Critical Review of Methods available for Microarray Data Analysis

Krithika Ramamoorthy

Introduction

The invention of DNA microarrays has triggered numerous efforts to analyze relative mRNA expression information from different cellular subsystems across a variety of experimental conditions. These conditions include internal cellular physiology from different cell lines, diverse physiological conditions in an intact organism, pathological tissue specimens from patients and serial time points following a stimulus to the cell or organism (1). A typical microarray experiment contains expression levels of thousands of genes across hundreds of different conditions. This results in the creation of enormous amounts of information which while being potentially useful in a variety of ways (discussed below) generates a need for high quality methods for determining statistical and biological significance. For instance, in Stanford and Rosetta alone, more than 30 million independent gene expression measurements were made over a span of two years (2). Potential uses for data generated from microarray experiments include (i) analysis of gene expression patterns in normal cells and tissues (ii) differential gene expression patterns (biomarker determination) in disease (iii) gene expression in model systems (iv) gene expression patterns in pathogens (v) gene expression in response to drug treatment i.e. dose response studies, mechanism of drug action studies etc (3) (vi) study of toxicogenomics or finding gene expression patterns in a model organism or tissue exposed to a compound and their use as early predictors of adverse events in humans (vii) evaluation of target selectivity by comparing the gene expression patterns in the target tissue with other compounds (viii) design of prognostic tests by finding a set of genes that accurately and adequately distinguishes one disease from another and determines sub classes. (4)

The analysis of microarray data can be performed at different levels of complexity: (i) at the level of single genes (probe level), where one seeks to determine whether a particular gene is differentially expressed under control and experimental conditions (ii) at the level of multiple genes, where one seeks to classify genes into known classes or to identify new and unknown classes (iii) at the systemic or genomic level, where one seeks to identify the underlying gene and protein networks responsible for the gene expression patterns observed (5). In this review, I will briefly discuss some of the aspects of the first and third kind of analysis and will focus more on the analysis at the multivariate gene level. I will review the issues that arise when evaluating algorithms that group together elements of these large data sets and for removing features that are redundant and provide little additional information.

Experimental protocols

The two most commonly used microarray techniques include the use of the high density oligonucleotide microarray technology as provided by the Affymetrix GeneChip technology and spotted cDNA arrays. The GeneChip contains oligonucleotides of 25 base pairs in length (or lesser) to probe for genes. The chip contains two kinds of probes – reference probes that match the target sequence exactly (perfect match) and partner probes that differ from the reference probes by only one base pair at the center (Mismatch probe). Typically 16-20 of these probe pairs, each hybridizing with a different sequence on the gene make up the complete probe set and are located at different pre-determined locations on the chip. In the spotted cDNA arrays, the DNA sequences are attached to a glass slide or other surface at multiple locations using a robotic arm (6). These sequences are laid out as spots with one DNA sequence per spot. Both technologies use similar experimental protocols. mRNA is isolated from experimental and control conditions and reverse transcribed in presence of a flurophore to generate differentially labeled cDNA samples. Traditionally experimental cDNAs are labeled with Cy5 while the control cDNAs are labeled with Cy3. Purified cDNAs are mixed in a 1:1 ratio and competitively hybridized to the microarray. Slides are washed to remove excess sample and read using a flourimeter. The fluorescent intensity of each spot is read separately for the control (green) and experimental (red) samples. Spots that appear yellow have

approximately equal amounts of control and experimental samples bound, while the red or green spots have increased levels of the experimental and control sample respectively. Black spots correspond to genes that are not differentially expressed, while red spots indicate upregulation of the gene (increase in expression levels) and green spots indicate downregulation (7). Software is available to for the image analysis and data normalization and these methods will not be discussed in this review. After normalization, the expression ratio of the experimental to control value is calculated for each spot and is typically recorded as a $\log_2$ [Cy5/Cy3] ratio in an n-dimensional expression matrix, where n is the total number of genes in the experiment. In this matrix, every row represents a different gene, every column represents a different condition, all the elements of a row or column represent a profile of the experiment that we try to analyze and the individual expression measurements within the profile represent the features (1). Logarithms are used rather than the ratios themselves because they are easier to model and interpret. A gene that is upregulated by a factor of 2 has a log ratio of 1, a gene that is downregulated by 2 has a log ratio of -1, and a gene expressed at a constant level has a log ratio of 0. One can use log ratios of other bases as well as long as one is consistent (5).

