VOL.2 NO.2September 2013 - ISSN 2233 – 1859

Operations Research Society in Bosnia and Herzegovina

Southeast Europe Journal of Soft Computing

Available online:www.scjournal.com.ba

Gene Expression Data Clustering

Betul Akcesme

International University of Sarajevo, Faculty of Engineering and Natural Sciences, HrasnickaCesta 15, Ilidža71210 Sarajevo, Bosnia and Herzegovina

Article Info

Article history:
Received 17 Sep.2013
Received in revised form 17 Oct 2013

Keywords:
Gene expression, Clustering, Hierarchical Clustering, Genetic Algorithms, K-Means Clustering, microarrays

Abstract

Gene expression analysis is becoming very important in order to understand complex living organisms. Rather than analyzing genes individually, there is more powerful approach, microarray technology to analyze the genes expression in high throughput. This new approach brings new analyses problems that make the interpretation difficult. To understand the correlated gene expression analysis easier some clustering methods are applied to the gene expression analysis. In this paper, different approach is represented to start to cluster with using some computational strategies.

1. Introduction

A. Introduction to microarray technology

Gene expression, the flow of genetic information, is the process by which information from a gene stored in DNA is used to produce a functional gene product including, functional RNA (ribosomal RNA, transfer RNA) and protein. Since the genotype of an organism gives rise to the phenotype, gene expression is the fundamental link between genotype and phenotype in a species (http://www.plexpress.com) Therefore, gene expression analysis is becoming very important in order to understand complex living organisms. Gene expression analysis measures the amount of mRNA in order to estimate protein level. It assumes that the amount of mRNA is correlated with the amount of its protein produced by the cell. It should be underlined that a number of processes affect the production of proteins in the cell including, transcription, splicing, translation, post-translational modifications. However, this correlation between amount of mRNA and protein is still significant and intermediate steps can be slightly ignored (Jones and Pevzner, 2004).
Traditional analyses of gene expression were only able to manipulate a single gene or few genes and then monitor impact of these genes in different conditions. These methods reveal very little information about other genes or how single genes may interact with the collective gene product in a living organism (Jiang, 2004). With advances in technology in recent years, there are new approaches to design an experiment that is not only dealing with few genes but wide range of genes at the same time. The omics technology gives opportunity to scientist to do high throughput analysis. For instance, the microarray is a powerful genomic tool, which allows researchers to examine expression levels of thousands of genes simultaneously under many time points and conditions such as a particular point in the cell cycle, after an interval response to some environmental change, RNA isolated from a tissue exhibiting certain phenotypic characteristics and to reveal which genes are switched on and switched off in the cell. Researchers can estimate the relationships between genes and gain a better understanding of how genes interact within an organism by interpreting microarray data.

1. General steps of Microarray

There are two main types of microarrays. Glass complementary DNA (cDNA) microarray was the first type of DNA microarray technology developed. It was pioneered by Patrick Brown and his colleagues at Stanford University and is produced by using a robotic device, which deposits (spots) a nanoliter of DNA (50-150 µm in diameter) onto a coated microscope glass slide surface in serial order with a distance of approximately 200-250 µm

from each other, one spot-one gene. These moderate sized glass complementary DNA microarrays also bear about 10,000 spots or more on an area of 3.6 cm^2.(Schena, 1995). Second type of microarray, in situ (on chip) oligonucleotide array format is a sophisticated platform of microarray technology which is manufactured by using the technology of in situ chemical synthesis that was first developed by Stephen Fodor et al. (1991). However, the industry leader in the field of in situ oligonucleotide microarrays (Affymetrix) has further pioneered this type of technology to manufacture so-called GeneChips which refers to its high density oligonucleotide based DNA arrays (Lockhart et.al, 1996). Both types of micro array contain common procedure which is listed below (Tefferi et.al, 2002).

a. **Production of Chip:** A microarray is a small chip which is made of chemically coated glass, nylon membrane, or silicon. Thousands of DNA molecules called probes are attached in fixed cells on chip. Each cell is related to a specific DNA sequence.
b. **How to prepare, label and hybridize target**: Two mRNA samples which are test sample and control sample are reverse transcribed into cDNA. They are labeled using either fluorescent dyes or radioactive isotopics, and then hybridized with the probes on the surface of the chip.
c. **The scanning process:** Chips are scanned to read the signal intensity that is emitted from the labeled and hybridized targets.

2. Applications and importance of Clustering Gene Expression Data in biological Sciences

As we mentioned above, gene expression data sets, microarray contain large number of genes which need to be analysed at the same time. It is one of the difficulties of high throughput analysis and can be overcome by clustering techniques. There are clustering algorithms which group genes with similar expression patterns into clusters. It is thought that these clusters correspond to groups of functionally related genes or are involved in the same biological process.

