Microarray Data Analysis Methods Comparison : A Review

Biochemistry 218 Project

Hwangmin Ki

hmkey@stanford.edu

I. Introduction

Genome sequencing projects have provided us with static pictures of the genomes of many organisms. While it is possible to identify most of the genes in these genomes by analyzing the sequences, sequence analysis alone does not tell us what the genes do or how they are used. To observe a living genome in action, and to see the specific set of instructions it sends in a specific cell, under a specific set of conditions, we can use DNA microarrays to capture an image of the genome's instructions, by detecting and quantitating the transcripts from each gene.

Huge amounts of data are being generated using DNA microarrays and one of the

immediate challenges is to develop tools and interpretive framework that will enable us to make sense of this information. Because the measurements of gene expression obtained with DNA microarrays are systematic and quantitative, powerful mathematical and statistical methods can be applied to search for orderly features and logical relationships in genomic expression patterns.

A wide range of different methods have been proposed for the analysis of gene expression data including hierarchical clustering, self-organizing maps, and k-means approaches. Many of the proposed algorithms have been reported to be successful but no single algorithm has emerged as a method of choice. Most of the algorithms are based on heuristic methods, and the issues of determining the "correct" number of clusters and the choice of "best" algorithm has yet to be solved.[1] In this critical review, I will look at the three different clustering methods of DNA microarray data analysis, evaluate their strong points and weak points respectively, and suggest ways of modifying the methods for improvement.

II. Clustering Methods

A. Hierarchical clustering

Hierarchical clustering is a method familiar to most biologists through its application in generating phylogenetic trees. Relationship among genes are represented by a tree whose branch lengths reflect the degree of similarity between the genes. Such relationships are useful in their ability to represent varying degrees of similarity and more distant relationships among groups of closely related genes, as well as requiring few assumptions about the nature of the data.[2] The computed trees can be used to order genes in the original data table, so that genes or groups of genes with similar expression patterns are clustered as shown below.

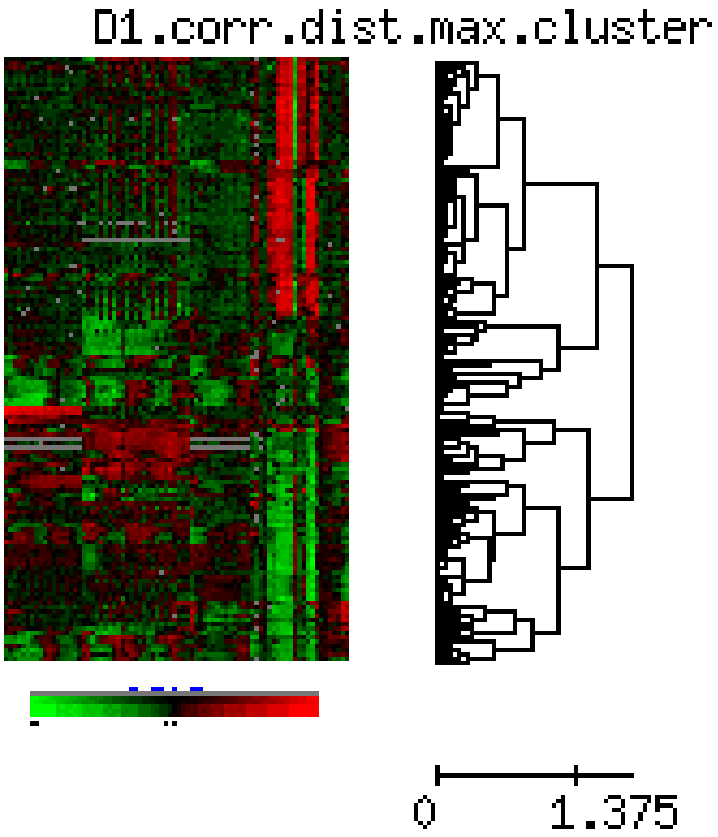

Figure1. Example of hierarchical clustering of DNA microarray data[3]

The advantage of the hierarchical method, as noted in the paper by Eisen et al[ii], is that a natural way of looking at complex data sets is first to scan the large-scale features and then to focus on the more interesting details, and this method enables you to take the same kind of intuitive approach in analyzing the genomic data. Although this approach is a general one that has no inherent specificity to the particular method used to acquire the genomic data, it has been proven to be very useful.

