International Journal of Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-4, Issue-3)

Research Article | **March 2015**

A Survey on Microarray Gene Expression Data sets in Clustering and Visualization Plots

T. Deepika*, Dr. R. Porkodi
Department of Computer Science
Bharathiar University, Coimbatore, Tamil Nadu, India

***Abstract*—**

In data mining, clustering techniques have been applied in cellular processes, gene regulation, sub types of cells and gene function. Clustering in microarray gene expression handles various experimental conditions in various algorithms by using different data sets. This paper focuses the study on the clustering of gene expression data using the data sets such as yeast data, yeast cell-cycle, serum, Arabidopsis and rat CNS. This paper also discusses the different visualization plots available to display the microarray gene expression data for easier understanding of clustering results. This paper also studies the description and its usage of the popular above mentioned microarray datasets in various literatures along with their visualization results.

Keywords*— *Clustering gene expression data, clustering bioinformatics, microarray technology, clustering.

I. INTRODUCTION

Data mining (DM) refers to extracting or mining, knowledge from large amounts of data. DM is the science of finding new interesting patterns and relationship in huge amount of data [5]. Clustering is fundamental and widely applied method for understanding and exploring a data set [1]. Clustering is an unsupervised learning or segmentation of descriptive method, where classification of patterns is divided into groups. Clustering occurs by observing only independent variable unlike supervised learning, analyse both independent and dependent variable that does not use class. Clustering is used to group objects into a specific number of clusters, so that the objects within a cluster have very high similarity and objects from different clusters have very low similarity in which similarity between two objects are measured by using their same attribute values that attributes may be geographical distance-based or size-based. Clustering algorithms are mainly categorized into two groups: Hierarchical clustering, Partitional clustering.

1) Hierarchical clustering: These algorithms divided into agglomerative clustering and divisive clustering algorithms as Minimum Spanning Tree algorithm (MST) and linkage clustering such as single link, average link and complete link.

2) Partitional clustering algorithms: These algorithms are categorized as Minimum Spanning Tree algorithm (MST), Squared Error Clustering algorithm, K-Means clustering, Nearest Neighbor algorithm, Partitioning Around Medoids algorithm (Clustering large applications (CLARA) and Clustering large applications based upon randomized search (CLARANS) clustering algorithms are based on PAM), Bond Energy algorithm (BEA), Clustering with Genetic algorithms and Clustering with Neural Networks.

3) Other clustering algorithms: These algorithms categorized as Balanced Iterative Reducing and Clustering using Hierarchies algorithm (BIRCH), Density-Based Spatial Clustering of Application with Noise algorithm (DBSCAN), Clustering Using Representatives algorithm (CURE), Robust Clustering using Links algorithm (ROCK), Sieving Through Iterated Relational Reinforcement (STIRR), Clustering Categorical Data Using Summaries (CACTUS).

The clusters using different algorithms have slight variations in their properties and understanding cluster models have differences between various algorithms. These cluster models are classified into connectivity models, centroid models, distribution models, density models, subspace models, group models and graph-based models.

Application domains of clustering are biology, marketing and economics such that plant and animal classification, disease classification, image processing, pattern recognition, document retrieval etc. Application of data mining in bioinformatics includes protein function domain detection, protein function inference, gene finding, disease diagnosis, function motif detection, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, protein sub-cellular location prediction and data cleansing.

Microarray technology is a study about the expression levels of a huge number of genes across different experimental conditions. The information on the gene expression process from a gene is used in the synthesis of functional gene products like proteins and functional RNAs for non-protein coding genes. Gene expression data are generally very huge in size and to search for useful patterns within this data, genes have to be grouped into clusters on the basis of similar features. This gene expression data in the form of a 2D matrix, which is *n x m* matrix where *n* is specified as the genes and *m* is specified as the experiments includes time points, tissues, conditions, cell lines and patients. Clustering gene expression data can be done through experimenting the samples or genes. The samples are grouped together from tissues of the patients who are similarly affected by the disease. The genes are grouped together functionally also by relating genes that are similarly affected by the disease and that respond similarly to an experimental condition. Clustering gene expression data applications infer unknown gene function, built regulatory networks, reduce dimensionality and discover subtypes of a disease.

A. Microarray Gene Expression

A group of small DNA spots attached to a solid surface is called microarray. It contains thousands of DNA spots, covering almost every gene in a genome. In microarray experiments, the signal collected from each spot is used to estimate the expression level of a gene. DNA microarray technology typically contains thousands of genes. These gene experiments shows two ways which is control or healthy cells and treated or diseased cells and the mRNA produced are extracted, labelled, mixed and hybridized to the chip.

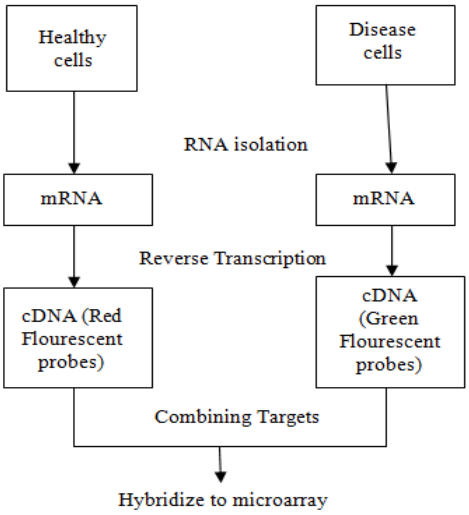

Fig. 1 Microarray experiment

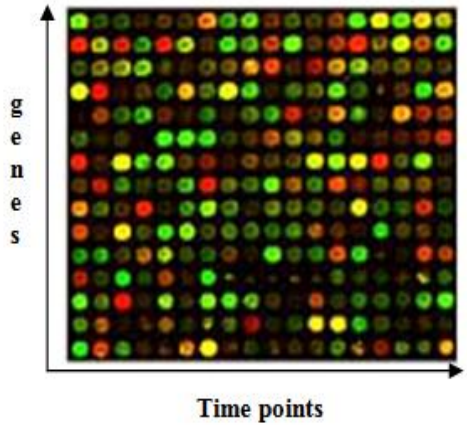

Fig. 2 2-D matrix representation of microarray Gene chip

The Fig. 1 shows microarray gene experiments, where gene chip is used for every gene experiment for each gene data.