Clustering techniques are revealed as helpful tools to understand gene function, gene regulation, and cellular processes. This approach may further understanding of the functions of many genes for which information has not been previously available (Tavazoie, 1999). Also, coexpressed genes in the same cluster are likely to be involved in the same cellular processes (Brazma, 2000). Clustering of gene expression data also help us to determine and understand the mechanisms of the transcriptional regulatory network. (Alizadeh et al, 2000).

3. Step Before Processing of Gene Expression Data

Various number of DNA sequences such as genes, cDNA clones, or expressed sequence tags [ESTs] which are under multiple conditions can be processed in a microarray experiment. These conditions may be a time series during a biological process or a collection of different tissue. DNA sequences will be considered as genes and all kinds of experimental conditions will be defined as samples in order to prevent any confusion. A gene expression data set from a microarray experiment can be represented by expression matrix. The rows represent the expression patterns of genes and the columns represent expression profiles of samples, and each cell represents the measured expression level of gene i in sample j.

After a scanning process, the original gene expression matrix consists of noise, missing values, and variations due to the experimental procedure. Before performing cluster analysis, preprocessing of data has to be done. Some problems of data preprocessing have themselves become interesting research topics. In the following discussion of clustering algorithms, it will not mentioned the details of preprocessing procedures and assume that the input data set has already been properly preprocessed(Jones and Pevzner, 2004; Jiang,2004)

B. Introduction to Clustering Techniques

Clustering, a common problem in biology is the process of dividing experimental data into clusters (groups) in which they display high similarity while experimental data in different clusters have high difference based on their certain characteristics (Jiang, 2004). Various algorithms have been developed in order to solve clustering problem which usually contains large amount of data set. However, there is no one particular approach (algorithm) over another for a particular clustering problem.

1. Types of Gene Expression Data Clustering

There are two straightforward ways how gene expression matrix can be studied:

1. Comparing expression profiles of genes by comparing rows in the expression matrix; in such gene-based clustering, the genes are treated as the objects, while the samples are the features (Einsen, 1998).

2. Comparing expression profiles of samples by comparing columns in the matrix. Such sample-based clustering regards the samples as the objects and the genes as the features

Additionally both methods can be combined. When comparing rows or columns, it canbe looked either for similarities or for differences. If two rows are similar, it is can hypothesized that the respective genes are co-regulated and possibly functionally related. By comparing samples, it can end which genes are differentially expressed and, for instance, study effects of various compounds (Brazma, 2000; Golub, 1999). In this article I will only focus on gene based clustering.

2. Proximity Measurement for Gene Expression Data

Many methods of cluster analysis depend on some measure of similarity (or distance) between the vectors to

be clustered. Although Euclidean distance is a popular distance measure for spatial data, the correlation coefficient is widely believed to be more suitable for pattern-discovery approaches because it measures the strength of the linear relationship between two vectors. This measure has the advantage of calculating similarity on the basis only of the pattern and not the absolute magnitude of the spatial vector (Tang, 2001).The formula of the correlation coefficient between two vectors:

$$\text{Pearson}(O_i, O_j) = \frac{\sum_{d=1}^{n}(O_{id} - \mu_{oi})(O_{jd} - \mu_{oj})}{\sqrt{\sum_{d=1}^{n}(O_{id} - \mu_{oi})^2 (O_{jd} - \mu_{oj})^2}}$$

The distance between objects Oi andOj in n dimensional space by Euclidian Distance is defined as:

$$\text{Eucledian}(O_i, O_j) = \sqrt{\sum_{d=1}^{n}(O_{id} - O_{jd})^2}$$

For gene expression data, the overall shapes of gene expression patterns (or profiles) are of greater interest than the individual magnitudes of each feature. Euclidean distance does not score well for shifting or scaled patterns (Wang, 2002)

