Clustering Algorithms for Gene Expression Data: A Review

Deepika Kumar[1], Dr Usha Batra[2]

[1]PhD Scholar, SoE, G D Goenka University, Gurgaon, India
[2]Associate prof, SoE, G D Goenka University, Gurgaon, India

***Abstract.* The DNA Microarray Technology has the capability to measure gene expression levels under various experimental conditions. It is an important task in Bioinformatics research to identify genes have similar patterns or characteristics in microarray or gene expression data analysis. It can be resolved by using clustering algorithms, which help in identifying distribution, natural structure and exploring the patterns of the given data. Clustering is an important technique in data mining process. It is the process of dividing genes into clusters so that genes within a cluster possesses similar features and share a common biological role. It is widely used method for the extraction of meaningful data from gene expression data analysis. This paper gives a brief introduction to DNA microarray technology and various clustering techniques. This paper also gives the detailed description of various clustering algorithms used for extracting meaningful information from gene expression data.**

***Keywords:* Microarray technology, clustering, gene expression data.**

I. INTRODUCTION

Gene expression analysis is very important in bioinformatics research, since any minor change in organism or cell effects the gene expression pattern as well. The DNA microarray technology [1] is an emerging technology and results high throughput when thousands of gene expression are analyzed. The main aim of gene expression data analysis is to identify different levels of gene expression, and genes with similar profile [2]. The DNA microarray data for gene expression is transformed into matrix form where row represent genes and column represent sample or expression level of gene in a particular sample [3]. Gene clustering is used for extracting meaningful information from different expression profiles. For example similar gene expression profile indicates that corresponding genes are interrelated to each other or they share a common biological role. It helps in understanding various gene function, relationship between genes, cellular processes etc. [1, 4]. Clustering is done for initial data analysis and data mining process. Many clustering algorithm were discussed in literature based on different approaches for extracting meaningful information from cluster or sample of gene expression data.

Microarray technology is used to measure gene expression levels. The two main types of microarray experiments are the cDNA microarray [6] and oligonucleotide arrays [7].They enable to view on the transcription levels of genes, under specific processes or conditions [5].It helps in understanding gene expression levels in different development stages, clinical condition, tissue types, and gene networks for disease prediction, diagnosis and medical treatment outcomes.

Clustering Algorithm is the initial step in gene expression data study and identifies gene cluster that exhibit similar features [8]. The aim is to divide the elements into clusters according to homogenous elements, where elements of a cluster kept according to their high and low similarity features. The group of elements have high similarity feature are in one and low in another.

The rest of the paper is organized as follows: section 2 describes the challenges of gene expression data. In section 3 evaluation criteria for clustering algorithm is discussed. Various gene based clustering approached are discussed in Section 4. Finally we conclude in Section 5.

II. CHALLENGES OF GENE EXPRESSION DATA

Gene expression data is analyzed to extract meaningful information from noisy data. A good clustering algorithm should be able to plot graphical representation of the clusters. They possess highly connected and embedded pattern .Therefore gene based clustering algorithm is used to effectively handle this. The main aim of biologists and researchers to find the relationship among various genes clusters or sub clusters. The challenges for the gene expression data clustering are [9]:

- The attributes used for gene clustering.
- The parameter defined for describing the true cluster or cluster structure.
- The choice of proximity measure used for clustering procedure.
- Gene expression algorithm should be capable of extracting meaningful information out of excessive noisy and huge data.

III. EVALUATION CRITERIA FOR CLUSTERING ALGORITHMS

The gene expression based clustering algorithm must have the following properties [35]:

- Robustness: It should be robust against huge amount of noisy data.
- Scalability and efficiency: It should be efficient and scalable so that it will be to handle large amount of data.
- Order insensitivity: It should not be dependent upon the ordering of input data .It should be independent of data order.