1) Gene chip: Various types of microarrays are used in biomedicine, which include DNA microarrays, protein microarrays, MMChips, tissue microarrays, peptide microarrays, cellular microarrays, antibody microarrays, chemical compound microarrays, carbohydrate arrays (glycoarrays), reverse phase protein microarrays, phenotype microarrays, etc. Fig. 2 shows the gene chip for microarray experiments.

Specifically, DNA microarrays have been used out of all the other microarrays. There are several synonyms of DNA microarrays such as DNA chips, gene chips, DNA arrays, gene arrays and biochips. Types of DNA chips or microarrays are cDNA microarrays, Oligonucleotide microarrays, BAC microarrays, SNP microarrays. cDNA is the widely used microarray and it is called cDNA chips or cDNA microarray or probe DNA.

This paper is organized as follows: section 1 describes the introduction on clustering microarray gene expression data, section 2 describes the literature review, section 3 describes the algorithms of clustering gene expression datasets, section 4 describes the visualization plots available for microarray data, section 5 gives the description of most frequently used popular microarray datasets in the literature and finally section 6 gives the conclusion.

II. REVIEW OF LITERATURE

In [6], clustering is widely used for microarray analysis tool. The genes with similar expression profiles which co-expressed genes should be put in a single cluster with similar cellular functions. It can be detected by using clustering techniques. Clustering methods divide a set of n objects into k partitions depending on some similarity and dissimilarity metric. Most of clustering algorithms optimized a single cluster quality measure which mirrors a single measure of the goodness of a partitioning solution. But sometimes a single cluster validity index is not applicable for determining appropriate partitioning for different types of dataset shaving different properties. Hence, it is required to optimize simultaneously multiple cluster quality measures which are responsible for capturing different cluster properties. Thus, here the clustering problem is defined as a multi objective optimization problem where multiple number of cluster quality measures are optimized simultaneously.

In [9], the microarray chip technology, large data sets is emerging containing the simultaneous expression levels of thousands of genes at various time points during a biological process. Biologists are attempting to group genes based on the temporal pattern of their expression levels. The use of hierarchical clustering with correlation distance has been the most common in the microarray studies. At the moment, there do not seem to be any clear-cut guidelines regarding the choice of a clustering algorithm to be used for grouping genes based on their expression profiles. Cluster analysis programs are routinely run as a first step of data summary and grouping genes in a microarray data analysis.

In [11], in the gene expression context, the elements are the genes and assume that there exists some correct partition of the genes into true clusters. Alternatively, the elements are the conditions or tissues that are assumed to belong to one of several categories, as a tumor or normal tissues. The goal in a clustering problem is to partition the set of elements into homogeneous and well-separated clusters. That requires elements from the same cluster will be highly similar to each other, while elements from different clusters will have low similarity to each other. Clustering problems and algorithms are often represented in graph-theoretic terms and use this representation.

In [12], gene expression provides a powerful tool for studying about genes collectively react to changes in environments, provide hints about the structures of the involved genes. Basic problem in interpreting the observed

expression data is to cluster genes with correlated expression patterns over some time series and under different conditions. Unlike a continuous optimization problem, finding a globally optimal solution for a combinatorial optimization problem is often possible.

In [13], the biological interpretation of a cluster of genes can be best assessed by studying the functional annotation of the genes of that cluster. The two-stage clustering algorithm for clustering gene expression data using the idea of points having significant membership in multiple classes. The clustering solutions is evaluated both quantitatively and using some gene expression visualization tools. Statistical tests have also been conducted in order to establish the statistical significance of the results produced by the proposed technique.

In [14], the wide application of microarray technologies now generates very large amounts of data. As a result, there is an increasing need for technology that can extract useful and rational, fundamental patterns of gene expression from the data. Clustering technology is one of the most useful and popular methods for identifying these patterns. Generally, there are two classes of cluster algorithms, hierarchical and non-hierarchical. Both have been successfully used in the analysis of gene expression data.

III. CLUSTERING

Clustering is one of the methods used to gain insight into biological processes, particularly in the genomics level. Clearly, clustering can be used in many areas of biological data analysis. A good clustering approach may detect patterns or relationships in expression data. A clustering algorithm used for group together genes based on their expression profiles. In clustering, group together similar expression profiles of genes as expression data analysis are to identify the changing and unchanging levels of gene expression and to correlate changes to identify sets of genes with similar profiles.

Clustering results based on an observed data set. Clustering parameters can be viewed as any other parameter in statistics. The output of a clustering algorithm applied to a data set is an estimate of the underlying parameter. There will not, generally closed formula for the variance of the parameter estimate. Re-sampling methods which is parametric or non-parametric bootstrap can be used to estimate the reliability and repeatability of clustering results. The general steps of clustering gene expression algorithm of given input and expected output defined as follows,

Input:
Large number of gene data G //data points = genes
Measure distance between genes g_{ij} // data points d_{ij}=gene points g_{ij}
Output:
Similar expressions of the gene grouped together with K //clustering
Sometimes an objective measure also defined (obtained clustering search to minimize).

Where *i* is the number of genes as rows and *j* is the time points of gene expression data. Clustering procedure represents the clustering process [15] as follows,

1) At each step, select the two closest sequences and join them into one cluster.
2) Then replace the two just joined sequences with their ancestor.
3) After reducing the size of the data matrix by one.
4) Finally, compute the distances from the new ancestor to the remaining sequences.