2. Clustering Algorithms

A. *K means*

K means algorithms divide a data set into k, number of subsets, by assigning random centers to each subset. It is one of the partition-based clustering methods. Firstly, it randomly identifies k points as cluster centers. Each member of data set is assigned to one of these predetermined centers based on minimization the sum of the distance between all points and their centers (Jones, 2004). After this distribution of each point in data set to a particular center, sum of the distance of each member of subset to the rest of the members of subset is calculated. And the smallest distance is assigned as new center of the subset. Algorithm iterates to find improved positions for the cluster centers. The K-means algorithm is simple and fast. However, it has some drawbacks. One of the main problems with this method is that the number of clusters, k, must be specified before running the algorithm. The number of gene clusters in a gene expression data set is usually unknown in advance. The algorithm should be run with different number of k and compare the clustering results in order to find the optimal and the most efficient number of clusters. Second drawback is that K means algorithms force each gene into a particular cluster even though there is huge amount of noise due to large date set. Therefore, algorithm will be sensitive to noise. Various research groups have proposed different algorithms to solve these problems. Ralf-Herwig et al. have described new approach to K means algorithm. They presented a sequential k-means approach that has been introduced by MacQueen and further described by Mirkin that finds the number of different clusters from data itself and is independent of a prespecified number of centers. The sequential structure of their algorithm allows the analysis of even larger data sets in a reasonable amount of time by splitting the data in smaller data sets, clustering these sets in parallel using multiple processors, and then reclustering the calculated centers. This procedure reduces computation time to a high degree because the time-consuming step of finding the centers in the data can be computed in parallel.

B. *Hierarchical clustering*

Unlike K-means clustering, which directly partitions the data set into particular separated clusters with assigned centers, hierarchical clustering generates a hierarchical series of nested clusters from each data point. Hierarchically clustered data sets are also represented by a graph (tree) called dendogram. Leaves of the tree represent the genes. Edges of the trees are assigned lengths and distances between leaves correlate with entries in the distance matrix. The branches of a dendrogram not only record the formation of the clusters but also indicate the similarity between the clusters (Jones, 2004) A specified number of clusters can be obtained by cutting the dendrogram at some level. The data set can be arranged with similar genes placed together by reordering the genes such that the branches of the corresponding dendrogram do not cross. There are two different ways of constructing hierarchical clustering dendogram: agglomerative approaches and divisive approaches. Agglomerative algorithms (bottom-up approach) initially regard each data object as an individual cluster, and at each step, merge the closest pair of clusters until all the groups are merged into one cluster. For agglomerative approaches, different measures of cluster proximity, such as single link, complete link, and minimum-variance derive various merge strategies.

Eisen et al. proposed an agglomerative algorithm called UPGMA (Unweighted Pair Group Method with Arithmetic Mean) and adopted a method to graphically represent the clustered data set. In this method, each cell of the gene expression matrix is colored on the basis of the measured fluorescence ratio and the rows of the matrix are reordered based on the hierarchical dendrogram structure and a consistent node-ordering rule. After clustering, the original gene expression matrix is represented by a colored table where large contiguous patches of color represent groups of genes that share similar expression patterns over multiple conditions.

Divisive algorithms (top-down approach) starts with one cluster containing all the data objects and, at each step split, only singleton clusters of individual objects remain. For divisive approaches, the essential problem is to decide how to split clusters at each step. Some are based on heuristic methods such as the deterministic annealing algorithm [19], while many others are based on the graph theoretical methods which we will discuss later. The

history of the data splitting is used to construct a binary tree. Alon et al.also include a node-switching algorithm to order the branches in a somewhat optimal manner that is similar, in concept, to the algorithm implemented in for agglomerative clustering.

Hierarchical clustering not only groups together genes with similar expression pattern but also provides a natural way to graphically represent the data set. The graphic representation allows users a thorough inspection of the whole data set and obtains an initial impression of the distribution of data. Eisen's method is much favored by many biologists and has become the most widely used tool in gene expression data analysis. However, hierarchical clustering has a number of shortcomings for the study of gene expression. Hierarchical clustering has been noted by statisticians to suffer from lack of robustness, nonuniqueness, and inversion problems that complicate interpretation of the hierarchy. Another drawback of the hierarchical approach is its high-computational complexity. Furthermore, for both agglomerative and divisive approaches, the "greedy" nature of hierarchical clustering prevents the refinement of the previous clustering. If a "bad" decision is made in the initial steps, it can never be corrected in the following steps (Tamayo, 1999)

C. *Graph-Theoretical Approaches*

Graph-theoretical clustering techniques are explicitly presented in terms of a graph, thus converting the problem of clustering a data set into such graph theoretical problems as finding minimum cut or maximal cliques in the proximity graph G. Proximity matrix, P can be obtained from a given data set X. A weighted graph G(V,E) called proximity graph is driven form P[i,j]=proximity(Oi,Oj). Where genes are represented by vertex and distance between them is edge. (Jiang, 2004)