- Irregular shape: Algorithm should be able to detect irregular non spherical clusters as well as dense set of cluster points.
- Cluster number: The number of clusters or sub clusters in the data should be identified by the clustering algorithm only.
- Parameter estimation: It should be able to identify data set parameters. A prerequisite information about the dataset should not be asked by the user.
- Dimensionality: It should be capable of handling high dimensional data.
- Stability: The algorithm should be stable.
- Incrementability: It should be able to handle incremental data set. It should be able to add new element or remove old element from old dataset without running the complete algorithm again for the new dataset.

IV. GENE BASED CLUSTERING APPROACHES : A REVIEW

In this section we discuss the various gene based clustering approaches. The aim of gene based clustering is to identify group of co-expressed gene. There are many clustering algorithms available for gene expression data analysis. They are broadly classified into:

Partition based approach

Partition based approach are broadly divided into centroid based and mediod based. In centroid based approach a cluster is represented by using the gravity centre of the instances and mediod based represents cluster by means of the instances to the gravity centre.

K-means algorithm [10] is the common method used for clustering gene expression data. It is a typical partition based algorithm. It is simple and fast algorithm used for large amount of gene expression dataset. Gene expression data have an enormous amount of noisy data and it forces each gene is to be included in a single cluster which can create biologically irrelevant clusters. However, it may be found incapable of detecting arbitrary shape clusters. It initially takes number of k known clusters and minimize the distance between the centroid of given clusters. It divides the dataset into k disjoint subsets. The main disadvantage of this algorithm is prior knowledge of number of gene cluster for gene expression data is required.

k-means clustering algorithm is widely used for clustering gene expression data because it is simple and easy to use. It also perform well when it is compared with new clustering algorithm. The various application of k means algorithm for clustering gene expression data is also discussed in literature [41, 42, 43, 48, 49]. K-modes [35] is an extension of k means algorithm. It is used to handle categorical data. In this algorithm k-means is replaced with k-mode and uses frequency basedmethod to update modes. It can be used only when numerical data is converted into categorical data. PAM and CLARA are the two early version of k-medoid approach. PAM uses dissimilar values and iterative approach for identifying a cluster point called medoid. Then non selected group is merged with most similar mediod. The efficiency of algorithm is measured by the average dissimilarity between cluster points and medoid of its cluster.

CLARANS [44] uses a random approach to find mediods that represents clusters. It takes maxneighbor and num local as input parameter. It select current node as arbitrary node and find group of neighbors of the node. It calculates the cost differential of the two nodes and identifies the better neighbor. The current node is compared with available maximum number of nodes and it is declared as maxneighbor and if it also have lowest cost among them then it is declared as local minimum. The local minimum cost is compared with rest of the lowest cost obtained till now. The lower of the two cost is stored as mincost. The algorithm starts search again for the local minima until numlocal is found.

Hierarchical clustering

Hierarchical clustering [14] is extensively used in of gene expression data analysis. This type of clustering is divided into two methods, agglomerative (top down) and divisive (bottom up).The top down approach uses different points as individual clusters and then merges with the closest pair of cluster. The bottom up approach uses inclusive cluster and splits cluster until each cluster have a point. It should be decided which cluster should be divided. It creates hierarchical series of nested clusters which are represented as tree called dendrogram. The leaves of dendrogram shows the similarity between the clusters and formation of clusters. UPGMA (Unweighted Pair Group Method with Arithmetic Mean) uses agglomerative method for the graphical representation of cluster data. DCCA (Divisive Correlation Clustering Algorithm) [15] uses Pearson's correlation as the similarity measure. Hierarchical clustering approach perform very good results for clustering gene expression data and it is also discussed in literature [2, 21, 43, 57, 58].

CURE [37] is more robust algorithm .It identifies dissimilar and non-spherical shapes. Each cluster is generated by identifying dispersed points from the cluster represented by fixed number of points. Then they shrink towards the center of the cluster. It uses random sampling and partitioning for handling large amount of data.

