Microarray Gene Expression Data Mining: Clustering Analysis Review

Erfaneh Naghieh and Yonghong Peng
Department of Computing, University of Bradford
E.Naghieh@Bradford.ac.uk, Y.H.Peng@Bradford.ac.uk

Abstract- *After genome sequencing, DNA microarray analysis has become the most widely used functional genomics approach in the bioinformatics field. Biologists are vastly plagued by the enormous amount of unprecedented qualities of genome-wide data produced by the DNA Microarray experiment. Clustering is the process of grouping data objects into set of disjoint classes called clusters so that objects within a class are highly similar with one another and dissimilar with the objects in other classes. It is presently the far most used method for gene expression analysis which provides a divide-and–conquer strategy to extract meaningful information from expression profile. This paper presents a review on the recently development of microarray clustering techniques. In this paper, the procedures of clustering analysis are highlighted followed by the different categories of gene expression data clustering with some conventional approaches, to provide a framework for an enhanced general understanding of related methods for further development.*

I. INTRODUCTION

The eminence of DNA microarray technology [1] is the aptitude to be used to simultaneously monitor and study the expression levels of thousands of genes, relationship between genes, their functions and classifying genes or samples that perform in a parallel or synchronized manner during imperative biological processes.

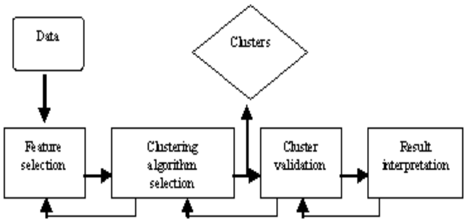
Functional genomics can be better implicit when the veiled patterns in gene expression data is elucidated, however, it is very challenging to comprehend and construe this due to the complexity of biological networks and large number of genes.

The most important area of microarray bioinformatics is possibly the data clustering analysis. Clustering is an exceptional preference for initial data analysis and data mining processes.

To perceive and identify appealing patterns of expression across multiple genes and experiments, reveal natural structures and compress high-dimensional array data clustering must be ascertained to allow easier management of data set. This data reduction method is a simple tool yet powerful method of organising genes based on their interdependence behaving similarly over the different conditions in different mutants, patients or at different time points in a time series during an experiment with similar expression patterns and properties into a set of disjoint groups based on specific features so that the underlying structures can be acknowledged and explored.

II. PROCEDURES OF CLUSTERING ANALYSIS

The procedures of cluster analysis are the feature selection, cluster algorithm selection, cluster validation and result interpretation. [2] The intimately connected steps of cluster analysis with feedback pathways is shown in the following figure.



A. Feature selection

Microarray experiments provide a expression information of large number of genes (from 10^3 to 10^4 or10^5). It is essential to consider which feature (gene) subset will be employed in clustering analysis, by eliminating the least interesting and highlight the most interesting genes. Distinctive features from a set of candidates are neatly selected, while feature extraction exploits some alteration to produce useful and novel features from the original ones which are very essential to the efficiency of clustering purpose [2].

B. Cluster algorithm design or selection

Different clustering algorithms and methods have been developed to improve the preceding ones, unravelling the problems and fit for specific fields [3]. There is no absolute clustering method that can be universally used to solve all problems. So in order to select or generate a suitable clustering strategy, it is vital to investigate the features of the problem.

As Xu and Wunsch [2] revealed the step is usually combined with the selection of a corresponding proximity measure and the construction of a criterion function. Patterns are grouped according to whether they resemble each other. Once a proximity measure is chosen, the construction of a clustering criterion function makes the partition of clusters an optimizing problem.

C. Clustering validation

Finding the number of clusters in a dataset and many of this method have been proposed some of which are silhouette index, Dunn's index, and Davis-Bouldin index [4] for gene expression data which evaluates the partitions generated using clustering algorithm and find the pre-eminent partition based on intra-cluster and inter-cluster distance.

It is vital to evaluate diverse clustering results, the quality and reliability of clusters before deciding on the finest data distribution.

D. Result explanation

Assessing the results and interpreting the clusters found are as significant as generating the clusters [5]. The objective of clustering is to solve the encountered problem efficiently and offer the users with significant understanding of their original data.

III. CLUSTERING TECHNIQUES

A. Distance metric

In order to group together similar objects, the meaning and measure of similarity has to be defined which is referred to as the distance metric which clustering is highly dependant on. A distant metric is a function that takes two points in a dimensional space where this should be symmetrical, positive and triangle unequal [6].

