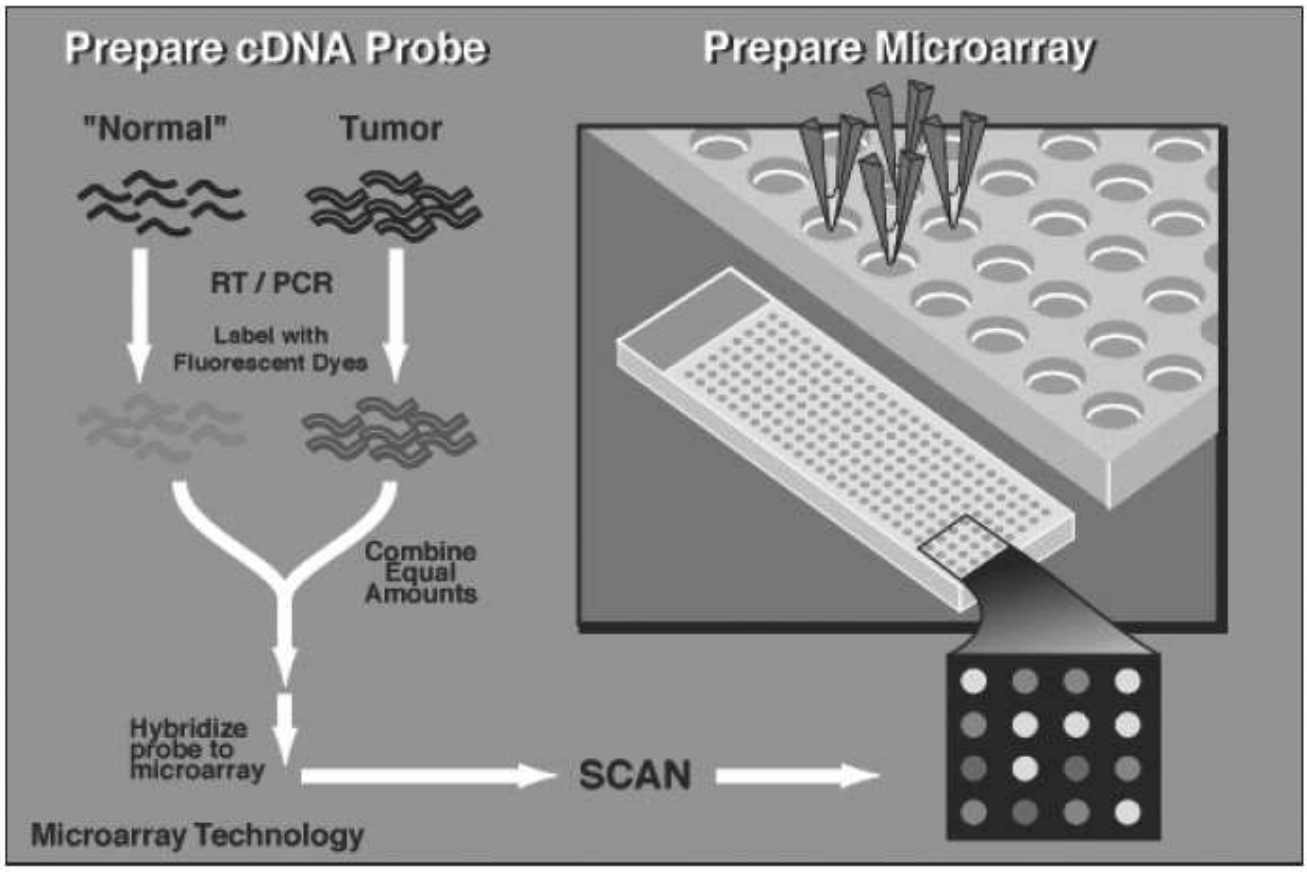

Figure 2.2: Figure Explaining the micro array process

The microarray process is shown in Figure 2.2. The DNA sample (which may be several thousands) are fixed to a glass slide, each at a known position in the array. A

target sample and a reference sample are labeled with red and green dyes, respectively, and each is hybridised on the slide. Using a fluorescent microscope and image analysis, the log(green/red) intensities of mRNA hybridising at each site is measured. The result is a few thousand numbers, typically ranging form -4 to 4, measuring the expression level of each gene in the experimental sample relative to the reference sample. Positive values indicate higher expression in the target versus the reference, and vice verses for negative values.

According to *Rui Xu, and Donald Wunsch* [27], data analysis plays an indispensable role for understanding various phenomena. Cluster analysis, primitive exploration with little or no prior knowledge, consists of research developed across a wide variety of communities. Cluster analysis is not a one-shot process. It needs a series of trials and repetitions. Moreover, there are no universal and effective criteria to guide the selection of features and clustering schemes. Validation criteria provide some insights on the quality of clustering solutions. But how to choose the appropriate criterion is still a problem, which requires more efforts. Clustering has been applied in a wide variety of fields, ranging from engineering (machine learning, artificial intelligence, pattern recognition, mechanical engineering, electrical engineering), computer sciences (web mining, spatial database analysis, textual document collection, image segmentation), life and medical sciences (genetics, biology, micro biology, paleontology, psychiatry, clinic, pathology), to earth sciences (geography. geology, remote sensing), social sciences (sociology, psychology, archeology), and economics (marketing, business).

Accordingly, clustering is also known as numerical taxonomy, learning without a teacher (or unsupervised learning), typological analysis and partition. The diversity reflects the important position of clustering in scientific research. On the other hand, it causes confusion, due to the differing terminologies and goals. Clustering algorithms developed to solve a particular problem, in a specialized field, usually make assumptions in favor of the application of interest. These biases inevitably affect performance in other problems that do not satisfy these premises.

According to *Daxin Jiang, Chun Tang, and Aidong Zhang* [3], Elucidating the patterns hidden in gene expression data offers a tremendous opportunity for an enhanced understanding of functional genomics. Many conventional clustering algorithms have been adapted or directly applied to gene expression data, and also new algorithms have recently been proposed specifically aiming at gene expression data. These clustering algorithm relevant groups of genes and samples. In this thesis paper it is explained that gene expression clustering is divide into gene based clustering and sample based clustering. It is explained that K-Means a partition based clustering where no of clusters has to be mentioned earlier, in hierarchical clustering it produces a dendogram, in SOM it requires the grid structure earlier before clustering, and for CAST it requires the affinity threshold as input parameter.

According to *Bendor et.al*[2] Current approaches to clustering gene expression patterns utilize hierarchical methods (constructing phylogenetic trees) or methods that work for Euclidean distance metrics (e.g K-Means). We take a graph theoretic approach, and make no assumptions on the similarity function or the number of clusters sought. The cluster structure is produced directly, without involving an intermediate tree stage.

Chapter 3

Cluster formation algorithm

3.1 Introduction

Clustering plays a vital role in the Gene Expression Analysis. In this chapter we will first discuss the concepts of *clustering*, and later discuss the various algorithms used such as K-Means, SOM, hierarchical clustering algorithms and their pros and cons.