1. CLICK

Sharan et al developed a clustering algorithm which is called CLICK (CLuster Identification via Connectivity Kernels) (Shamir, Sharan, 2000). Algorithm does not make any assumption about the number of clusters at the beginning. It looks for the identification of highly connected components in the proximity graph as clusters. The clustering process of CLICK iteratively finds the minimum cut in the proximity graph and recursively splits the data set into a set of connected components from the minimum cut. The authors compared the clustering results of CLICK on two public gene expression data sets with those of GENECLUSTER (Tamayo, 1999) (a SOM approach) and Eisen's hierarchical approach (Einsen,1998) respectively. In both cases, clusters obtained by CLICK demonstrated better quality in terms of homogeneity and separation. However, CLICK has little guarantee of not going astray and generating highly unbalanced partitions, e.g., a partition that only separates a few outliers from the remaining data objects. Furthermore, in gene expression data, two clusters of coexpressed genes, C1 and C2, may be highly intersected with each other. In such situations, C1 and C2 are not likely to be split by CLICK, but would be reported as one highly connected component.

2. CAST

The idea of corrupted clique graph data model was introduced (Ben-Dor et al., 1999). The input data set is assumed to come from the underlying cluster structure by contamination with random errors caused by the complex process of gene expression measurement. Specifically, it is assumed that the true clusters of the data points can be represented by a clique graph H, which is a disjoint union of complete subgraphs with each clique corresponding to a cluster. The similarity graph G is derived from H by flipping each edge/nonedge with probability. Therefore, clustering a data set is equivalent to identifying the original clique graph H from the corrupted version G with as few flips (errors) as possible. Ben-Dor et al. presented both a theoretical algorithm and a practical heuristic called CAST (Cluster Affinity Search Technique).

D. A Density-Based Hierarchical Approach: DHC

Ben Dorproposed a new clustering algorithm, DHC (a density-based, hierarchical clustering method), to identify the coexpressed gene groups from gene expression data(Jiang, and Pei, 2003). DHC is developed based on density and attraction of data objects. The basic idea is to consider a cluster as a high-dimensional dense area, where data objects are attractedwith each other. At the core part of the dense area, objects are crowded closely with each other and, thus, have high density. Once the density and attraction of data objects are defined, DHC organizes the cluster structure of the data set in two-level hierarchical structures.

E. *Difficulties of Gene Clustering*

There are several essential challenges that need to be pointed out in cluster analysis.Firstly, a good clustering algorithm should not depend on prior knowledge, which is usually not available before cluster analysis. For instance, a clustering algorithm which can find accurate number of clusters are considered better than an algorithm requiring predetermined cluster number.

Secondly, gene expression data often contains a huge amount of noise. Therefore, clustering algorithms for gene expression data should be capable of extracting useful information from a high level of background noise.

Thirdly, gene expression data are often connected (Jiang, 2003) and clusters may be intersected with each other (Kaufman, 1990). Therefore, algorithms for gene-based clustering should be able to effectively solve this situation.

Finally, researchers who use microarray might be interested in the relationship between the clusters and the

relationship between the genes within the same cluster. A clustering algorithm provides some graphical representation of the cluster structure would be more favoured by the biologists.

3. Material and Methods

First of all 50 x 50 matrixes were created which lay down to the 2D array. This matrix is filled by 0 and 1 digits randomly. The rows of this matrix were summed up to other one dimensional array which is 50 x 1. This is applied also for the column. These rows and columns are summed up to other 50 x 1 array. The highest 3 numbers of the last array are taken as a clustering centers. The number of the centers can be determined by the user. The coordinates of these 3 highest points are determined via 3 matrixes projections. The other points (genes) are clustered according to these center points. K means clustering algorithm was applied.

4. Results and Discussions

In this section, I have summarized a series of approaches to gene clustering. The purpose of clustering genes is to identify groups of highly coexpressed genes from noisy gene expression data. Clusters of coexpressed genes provide a useful basis for further investigation of gene function and gene regulation. Some conventional clustering algorithms, such as K-means, hierarchical approaches (UPGMA), were applied in the early stage and have proven to be useful. However, those algorithms were designed for the general purpose of clustering, and may not be effective to address the particular challenges for gene- based clustering. There are several clustering algorithms, such as CLICK, CAST, and DHC have been proposed specifically to aim at gene expression data. The experimental study (Shamir, and Sharan, 2000; Jiang, and Pei, 2003) has shown that these new clustering algorithms may provide better performance than the conventional ones on some gene expression data. However, different clustering algorithms are based on different clustering criteria and/or different assumptions regarding data distribution. The performance of each clustering algorithm may be greatly with different data sets and there is no absolute “best” among the clustering algorithms.

