A Review of DNA Microarray Data Analysis

Background:

The mystery of life for a living organism resides in the function of thousands of genes and their products. The striking question of how to get the whole system of an organism in one picture, has been pondering for years. Traditional methods work on one gene at a time, which is time consuming and costly. The ingenuous idea of DNA microarrays created a solution to this problem. DNA microarrays are a new technology that allows the whole genome to be monitored on a single chip so that a better picture of the interactions among thousands of genes can be observed simultaneously (Brazam et al. 2000).

A DNA microarray is composed of pieces of DNA ranging from 20-5000 base pairs concentrated into specific areas on a solid support such as a glass chip (Schena 1999). Thousands of same oligonucleotides are attached in a specific location on the support and gene expression can be observed by counting the amount of oligonucleotides that are bound. Therefore, the array works as a capturing device for specific complementary gene products.

The experiment using DNA microarrays starts with the process of PCR, used to amplify genes of interest. The PCR products are purified and then placed on glass microscope slides using a robot. Reverse transcription is used to label both the test and reference total RNA using two different fluorescent dyes. The two fluorescent targets are added together and allowed to hybridize to the microarray. Laser exicitation of each hybridized target causes a specific emission with fingerprint-like spectra being produced.

The spectra is measured by scanning confocal laser microscope, which is then imported into software that merges the two images and gives them specific colors. The gene expression of each target is given a value, and these values are imputed into data sets (Brown et al. 1999).

Introduction:

The data from microarray experiments is usually in the form of large matrices of expression levels of genes (row) under different experimental conditions (columns) (Brown et al.1999). Clustering methods are used to arrange the genes in a natural order, where similar genes are placed close together. Hierarchical clustering and k means clustering, are two most frequently used methods. Hierarchical clustering takes a bottom-up approach, which starts with each gene in its own cluster. K means clustering takes a top-down approach, which starts with a specified number of clusters and initial positions for the cluster centers (Tibshirani et al.1999). The first part of this paper will compare these methods, exploring the disadvantages and advantages of each method.

One major disadvantage with these methods is that they require a full array of gene expression values and are not robust to missing values. Missing values occur due to insufficient resolution, image corruption, dust or scratches on slides, or even robotic methods can create missing values (Troyanskaya et.al 2001). The easiest solution to missing values is to re-do the experiment, but life is not this simple. This can be very expensive and un-realistic.

Methods have been created in order to deal with missing values in a more realistic approach. Most commonly used methods are replacing missing data with zeros or by an average of expression in a row. These methods do not take into consideration the

correlation structure of data, and is therefore not very precise. The second part of this paper will compare k-nearest neighbors (KNN) and singular value decomposition (SVD), which do take into consideration the correlation of data. These approaches have not been researched greatly in the field of DNA microarray analysis, but have been used in many other fields. The two methods are commonly used as statistical and mathematical methods for classification, such as Text Categorization and face recognition (Jiangsheng 2002).

Definition of Clustering Methods:

Hierarchical clustering is a familiar method used in sequence and phylogenetic analysis. As applied to DNA microarray analysis, a tree represents relationships amongst genes in which, branch lengths represent degrees of similarities. This method is useful in it's ability to represent varying degrees of similarity and distant relationships among groups of closely related genes (Eisen et al 1998). The computed tree (called a "dendogram") can then be used to organize genes in the original data table, so that genes with similar expression patterns are adjacent.

The general procedure for hierarchical clustering follows in two steps, 1. Find the closest points (clusters) and merge them, and 2. Proceed until you have a single cluster (all the points). There are two prerequisites for this procedure, 1. The distance measure between two points, and 2. The distance measure between clusters. There are various methods used to calculate these distances.

K-means clustering is a top down (non-hierarchical) approach, where it starts with a specified number of clusters and initial positions for the cluster centers. The procedure is represented as follows, 1. Pick k arbitrary centroids, 2. Assign each gene to its

"closest" centroid, 3. Adjust the centroids to be the means of the examples assigned to them, and 4. Repeat to step 2 until no change.

Comparison of Clustering Methods:

The main strength of hierarchical clustering is that it forms a hierarchy of clusters enabling small groups of co-expressed genes to be identified and it can distinguish between ball shape compact clusters, as well as long chain-like clusters (Razaz 2000). This method is preferable because it is conceptually simple and the theoretical properties of the method are very well understood, which is very appealing. Also, when clusters are merged/split, the decision is permanent. This reduces the number of different alternatives that need to be examined.