A. Distance measure and parameters of clustering

1) Distance measure: Each distance measure quantifies a different notion of what it means for two vectors to be close to each other. There are various techniques for calculating the distance to new ancestral sequence *a* joins sequences *n* and *m*. The distance measurements as follows,

Single linkage clustering as,

$$d(x,a) = min\,[d(x,m), d(x,n)] \quad \ldots\ldots (1)$$

Complete linkage clustering as,

$$d(x,a) = max\,[d(x,m), d(x,n)] \quad \ldots\ldots (2)$$

UPGMA- Unweighted Pair Group Method with Arithmetic Mean algorithm as follows,

$$d(x,a) = \frac{d(x,m)+d(x,n)}{2} \quad \ldots\ldots (3)$$

WPGMA- Weighted Group Method with Arithmetic Mean algorithm as follows,

$$d(x,a) = \frac{s(m)d(x,m)+s(n)d(x,n)}{s(m)+s(n)} \quad \ldots\ldots (4)$$

Where s(n) counts the number of actual sequences represented by node n [15].

Euclidean distance in D-dimensions calculated by using following equation as,

$$D(x,y) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2} \quad \ldots\ldots (5)$$

2) Cluster parameter: The cluster parameters define the cluster that is common in all clusters. The cluster parameter must be unique to each dataset. Parameters of clustering also include number of clusters, cluster labels, fuzzy clustering memberships, hierarchical tree etc. Parameter for centroid as follows,

$$C_m = \frac{\sum_{i=1}^{N}(t_{mi})}{N} \quad \ldots\ldots (6)$$

Where middle of the cluster called centroid and it need not be an actual point in the cluster. Cluster parameter for radius as follows,

$$R_m = \sqrt{\frac{\sum_{i=1}^{N}(t_{mi}-c_m)^2}{N}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (7)$$

Where radius is the square root of the average mean squared distance from any point in the cluster to the centroid. Cluster parameter for diameter as follows,

$$D_m = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(t_{mi}-t_{mj})^2}{(N)(N-1)}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (8)$$

Where the diameter is the square root of the average mean squared distance between all pairs of points in the cluster.

B. Clustering algorithms for microarray gene expression data

1) DIANA: Divisive Analysis clustering (DIANA) is a divisive hierarchical algorithm that initially starts with all observations in a single cluster, and successively divides the clusters until each cluster contains a single observation.

In [9], Model-based clustering has several modeling and optimization options, and it worked reasonably well in the simulated data. Hierarchical clustering with Euclidean distance is able to distinguish the target patterns. Model based and Diana are slightly better than K-means and Fanny.

2) MST: Minimum Spanning Tree (MST) of a weighted graph is the minimum weight spanning tree of that graph. The cost of constructing a minimum spanning tree is O(m log n), where m is the number of edges in the graph, n is the number of vertices. MST is used to represent a set of expression data and their significant inter-data relationships to facilitate fast rigorous clustering algorithms. The MST clustering algorithm is known to be capable of detecting clusters with irregular boundaries. Unlike traditional clustering algorithms, the MST clustering algorithm does not assume a spherical shaped clustering structure of the underlying data.

In [12], based on the data representation framework and the MST-based clustering algorithms, they have developed computer program EXCAVATOR for gene expression data clustering. The various unique features of EXCAVATOR make the program highly useful tool in mining large-scale gene expression data in a reliable meaningful way.

3) K-means: K-means assumes that the number of clusters k is known. It minimizes the distances between elements and the centroids of their assigned clusters. It is an iterative method which minimizes the within-class sum of squares for a given number of clusters. The algorithm starts with an initial guess for the cluster centers, and each observation is placed in the cluster to which it is closest. The cluster centers are then updated, and the entire process is repeated until the cluster centers no longer move.

In [9], K-means clustering used and model based and Diana being slightly better than K-means and Fanny.

In [11], the data set was clustered using four methods one of the methods is K-Means.

4) SOM: Self-organizing map (SOM) is an unsupervised learning technique that is popular among computational biologists and machine learning researchers. SOM is based on neural networks, and is highly regarded for its ability to map and visualize high-dimensional data in two dimensions. SOM is a discrete grid of map units. Each map unit can represent certain kinds of data, and the units represent genes expressed in similar ways in a chosen set of treatments. The input data is represented in an ordered fashion on the map: map units close-by on the grid represent more similar expression profiles and units farther away represent progressively more different profiles. The map is a similar diagram that presents an overview of the mutual similarity of the large number of high-dimensional expression profiles. Each of the nodes is associated with a reference vector of the same dimension as the expression patterns.

In [6], AMOSA has been a recently developed simulated annealing based multi objective optimization method, is utilized to optimize these indices so that it can enable the proposed algorithm to automatically detect the appropriate number of clusters as well as the appropriate partitioning from different datasets. The proposed technique applied to extracting clusters from gene expression data. Results on five real life gene expression datasets with the performance of the proposed technique has been compared with some algorithms include SOM clustering method.

In [13], the number of clusters in a gene expression data set is automatically evolved in the proposed SiMM-TS clustering technique. The performance of the proposed clustering method has been compared with that of the other combinations of algorithms in the two stages as well as with some algorithms includes SOM to show its effectiveness on an artificial and three real life gene expression data sets.

5) CLICK: Cluster Identification via Connectivity Kernels (CLICK) uses a graph theoretic approach to clustering. The input data are represented as a weighted graph, in which each gene is represented by a vertex, and the similarity between the expression patterns of each two genes is used to calculate the weight of the edge connecting their vertices.

In [11], the data set was clustered using four methods one of them is CLICK.

6) FANNY: This algorithm performs fuzzy clustering, where each observation can have partial membership in each cluster. Each observation has a vector which gives the partial membership to each of the clusters. A hard cluster can be produced by assigning each observation to the cluster where it has the highest membership.

In [9], Fanny used and performs quite well with Model based and Diana being slightly better than K-means and Fanny.

7) DGSOT: The DGSOT is a tree structure self-organizing neural network designed to discover the proper hierarchical structure of the underlying data. The DGSOT grows vertically and horizontally. In each vertical growth, the DGSOT adds two children to the leaf whose heterogeneity is greater than a threshold and turns it to a node. In each horizontal growth, the DGSOT dynamically finds the proper number of children (sub clusters) of the lowest level nodes.

In [14], the biological functionality enrichment, the clustering result of DGSOT is considerably higher than the clustering result of SOTA and the *K*-means. They believe that DGSOT can be a robust and accurate framework for the study of patterns among large sets of gene microarray expression data.