Description	Methods
6220 ORFs in S. cerevisiae with 15 time points	Kmeans,(Tavazoie,1999) CLICK (Shamir, Sharan, 2000) DHC (Jiang, Pei, 2003)
9800 cDNAs with 12 time points	Agglomerative hierarchical (Eisen,1998), CLICK (Shamir, Sharan, 2000), DHC (Jiang, Pei, 2003)
112 rat genes during 9 time points	CAST .(Ben-Dor, 1999)
6178 ORFs S. cerevisiae during 4 time courses	Agglomerative hierarchical (Eisen,1998)
6500 human genes in 40 tumor and 22 normal colon tissue sample	Divisive hierarchical (Alon et al, 1999)
1246 genes in 146 experiments	CAST (Ben-Dor, 1999)

REFERENCES

Gene expression analysis: A review, http://www.plexpress.com/

N. C. Jones, P. A. Pevzner, (2004) An Introduction to Bioinformatics Algorithms, The MIT Press Cambridge, Massachusetts London, England,

D. Jiang, C. Tang, A. Zhang, (2004) Cluster Analysis for Gene Expression Data: A Survey, IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 11.

M.D. Schena, R. Shalon, R. Davis, and P. Brown, (1995) “Quantitative Monitoring of Gene Expression patterns with a Compolementatry DNA Microarray,” Science, vol. 270, pp. 467-470,

D. Lockhart et al., (1996) Expression Monitoring by Hybridization to High-Density Oligonucleotide Arrays, Nature Biotechnology, vol. 14, pp. 1675-1680.

A. Tefferi, E. Bolander, M. Ansell, D. Wieben, and C. Spelsberg, (2002) Primer on Medical Genomics Part III: Microarray Experiments and Data Analysis, Mayo Clinic Proc., vol. 77, pp. 927-940,

S. Tavazoie, D. Hughes, M.J. Campbell, R.J. Cho, and G.M. Church, (1999) “Systematic determination of Genetic Network Architecture,” Nature Genetics, pp. 281-285,

A. Brazma and J. Vilo, (2000) Minireview: Gene Expression Data Analysis, Federation of European Biochemical Soc., vol. 480, pp. 17- 24. .

A.A. Alizadeh et al., (2000) Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling, Nature, vol. 403, pp. 503-511.

M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein.(1998) Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America, 95:14863–14868,

T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gassenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, D.D. Bloomfield, and E.S. Lander, (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring,” Science, vol. 286, no. 15, pp. 531-537.

C. Tang, L. Zhang, A. Zhang, and M. Ramanathan, ,(2001) Interrelated Two-Way Clustering: An Unsupervised Approach for Gene Expression Data Analysis, Proc.

BIBE2001: Second IEEE Int'l Symp. Bioinformatics and Bioeng., pp. 41-48.

H. Wang, W. Wang, Y. Wei, J. Yang, and P.S. Yu, (2002) Clustering by Pattern Similarity in Large Data Sets, SIGMOD, Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 394-405.

D. Jiang, J. Pei, and A. Zhang, (2003) Interactive Exploration of Coherent Patterns in Time-Series Gene Expression Data, Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '03).

L. Kaufman and P.J. Rousseeuw, (1990) Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons,

P.A. Ralf-Herwig, C. Muller, C. Bull, H. Lehrach, and J. O'Brien, (1999) Large-Scale Clustering of cDNA-Fingerprinting Data," Genome Research, vol. 9, pp. 1093-1105.

MacQueen, J.B.(1967) Some methods for classification and analysis of multivariate observations. In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability (ed. L.M. LeCam and J. Neyman), vol. 1, pp. 281–297. University of California Press, Los Angeles, CA., 1967.

Mirkin, (1996) B. Mathematical classification and clustering. Kluwer Academic Publishing, Dordrecht, The Netherlands.

U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine, (1999) Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Array," Proc. Nat'l Academy of Science, vol. 96, no. 12, pp. 6745-6750.

P. Tamayo, D. Solni, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, and T.R. Golub, (1999) Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods andApplication to Hematopoietic Differentiation," Proc. Nat'l Academy of Science, vol. 96, no. 6, pp. 2907-2912.

R. Shamir and R. Sharan, (2000) Click: A Clustering Algorithm for Gene Expression Analysis," Proc. Eighth Int'l Conf. Intelligent Systems for Molecular Biology (ISMB '00),

A. Ben-Dor, R. Shamir, and Z. Yakhini, (1999) Clustering GeneExpression Patterns," J. Computational Biology, vol. 6, nos. 3/4, pp. 281-297.

D. Jiang, J. Pei, and A. Zhang, (2003) DHC: A Density-Based Hierarchical Clustering Method for Time-Series Gene Expression Data," Proc. BIBE2003: Third IEEE Int'l Symp. Bioinformatics andBioeng.