3.2 Cluster formation in Data Mining

Clustering is the process of grouping data objects into a set of disjoint classes, called *clusters*, so that objects within the same class have high similarity to each other, while objects in separate classes are more dissimilar. Clustering is an example of *unsupervised classification*. "Classification" refers to a procedure that assigns data objects to a set of classes. "Unsupervised" means that clustering does not rely on predefined classes and training examples while classifying the data objects. Thus, clustering is distinguished from pattern recognition or the areas of statistics known as discriminate analysis and decision analysis, which seek to find rules for classifying objects from a given set of pre-classified objects.

3.3 Categories of Gene Expression Data Clustering

Recently, a typical micro array experiment contains 10^3 to 10^4 genes, and this number is expected to reach the order of 10^6. However, the number of samples involved in micro array experiment is generally less than 100. One of the characteristics of gene expression data is that it is meaningful to cluster both genes and samples. Clustering gene expression data can be categorized into two groups.

3.3.1 Gene based clustering

In this type of clustering genes are treated as the objects, while samples as the features. The purpose of gene-based clustering is to group together co expressed genes which indicate co-function and co-regulation.

3.3.2 Sample based clustering

In this type of clustering samples are the objects and genes are features. Within a gene expression matrix, there are usually particular macroscopic phenotypes of samples related to some diseases or drug effects, such as diseased samples, normal samples or drug treated samples. The goal of sample based clustering is to find the phenotype structures or sub-structures of the sample.

The distinction of gene based clustering and sample based clustering is based on different characteristics of clustering tasks for gene expression data [3].

In this thesis work it has been discussed the clustering algorithms used for grouping the genes, i.e., we have studied gene based clustering.

3.4 Proximity measurement for gene expression data

Proximity measurement measures the similarity(distance) between two data objects. Gene expression data objects, no matter genes or samples, can be formalized as numerical vectors $\vec{O_i} = \{o_{i,j} | 1 \leq j \leq p\}$, where $o_{i,j}$ is the value of the j^{th} feature for the i^{th} data

object and p is the number of features. The proximity between two objects O_i and O_j is measured by a *proximity function* of corresponding vectors $\vec{O_i}$ and $\vec{O_j}$.

3.4.1 Euclidean Distance

EuclideanDistance is one of the most commonly used methods to measure the distance between two data objects. The distance between objects $\vec{O_i}$ and $\vec{O_j}$ in p-dimensional space is defined as

$$Eucledian(O_i, O_j) = \sqrt{\sum_{d=1}^{p}(o_{id} - o_{jd})^2} \tag{3.1}$$

However, for gene expression data, the overall shapes of gene expression patterns are of greater interest than the individual magnitudes of each feature.

3.4.2 Pearson's correlation coefficient

Pearson's correlation coefficient, which measures the similarity between the shapes of two expression patterns. Given two data objects O_i and O_j, pearson's correlation coefficient is defined as

$$pearson(O_i, O_j) = \frac{\sum_{d=1}^{p}(o_{id} - \mu_{oi})(o_{jd} - \mu_{oj})}{\sqrt{\sum_{d=1}^{p}(o_{id} - \mu_{oi})^2}\sqrt{\sum_{d=1}^{p}(o_{jd} - \mu_{oj})^2}} \tag{3.2}$$

where μ_{oi} and μ_{oj} are the means for $\vec{O_i}$ and $\vec{O_j}$ respectively. Pearsons correlation coefficient views each object as a random variable with p observations and measures the similarity between two objects by calculating the linear relationship between the distributions of the two corresponding random variables[16].

Hence, in this entire thesis work Euclidean distance is used as the proximity measure.

3.5 Clustering Paradigms

Based on the method of clustering, the clustering algorithms are divided into two paradigms [1].

- Partitioning clustering: in which the database is partitioned into a predefined number of clusters.
  for example:K-Means, K-mediods etc.

- Hierarchical clustering do sequence of partitions, in which each partition is nested into the next partition in the sequence Based on the approach Hierarchical clustering is further divided into two types

  - *Agglomerative clustering* technique starts with as many clusters as there are records, with each cluster having only one record. Then pair of clusters successively merged.This is also called as bottom up approach.
    for example:single linkage hierarchical algorithm, complete linkage hierarchical algorithm etc.

  - *Divisive clustering* takes the opposite approach from agglomerative techniques.In this approach the algorithm starts with one cluster containing all the data objects, and at each step split a cluster until only singleton clusters of individual objects remain.
    for example:Graph theoretical algorithms, CAST etc.

3.6 Clustering Algorithms

In this section different algorithms that is studied in this thesis work is discussed in brief.

3.6.1 K-Means

The K-Means algorithm is a typical partition-based clustering method. Given a prespecified number K, the algorithm partitions the data set into K disjoint subsets which optimize the following objective function:

$$E = \sum_{i=1}^{k} \sum_{O \epsilon C_i} |O - \mu_i|^2. \tag{3.3}$$

Here, O is a data object in cluster C_i and mu_i is the centroid (mean of objects) of C_i. Thus, the objective function E tries to minimize the sum of the squared distances of objects

from their cluster centers.

Algorithm

1. The K-Means algorithm accepts the "number of clusters" to group data into, and the dataset to cluster the input values.
2. The K-Means algorithm then creates the first k initial clusters from the data set
3. The K-Means algorithm calculates the arithmetic mean of each cluster formed in the data set. The arithmetic mean is the mean of all the individual records in the cluster.
4. Next K-Means assigns each record in the dataset to only one of the initial clusters. Each record is assigned to the nearest cluster using proximity measure like Euclidean distance.
5. K-Means reassigns each record in the dataset to the most similar cluster and recalculates the arithmetic mean of the clusters in the dataset.
6. K-Means reassigns each record in the dataset to only one of the new clusters formed
7. The preceding steps are repeated until "stable clusters" are formed and the K-Means clustering is completed

The K-Means algorithm is simple and fast. The time complexity of K-Means is O(l*m*n), where l is the number of iterations and K is the number of clusters, m is the number of genes and n is the number of samples. Our empirical study has shown that the K-Means algorithm typically converges in a small number of iterations. However, it also has several drawbacks as a gene-based clustering algorithm. First, the number of gene clusters in a gene expression data set is usually unknown in advance. To detect the optimal number of clusters, users usually run the algorithms repeatedly with different values of K and compare the clustering results.

For a large gene expression data set which contains thousands of genes, this extensive parameter fine-tuning process may not be practical. Second, gene expression data typically contain a huge amount of noise; however, the K-Means algorithm forces each gene into a cluster, which may cause the algorithm to be sensitive to noise.

3.6.2 SOM

The Self-Organizing Map (SOM) was developed by Kohonen in 1997[12], on the basis of a single layered neural network. The data objects are presented at the input, and the output neurons are organized with a simple neighborhood structure such as a two dimensional p*q grid. Each neuron of the neural network is associated with a reference vector, and each data point is "mapped" to the neuron with the "closest" reference vector. In the process of running the algorithm, each data object acts as a training sample which directs the movement of the reference vectors towards the denser areas of the input vector space, so that those reference vectors are trained to fit the distributions of the input data set. When the training is complete, clusters are identified by mapping all data points to the output neurons.