To calculate the distance between clusters different distance linkages are involved which affects the complexity and performance of the clustering.

Single linkage calculates the distance between the closest neighbours. Complete linkage calculates the distance between the furthest neighbours. Centriod linkage defines the distance between two clusters.

Average linkage measures the average distance between members of different clusters. Average and complete linkage are the preferred methods for microarray data analysis [7].

B. Clustering Techniques

Many diverse clustering techniques have extensively been under development [3]. The most widely used techniques in analysis of gene expression data which are applied in the early stages and proven to be useful are Hierarchical clustering [7], K-means clustering [8] and Self-organized maps (SOM) [9].

B.1) Hierarchical clustering

Hierarchical clustering [7] is the first and most common clustering method applied to gene expression data which is developed on the basis of a single layered neural network. A hierarchical series of nested clusters are generated by grouping genes with similar pattern of expression across a range of samples located near each other. Hierarchical clustering calculates all pairs-wise distance relationships between genes and experiments to merge pairs of values that are most similar for the formation of a node. The inter-cluster distance groups together these clusters to make a higher level cluster which can be graphically illustrated by a tree, called dendrogram representing the clusters and relationship between them. This is repeated, comparing genes or new clusters until all clusters are joined.

These methods are either agglomerative algorithms (bottom-up approach) which joins clusters in a hierarchical manner or the more rapid dividing algorithms (to-down approach) which splits clusters hierarchically.

The drawbacks of this method are its high-computational intricacy, lack of robustness, vagueness of termination criteria and failure with large number of genes as data sets grow in complexity.

B.2) K-means Clustering

K-means clustering [8] is a simple and fast method used commonly due to its straightforward implementation and small number of iterations. This algorithm divides the data set into k disjoint subsets. An estimation of the number of clusters (*k*) is made by the user and calculated as an input where the computer randomly assigns each gene to one of the k clusters.

The distance between each gene and the centre of each cluster is promptly calculated resulting in an optimal grouping of data to clusters where the genes inside every cluster are as close to the centre of the cluster as possible while at the same time there is maximal distance between genes of different clusters. This method is useful if different values of *k* are attempted and it only gives the number of clusters not the relationship between them like hierarchical clustering.

The drawbacks of this method are the lack of prior knowledge of the number of gene clusters in a gene expression data which results in the changing of results in the altering of results in successive runs since the initial clusters are selected randomly and the quality of the attained clustering has to be assessed.

B.3) Self-Organised Maps Clustering (SOM)

SOM [9] is a reasonably fast and easy to implement method, scalable to large data sets. It is intimately related to multidimensional scaling and its objective is to represent all data points in the source space by points in a target space where the distance and proximity relationships are preserved.

At the input, the data objects are presented and output neurons are organised with a sample neighbourhood grid structure. The remarkable features of SOM is that it generates an intuitively appealing map of a high-dimensional data set and places similar clusters near each other so that the neighbouring clusters in this grid are more related than clusters that are not neighbours.

SOM is trained through competitive learning for the distribution of the input data set which provides a relatively robust approach than k-means in the clustering of highly noisy data. However SOM requires users to input the number of clusters and the grid structure of the neuron map. After the completion of the training, clusters are identified by mapping all data points to the output neurons.

The drawbacks of this method is that it is not effective since the main interesting patterns may be merged into only one or two clusters and cannot be identified.

C. Supervised and Unsupervised clustering

Supervised sample-based clustering is extensively practical, reasonably easy to get high clustering accuracy rate of extracted data with the presence of a teacher signal. Supervised methods are used for analysis when trying to classify objects into known classes and finding genes that is mainly applicable to label classification [6].

Unsupervised sample-based clustering mines through data, congregating into a precise partition of the samples and a set of informative genes extracting relevant information without the presence of a teacher signal.

The main input a typical unsupervised clustering algorithm takes is the number of classes it should find. Unsupervised approach is more complex than supervised since no reference training set of samples can be developed to guide informative gene selection. Common unsupervised methods include hierarchical clustering, k-means clustering and self-organised maps etc.

Unsupervised methods such as hierarchical clustering are useful for exploring data sets to find the unexpected, or when additional information is not available about samples. Although unsupervised methods may seem to be less biased than supervised techniques, the ability to identify useful molecular classes may be enhanced if all of the available information is used. To provide robust and reliable conclusions, training and test should be largely independent [10].