IV. DATA VISUALIZATION

In data mining, data visualization tools for visualizing various kinds of datasets in massive amount, of multidimensional space and easy-to-understand visual forms [16]. It is essential for users to visualize raw data (tables, images, point information, textual annotation, and other metadata) [17]. There are some methods to visualize data sets for various kinds of patterns and knowledge visualizing sequences, time series data, phylogenetic trees, graphs, networks and web [16]. Visualization utilizes the capabilities of the human visual system to aid data comprehension with the help of computer generated representations [17]. The visualization tools are composed by following techniques as,

- Visualization techniques classified based on tasks, data structure or display dimensions.
- Visual perception type such as selection of graphical primitives, attributes, attribute resolution and the use of color in fusing primitives.
- Display techniques such as static or dynamic interaction; representing data as a line, surface or volume geometrics; showing symbolic data as pixels, icons, arrays or graphs.

Clustering visualization is the most frequent high-dimensional data visualization because of its direct use in studies searching for similarities and differences in biological materials. Performance metrics of clustering algorithm visually evaluate the clusters. Two familiar cluster visualization tools are Eisen plot and cluster profile plot. These two tools are mostly used in dataset experiment visualization. Data visualization represents the range of generic tools as follows,

1) Line graphs and Bar charts: These tools are used to compare unconnected column values over a continuous column. A line graph shows the relationship of one data point to another. They are most often used to track changes or trends over time which shown in Fig. 3. Line charts are useful when comparing multiple items over the same time period. Bar charts are most commonly used for comparing the quantities of different groups. Values of a group are represented using the bars, and they can be configured with either vertical or horizontal bars with the length or height of each bar representing the value shown in Fig. 4.

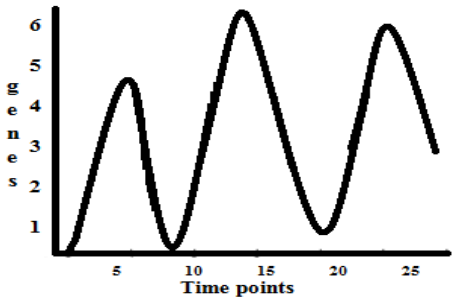

Fig. 3 Line graph visualization

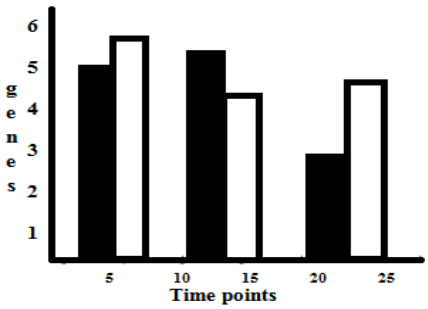

Fig. 4 Bar chart visualization

2) Box plots and Histograms: These tools are used to compare the distribution of distinct values for one or more unconnected column, which shown in Fig. 5 and Fig. 6.

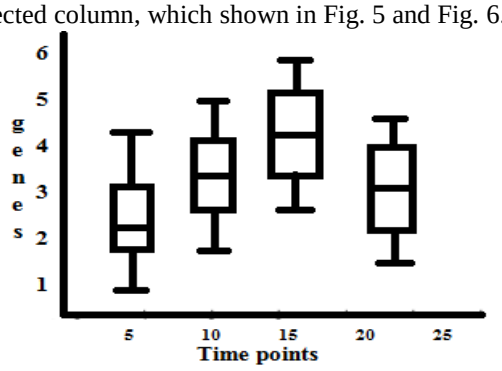

Fig. 5 Box plots visualization

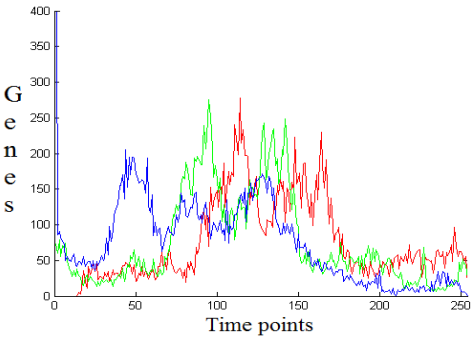

Fig. 6 Histogram visualization

3) Tree graph and Scatter plots: Tree tool represents a data set in the form of a tree and each level of the tree branches or splits based upon the values of different attribute (hierarchy in the data set). This is also called as a dendrogram in cluster visualization. Tree map (or dendrogram) uses in the hierarchical clustering, which shown in Fig. 7.

A scatter plot also called as X-Y plot which is a two-dimensional plot that shows the joint variation of two data items. In a scatter plot, each data point symbols such as dots, squares and plus signs represents an observation. The data point position indicates the value for each observation and it supports grouping. When there assign more than two measures, a scatter plot matrix is produced and is a series of scatter plots that displays every possible pairing of the measures that are assigned to the visualization. This tool is used to investigate the relationship between two or more columns, which shown in Fig. 8.

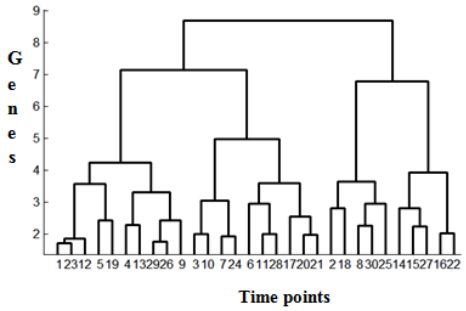

Fig. 7 Dendrogram (or tree map) visualization

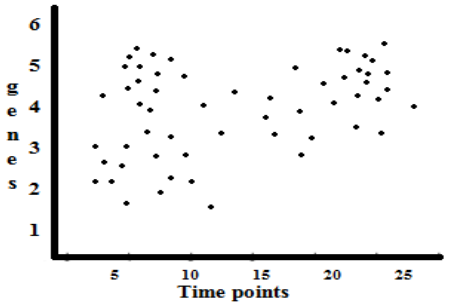

Fig. 8 Scatter plot visualization

4) Eisen plot: In microarray experiment, gene expression levels of data sets are easier to understand by using this visualization tool. Gene expression displays by using heat map and coloring values as healthy cells denoted by the green color, disease cells denoted by the red color and absence of differential expression values denoted by black as shown in Fig. 9.