However hierarchical clustering has a number of shortcomings for studying gene expression. Hierarchical trees are not designed to reflect the multiple distinct ways in which expression patterns of genes can be similar, and as the size of the data becomes larger this problem is exacerbated.[4] Statisticians have noted that hierarchical clustering suffers from lack of robustness, nonuniqueness and inversion problems that complicate interpretation of the hierarchy[5]. Also the deterministic nature of hierarchical clustering can result in local grouping of the results with no opportunity to evaluate the clustering.

B. Self-Organizing Maps

The self-organizing map is a method for producing ordered low-dimensional representations of an input data space. Typically such input data is complex and high-dimensional with data elements being related to each other in a nonlinear fashion. These maps can successfully approximate high-dimensional input space by extracting invariant features of the input signals and maintaining topological relationships between

them in lower dimensions. A multiple dimension DNA microarray data matrix is constructed, each node representing a point on the DNA microarray. Then a random node is selected and iteratively adjusted in the n-dimensional space according to the pattern of expression. So, self-organizing maps impose structure on the data with neighboring nodes tending to define related clusters. These clusters become nodes for the lower-dimension matrix.

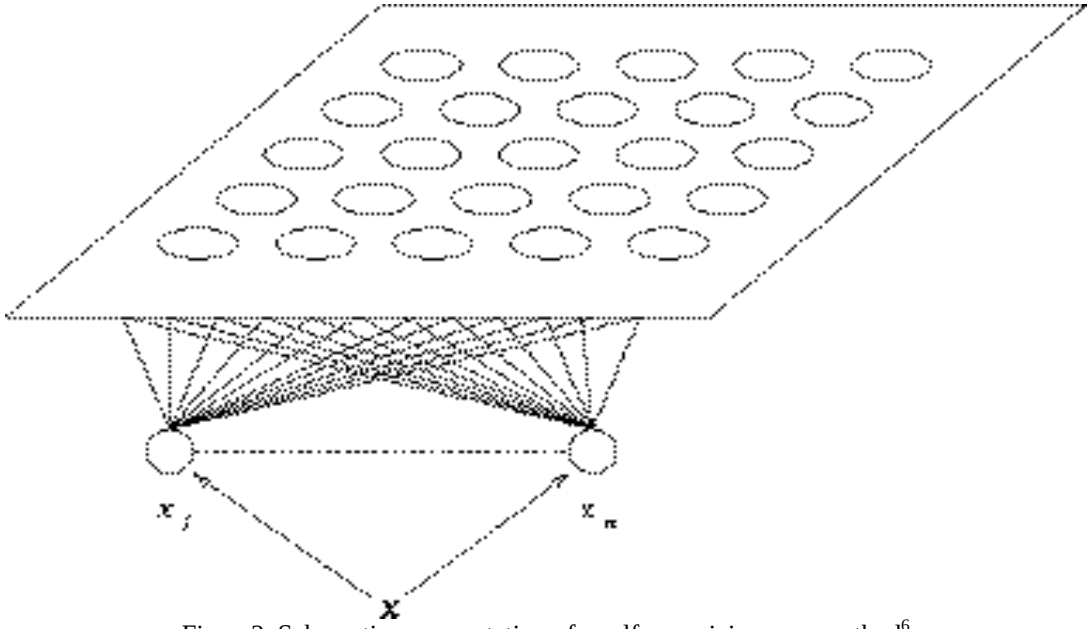

Figure2. Schematic representation of a self-organizing map method[6]

The self-organizing map method is ideally suited for explorative data analysis where prior information about the distribution of the data is not available. Also the computational algorithms are relatively easy to implement, fast, and scalable to large data sets. The results are easy to visualize and interpret. And the paper by Mangiameli et al[7] applied self-organizing maps and hierarchical methods to 252 "messy" data sets with real-world data imperfections such as dispersion, irrelevant variables, outliers, and non-uniform densities and found self-organizing maps to be significantly superior in both robustness and accuracy.