CHAMELEON [38] is a dynamic approach used for clustering which identifies similarity between two clusters .It uses a graph method to partition the data and uses a method which is based upon k-nearest neighbour. It combines the sub clusters and uses agglomerative hierarchical clustering algorithm for finding real clusters.

ROCK [39] also an agglomerative hierarchical clustering algorithm. It uses links to measure the similarity/proximity between a pair of data points in a cluster. Then it merge the data points of a cluster.

BIRCH [40] (Balanced Iterative and Clustering using Hierarchies) is suitable for large database. It is an incremental and dynamic clustering algorithm for input data. It results best clustering with available memory and time constraints. It uses multi scanning technique, a single scan outputs good results which can be maximized by multi scanning. It was the first algorithm which is able to handle noise.

AMOEBA [45] uses Dalaunay diagram to incorporate spatial proximity. There is no prior knowledge of dataset and parameters from user are required. Multilevel clustering is used and it also able to construct a tree graph that gives better understanding of hierarchy of clusters.

Model Based Approaches

Model based clustering approach [16] give a statistical framework for modelling gene expression data. The data set is supposed to come from finite distribution where each value corresponding to different cluster. The Expectation Maximization (EM) algorithm [17] determines good values for its parameters iteratively. It is able to handle different shapes of cluster, and lots of iteration are required that makes this algorithm costly. A signal shape similarity method is used to cluster genes in Variational Bayes algorithm18. A Model Based approach gives an approximate value that shows data points belong to a specific cluster. It results high correlation between two different gene clusters. The literature of the model based clustering approaches for gene expression data is discussed in [2, 8, 21, 59].

Self-Organizing Map (SOM) [11] is easy to implement, fast and scalable for large gene expression dataset. It is based on a single layered neural network. It is represented in a two dimensional m*n grid where data points are taken as input and output neuron. Then neurons are represented as simple neighborhood structure. A reference number is attached with each neuron, and each data point is mapped to the nearest reference vector. Each data point is act as training sample which leads the movement of reference vectors towards the deeper input space so that it will is distributed to input dataset. Clusters are identified by mapping all data points to the output neuron after the completion of training process. The Self organizing map clustering algorithm starts with the initialization of the reference vector followed by randomly selection of data points. Then nearest reference vector to the current data point is determined and finally reference vector and neighboring reference vectors are updated. SOM is very efficient method used for gene expression data clustering and it is also discussed in literature [5, 11, 12, 21, 41, 60].

AutoClass [47] uses the Bayesian approach and randomly initialize the parameters. It is used to find class description which predict the data. There is no need to specify number of classes by the user. It uses the heterogeneous data means both real & discrete valued data. It is also able to handle large dataset and missing values. The use of AutoClass for gene expression data is discussed in [50].

Soft Computing Approaches

Fuzzy clustering [19] algorithm connect each gene to all clusters with a real valued index vectors. The indexed vector states that the membership of a gene with respect to other clusters. The values of components lie between 0 and 1.If a gene value is close to 1 it shows a strong association to the cluster. If it is close to 0 it shows weak association to the cluster. Fuzzy c-means and Genetic algorithm [20, 21] are used for gene expression data clustering. The Genetic Algorithm are highly dependent on input parameter. FCM uses Euclidean distance and encounter error in identification of initial partitioning, shapes of all clusters. The various application of FCM is discussed in [20, 56].

GenClust [52] is a genetic algorithm for gene expression clustering. It uses two main features: a) a novel coding for search space, b) use of internal data driven validation methods. It display data in a very simple way. It then finds the local optimum value and identify meaningful clusters. GenClust works in stages and gives a sequence of partition, each have classes until termination condition is reached. It assembles the clusters according to cluster number.

A Genetically Based Clustering Algorithm (GCA) is proposed in [55].It works on split and merge technique for finding clusters. The complete dataset is divided into large number of clusters, then these clusters are merged using HCMA (Hierarchical Cluster Merging Algorithm).It uses several cycles until k clusters are obtained. This merging cluster technique is based on genetic algorithm.