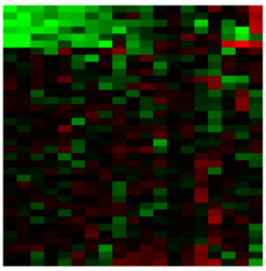

Fig. 9 Microarray gene expression visualization using Eisen plot

From the color representations of the Eisen plot, similar colors are grouped together, denoting that the expression profiles of the genes of a cluster are similar to each other as they produce similar color patterns.

5) Cluster profile plot: Gene expression levels of data sets are visualizing x, y matrix representation plot of time points by using the cluster profile plot visualization tool. Each cluster normalized expression values (light green) of the genes in that cluster with respect to the time points as shown in Fig. 10. Average expression values of genes of the cluster over different time points together with the standard deviations within the cluster at each time point.

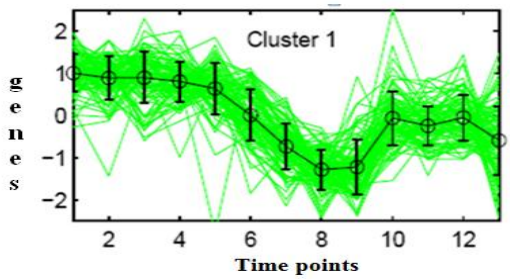

Fig. 10 microarray gene expression visualization using cluster profile plot

The cluster profile plots draw based on the obtained clustering results of the proposed algorithm. These plots also give the expression profiles for the different groups of genes differ from each other, while the profiles within a group are reasonably similar.

V. DATASETS

This part of the paper discussed the microarray datasets such as yeast sporulation data, yeast cell-cycle, Arabidopsis thaliana, human serum and rat CNS used in various clustering algorithms along with their experimental results reported in the literature.

A. Yeast sporulation data

Yeast typically grows a sexually by budding. A small bud which will become the daughter cell is formed on the parent (mother) cell, and enlarges with continued growth. As the daughter cell grows, the mother cell duplicates and then segregates its DNA. The nucleus divides and migrates into the daughter cell. Once the bud contains a nucleus and reaches a certain size it separates from the mother cell. Under adverse conditions, one yeast cell transforms itself into spores, tetrad of cells with tough cell wall, goes dormant. Yeast is ordinarily diploid, spores are haploid i.e., genetically, sporulation is analogous to formation of egg/sperm in most sexual organisms, 2 rounds of meiotic (not mitotic) cell division. Many of the genes/proteins involved in this are recognizably similar to human genes/proteins.

Yeast sporulation used the algorithms such as CRC, SOM, average link, MO-fuzzy, average link, IFCM, VGA, UPGMA, SiMM-TS, MOGA, SGA, MST, K-means, Fanny, Diana, Model based and Partial least squares.

In [9], consider microarray data on the transcriptional program of sporulation in budding yeast collected and analysed by Chu et al. (1998). They used DNA microarrays containing 97% of the known and predicted genes involved, 6118 in total. The mRNA levels were measured at seven time points during the sporulation process. The ratio of each gene's mRNA level (expression) to its mRNA level in vegetative cells just before transfer to sporulation medium is measured, and the ratio data are then log-transformed. Chu et al. (1998) determined by using a threshold level of 1.13 for the root mean squares of the log2-transformed ratios. Overall, 513genes were (positively) expressed during the process.

In [12], a set of gene expression data in the budding yeast S. cerevisiae (Eisen et al., 1998), with each gene having 79 data points (or 79dimensions) was used as input data set. Then selected four clusters (68genes in total). These are protein degradation, glycolysis, protein synthesis, and chromatin. Genes in each four cluster share similar expression patterns and are annotated to be in the same biological pathway. The goal of this application is to compare clustering results with known cluster information.

In [13] [6], Yeast Sporulation (Chu et al., 1998) was used as shown in Fig. 11. This data set consists of 6118 genes measured across 7 time points (0, 0.5, 2, 5, 7, 9 and 11.5 h) during the sporulation process of budding yeast. The data are then log-transformed. Among the 6118 genes, the genes whose expression levels did not change significantly during the harvesting have been ignored from further analysis. This is determined with a threshold level of 1.6 for the root mean squares of the log2-transformed ratios. The resulting set consists of 474 genes.

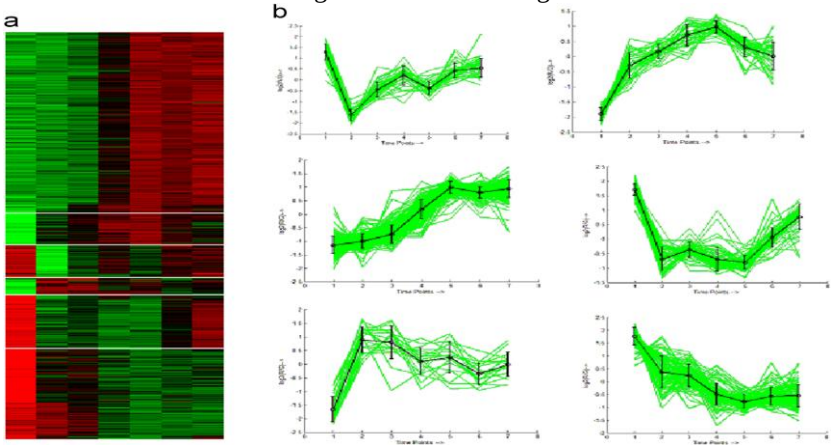

Fig. 11 yeast sporulation data (Chu et al., 1998) clustering using (a)Eisen plot (b)Cluster profile plot visualization tools

Yeast data: TABLE I shows the clustering of yeast sporulation data set.