Normalization and noise

Before data from multiple microarray experiments can be pooled into a single analysis, the data must first be normalized and corrected for possible sources of noise. The difficulties arise from numerous potential sources of random and systematic measurement error and from the small number of samples or replicates relative to the large number of genes or probes and conditions tested for (8). Normalization methods might include simple methods such as adjusting the overall brightness of each scanned microarray image (assuming that the quantity of RNA is equal), using expression levels of housekeeping genes whose expression levels are assumed to be constant across the experimental conditions considered (not always valid) and the use of other more sophisticated nonlinear techniques that are reviewed elsewhere (9). There are several sources of noise in microarray data. Inter and intra microarray variations can markedly skew the interpretation of the expression data. Improving the reliability of the expression measurements starts with proper experimental design such as pooling biological samples before hybridization to ensure true replicates. Scanned hybridization images need to be inspected for artifacts such as scratches and bubbles. There is substantial heterogeneity of gene expression in cell subpopulations of most organs and disease states. Failure to account for such variation could lead to over-interpretation or spurious functional gene associations (4). All the abovementioned sources of noise need to be incorporated directly into the analytical tools that interpret the data in order to get more reliable estimates of clinical and biological data. Also, differences between the two microarray technologies need to be taken into account. Specifically, oligonucleotide microarrays report absolute expression levels while spotted cDNA microarrays report relative differences in gene expression between samples. Different normalization techniques need to be used in both cases as the assumptions made about the data are different and data from the two different assays cannot be directly combined. For instance, if we assume that in any given experiment, most genes do not change in expression levels and that equal numbers of genes are upregulated and downregulated (not always a valid assumption), then differential expression measurements from spotted arrays might be found to be normally distributed while measurements from oligonucleotide microarrays will not have the same distribution. Furthermore there has been found to be striking non-correlation between quantitative measurements in Affymetrix arrays and ratios of intensities from spotted arrays because the two technologies measure gene expression differently (4). In addition to random and systematic errors (that include biases), outliers in the data reduce both specificity (measure of false positives) and sensitivity (measure of false negatives). Outliers can be caused by factors such as uncorrected image artifacts, improper or failed hybridizations etc (8).

Initially measurements of differential expressions were assessed by comparing the ratio of expression levels between two conditions, a method known as fold change approach. Genes with ratios above a fixed cutoff k were said to be differentially expressed. However, this method has been proved to be unreliable because it fails to take into account measurement errors. For example, an excess of low intensity genes

might be mistakenly identified as differentially expressed because their fold change values have a larger variance than the fold change values of high intensity genes. A more sophisticated method proposed by Li and Wong, fits the data to a model that accounts for random, array and probe specific noise and then evaluates whether the 90% confidence interval for each gene's fold change excludes 1.0. This model incorporates measurement variability but does not perform well when the data set is too small or heterogeneous (5).

Significance and errors of inference

All statistical inferences are associated with a probability of being incorrect. False positives are incorrect expressions of differential expression. False negatives are failures to detect true differential expression. Regardless of the test statistic used (false negative rate or false positive rate), one needs to convert it to a p value to determine its significance. Standard methods for estimating p values use statistical distribution tables. However, these tables rely on the assumption that the data is sampled from normal populations with equal variances. This is not true when considering gene expression data from different conditions such as tumor and normal cells. Permutation tests, which are carried out by repeatedly shuffling the sample's class labels and computing t statistics for the genes in the shuffled data enable one to assess significance without assuming normality. However, these permutation tests are time consuming, complicated and require that the data set be large enough so that different permutations are possible.

The issue of multiple testing is crucial in microarray analyses because most microarray experiments require that the expression levels of thousands of genes be monitored across numerous conditions requiring that thousands of statistical tests be computed. If we assume a standard p value for each experiment, we run the risk of accumulating a large number of false positives. For instance, if the p value used is 0.01, which means that 1% of all results are false positives, while this might be an acceptable statistic in most biological studies, a microarray experiment that studies 10000 genes will have a false positive level of 100 genes which is clearly unacceptable. Two methods that have been proposed to address the problem of multiple testing include Family-wise error-rate control (FWER) and False-discovery rate control (FDR). FWER is the overall probability that at least one gene is incorrectly identified as over expressed over a number of statistical tests. One way to control the FWER is to increase the stringency of each individual test. The single step Bonferroni correction is the best known procedure to control the FWER (and FDR) and defines an effective rate as the standard false positive rate divided by the number of tests conducted (e.g. 0.01/10000). This procedure ensures that the probability of making at least one false positive error among an entire set of statistical tests is no more than 0.01. This is an extremely stringent control that drastically increases the false negative rate (8). Other methods have been proposed such as the step down correction method, permutation based one step correction method etc. These latter tests perform better compared to the standard Bonferroni but are more computationally complex.

The FDR is the probability that a given gene identified as differentially expressed is a false positive. The FDR is a post measure of confidence and uses information available in the data to estimate the proportion of false positives that have occurred. A simple method for bounding the FDR proposed by Benjamini and Hochberg assumes independent tests and sets an upper bound for the FDR by a step up or step down procedure applied to individual p values. In this method, the calculated p values of each individual test are ordered from the most significant p (1), to the least significant p (n). A rule R is then formulated that will specify when the null hypothesis (a gene is not differentially expressed) is rejected. The FDR of R is the expected proportion of hypotheses, h (i) that is actually true. Benjamini and Hochberg identified an algorithm that allows the specification of a preset value of α which serves as the upper bound of FDR or R, where FDR (R α) $\leq \alpha$. The analyst can then use R as a measure of significance and be assured that the FDR using R will be less than or equal to the preset value α derived from Benjamini and Hochberg's algorithm (5).