Hierarchical clustering can be performed in three different ways: single-link, average-link or complete-link. Single-link is the most commonly used method, which has a weakness. If two points from disjoint clusters happen to be near each other, the distinction between clusters will be lost. But, average-link and complete-link also have a weakness. Both of these methods are biased towards spherical clusters. Hierarchical clustering does not produce clusters, so the user must decide where to split the tree into groups. Another weakness is that it is sensitive to noise and outliers.

An advantage of the k-means clustering method is that it is relatively scalable in processing large data sets. It is also relatively efficient: O(tkn) (where n is the number objects, k is the number of clusters, and t is the number of iterations), normally $k,t < n$. This method also often terminates at a local optimum, the global optimum can be found using techniques such as genetic algorithms.

K-means clustering also has weaknesses, as does hierarchical clustering. One is that the parameter k must be chosen in advance. Another is that the data must be numerical and must be compared via Euclidean distance. The k-means algorithm works the best on data, which contains spherical clusters, and clusters with other geometry may not be found. This method is sensitive to outliers (points that do not belong in any clusters), as hierarchical clustering is. These outliers can distort centroid positions and ruin the clustering.

The most crucial disadvantage of both methods is that they are not robust to missing values/data in the matrix. These algorithms of these methods can only be calculated if the data sets are complete. The second part of this paper will discuss different methods in dealing with missing values.

Definition of K-Nearest Neighbors and Singular Value Decomposition:

The first method in dealing with missing values is a nonparametic approach to classification, called k-nearest neighbors (Jiangsheng 2002). The classification of records from the given dataset takes place in several steps. First, store all input/output pairs in the training set. For each pattern in the test set the following steps should be done. Search for the k nearest patterns to the input patterns using Euclidean distance measure. For classification, compute the confidence for each class as C_i/K, where C_i is the number of patterns among the k nearest patterns belonging to class i. The classification of the input pattern is the class with the highest confidence. For estimation, the output value is based on the average of the output values of the k nearest patterns.

For DNA microarray missing value analysis, the KNN based method can select genes with expression profiles similar to the gene of interest to impute missing values.

For example, consider gene 1 that has one missing value in experiment 1, this method would find K other genes, which have a value present in experiment 1, with expression most similar to gene 1 in experiments 2-N. A weighted average of values in experiment 1 from the K closest genes is then used as an estimate for the missing value in gene 1 (Troyanskaya et. al 2001).

The second method for dealing with missing values, as described by Alter et al., is a linear transformation of expression data from genes x arrays space to reduced "eigengenes" x "eigenarrays" space, called singular value decomposition. These mutually orthogonal expression patterns can be linearly combined to approximate the expression of all genes in the data set, and this is referred to as "eigengenes". The equation used here is: $A_{mxn} = U_{mxm} \Lambda_{mxn} V^T_{nxn,}$ where Λ is diagonal, and U and V are orthonormal (Alter et al. 2000). For microarray data, it is assumed that m>n as there are m rows of genes, and n rows of experiments. The columns are assumed as linearly independent. The matrix V^T contains eigengenes, which is quantified by corresponding eigenvalues on the diagonal matrix Λ. The most significant eigengenes are then selected by sorting eigengenes based on their eigenvalue (Troyanskaya et al. 2001). Troyanskaya et al. estimated missing value j in gene i by regressing this gene against k eigengenes and then used the coefficients of the regression to reconstruct j from a linear combination of the k eigengenes.

Comparison of KNN and SVD:

According to Troyanskaya et al., both the k nearest neighbor method and singular value decomposition method are proven to be better than replacing missing values by either zero or by the row average. They also found that KNN was a very accurate method

for estimating missing values showing only 6-26% average deviation from the true values. KNN is a great tool for estimating missing values for genes that are expressed in small clusters, whereas SVD is more likely to be inaccurate in this case because small clusters do not contribute significantly to global parameters (which SVD relies on). SVD showed quick deterioration in performance when a non-optimal fraction of missing values was used, whereas KNN showed to have less deterioration in performance. The final conclusion in this experiment was that KNN provides a robust and sensitive approach to estimating missing data for microarrays.