TABLE I CLUSTERING YEAST SPORULATION DATA

Algorithms used	Number of Genes	Time points	Threshold (log2-transform)	Significant genes/time points
Hierarchical, K-means, Diana, Model-based, partial least squares[9]	6,118	7	1.13	513/7
MO-fuzzy, MOGA, SGA, CRC, SiMM-TS, IFCM, VGA, SOM, Average Linkage[13] [6]	6,118	7	1.6	474/7
MST[12]	68	79	-	-

B. Yeast cell-cycle

The series of events that occur in a cell and lead to duplication and division are referred to as the cell cycle. The cell cycle consists of four distinct phases (G1, S, G2 and M) and is regulated similar to that of the cell cycle in larger eukaryotes. As long as adequate nutrients such as sugar, nitrogen and phosphate are present yeast cells will continue to divide asexually.

Yeast cell-cycle data used the algorithms such as Mo-fuzzy, MOGA, SGA, CRC, DGSOT, and CLICK.

In [11], yeast cell cycle dataset of Cho et al. (1999) contains the expression levels of 6, 218 S. cerevisiae putative gene transcripts (identified as ORFs) measured at 10-minute intervals over two cell cycles (160 minutes). The filtering removes genes which do not change significantly across samples, resulting in a set of 826 genes.

In [6], The Yeast Cell Cycle data set was determined from a dataset which shows the variation of expression patterns of approximately 6000 genes over two cell cycles (17 time points). 384 genes out of total 6000 genes have been chosen to be cell-cycle regulated as shown in Fig. 12.

In [14], applied the DGSOT to the yeast gene cell cycle gene expression profiles published by Cho *et al*. (1998) was used as input data set. They used Affymetrix oligonucleotide microarrays to test the expression of more than 6000 genes of yeast Saccharomyces cerevisiae. The gene expressions of yeast cell were examined in 17 time points at 10 min intervals, which covered over two cell cycles. First, normalized the data to have unit variance and zero mean (*Z*-score normalization). Second selected 3000 genes with the most relative variation (standard variance over the mean) in expression level across 15 time points for clustering.

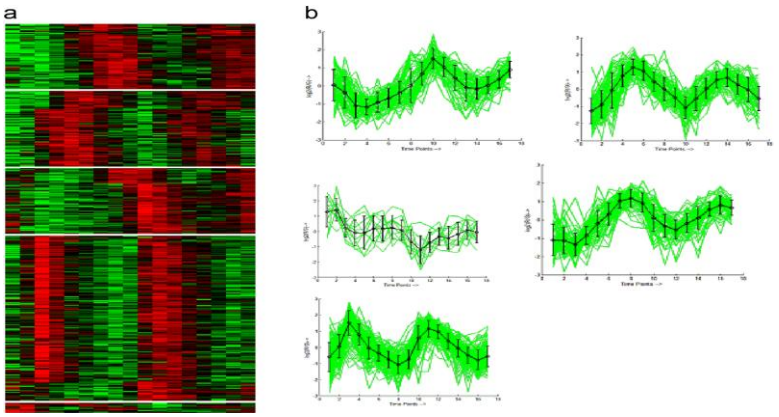

Fig. 12 Yeast cell cycle data (Cho et al., 1998) clustering using (a)Eisen plot (b)Cluster profile plot visualization tools

Yeast cell-cycle: TABLE II shows the clustering yeast cell-cycle data set in which determined over two cell-cycles.

TABLE IIII CLUSTERING YEAST CELL-CYCLE

Algorithms used	Number of Genes	Time points	Significant genes/time points
CLICK, GeneClust[11]	6,218	10	826/10
MO-fuzzy, MOGA, SGA, CRC[6]	6,000	17	384/17
DGSOT[14]	6000	17	3000/17

C. Arabidopsis Thaliana

Arabidopsis thaliana is a small flowering plant native to Eurasia. Arabidopsis thaliana is edible by humans and, as with other mustard greens, is used in salads or sautéed, like many species in the Brassicacea. Considered a weed, it is found by roadsides and in disturbed lands. A winter annual with a relatively short life cycle, Arabidopsis is a popular model organism in plant biology and genetics. Arabidopsis thaliana was the first plant which is genome sequenced and is a popular tool for understanding the molecular biology of many plant traits such as flower development and light sensing.

Arabidopsis Thaliana used the algorithms such as MO-fuzzy, MOGA, SGA, CRC and MST.

In [6], Gene expression values of 138 genes of Arabidopsis Thaliana are contained in this data set as shown in Fig. 13. It consists of expression levels of the genes over 8 time points viz., 15 min, 30 min, 60 min, 90 min, 3 h, 6 h, 9 h, and 24h.

In [12], a set of gene expression data of Arabidopsis in response to chitin elicitation (Ramonel et al., 2001) was used. The data was averaged over two experiments. Each gene had six data points (collected at 10min, 30min, 1h, 3h, 6h and 24h). 68 genes were selected for clustering, each containing at least one data point with a 3-fold change of expression level by chitin elicitation.

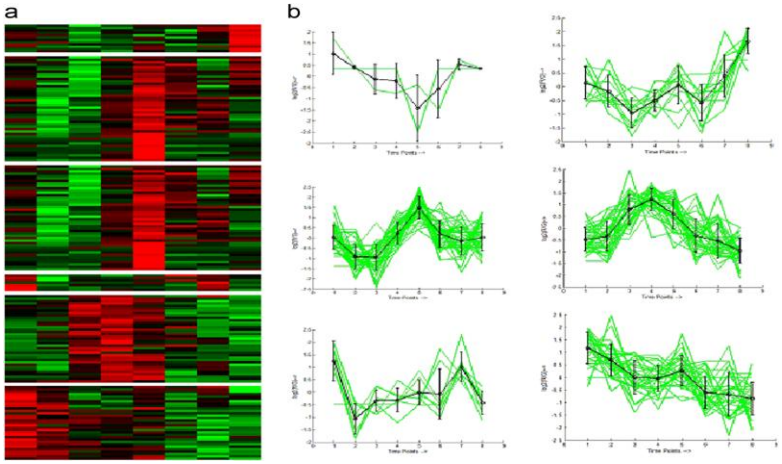

Fig. 13 Arabidopsis Thaliana data clustering using (a)Eisen plot (b)Cluster profile plot visualization tools

Arabidopsis Thaliana: TABLE III shows the clustering Arabidopsis thaliana data set.