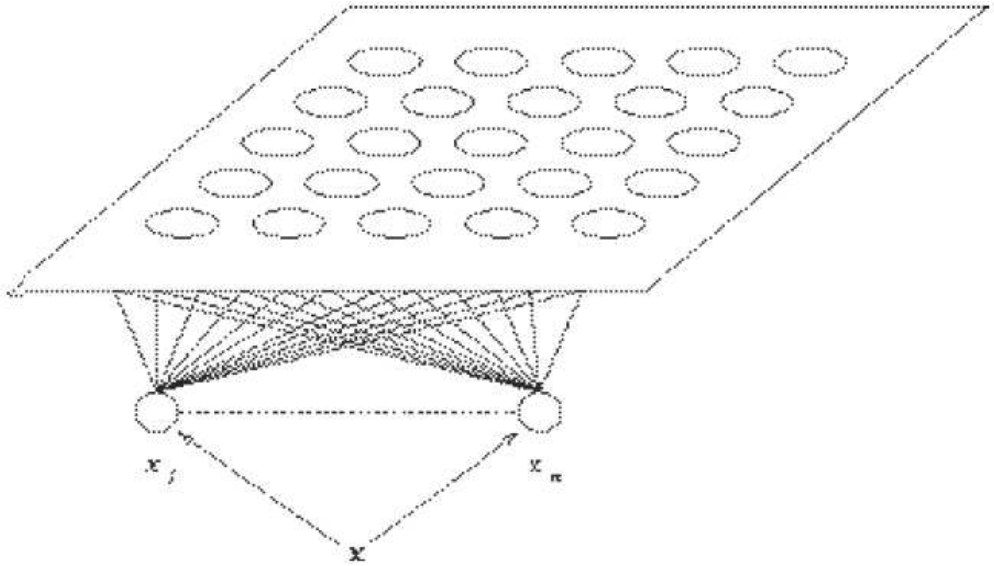

Figure 3.1: Schematic representation of a self-organizing map method

The neuron training process of SOM provides a relatively more robust approach than K-Means to the clustering of highly noisy data [12]. However, SOM requires users to input the grid structure of the neuron map. This parameter is preserved through the training process;hence, improperly-specified parameter will prevent the recovering of the natural

cluster structure. Furthermore, if the data set is abundant with irrelevant data points, such as genes with invariant patterns, SOM will produce an output in which this type of data will populate the vast majority of clusters. In this case, SOM is not effective because most of the interesting patterns may be merged into only one or two clusters and cannot be identified.

3.6.3 Hierarchical Clustering

Hierarchical clustering generates a hierarchical series of nested clusters which can be graphically represented by a tree, called *dendrogram*. The branches of a dendrogram not only record the formation of the clusters but also indicate the similarity between the clusters. By cutting the dendrogram at some level, we can obtain a specified number of clusters. By reordering the objects such that the branches of the corresponding dendrogram do not cross, the data set can be arranged with similar objects placed together.

The hierarchical clustering scheme:

Let S=$\{S_{i,j}\}$ is the input similarity matrix, where $S_{i,j}$ indicates similarity between two data objects based on Euclidean distance.

Algorithm:

1. Find a minimal entry s(i, j) in S, and merge clusters i and j.
2. Modify S by deleting rows and columns i, j and adding a new row i and column j, with their dissimilarities defined by:

$$s(k, i \cup j) = s(i \cup j, k) = \alpha_i s(k, i) + \alpha_j s(k, j) + \gamma |s(k, i) - s(k, j)| \tag{3.4}$$

3. If there is more than one cluster, then go to Step 1.

Common variants of this scheme, obtained for appropriate choices of the $\alpha - s$ and γ parameters, are the following:

$$singlelinkage : s(k, i \cup j) = min = \{s(k, i), s(k, j)\} \tag{3.5}$$

$$completelinkage : s(k, i \cup j) = max\{s(k, i), s(k, j)\} \tag{3.6}$$

$$averagelinkage : s(k, i \cup j) = (n_i d(k, i) + n_j d(k, j))/(ni + nj), \tag{3.7}$$

where n_i denotes the number of elements in cluster i.

Hierarchical clustering not only groups together genes with similar expression pattern but also provides a natural way to graphically represent the data set. The graphic representation allows users a thorough inspection of the whole data set and obtain an initial impression of the distribution of data. However, the conventional agglomerative approach suffers from a lack of robustness [19] , i.e., a small perturbation of the data set may greatly change the structure of the hierarchical dendrogram. Another drawback of the hierarchical approach is its high computational complexity. To construct a complete dendrogam (where each leaf node corresponds to one data object, and the root node corresponds to the whole data set), the clustering process should take $\frac{n^2-n}{n}$ merging (or splitting) steps. The time complexity for a typical agglomerative hierarchical algorithm is $O(n^2 logn)$ [17]. If a wrong decision is made in the initial steps, it can never be corrected in the subsequent steps.

3.7 Conclusion

In this chapter clustering algorithms like K-Means, SOM, and hierarchical algorithms were studied and observed that K-Means requires number of clusters before clustering where it is not known earlier for gene expression data. For SOM, grid structure of the neuron map has to be mentioned earlier. The time complexity of hierarchical clustering is very high. To overcome these problems, the next chapter discusses a divisive graph theoretical algorithm CAST and how its faults are overcome by the proposed algorithm of this thesis work.

Chapter 4

Proposed Algorithm

4.1 Introduction

Several data mining solutions have been presented for Bioinformatics [22], [23], and [5]. Cluster analysis received significant attention in the area of gene expression. It allows the identification of groups of similar objects in multidimensional space. In this chapter it has been discussed a graph-based clustering algorithm, CAST, its disadvantages, and how they are overcome in the proposed algorithm .

4.2 Cluster Affinity Search Technique(CAST)

On study of clustering algorithms with an emphasis on graph theoretic approaches[21], it is observed that for any micro array data analysis of gene expression patterns with clustering algorithms involve the following steps:

- **Determination of gene expression data**: The data can be represented by a real-valued *expression matrix* I where I_{ij} is the measured expression level of gene i in experiment j. The i^th row of the matrix is a vector forming the *expression pattern* of gene i.

- **Calculation of a similarity matrix S**: In this matrix, the entry S_{ij} represents the similarity of the expression patterns for genes i and j. Many possible similarity

measures can be used here. A good choice of measure depends on the nature of the biological question and on the technology that was used to obtain the data.

- **Clustering the genes based on the similarity data or the expression data**: Genes that belongs to the same cluster should have similar expression patterns, while different clusters should have distinct, well-separated patterns.

- Representation of the constructed solution.