GA Based Clustering are based upon evolution & natural genetic which uses random and optimized search technique. There are several GA based clustering algorithm are discussed in literature [53, 54, 55].

Incremental Approach

In [22] an incremental clustering algorithm based on DBSCAN is presented. In [25] Incremental Genetic k-means algorithm is proposed which calculates total objective value within cluster variation (TWCV) that increments to cluster centroid whenever its mutation probability is small. HIREL (Hierarchical Incremental Relational Clustering) [23] Algorithm is proposed for clustering interval datasets. It restrict to the shape of the derived clusters and minimizes the total number of distance computation. The whole data is required to scan only once because clusters are updated when new data is inserted into it. An incremental rough Set theory is used for clustering interval datasets in [24]. An incremental gene selection algorithm is proposed in [51]. It uses a wrapper based method and works on directly ranking system that minimizes search space complexity.

Density Based Approach

Density based clustering [26] identifies clusters which has highly dense areas separated by sparsely dense areas. Density based hierarchical clustering method was proposed to identify co expressed gene groups [29]. It identifies the outliers and embedded clusters in the dataset and internal structure of the cluster. However density based clustering technique depends on input parameter and have high computational complexity.

DBSCAN [26] (Density Based Spatial Clustering of Application with Noise) was designed to identify the clusters and the noise in database. It depends on a density based notion of clusters which identifies the arbitrary shape of clusters. It requires only one input parameter. An experiment was performed for checking the effectiveness and efficiency of algorithm using real data of SEQUOIA 2000 and results shows: a) DBSCAN is better than CLARNS in discovering arbitrary shapes. b) DBSCAN perform better than CLARNS by a factor of 100 in terms of efficiency.

Gene Clus Tree [28] identifies clusters over gene expression data. It used tree based density approach to find all clusters over subspaces. It scans complete database in minimum time. This technique is not restricted to use proximity measure and effectiveness of this is measured in zscore and p-value over real time datasets. The p-value analysis shows that Gene Clus Tree is capable of identifying biological clusters from gene expression data.

DBCLASD30 (Distribution Based Clustering of LAarge Database) algorithm identifies cluster of points belongs to spatial point. It is nonparametric in nature and discovers good quality of arbitrary shape clusters. DBCLASD assigns a point in cluster without identifying the complete cluster or database. It increments initial cluster by its neighboring points until nearest neighbor distance of the resulting cluster fits the expected distance.

OPTICS [31] (Ordering Points To Identify The Clustering Structure) algorithm creates a augmented ordering of the database which represents its density based clustering structure. It is very good algorithm for both interactive and automatics cluster analysis. It extracts both traditional clustering information (cluster points, arbitrary shapes of cluster) and inherent clustering structure. OPTICS creates ordering of a database and stores two values of an object: core distance and reachability distance.

DENCLUE [32] is generalization of hierarchical, partitioning, density based clustering methods. Pre-clustering is done by creating a map of the active portion of dataset which speeds up the calculation. It identifies the density attractors and their corresponding cluster points. DENCLUE identify clusters of arbitrary shape and have good clustering properties in presence of noise.

DHC [46] Density based Hierarchical Clustering algorithm effectively handle the time series gene expression data. The result is shown in the form of a density tree which is able to show embedded cluster in the dataset. All the objects in a dataset are organized into an attraction tree according to their density based connectivity. Then the number of clusters and dense are identified.

Graph based Approach

Graph based approach work with data represented in terms of graph. Graphs are built as combination of nodes, edges and classified using graph based algorithm.

CLICK [33] uses graph-theoretic and statistical technique to identify group of similar elements (kernels) then many heuristic measures are applied to expand the kernels into clustering. It defines the weight and vertices of an edge within the same cluster and then minimum cut in the graph is identified. It divide the dataset into a set of connected components on the basis of a predefined threshold value. CLICK has been tested on different biological datasets like gene expression cDNA olino-fingerprinting to protein sequence similarity. CLICK is very fast and generates good quality cluster over gene expression dataset in terms of similarity and separation. CLICK better recognizes the intersecting clusters. It performs well for the gene expression data clustering in terms of homogeneity and separation of clusters [5, 8, 51].