IV. MICROARRAY DATA CLUSTERING ANALYSIS

Clustering gene expression data can be categorised into the three groups, 1) gene-based, 2) sample-based and 3) subspace clustering as both genes and samples is required to be clustered significantly.

A. Gene-based clustering

The gene-based clustering intends to group together coexpressed genes which indicate cofunction, coregulation and reveals the natural data structures [11]. Genes are treated as the object, while the samples are the features. Clustering algorithms for gene expression data should be competent of extracting useful information from a high level of background noise. A good clustering algorithm should depend as little as possible on prior knowledge also provide graphical representation of the cluster structure other than partitioning the data.

B. Sample-based clustering

To find the substructure of the sample, regards the samples as the objects and the genes as the features.

Samples are generally related to various disease or drug effects within a gene expression matrix. Only a small subset of genes whose expression levels strongly correlate with the class distinction, rise and fall coherently and exhibiting fluctuation of a similar shape under a subset of conditions, called the informative genes that participate in any cellular process relevant. The remaining genes are regarded as noise in the data as they are irrelevant to the sample of interest. By focusing on a subset genes and conditions of interest, the noise levels induced by other genes and conditions can be lowered which is characterised by co-clustering. Therefore to identify informative genes and reduction of gene dimensionality for clustering samples to detect their substructure particular methods should be applied.

C. Subspace clustering

To find subset of objects such that the objects emerge as a cluster in a subspace created by a subset of the feature [11]. The subset of features for different subspace clusters can be unlike in a subspace clustering.

Genes and samples are treated symmetrically such that either genes or samples can be regarded as objects or features. A single gene may participate in multiple pathways that may or may not be coactive under all conditions

Subspace clustering [12] techniques confine coherence exhibit by the blocks within gene expression matrices. A block is a sub-matrix defined by a subset of genes on a subset of samples.

C.1. Biclustering

Biclustering [13] performs simultaneous clustering on the row and column dimension of the data matrix where the gene exhibits highly correlated activities for every condition instead of clustering these two dimensions separately which distinct classes of clustering algorithms that perform simultaneous row-column clustering to identify submatrices, subgroups of genes and subgroups of conditions.

Clustering derives a global model while Biclustering produces a local model. Unlike clustering algorithms, Biclustering algorithms identify groups of genes that show similar activity patterns under a specific subset of the experimental conditions, each gene and condition in a bicluster are only a subset of the gene and condition. In biclustering, if some points are similar in several dimensions they will be clustered together in that subspace proved of great value of finding the interesting patterns in the microarray expression data.

C.2. Triclustering

Triclustering [14] is mining coherent clusters in three-dimensional 3D gene expression datasets. It mines arbitrary positioned and overlapping clusters and depending on different parameter values which mines diverse variety of clusters, together with those with constant or similar values along each dimension as well as scaling and shifting expression patterns. Tricluster relies on graph-based approach to mine all valid clusters and merge/delete some clusters having large overlaps. Tricluster can find significant triclusters in the real microarray datasets.

Once the data is obtained using the different methods usually it is transported to an Excel file where analysis takes place which embraces Gene Ontology, classification of genes into diverse functional groups and inputting the genes into pathways from diverse databases.

V. CONCLUSIONS

Clustering methods are rather effortless to implement and have a reasonable computational complexity yet fail to represent the genuine clustering of data.

The performance of every clustering algorithm may vary significantly with diverse data sets, and there is no absolute finest algorithm among the clustering algorithms.

The significant disadvantage of clustering algorithms are the fact that time variation is not considered in its calculations, variations of densities in the data space resulting in overlapping clusters, cluster validation, presence of irrelevant attributes, high level of background noise, no prior knowledge and the dimensionality curse.

To overcome the problems with cluster analysis, in addition to these long-established traditional algorithms new developed approaches can be used to depict the underlying structure of the genetic network.

So far Biclustering and Triclustering algorithms have proven significant improvement to the weaknesses and inadequacy of clustering algorithms, but there’s always room for improvement.

References

[1] M.B. Eisen and P.O. Brown, “DNA arrays for analysis of gene expression”, Methods Enzymol, vol. 303, pp. 170-205, P.O. 1999

[2] R. Xu and D. Wunsch, “Survey of clustering Algorithms”, IEEE Trans on Neural Networks. Vol. 16, no. 3, pp.645-678, 2005.