Several techniques were previously used in clustering gene expression data such as Hierarchical clustering techniques[13], Self-Organizing Maps used by Tamayo et. al [12], and K-Means [1]. In this thesis work it has been discussed a novel algorithm for the problem of clustering gene expression patterns. Unlike the hierarchical approaches mentioned above, our algorithm doesn't build a tree of clusters. Clusters are built and portrayed as unrelated entities. In contrast to self-organizing maps, it does not assume an initial spatial structure, but determines the cluster and structure based on the data. Unlike K-Means it doesn't require the number of clusters earlier before clustering.

4.2.1 Experimental Representation of Data set

Formally, a set of genes can be viewed as a set of vectors $V = \{v_i, v_2, v_3, ..., v_m\}$ with each expression level of a given experiment, x_j, being the components in the vector $v_i = (x_1, x_2, x_3, , x_n)$, where m is the number of genes in the experiments and n is the number of experiments.Figure 4.1is an example gene expression matrix.(This works equally well when the experiments form the vectors). These vectors can then be viewed as points in n dimensional space and a similarity measurement between points can be calculated and stored in a m *by* m similarity matrix M. Where M_{ij} is the distance (similarity) measure between gene i and gene j. There are several similarity measures, e.g., Euclidean distance and Pearson correlation. Figure 4.2 shows the similarity matrix of the given gene expression matrix generated using Euclidean distance. Then algorithms used for clustering is run on the similarity matrix to group the members of V into clusters, which attempts to maximize the intra-cluster similarity and minimize the inter-cluster similarity.

	exp1	exp2	exp3	exp4	exp5	exp6	exp7	exp8	exp9	exp10
gene1	0	0.39127	2.5988	1.2616	0	0	3.6528	0	0	0.4636
gene2	0.7589	0.19452	0	3.2465	0	1.2828	1.019	0	0	1.1501
gene3	0	0	2.1117	0	0	3.8196	0	0	0.55593	0.22982
gene4	0.5777	0	2.8689	0	0	0.4377	1.063	2.6607	0.31721	0.8756
gene5	0	0.3548	0	0	0.68914	0	0	0	3.5689	2.4567
gene6	2.1678	0.7364	0	3.4144	1.392	0.50472	1.4681	4.7503	1.8895	0
gene7	0	0	0	1.3995	0.21102	0.48775	1.2247	0	0	0.18321
gene8	0	0.2165	0	1.2187	0	0.7198	2.1415	0	2.1826	0.69041
gene9	3.4589	0	0.6535	0	3.6041	0	0	1.1456	0	0.42756
gene10	3.6578	0	2.5754	1.8249	0.39757	0	0	1.3696	0	1.0123

Figure 4.1: A Gene Expression Matrix

	gene1	gene2	gene3	gene4	gene5	gene6	gene7	gene8	gene9	gene10
gene1	0	4.5124	5.5025	4.0497	6.2349	7.0372	3.6308	3.7951	6.7186	5.434
gene2	4.5124	0	4.9236	5.1781	5.3614	5.7017	2.3831	3.3526	5.9835	4.6725
gene3	5.5025	4.9236	0	4.5856	5.8039	7.7998	4.4016	4.8004	6.5825	5.871
gene4	4.0497	5.1781	4.5856	0	5.5353	5.7158	4.2726	4.6832	5.4879	4.0211
gene5	6.2349	5.3614	5.8039	5.5353	0	7.1294	4.6858	3.4821	6.2595	6.3445
gene6	7.0372	5.7017	7.7998	5.7158	7.1294	0	6.0766	5.9461	6.1923	5.5986
gene7	3.6308	2.3831	4.4016	4.2726	4.6858	6.0766	0	2.4575	5.3826	4.9527
gene8	3.7951	3.3526	4.8004	4.6832	3.4821	5.9461	2.4575	0	6.1775	5.6949
gene9	6.7186	5.9835	6.5825	5.4879	6.2595	6.1923	5.3826	6.1775	0	4.2116
gene10	5.434	4.6725	5.871	4.0211	6.3445	5.5986	4.9527	5.6949	4.2116	0

Figure 4.2: Similarity Matrix for the above gene expression matrix

4.2.2 CAST

The Cluster affinity search technique (CAST) developed by Ben-Dor et. al., 1999 [2] takes a graph theoretic approach that relies on the concept of a clique graph and uses a divisive clustering approach. A clique graph is an undirected graph that is the union of disjoint complete graphs. Thus, the model assumes that there is a "true biological partition of the genes into disjoint clusters bases on the functionality of the genes. The clique graph would then be composed of clusters (cliques) of genes (vertices) whose interconnections (edges) are present or not present corresponding to their respective similarity measures (i.e. if two genes are similar there is an edge between them). So, ideally, the genes would form

subgraphs (cliques) where every gene would be completely similar to every other gene in the clique and completely dissimilar to every gene not in the clique. Thereby, producing a clique graph G of $U = \{u_1, u_2, , u_n\}$ vertices partitioned such that every clique S_i contains edges connecting every vertex $u \epsilon S_i$ to every other $u \epsilon S_i$ and no edges connecting any $u \epsilon S_i$ to any $u \epsilon U \setminus S_i$. This, model can be applied just as easily to experiments instead of genes. Where, the experiments become the vertices and one experiment is linked to another based on the similarity of their respective patterns.

It is very probable that a set of gene (or experiment) vectors will tend to have a similarity gradient across other vectors and the high incidence rate of errors in micro-array technology, the ideal clique graph would be impossible to generate, or, at the very lease, would create very small clusters. So small, in fact, that many would contain single data points, and therefore defeat the purpose of the algorithm. Thus, an approximation of the preceding model is called for.

The CAST algorithm takes as input an $n-by-n$ similarity matrix S where $(S(i,j) \epsilon [1,0])$ and an affinity threshold T is to be defined by the user. T is used to determine node membership to a cluster. The pseudo code for both CAST and proposed algorithm is shown in fig 4.3, 4.4, 4.5, 4.6.

4.3 Proposed Algorithm

It is studied that the main draw back of CAST algorithm is taking *affinity threshold* as input, which determines the size and number of clusters produced. In this thesis work we have proposed an *affinity threshold* by taking the mean of affinity values of all the elements in the dataset.The proposed algorithm may be named as CAST+. Let us see the main terminology used

Definition 1: The affinity of a node x to a cluster C is defined as follows:

$$a(x) = \sum_{k \epsilon C} S(x,k) \tag{4.1}$$

Definition 2: The connectivity threshold,χ , of a cluster C is:

$$\chi = T \mid C \mid \tag{4.2}$$

where $\mid C \mid$ is the cardinality of C.

Definition 3: A high connectivity node is a node that will be included in a cluster. Its affinity satisfies the following.

$$a(i) \geq \chi \tag{4.3}$$

where a(i) is the affinity of i.

Definition 4: A low affinity node will be removed from a cluster if its affinity satisfies the following:

$$a(i) < \chi \tag{4.4}$$

where a(i) is the affinity of i.