C. K-means

This nonhierarchical method initially takes the number of data points on the microarray equal to the final required number of clusters. In this step itself the final required number of clusters is chosen such that the points are mutually farthest apart. Next, it examines each point in the microarray and assigns it to one of the clusters depending on the minimum expression distance. The centroid's position is recalculated every time a data point is added to the cluster and this continues until all the data points are grouped into the final required number of clusters. This method thus requires prior knowledge of the genes that are analyzed. Tavazoie et al[8] showed that this method is effective in analyzing data in *Saccharomyces cerevisiae* gene expression where 30 clusters were chosen according to their biological expression diversity. But they also admit that they

erred on the side of over-classification to avoid missing significant expression classes. One of the most significant downsides of this method is that you may get different results from the same data if the starting conditions are varied as illustrated below.

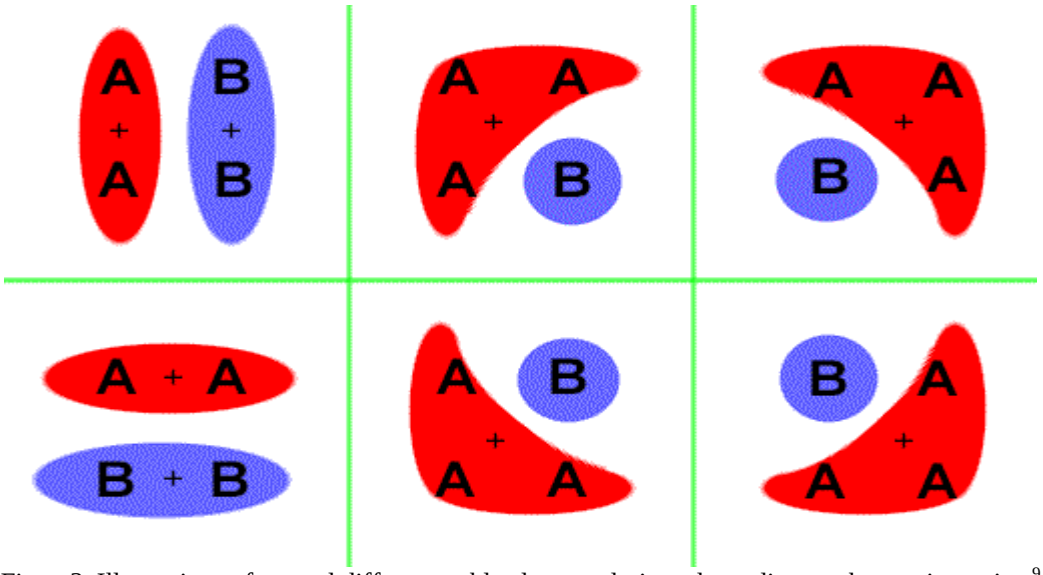

Figure3. Illustrations of several different stable cluster solutions depending on the starting points[9]

Some other aspect of this method to note is that depending on the "distance" used for clustering, some data points may not converge and oscillate indefinitely, and the cluster solution can be influenced by the order of the cluster cases.