CAST [34] works on the concept of clique graph and uses divisive clustering approach. It is an undirected graph which is the union of disjoint complete graphs. Therefore it is assumed that it has true biological partition of the genes into disjoint clusters which are based upon gene functionality. The clique graph is composed of clusters (cliques) of genes (vertices) whose edges (interconnections) presence depends upon their respective similarity measures. Then similar gene is kept in same subgraph and genes which are not similar to each other not kept in clique. The algorithm make one cluster at a time and literature of CAST is discussed in [5, 8, 34].

Computational Complexity of Clustering Algorithms

Table 1. Clustering Algorithms computational complexity

Clustering Algorithm	Complexity	Capable of handling high dimensional data

K-means[10]	O(NKd) (time)
	O(N + K) (space)
		No

Fuzzy c- means [20]	Near O(N)	No
Hierarchical Clustering [14]	$O(N^2)$ (time) $O(N^2)$ (space)	No
ROCK [39]	$O(n^3)$	No
CHEMLEON [38]	$O(m^2 \log m)$	No
BIRCH [40]	$O(N)$ (time)	Yes
DBCLASD [30]	O(nlogn)	No
DBSCAN [26]	$O(NlogN)$ (time)	No
CURE [37]	$O(N^2_{sample} \log N_{sample})$ (time) $O(N_{sample})$ (space)	Yes
DENCLUE [32]	$O(NlogN)$ (time)	Yes

V. CONCLUSION

Clustering of gene expression data or biological data is relevant for biologist and researchers. Former one uses an existing clustering algorithm to solve the biological problem whereas later one is in consistent pursuit to improve existing algorithm thereby efficiently solving the basal biological problems. A clustering algorithm depends upon certain features like speed, sturdiness to noise and outliers, overabundance, minimum number of input parameter, independence of object order input. Normally it is beyond the bounds of possibility that all validity measures matches to all gene datasets, hence an efficacious choice becomes imperative for validity measure.

An extensive survey of existing clustering algorithm in context of rational and systematic pattern establishment in gene expression data is done in this review. Organized genes indicate co-regulation and show similar functional classification. It was discussed that different clustering algorithm require different type of input parameter and their results are dependent on values of parameter. Clustering algorithm are dependent to proximity measure chosen. This paper strives to proffer analysis of cogency of algorithm that belongs to explicit approach in gene expression data mining. At the end the advantages and disadvantages of different clustering techniques are concluded in table 2.

Table 2. Advantages and disadvantages of various type of clustering approach.

Approach	Advantage	Disadvantage
Partition Based Approach	• separate clusters in context of data mining	• No of clusters not known apriori. • Proximity measure used are not sufficient because of high dimensional gene expression data. • Embedded and intersected gene pattern are not detectable.
Hierarchical Clustering	• Show cluster result very effectively • Proximity measure used are sufficient. • Capable of handling high dimensional data.	• Represent intersected clusters pattern in inadequate manner.
Model Based Approach	• More relevant to gene expression data.	• Sensitive to input parameter. • Algorithm are expensive. • It is required to give number of cluster and grid structure as input.
Soft computing Approach	• Algorithm used are efficient, fast and exhibit constant learning and pattern detection	• Local optima problem. • Number of clusters should be declared first.
Incremental Approach	• Store cluster information regularly.	• Algorithm effectiveness depends upon proximity measure and differing density.
Density Based Approach	• Able to handle two dimensional data with uniform density distribution. • Capable of handling high dimensional numeric gene expression data. • Recognize uniform clusters as well as embedded and intersected cluster pattern. • Identify arbitrary shape cluster with varying size.	• Unable to handle high dimensional data. • Sensitive to input parameter.
Graph Based Approach	• Able to identify intersected and embedded pattern.	• Results are dependent upon proximity measure. • Result shown maybe in non-realistic.