Each cluster is formed by alternating between adding and removing nodes from the current cluster until such time that changes no longer occur or a maximum of iterations executed:

- **Node Addition:** Add nodes with high connectivity to the nodes in the open cluster. **For CAST+** Before performing this step we check the node with existing clusters and adds to the cluster which is highly connected.
- **Node Removal:** Remove any nodes in the open cluster with low connectivity to the other nodes in the cluster.
- **Cluster Cleaning:** Make sure all nodes are in clusters with highest affinity. **For CAST+** this step is not required.

CAST algorithm relies on the *affinity threshold*, T, being an input variable defined by the user before initiating the clustering process. This is the problem because the size and the

quantity of the clusters produced by the algorithm is directly affected by this parameter[2]. Implying that some knowledge of the data set is required before the clustering can be performed. We have enhanced the algorithm to calculate this threshold. The threshold parameter, T, is calculated based on the similarity values of the nodes in the data set. The threshold is computed as follows:

$$T = \frac{\sum_{i,j \leq n} S(i,j)}{n^2} \tag{4.5}$$

The following figure 4.3 provides the pseudo code for threshold value, figure 4.4 provides the pseudo code for Node addition, and 4.5 provides the pseudo code for Node Removal.

```
Threshold
// T is an input parameter

CAST:
T = fixed value (for example 0.5 or 7.6)

CAST+:
sum = 0;
count = 0;
for all i, j ? n {
sum = sum+S (i, j);
count++ ;
}
T = sum/count;
```

Figure 4.3: Pseudo code for finding threshold Value

The threshold assignment and affinity check with existing clusters in the step of node addition, obviate the need for the "cleaning " step as proposed in the original CAST

Cluster formation:

While (U ? Ø) {
    for all u ∈ U set a(u) = 0
    create empty cluster C_{open}
Pick an element u ∈ U such that S (u, x) = max {S (w, x)/w and x ∈ U}
Copen = Copen U u;
U = U\u
For all x ∈ U set a(x) = a(x) + S(x, u)
while (changes in copen occur) or (iterations < max iterations) {
 //Addition S tep
 CAST: while max {a (w) |w ∈ U} ? χ {
CAST+: while max {a (w) |w ∈ U} ? χ for all clusters open {
    Pick an element u ∈ U such that a (u) = max { a (w) |w ∈ U}
    Copen? Copen U {u}
    U? U | {u}
    // Update affinity of all nodes
    For all x ∈ U U Copen set a(x) + S (x, u)
}

Figure 4.4: Pseudo code for node addition

algorithm. The cleaning step is used to move any vector from its current cluster to one that it may have a higher affinity for and has a time complexity on the order of $O(n^2)$. The output of the gene expression matrix of figure 4.1 is given in figure 4.7

4.4 Analysis of Clustering Solutions

Different clustering algorithms yield different solutions on the same data and also same algorithm gives different solutions for different parameter settings.

Different measures for the quality of a clustering solution are applicable in different situations, depending on the data and on availability of the true solution.

```
//Removal step
While min{ a (w) |w ∈ Copen} < χ{
    Pick an element u ∈ Copen such that a (u)
= min{ a (w) |w ∈ Copen}
    Copen?  Copen\ {u}
    U?  U U {u}
    //Update affinity of all nodes
    For all x ∈ U U Copen set a(x) =a(x)-S(x)
}
```

Figure 4.5: Pseudo code for node deletion

In case true solution is known, and we wish to compare it to another solution, one can use Minkowski Measure or Jaccards Coefficient method.

4.4.1 Jaccards Coefficient

Jaccards Coefficient is a static measure used for comparing the similarity and diversity of sample sets, by everitt(1993)[13].
The jaccards coefficient is defined as the size of the intersection divided by the size of the union of the sample sets.

$$J(A, B) = \frac{\mid A \cap B \mid}{\mid A \cup B \mid} \tag{4.6}$$

where A indicates the true solution, and B indicates the solution generated by the algorithm.

4.4.2 Minowski Measure

: A clustering solution of n elements is represented by a $n - by - n$ similarity matrix C, where $C_{ij} = 1$ if i and j belong to the same cluster and $C_{ij} = 0$ otherwise.

Given such matrix representation of the true clustering T and any clustering C of the same data set, the minowski measure for the quality of C is the normalized distance between

Cleaning Step

While (changes in any C_i occur) or (iterations < max iterations) { //cleaning step may not converge

for each $c \in C_i$ and $C_i \in C$ and $C_i \in C$ {

Compute a normalized affinity of c to each cluster Cj such that $a_i(c) = (\Sigma_{k \in C_i} S(c, k))/(|Cj|)$

}

If max $\{a_i(c)\} > a_{i,}$ for all $C_i \in C$ and i ? j {

$C_i = C_i \backslash c$

$C_i = C_i \cup c$

}

}

Figure 4.6: Pseudo code for Cleaning or pruning

the two matrices.

$$MinowskiMeasure = \frac{| T - C |}{| T |} \tag{4.7}$$

where $| T | = \sqrt{\sum_i \sum_j T_{ij}^2}$ as developed by Sokal(1977) [33].

Since the matrices are binary, this is simply the number of pairs on which the two solutions disagree and normalized according to the true solution. A perfect clustering would thus obtain the score zero.

4.4.3 Homogeneity Vs Separation

When the true solution is not known the algorithms are analyzed by presenting a curve of homogeneity versus separation.The intra cluster is called as Homogeneity value, and inter cluster distance is called as Separation value. Such a curve can show that one algorithm dominates another if it provides better homogeneity for all separation values.

4.4.4 Silhouette Width

The Silhouette validation technique (Rousseau w, 1987) [33] calculates the silhouette width for each sample, average Silhouette width for each cluster and overall average silhouette

	no of the cluster
gene1	1
gene2	2
gene5	2
gene6	2
gene7	2
gene8	2
gene3	3
gene4	4
gene9	4
gene10	4

Figure 4.7: Sample result of the proposed algorithm

width for a total data set. Using this approach each cluster could be represented by so called silhouette width, which is based on the comparison of its tightness and separation.

The average silhouette width could be applied for evaluation of clustering validity and can also be decide how good is the number of selected clusters.

To construct the silhouettes S(i) the following formula is used:

$$S(i) = \frac{(b(i) - a(i))}{max\{a(i), b(i)\}} \tag{4.8}$$

where

a(i)=average dissimilarity of object i to all other objects in the same cluster.

b(i)= minimum of average dissimilarity of object i in other cluster (in the closest cluster).

The largest overall average silhouette indicates the best clustering. Therefore, the number of clusters with maximum overall average silhouette width is taken as the optimal number of the clusters.