[3] B. Everitt, “Cluster analysis” 1st ed. Heinemann, London, 1980

[4] N. Bolshakova, F. Azuaje, “Cluster validation techniques for genome expression data”. Signal processing, 83, pp.825-833, 2003

[5] A.K. Jane and R.C. Dubes, “Algorithm of clustering data”, Prentice Hall, Englewood Cliff, NJ, 1998

[6] E. Shay, “Microarray cluster analysis and applications”, Available at: http://www.science.co.il/enuka/Essays/Microarray-Review.pdf, Jan, 2003.

[7] M.B. Eisen, T.P. Spellman, P.O. Brown, and D. Botstein, “Cluster analysis and display of genome-wide expression patterns”, Proc. Natl. Acad. Sci. USA, 95(25): 14863-14868, December 1998.

[8] S. Tavazoie, D. Hughes, M.J. Campbell, R.J. Cho, G.M. Church, “Systematic determination of genetic network architecture”, Nature Genet, pp. 281-285, 1999.

[9] T. Kohonen, “Self-organising maps.” Springer, Berlin, 1995.

[10] D. Bowtell, J. Sambrook, A DNA Cloning Manual: DNA Microarray. Cold Spring Harbor Laboratory Press, 2003.

[11] D. Jiang, C. Tang, A. Zhang, “Cluster Analysis for Gene Expression Data: A Survey”, IEEE, vol. 16, no. 11, Nov. 2004.

[12] R. Agrawal, J. Gehrke, D. Gunopulos, P. Raghavan, “Automatic subspace clustering of high dimensional data for data mining applications”, Proc. ACM SIGMOD international conference on Management of Data, pp. 94-105, 1998.

[13] S.C. Madeira and A.L. Oliveria, “Biclustering Algorithms for Biological Data Analysis: A survey”, IEEE, vol. 1, no. 1, Jan-March 2004.

[14] L. Zhao, M.J. Zaki, “Tricluster: An Effective algorithm for Mining coherent clusters in 3D Microarray Data”, SIGMOD, ACM press, USA, June 2005.

[15] P.A. Ralf-Herwig, C. Muller, C. Bull, H. Lehrach, and J. O’Brien. “Large-scale Clustering of cDNA-Fingerprinting Data,” Genome Research, vol. 9, pp. 1093-1105, 1999.

[16] P. Toronen, “Analysis of gene expression data using clustering analysis and functional classification”, Kuopio University Publications G. A.I. Virtanen Institute for Molecular Sciences, pp. 22-65 2004.

[17] C. Tang, A. Zhang, and J. Pei, “Mining Phenotypes and informative genes from gene expression data,” Proc. Ninth ACM AIGKDD Int’l Conf. Knowledge Discovery and Data mining (SIGKDD ’03), 2003.

[18] P. Baldi, G. W. Hatfield, “DNA Microarray and Gene Expression: From Experiments to Data Analysis and Modelling.” Cambridge press, 2002.

[29] S. Knudsen, Guide to Analysis of DNA Microarray Data. 2nd ed., Wiley-Liss, 2004.

[20] F.D. Smet, J. Mathys, K. Marchal, G. Thijs, M. Moor, D. Bart, and Y. Moreau. “Adaptive quality based clustering of gene expression profiles,” Bioinformatics, col. 18, pp. 735-746, 2002.

[21] K.Y. Yeung, D.R. Haynor and W.L. Ruzzo, “Validating clustering for gene expression data”, Oxford press, vol. 17, no. 4, pp. 309-318, 2001.

[22] H. Wang, W. Wang, J. Yang, P. Yu, “Clustering by pattern similarity in large data sets”, Proc. ACM press, NY

[23] D. Stekel, Microarray Bioinformatics. Cambridge press, 2003

[24] E. P. Xing and R.M. Karp, “Cliff: Clustering of high-dimensional microarray data via irerative feature filtering using normalisation cuts,” Bioinformatics, vol. 17, no. 1, pp. 306-315, 2001.

[25] L.J. Heyer, S. Kruglyak, and S. Yooseph, “Exploring expression data: Identification and analysis of coexpressed genes,” Genome Research, vol. 9, no.11, pp. 1106-1115, 1999.

[26] P. Tamayo, D. Solni, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S Lander, and T.R. Golub, “ Interpreting patterns of gene expression with Self-Organizing Maps: Methods and Application to Hamatopoietic Differentiation,” Proc. Nat’l Academy of science, vol. 96, no. 6, pp. 2907-2912, Mar. 1999.

[27] J. Herrero, A. Valencia and J. Dopazo, “A Hierarchical Unsupervised Growing Neural Network for clustering gene expression patterns,” Bioinformatics, vol. 17, pp. 126-136, 2001.