References

[1]. Van Hal NL, Vorst O, van Houwelingen AM, Kok EJ, Peijnenburg A, Aharoni A, van Tunen AJ, Keijer J. The application of DNA microarrays in gene expression analysis. Journal of Biotechnology. 2000 Mar 31;78 (3):271-80.

[2]. Datta S, Datta S. Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics. 2003 Mar 1; 19 (4):459-66.

[3]. Brazma A, Vilo J. Gene expression data analysis. FEBS letters. 2000 Aug 25; 480 (1):17-24.

[4]. Quackenbush J. Computational analysis of microarray data. Nature reviews genetics. 2001 Jun 1;2 (6):418-27.

[5]. Sharan R, Elkon R, Shamir R. Cluster analysis and its applications to gene expression data. InBioinformatics and Genome Analysis 2002 (pp. 83-108). Springer Berlin Heidelberg.

[6]. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995 Oct 20; 270 (5235):467.

[7]. Govindarajan R, Duraiyan J, Kaliyappan K, Palanisamy M. Microarray and its applications. Journal

of pharmacy & bioallied sciences. 2012 Aug; 4 (Suppl 2):S310.

[8]. Jiang D, Tang C, Zhang A. Cluster analysis for gene expression data: a survey. IEEE Transactions on knowledge and data engineering. 2004 Nov; 16 (11):1370-86.

[9]. Bryan J. Problems in gene clustering based on gene expression data. Journal of Multivariate Analysis. 2004 Jul 31; 90 (1):44-66.

[10]. Zhang C, Xia S. K-means clustering algorithm with improved initial center. In Knowledge Discovery and Data Mining, 2009. WKDD 2009. Second International Workshop on 2009 Jan 23 (pp. 790-792). IEEE.

[11]. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proceedings of the National Academy of Sciences. 1999 Mar 16; 96 (6):2907-12.

[12]. Tomida S, Hanai T, Honda H, Kobayashi T. Analysis of expression profile using fuzzy adaptive resonance theory. Bioinformatics. 2002 Aug 1; 18 (8):1073-83.

[13]. Yedla M, Pathakota SR, Srinivasa TM. Enhancing K-means clustering algorithm with improved initial center. International Journal of computer science and information technologies. 2010 Jun; 1(2):121-5.

[14]. Luo F, Tang K, Khan L. Hierarchical clustering of gene expression data. InBioinformatics and Bioengineering, 2003. Proceedings. Third IEEE Symposium on 2003 Mar 10 (pp. 328-335). IEEE.

[15]. Bhattacharya A, De RK. Divisive Correlation Clustering Algorithm (DCCA) for grouping of genes: detecting varying patterns in expression profiles. Bioinformatics. 2008 Jun 1; 24 (11):1359-66.

[16]. Ghosh D, Chinnaiyan AM. Mixture modelling of gene expression data from microarray experiments. Bioinformatics. 2002 Feb 1; 18 (2):275-86.

[17]. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological). 1977 Jan 1:1-38.

[18]. Hestilow TJ, Huang Y. Clustering of gene expression data based on shape similarity. EURASIP Journal on Bioinformatics and Systems Biology. 2009 Mar 4; 2009 (1):1.

[19]. Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, *10*(2-3), 191-203.

[20]. Dembélé D, Kastner P. Fuzzy C-means method for clustering microarray data. Bioinformatics. 2003 May 22;19 (8):973-80.

[21]. Bandyopadhyay S, Mukhopadhyay A, Maulik U. An improved algorithm for clustering gene expression data. Bioinformatics. 2007 Nov 1; 23 (21):2859-65.