4.5 Results

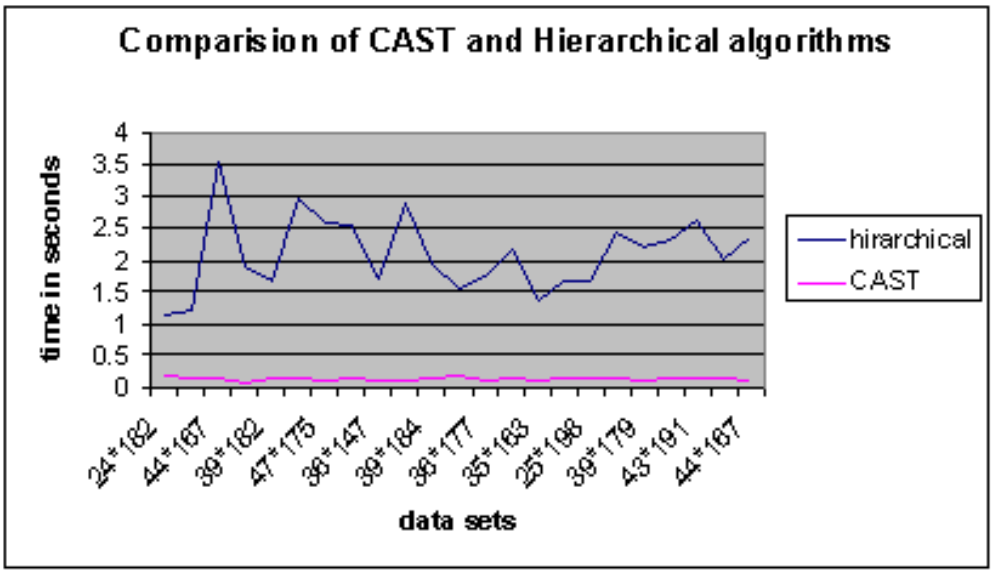

Figure 4.8: Figure showing the comparison between the CAST and Hierarchical algorithm

Data Set:

The synthetic random datasets in our simulation provides randomly generated classes in a two dimensional Euclidean space. These data sets are used to evaluate the performance of the algorithms. It is supposed to provide data with low noise. Twenty two different synthetic random data sets are being generated which yields similarity matrix using Euclidean distance in interval [0, 1].

Figure 4.8 shows that the time of execution of hierarchical algorithms is comparatively very high when compared to CAST algorithm.

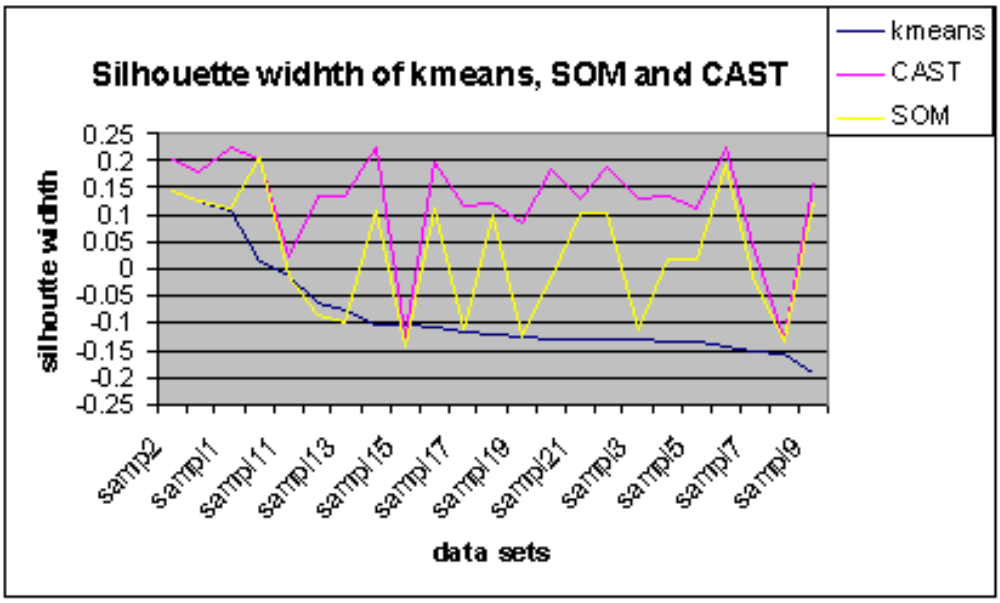

Figure 4.9: Figure showing the comparison between the K-Means, SOM, and CAST based on silhouette width

Figure 4.9 explains that K-Means, SOM, and CAST are executed on 22 different Data Sets and is observed that SOM, and CAST performs better than K-Means algorithm based on Silhouette width. Out of twenty two data sets sixteen data sets showed better result using CAST than SOM. For the remaining six data sets also SOM reached the performance of CAST but not exceeded. Hence it can be inferred that CAST performs better than SOM.

Figure 4.10 explains that CAST, ECAST, and proposed CAST+ algorithms are tested on twenty two different data sets and result is analyzed using silhouette width. It is observed that overall 22 data sets both ECAST and CAST+ algorithms performed better

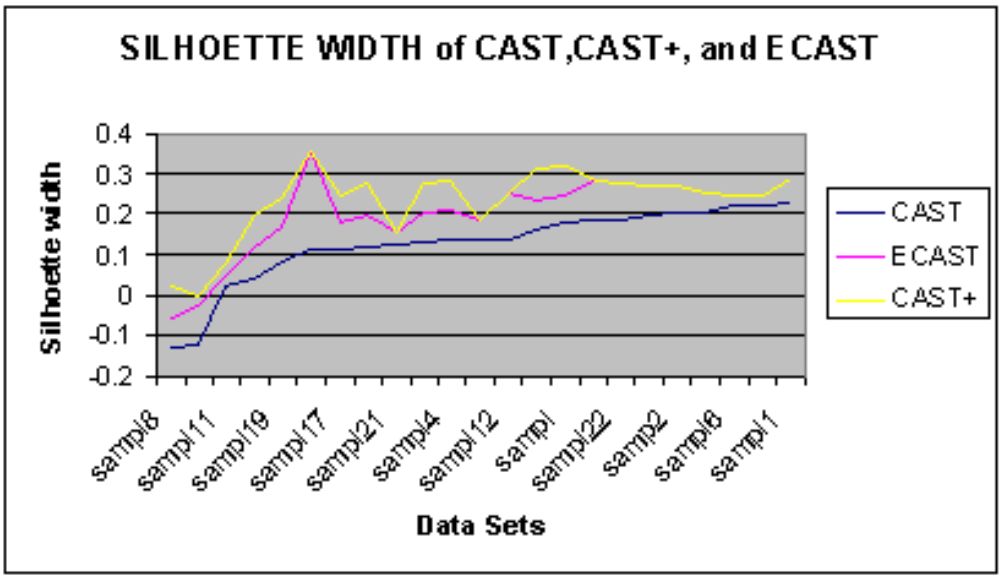

Figure 4.10: Figure showing the comparison between the CAST, ECAST and Proposed algorithm

than CAST. But CAST+ has performed better on fourteen datasets in comparison with ECAST algorithm. In remaining data sets both have performed comparatively well. i.e. CAST+ showed 60% better performance over ECAST and 100

Figure 4.11 explains the that K-Means, SOM, and CAST algorithms are performed on twenty two different data sets and respective homogeneity and separation values are calculated. It is observed that CAST has shown the best result over SOM and K-Means as indicated in figure 4.11.