Alter et al. discusses the use of singular value decomposition for genome-wide expression data processing and modeling. SVD gives a global picture of the dynamics of gene expression by sorting data according to the correlations of the genes with eigengenes. Through a thorough analysis of the mathematical framework and biological data analysis of SVD, this paper concludes that SVD provides a useful mathematical framework for processing and modeling genome-wide expression data, in which biological meaning can be found. This method is not compared to other missing data estimation methods, so it cannot be inferred if SVD is in fact better than the KNN methodology as stated in Troyanskaya et al.'s paper.

The method of K nearest neighbors is shown to be an easy and efficient way of solving for missing data in Jiangsheng's paper because of its "perfect mathematical theory". It is reported that the KNN method has proven to work very well in many applications of classification.

Discussion:

The combination of avoiding missing values in data matrices and improvement of clustering methods will increase the validity of gene expression interpretation. I believe that the k-nearest neighbor missing value estimation is the most robust and sensitive approach to estimating missing data for microarrays. It is proven to be the most effective, and precise method through Troyanskaya et al's research. A complete data set matrix is the first step in improvement of clustering methods. Once this method is applied, we can move on to further improvement in clustering methods.

One of the major problems with the current clustering methods (hierarchical and k-means clustering) is that the application of clustering methods partitions a data set into clusters or classes, where similar data are assigned to the same cluster whereas dissimilar data should belong to different clusters. In real applications there is very often no sharp boundary between clusters. Fuzzy clustering is an approach that seems like a promising solution to this problem (Abunawass 1998).

Since, it is almost impossible to completely get rid of the noise in data, fuzzy clustering dispenses with unambiguous mapping of the data to classed and clusters, and instead computes degrees of membership that specify to what extend data belong to clusters (Delalin 2001). This fuzzy algorithm if applied to the current clustering methods, can smoothen out the expression level boundaries.

Conclusion:

We have already come so far in solving the mystery of life. DNA Microarray technology holds a great promise, in which only a few more refinements need to be

implemented. DNA microarray technology has already found many discoveries in the field of gene discovery, disease diagnosis, drug discovery, and toxicology research. With the suggested improvements mentioned in this paper, I see an optimistic future. Although, the biological value of gene expression should not be assumed and should be thoroughly researched before making any final conclusions.

References:

O. Alter, P. Brown, D. Botstein. Singular Value Decomposition for Genome-Wide Expression Data Processing and Modeling. Proc. Natl. Acad. Sci. USA, 97:10101-10106, 2000

A. Brazma, A.Robinson, G. Cameron, M. Ashburner. One-stop Shop for Microarray Data. Nature 403: 699-700, 2000

Y. Jiangsheng. Method of k-Nearest Neighbors. Institute of Computational Linguistics, Peking University, China 2002

O. Troyankskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R. Altman. Missing Value Estimation Methods for DNA Microarrays. Bioinformatics 17: 520-525, 2001

Eisen, Spellman, Brown. Cluster Analysis and Display of Genome-Wide Expression Patterns. Proc. Natl. Acad. Sci. USA, 95: 14863-14868, 1998

I. Davidson. Understanding K-means Non-Hierarchical Clustering. Albany Tech. Report: 02-2, 2002

P. Brown, D. Botstein. Exploring the New World of the Genome with DNA Microarrays. Nature Genetics 21: 33-37, 1999

R. Tibshirani, T. Hastie, M. Eisen, D. Ross, D. Bostein, P. Brown. Clustering Methods for the Analysis of DNA Microarray Data. Stanford University 1999

M. Schena. Microarrays: Biotechnology's Discovery Platform for Functional Genomics. TIBTech 16: 301, 1998

M. Razaz, B. Durrant. Intelligent Analysis of Genetic Data from DNA Chips. Bioinformatics Lab, UEA 2000

H. Delalin, J. Legar, G. Ranstein. A Fuzzy Algorithm for Gene Expression Analysis. IRIN, Ecole Polytech. Univ. de Nantes, 2001

A. Abunawass, G. Bhella, M. Ding, W. Li. Fuzzy Clustering Improves Convergence of the Backpropogation Algorithm. ACM Symposium on Applied Computing, 1998