[22]. Ester M, Kriegel HP, Sander J, Wimmer M, Xu X. Incremental clustering for mining in a data warehousing environment. InVLDB 1998 Aug 24 (Vol. 98, pp. 323-333).

[23]. Li T, Anand SS. Hirel: An incremental clustering algorithm for relational datasets. In2008 Eighth IEEE International Conference on Data Mining 2008 Dec 15 (pp. 887-892). IEEE.

[24]. Asharaf S, Murty MN, Shevade SK. Rough set based incremental clustering of interval data. Pattern Recognition Letters. 2006 Apr 15; 27 (6):515-9.

[25]. Lu Y, Lu S, Fotouhi F, Deng Y, Brown SJ. Incremental genetic K-means algorithm and its application in gene expression data analysis. BMC bioinformatics. 2004 Oct 28; 5 (1):1.

[26]. Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. InKdd 1996 Aug 2 (Vol. 96, No. 34, pp. 226-231).

[27]. Shu G, Zeng B, Chen YP, Smith OH. Performance assessment of kernel density clustering for gene expression profile data. Comparative and Functional Genomics. 2003 Jun 1; 4(3):287-99.

[28]. Sarmah S, Sarmah RD, Bhattacharyya DK. An effective density-based hierarchical clustering technique to identify coherent patterns from gene expression data. In Pacific-Asia Conference on Knowledge Discovery and Data Mining 2011 May 24 (pp. 225-236). Springer Berlin Heidelberg.

[29]. Jiang D, Tang C, Zhang A. Cluster analysis for gene expression data: a survey. IEEE Transactions on knowledge and data engineering. 2004 Nov; 16(11):1370-86.

[30]. Xu X, Ester M, Kriegel HP, Sander J. A distribution-based clustering algorithm for mining in large spatial databases. In Data Engineering, 1998. Proceedings. 14th International Conference on 1998 Feb 23 (pp. 324-331). IEEE.

[31]. Ankerst M, Breunig MM, Kriegel HP, Sander J. OPTICS: ordering points to identify the clustering structure. InACM Sigmod Record 1999 Jun 1 (Vol. 28, No. 2, pp. 49-60). ACM.

[32]. Hinneburg A, Keim DA. An efficient approach to clustering in large multimedia databases with noise. InKDD 1998 Aug 27 (Vol. 98, pp. 58-65).

[33]. Sharan R, Shamir R. CLICK: a clustering algorithm with applications to gene expression analysis. InProc Int Conf Intell Syst Mol Biol 2000 Aug 19 (Vol. 8, No. 307, p. 16).

[34]. Ben-Dor A, Shamir R, Yakhini Z. Clustering gene expression patterns. Journal of computational biology. 1999 Oct 1; 6 (3-4):281-97.

[35]. J. Han and M. Kamber,Data Mining: Concepts and Techniques,The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers, August 2000. ISBN 1-55860-489-8.

[36]. Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values.

Data mining and knowledge discovery. 1998 Sep 1; 2 (3):283-304.

[37]. Guha S, Rastogi R, Shim K. CURE: an efficient clustering algorithm for large databases. In ACM SIGMOD Record 1998 Jun 1 (Vol. 27, No. 2, pp. 73-84). ACM.

[38]. Karypis G, Han EH, Kumar V. Chameleon: Hierarchical clustering using dynamic modeling. Computer. 1999 Aug; 32 (8):68-75.

[39]. Guha S, Rastogi R, Shim K. ROCK: A robust clustering algorithm for categorical attributes. In Data Engineering, 1999. Proceedings, 15th International Conference on 1999 Mar 23 (pp. 512-521). IEEE.

[40]. Zhang T, Ramakrishnan R, Livny M. BIRCH: an efficient data clustering method for very large databases. InACM Sigmod Record 1996 Jun 1 (Vol. 25, No. 2, pp. 103-114). ACM.

[41]. J. H. Do and D. -K. Choi, "Clustering Approaches to Identifying Gene Expression Patterns from DNA Microarray Data," *Molecular Cells,* vol. 25, no. 2, pp. 1-1, 2007.