Figure 4.12 shows that over 78% of the data sets show better result to CAST+ over CAST and ECAST. The data sets whose homogeneity value is vast, they show same result of the CAST. Overall it is observed that the proposed CAST+ algorithm shows better result than other algorithms.

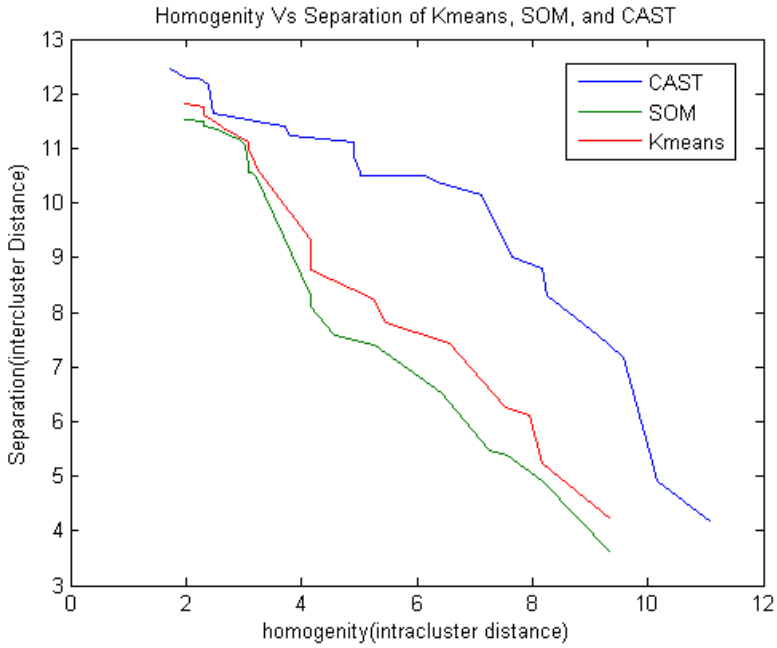

Figure 4.11: Figure showing the comparison between the K-Means, SOM, and CAST

4.6 Conclusion

In this chapter a graph theoretic divisive algorithm called CAST is studied and overcome the drawback what it is having by using the proposed algorithm. Comparing the result of the proposed algorithm with the existing algorithms and it is observed that proposed algorithm performed better than all other algorithms.

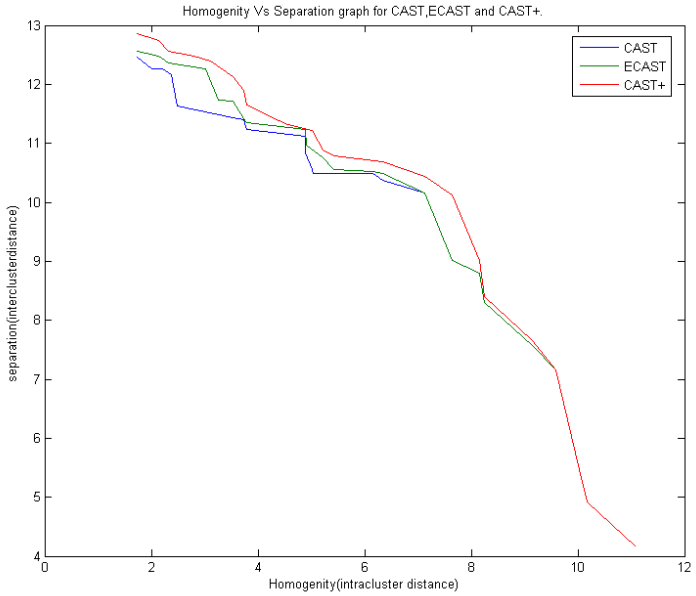

Figure 4.12: Figure showing the comparison between the CAST, ECAST, and Proposed algorithm

Chapter 5

Conclusion and Future Work

5.1 Conclusion

The introduction of new technologies such as computers, satellites, new mass storage media and many others during 1980's have lead to an exponential growth of collected data. Traditional data analysis techniques often fail to process large amounts of -often noisy- data efficiently, in an exploratory fashion. The scope of data mining is the knowledge extraction from large data amounts with the help of computers. It is an interdisciplinary area of research, that has its roots in databases, machine learning, and statistics and has contributions from many other areas.

Recently, the collection of biological data has been increasing at explosive rates due to improvements of existing technologies and the introduction of new ones such as the microarrays. These technological advances have assisted the conduct of large scale experiments and research programs.The explosive growth in the amount of biological data demands the use of computers for the organization for its maintenance and the analysis. This led to the evolution of bioinformatics, an interdisciplinary field at the intersection of biology, computer science, and information technology.

DNA microarray technology has now made it possible to simultaneously monitor the expression levels of thousands of genes during important biological processes and across collections of related samples. Elucidating the patterns hidden in gene expression data

offers a tremendous opportunity for an enhanced understanding of functional genomics. A first step towards addressing this challenge is the use of clustering techniques, which is essential in the data mining process to reveal natural structures and identifying interesting patterns in the underlying data.

Cluster analysis seeks to partition a given data set into groups based on specific features so that points within a group are more similar to each other than the points in different groups. Many conventional clustering algorithms have been adapted or directly applied to gene expression data. But each method has short comings. These shortcomings include problems of cluster boundaries, as for hierarchical techniques, where the output is a tree depicting the relation of each object to every other object in the data set. The requirement for knowing the expected number of clusters, as for K-Means, and knowing the grid structure for SOM are the underlining problems under different algorithms developed so far.

CAST algorithm[2] had overcome the above problems but takes the affinity threshold, which determines the size and number of clusters, as input value. Also after the clustering is over cleaning step has to be performed, which takes an additional time, resulting in increase of time complexity.

But the proposed algorithm overcomes the problem of taking threshold affinity as input by finding the affinity value as the mean of the affinity of all the genes in the data set i.e. gene expression array. It also overcomes the problem of cleaning or external pruning by performing the affinity check while performing the Node addition itself. The experimental results also show that the proposed algorithm performs better than all other existing algorithms we have studied in this thesis work.

5.2 Future Work

This work can be extended as follows:

1. Improve the performance of the algorithm using soft computing techniques.

2. Perform theoretical analysis of the determination of the threshold parameter.

3. Explore further improvements to Proposed CAST+ Algorithm.

4. Clustering is generally recognized as an "un supervised" learning problem. Prior to undertaking a clustering task, "global" information regarding the data set, such as the total number of clusters and the complete data distribution in the object space, is usually unknown. However, some "partial" knowledge is often available regarding a gene expression data set. For example, the functions of some genes have been studied in the literature, which can provide guidance to the clustering. Furthermore, some groups of the experimental conditions are known to be strongly correlated, and the differences among the cluster structures under these different groups may be of particular interest. If a clustering algorithm could integrate such partial knowledge as some clustering constraints when carrying out the clustering task, we can expect the clustering results would be more biologically meaningful. In this way, clustering could cease to be a "pure" un supervised process and become an interactive exploration of the data set.