III. Suggestions

Clustering algorithms based on probability models offer a principled alternative to heuristic-based algorithms as discussed above. The model-based approach assumes that the data is generated by a finite mixture of underlying probability distributions. With this approach the problems of determining the number of clusters and of choosing the appropriate clustering algorithm becomes a statistical model choice problem.[i] This is a great advantage over heuristic methods since heuristic methods do not have an established method of determining the number of clusters and the clustering algorithm to use. One of the most common probability model is the Gaussian mixture model, and it has been shown to be effective and powerful in applications. The same model can be applied for DNA microarray data clustering. In this model each data element is assumed to be generated by a mixture of underlying probability distribution of its components. And the methods can be further differentiated by the constraints on the components. (See K.Y. Yeung et al[i] for five such models) In each case, each component is modeled by multivariate normal distribution with mean vector and covariance matrix. The covariance matrix determines the shape, volume, orientation of each element and a combination of the covariance matrices and a different number of clusters is used to determine the corresponding probabilistic model. Hence, the probabilistic approach enables us to choose the optimal clustering algorithm and the optimal number of clusters to use. This is important because there is a tradeoff the corresponding probability model and number of clusters. For instance if a complex

probability model is used a small number of clusters may suffice, while if a simple model is used, a larger number of clusters may be needed to fit all the data appropriately. The model-based approach has been shown to be successful (See K.Y. Yeung et al[i] , Y. Barash & N. Friedman[10], I. Holmes and W.J. Bruno[11]) showing the possibility of incorporating additional knowledge to the analysis and achieving comparable to better results when analyzing DNA microarray data itself. Another probabilistic method is the graph-theory method as shown in A. Ben-Dor et al.[12] Probabilistic model is used to represent each gene and the genes become "nodes", and "edges" connect the genes that have a similar expression pattern. A "clique graph" is produced from this result, a disjointed union of complete graphs. A single clique represents a single cluster of corresponding genes. This algorithm is relatively easy implement and is shown to be powerful in Ben-Dor et al's paper.

Apart from the powerful clustering abilities of the modeling methods there are also other benefits from using a model-based approach. New insights about the genes and across different organisms, across different times can be obtained. And also standardization of the DNA microarray results is possible. This is a very important point especially when so much data without compatibility is being generated at a phenomenal speed.

IV. Conclusion

Genome sequencing projects and DNA microarray data has provided us with better insights to biology and respective organisms. But there is still a lot more interpretations of the data to conduct. And being able to construct a model of an organism or its gene expression patterns will surely enhance our understanding of the organism and the underlying biology. Probability models of gene expression seem to be a good starting point for such a model. Better models in terms of robustness, accuracy, efficiency need to be developed, and as computing power increases such methods as neural networks could be implemented to achieve this goal. Also a combination of heuristic methods and probabilistic methods will be able to enhance the results in the future.

References

[1] K.Y. Yeung, C. Fraley, A.Murua, A.E. Raftery, and W.L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17:977~987, 2001.

[2] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Nat'l Acad Sci USA*, 95: 14863~14868, 1988.

[3] http://ep.ebi.ac.uk/EP/ , May, 2002

[4] P. Tamayo, D. Slonim, 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 hematopoietic differentiation. *Proc. Nat'l Acad Sci USA*, 96: 2907~2912, 1999.

[5] B.J.T. Morgan and A.P.G Ray. Non-uniqueness and inversions in cluster analysis. *Applied Statisics*, 44: 114~134.

[6] http://www.dcs.napier.ac.uk/hci/martin/msc/node17.html , May, 1996

[7] P. Mangiameli, S.K. Chen, & D. West. A comparison of SOM neural network and hierarchical clustering methods. *Eur. J. Oper. Res.* 93: 402~417, 1996.

[8] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho & G.M. Church. Systemaic determination of genetic network architecture. *Nature Genetics*, 22:281~285, 1999.

[9] http://www.clustan.com/k-means_critique.html , March, 2001

[10] Y. Barash & N. Friedman. Context-specific Bayesian clustering for gene expression data. *RECOMB 2001*, 12~21, 2001.

[11] I. Holmes and W.J. Bruno. Finding regulatory elements using joint likelihoods for sequence and expression profile data. *AAAI Press*, 202~210, 2000.

[12] A. Ben-Dor, R. Shamir, & z. Yakhini. Clustering gene expression patterns. *J. Comp. Bio.* 1999.