[42]. Do JH, Choi D. Clustering approaches to identifying gene expression patterns from DNA microarray data. Molecules and cells. 2008 Apr 30; 25 (2):279.

[43]. Thalamuthu A, Mukhopadhyay I, Zheng X, Tseng GC. Evaluation and comparison of gene clustering methods in microarray analysis. Bioinformatics. 2006 Oct 1; 22 (19):2405-12.

[44]. Ng RT, Han J. Efficient and Effective Clustering Methods for Spatial Data Mining. InProc. of 1994 Sep (pp. 144-155).

[45]. Estivill-Castro V, Lee I. AMOEBA: Hierarchical clustering based on spatial proximity using Delaunay diagram. In Proceedings of the 9th International Symposium on Spatial Data Handling. Beijing, China 2000 Aug 10.

[46]. Jiang D, Pei J, Zhang A. DHC: a density-based hierarchical clustering method for time series gene expression data. InBioinformatics and Bioengineering, 2003. Proceedings. Third IEEE Symposium on 2003 Mar 10 (pp. 393-400). IEEE.

[47]. Cheeseman P, Self M, Kelly J, Stutz J. Bayesian Classification, 1996.

[48]. Costa IG, de Carvalho FD, de Souto MC. Comparative analysis of clustering methods for gene expression time course data. Genetics and Molecular Biology. 2004; 27(4):623-31.

[49]. Borg A, Lavesson N, Boeva V. Comparison of Clustering Approaches for Gene Expression Data. In SCAI 2013 (pp. 55-64).

[50]. Achcar F, Camadro JM, Mestivier D. AutoClass@ IJM: a powerful tool for Bayesian classification of heterogeneous data in achcar biology. Nucleic acids research. 2009 May 27:gkp430.

[51]. Ruiz R, Riquelme JC, Aguilar-Ruiz JS. Incremental wrapper-based gene selection from microarray data for cancer classification. Pattern Recognition. 2006 Dec 31; 39 (12):2383-92.

[52]. Di Gesú V, Giancarlo R, Bosco GL, Raimondi A, Scaturro D. GenClust: A genetic algorithm for clustering gene expression data. BMC bioinformatics. 2005 Dec 7; 6 (1):289.

[53]. Cowgill MC, Harvey RJ, Watson LT. A genetic algorithm approach to cluster analysis. Computers & Mathematics with Applications. 1999 Apr 30; 37 (7):99-108.

[54]. Tseng LY, Yang SB. A genetic approach to the automatic clustering problem. Pattern Recognition. 2001 Feb 28; 34 (2):415-24.

[55]. Garai G, Chaudhuri BB. A novel genetic algorithm for automatic clustering. Pattern Recognition Letters. 2004 Jan 19; 25 (2):173-87.

[56]. Wang YF, Yu ZG, Anh V. Fuzzy C-means method with empirical mode decomposition for clustering microarray data. International journal of data mining and bioinformatics. 2013 Jan 1; 7 (2):103-17.

[57]. Yeung KY, Haynor DR, Ruzzo WL. Validating clustering for gene expression data. Bioinformatics. 2001 Apr 1;17 (4):309-18.

[58]. Xing B, Greenwood CM, Bull SB. A hierarchical clustering method for estimating copy number variation. Biostatistics. 2007 Jul 1;8(3):632-53.

[59]. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL. Model-based clustering and data transformations for gene expression data. Bioinformatics. 2001 Oct 1; 17 (10):977-87.

[60]. Sugiyama A, Kotani M. Analysis of gene expression data by using self-organizing maps and K-means clustering. In Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on 2002 (Vol. 2, pp. 1342-1345). IEEE.

[61]. Yun T, Hwang T, Cha K, Yi GS. CLIC: clustering analysis of large microarray datasets with individual dimension-based clustering. Nucleic acids research. 2010 Jun 6:gkq516