TABLE IIIII CLUSTERING ARABIDOPSIS THALIANA

Algorithms used	Number of Genes	Time points	Significant genes/time points
MO-fuzzy, MOGA, SGA, CRC[6]	138	8	138/8
MST[12]	68	6	68/6

D. Serum

Fibroblasts are a heterogeneous population of cells found in numerous tissues and are of mesenchymal origin. The most common products secreted by fibroblasts are components of the extracellular matrix (ECM) such as different types of collagen, fibronectin and proteoglycans. Three types of human fibroblasts, namely, cardiac, dermal, and pulmonary fibroblasts. The proliferation of fibroblasts in accordance to their different tissue origins to determine whether tissue-specific proliferation rates of this cell type can be found. Primary human fibroblasts derived from heart, lung, and skin. cDNA probes were made from Human fibroblasts treated with serum or serum-deprived cells. For each spot, the ratio of green to red fluorescence gives the ratio of expression for each gene in serum-treated versus serum-deprived cells.

Human Fibroblasts serum used the algorithms such as CRC, SOM, MO-fuzzy, average link, IFCM, VGA, SGA, MST, MOGA, SiMM-TS and CLICK.

In [11], the dataset of Iyer et al. (1999) contains the expression levels of 8,613 human genes obtained as follows: Human fibroblasts were derived of serum for 48hours and then stimulated by addition of serum. Expression levels of genes were measured at 12 time-points after the stimulation. An additional data-point was obtained from a separate unsynchronized sample. A subset of 517 genes whose expression levels changed substantially across samples was analyzed by the hierarchical clustering method of Eisen et al. (1998).

In [12], a set of temporal gene expression data in response of human fibroblasts to serum (Iyer et al., 1999) was used as input data set. The data set consists of 517 genes and each gene has 18 data points. In this problem, work with Euclidean distance as the distance measure.

In [6] [13], (Iyer et al., 1999) Expression levels of 8613 human genes are contained in this dataset as shown in Fig. 14. There are total 13 dimensions in this data set corresponding to 12 time points (0, 0.25, 0.5, 1, 2, 4, 6, 8, 12, 16, 20 and 24 h) and one unsynchronized sample. Thereafter total 517 genes have been selected for consideration where expression levels of these genes changed drastically across the time points. Log2-transformation is then applied on these 517 genes.

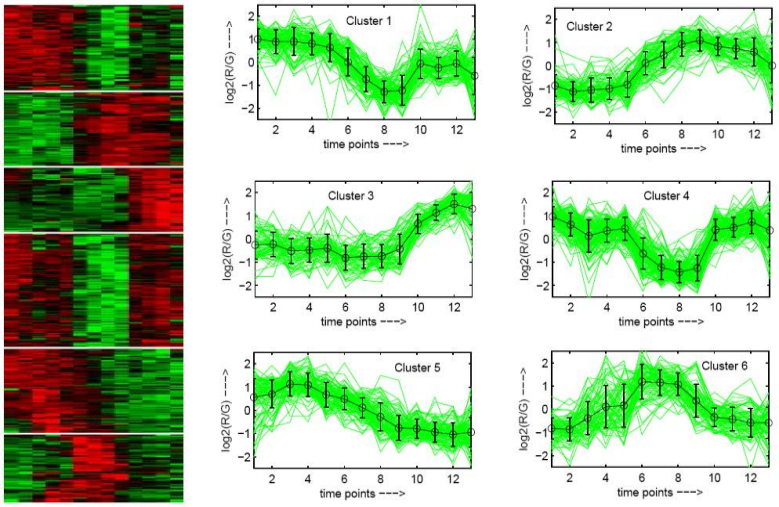

Fig. 14 Human Fibroblasts serum clustering using (a)Eisen plot (b)Cluster profile plot visualization tools

Serum: TABLE IV shows the experimental results of human fibroblasts data set.

TABLE IVV CLUSTERING HUMAN FIBROBLASTS SERUM

Algorithms used	Number of Genes	Time points	Significant genes/time points
MO-fuzzy, MOGA, SGA, CRC, CLICK, MST, SiMM-TS, IFCM, VGA, SOM, Average Linkage[6] [11] [12] [13]	8,613	12	517/12

E. Rat CNS

The central nervous system (CNS) is the part of the nervous system consisting of the brain and spinal cord. The CNS is so named because it integrates information it receives from, and coordinates and influences the activity of, all parts of the bodies of bilaterally symmetric animals. Sometimes experimental allergic encephalomyelitis (EAE) is brain inflammation model for animals. It is an inflammatory demyelinating disease of the CNS. It is mostly found with rodents.

Rat CNS was used in the algorithms such as CRC, SOM, MO-fuzzy, average link, IFCM, VGA, SGA, MOGA and SiMM-TS.

In [6] [13], (Wen et al., 1998) this data set is calculated using reverse transcription- coupled PCR, which is executed to determine the expression levels of a set of 112 genes during rat central nervous system development over 9 time points as shown in Fig. 15.

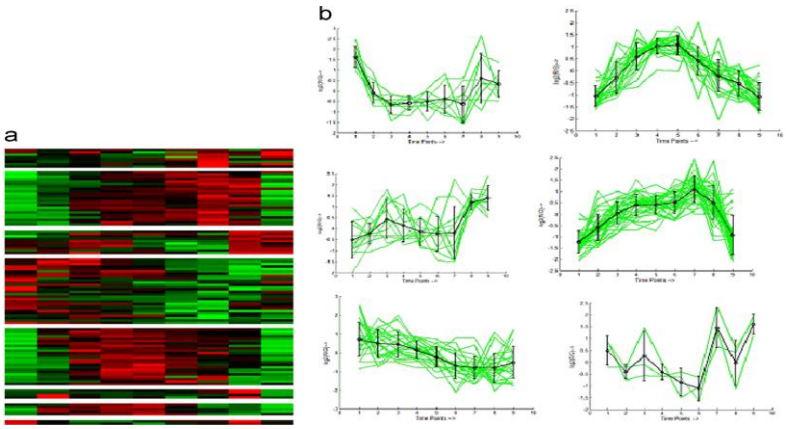

Fig. 15 Rat CNS data (Wen et al., 1998) clustering using (a)Eisen plot (b)Cluster profile plot visualization tools

Rat CNS: TABLE V shows the experimental results of ret CNS data set.