Bibliography

[1] Arun K Pujari, Data mining Techniques, University Press, Hyderabad, 2002.

[2] Ben-Dor A., Friedman N. and Yakhini Z, "Clustering gene expression patterns", *Journal of Computational Biology*, Vol. 6, No. (3/4), 1999, pp:281297.

[3] Daxin Jiang, Chun Tang and Aidong Zhang, "Cluster Analysis for Gene Expression Data: A Survey", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, No. 11, November 2004, pp:1370-1386.

[4] Sankar K.Pal, Sanghamitra Bandyopadhyay, and Shubra Sankar Ray, "Evolutionary Computation in Bioinformatics: A Review", *IEEE: Applications and Reviews*, Vol.36, No.5, September 2006, pp:601-615.

[5] Abdelghani Bellaachia, David Portnoy, Yidong Chen, and Abdel. G. Elkahloun, "E-CAST: A Data Mining Algorithm for Gene Expression Data", *Workshop on Data Mining in Bioinformatics* (with SIGKDD02 conference), Vol. 2, 2002, page:49-54.

[6] Dongsong Zhang and Lina Zhou,"Discovering Golden Nuggets: Data Mining in Financial Application", *IEEE Transactions on systems, man, and cybernetics-part c: applications and reviews*, Vol. 34, no.4, November 2004, pp:513-522.

[7] Sharan R, Elkon R, and Shamir R, "Cluster Analysis and its Applications to Gene Expression Data", *Ernest Schering workshop on Bioinformatics and Genome Analysis*, Springer Verlog, 2002.

[8] Dr. Diego Kuonen, "Challenges in bioinformatics for statistical data miners.", *Bulletin of the Swiss Statistical Society*, Vol.-46, October 2003, pp:10-17.

[9] Darlene R. Goldstein, Debashis Ghosh and Erin M. Conlon, "Statistical Isuues in the Clustering of Gene Expression Data", *Statistica Sinica*, Vol 12, 2002, pp:219-240.

[10] Gregory Piatetsky-Shapiro and Pablo Tamayo, "Microarray Data Mining: Facing the Challenges," *SIGKDD Explorations*,Vol. 5, 2003, No. 2, pp:1-5.

[11] Houle, J.L., Cadigan, W., Henry, S., Pinnamaneni, A. and Lundahl, S. "Database Mining in the Human Genome Initiative". Whitepaper, Biodatabases.com, Amita Corporation, March 2004.

[12] Herrero J., Valencia A. and Dopazo J, "A hierarchical unsupervised growing neural network for clustering gene expression patterns", *Bioinformatics*, Vol:17, 2001, pp: 126-136.

[13] Eisen, Michael B., Spellman, Paul T., Brown, Patrick O. and Botstein, David . "Cluster analysis and display of genome-wide expression patterns". Proc. *Natl. Acad. Sci. USA*, Vol.95, No.25, December 1998, pp:14863-14868.

[14] Haux, R., and Kulikowski,C. "Digital Libraries and Medicine", *IMIA Yearbook of Medical Informatics*, Vol. 2, 2001 , pp: 83-99.

[15] Brazma, Alvis and Vilo, Jaak. "Minireview: Gene expression data analysis," *Federation of European Biochemical societies*, Vol. 480, June 2000, pp:17-24.

[16] Heyer LJ., Kruglyak S., Yooseph S.expression data: identification and analysis of coexpressed genes", *Genome Res,* Vol. 9, No. 11, 1999, pp:11061115.

[17] Jain, A.K., Murty, M.N. and Flynn, P.J."Data clustering: a review". *ACM Computing Surveys*, Vol. 31, September 1999, No.3, pp: 254-323.

[18] Kjersti Aas, "Microarray Data Mining: A Survey", *Norwegian Regnesentral Computing Center*, vol. 23, February 2001, pp:114-149.

[19] Tamayo P., Solni D., Mesirov J., Zhu Q., Kitareewan S., Dmitrovsky E., Lander E.S. and Golub T.R. "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation". *Proceedings of National Academy of Science USA*, Vol. 96, No. 6, March 1999, pp: 2907-2912.

[20] Drowning in data. The Economist 1999 (26 June 1999).

[21] P.Hancen and B.Jaumard,"Cluster analysis and mathematical programming", *Mathematical programming*, Vol. 79, 1997, pp: 191-215.

[22] Judice L.Y.Koh1, Mong Li Lee, Asif M. Khan, Paul T.J. Tan1 and Vladimir Brusic,"Duplicate Detection in Biological Data using Association Rule Mining", *Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics*, 2004, pp:35-41.

[23] Jinyan Li and Hwee-Leng Ong, "Feature Space Transformation for Better Understanding Biological and Medical Classifications", *Journal of Research and Practice in Information Technology*,Vol. 36, No. 3, August 2004.

[24] Hideya Kawaji, Yosuke Yamaguchi, Hideo Matsuda, and Akihiro Hashimoto, "A Graph-Based Clustering Method for a Large Set of Sequences Using a Graph Partitioning Algorithm", *Genome Informatics*, Vol. 12, 2001, pp:93-102.

[25] Alex A. Freitas, "A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery", *advances in Evolutionary Computation,* A. Ghosh and S.S. Tsutsui, Ed. New York: Springer-Verlag, 2001.

[26] N.M. Luscombe, D.Greenbaum, M.Gerstein, "What is Bioinformatics? A proposed Definition and Overview of the Field", *Method inform Media*, Vol. 40, 2001, pp: 346-358.

[27] Rui Xu,Donald Wunsch, "Survey of Clustering Algorithms", *IEEE Transactions on Neural Networks*,Vol. 16, NO. 3, MAY 2005, pp:645-678.

[28] Vladimir B. Bajic, Vladimir Brusic, Jinyan Li, See-Kiong Ng, and Limsoon Wong, "From Informatics to Bioinformatics", *Conferences in Research and Practice in Information Technology*, Vol 19, 2003, pp: 45- 54.

[29] George Tzanis, Christos Berberidis and Ioannis Vlahavas, "Data Mining in Biological Data", *In Proceedings of the 10th Panhellenic Conference on Informatics*, 2002, pp:426- 434.

[30] Darlene R. Goldstein, Debashis Ghosh and Erin M. Conlon, "Statistical Isuues in the Clustering of Gene Expression Data", *Statistica Sinica*,Vol. 12, 2002, pp:219-240.

[31] Reichhardt T. "It's sink or swim as a tidal wave of data approaches." *Nature* . Vol. 399, No. 6736, 1999, pp: 517-520.

[32] G.Osuri, (2003), Bioinformatics, $http : //www.bioinformatics.org/tutorial/1 - 2.html$.

[33] $http : //gepas.bioinfo.cipf.es/cgi_bin/tutox?c = CAAT/caat.config$.