TABLE V CLUSTERING RAT CNS

Algorithms used	Number of Genes	Time points	Significant genes/time points
MO-fuzzy, MOGA, SGA, CRC, SiMM-TS, IFCM, VGA, SOM, Average Linkage[6] [13]	112	17	112/9

VI. CONCLUSIONS

Now a days, data mining and bioinformatics are fast growing research area in which some partial knowledge is often available for gene expression data sets. Clustering algorithms are useful for identifying biologically relevant groups of genes and samples. Clustering techniques is essential in the data mining process to reveal natural structure and identifying pattern in the data sets. Microarray is a technology that uses gene chip for the gene experiments. In clustering, this microarray technology is familiar used for cluster genes based on the experimental conditions.

In this paper, the clustering of microarray data sets are analyzed and discussed along with some of the distance measures and necessary parameter. To better understand the clustering results, different visualization plots are available and that has been discussed. Description and the use of microarray data sets that are popular have been discussed. All these have been seen in many literatures with their visualization results.

Finally, this paper definitely provides the clear picture on clustering used in microarray datasets, visualization tools used for microarray datasets, which may be really helpful for the researchers those who would like to do their research in microarray datasets.

REFERENCES

[1] Rahila H. Sheikh, M. M. Raghuwanshi, Anil N. Jaiswal, *Genetic Algorithm Based Clustering: A Survey*, IEEE Computer Society, DOI 10.1109/ICETET.2008.48.

[2] Margaret H. Dunham, *Data Mining Introductory and Advanced Topics*, Pearson education, 2003, ISBN:01300888923.

[3] IllhoiYoo, Patricia Alafaireet, MiroslavMarinov, *Data Mining in Healthcare and Biomedicine: A Survey of the Literature*, Springer Science-Business Media, LLC 2011.

[4] Anoop Kumar Jain, Prof. Satyam Maheswari, *Survey of Recent Clustering Techniques in Data Mining*, International Journal of Computer Science and Management Research, Vol 1 Issue 1 Aug 2012, ISSN 2278-733X.

[5] Khalid Raza, *Application of Data Mining in Bioinformatics*, Indian Journal of Computer Science and Engineering, Vol 1 No 2, 114-118, http://www.ijcse.com/docs/IJCSE10-01-02-18.pdf.

[6] SriparnaSaha, Asif Ekbal, Kshitija Gupta, SanghamitraBandyopadhyay, *Gene expression data clustering using a multiobjective symmetry based clustering technique*, Computers in Biology and Medicine 43 (2013), Elsevier Ltd, http://dx.doi.org/10.1016/j.compbiomed.2013.07.021.

[7] Chloé-Agathe Azencott and Karsten Borgwardt, Data Mining in Bioinformatics, *Day 7: Clustering in Bioinformatics, Clustering Gene Expression Data*, February 18 to March 1, 2013, Machine Learning & Computational Biology Research Group, Max Planck Institutes Tübingen and Eberhard Karls Universität Tübingen.

[8] Harun Pirim, BurakEks-io˘glu, Andy D. Perkins, Cetin Yuceer, *Clustering of high throughput gene expression data*, Computers & Operations Research 39 (2012)3046–3061, journal home page: www.elsevier.com/locate/caor.

[9] Susmita Datta and Somnath Datta, *Comparisons and validation of statistical clustering techniques for microarray gene expression data*, Vol. 19 no. 4, 2003, pages 459–466, DOI: 10.1093/bioinformatics/btg025.

[10] Marcilio CP de Souto, Ivan G Costa, Daniel SA de Araujo, Teresa B Ludermir and Alexander Schliep, *Clustering cancer gene expression data: a comparative study*, Published: 27 November 2008, BMC Bioinformatics 2008, 9:497 doi:10.1186/1471-2105-9-497,http://www.biomedcentral.com/1471-2105/9/497.

[11] R.Sharan, R.Elkon, R.Shamir, *Cluster Analysis and its Applications to Gene Expression Data*, http://www.bioinfo.uqam.ca/bif7001/articles/BIF7001-MA-ErnstSchering2002.pdf.

[12] Ying Xu, Victor Olman, Dong Xu, *Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees*, Bioinformatics, Vol. 18 no. 4 2002, pages 536-545.

[13] Sanghamitra Bandyopadhyay, Anirban Mukhopadhyay and Ujjwal Maulik, *An improved algorithm for clustering gene expression data*, Gene expression, Bioinformatics, Vol. 23 no. 21, 2007, pages 2859–2865, DOI: 10.1093/bioinformatics/btm418.

[14] Feng Luo, Latifur Khan1, FarokhBastani, I-Ling Yen andJizhong Zhou, A *dynamically growing self-organizing tree (DGSOT) for hierarchical clustering gene expression profiles*, Bioinformatics, Vol. 20 no. 16 2004, pages 2605–2617, DOI: 10.1093/bioinformatics/bth292.

[15] Sergei L Kosakovsky Pond, *Clustering in Bioinformatics*, (SPOND@UCSD.EDU), CSE/BIMM/BENG r8r, May 24, 2011.

[16] Jiawei Han and Jing Gao, *Chapter 1: Research Challenges for Data Mining in Science and Engineering*, University of Illinois at Urbana-Champaign, http://web.engr.illinois.edu/~hanj/pdf/ngdm09_han_gao.pdf.

[17] Jason T.L. Wang, Mohammed J. Zaki, Hannu T.T. Toivonen and Dennis Shasha (Eds), *Data Mining in Bioinformatics*, Springer-Verlag LondonLimited2005,ISBN:1852336714.