Several parametric and non parametric methods are available for analysis gene expression data at the level of a single gene. The parametric methods include the standard t test, variations of the t test and regression modeling. Nonparametric tests include Bayesian frameworks, mixed modeling approaches and simpler tests such as the Wilcoxon's test, Mann-Whitney U test etc. These methods will not be reviewed in detailed here and are available elsewhere (5, 9-11).

Supervised and unsupervised methods

Supervised methods require that the genes or conditions are associated with some external sources of information that provide pre-existing classifications. This information includes knowledge about gene function or regulation, disease subtype or tissue origin of a cell type. This classification information is used to drive the analysis of gene expression and hence the term supervised learning method. For example, consider the problem of classifying unknown genes as ribosomal or non-ribosomal. Because some genes are already known to be ribosomal, we can use these genes to build a model of ribosomal genes and to determine features that identify this set. If expression measurements are available over a variety of experimental conditions, we can assess whether these genes are ribosomal or not by comparing them with the expression profiles of the training set (1).

These methods are used primarily for two purposes: finding genes with expression levels that are significantly different between groups of samples and predicting characteristic(s) that completely define a sample or group. Significance can be evaluated in several ways as discussed above. When determining whether a particular gene is differentially expressed between two samples, there are four characteristics that need to be considered: absolute expression level (whether the gene is expressed at high or low intensity levels), degree of change between groups, fold change between groups or the ratio of expression levels across samples and finally the reproducibility of the measurement. All four characteristics are related. For example, genes measure at low expression intensities have poorer reproducibility across samples and have high fold changes that may not be biologically significant (4).

Like HMM based motif finding algorithms, supervised method suffer from over-fitting the training set. If true, the positive predictive value of the algorithms would be high for previously classified genes but potentially important genes would be misclassified. Another implicit assumption is that subtle differences in gene expression must be discerned which is not always true. The success of these algorithms depends heavily on the quality of the initial training set. Methods that fall in this category include nearest neighbor approach decision trees, neural networks and support vector machines. In this review, I will focus on two of the more popular supervised methods namely nearest neighbor technique and support vector machine.

Unsupervised methods require no additional information besides the gene expression data itself. These methods are driven towards discovering patterns or relationships in a data set. They are best used for exploratory tasks. With unsupervised techniques, there are three classes of techniques: feature determination or determining genes with interesting properties without looking for a particular pattern such as principal component analysis, cluster determination or looking for similar gene expression patterns such as nearest neighbor clustering, self organizing maps, k means clustering and finally network determination or determining graphs representing gene-gene or gene-phenotype interactions using Boolean networks, Bayesian networks and relevance networks. The final technique is used for genomic scale analyses. In this review, I will discuss the three most common unsupervised techniques namely hierarchical clustering, self organizing maps (analysis at the level of multiple genes) and relevance networks (analysis at the network level).

Dissimilarity measures

Dissimilarity measures indicate the extent of similarity between two genes. These measures are different from clustering methods which build on these dissimilarity measures to create groups with similar patters. The most commonly used dissimilarity measure is Euclidean distance (square root of the sum of the

squared differences between the corresponding features), for which each gene is considered a point in multidimensional space, each axis is a separate sample, and the coordinate on each axis is the amount of gene expression in that sample. There are several disadvantages to using this measure. One is that if the measurements are not normalized, correlation of measurements can be missed because the focus is on the overall extent of expression. The second disadvantage is that this measure does not identify negative associations such as gene interactions between tumor-suppressor genes. For example, the tumor repressor protein p53 acts as a transrepressor of several genes. This means that with high levels of p53, the expression of the other genes will be low. Negative interactions similar to this one are clearly different from no interactions.

A second dissimilarity method is the Pearson Correlation Coefficient (ranging from -1 to +1) which is measured between two genes that are treated as vectors of measurement. The disadvantages of using this method are that it assumes normal distribution of measurements which is not the case for most oligonucleotide microarray measurements; it assumes linear interaction between genes which is again not always the case in biology where a particular gene might best regulate other genes when it is in the middle of its expression. This method is also sensitive to outliers. A third dissimilarity method is mutual information which allow for any possible model of interaction between the genes and uses each measure equally regardless of the actual value. This method is not affected by outliers and allows for negative associations. However calculating mutual information requires using discrete measurements such as representing gene expressions as being high or low, and the measure depends on the exact number of bins used. It is also possible for gene-gene interactions having high mutual information to have complicated mathematical functions and biological interpretation is difficult.

Hierarchical Clustering

Hierarchical Clustering is the most frequently used method for analyzing microarray data analysis. This method compares similarity between gene expression vectors that are similar to those used for phylogenetic analysis. This technique builds clusters of genes with similar patterns of expression. An n dimensional matrix is obtained that represents the extent of mathematical similarity (pair-wise similarity scores) seen in the genes. Then the method builds a dendogram that assembles all the elements of the matrix into a single tree. This is done by iteratively grouping together genes which have a high correlation in terms of expression measures, then grouping the different groups themselves to form a tree. A node is generated for the highest scoring pair, its average gene expression vector is computed and the distance between the node and the remainder of the matrix is recalculated. This process is iterated n-1 times till all the gene expression profiles are incorporated into one single tree. Each leaf of the tree represents an expression profile for a single gene. Co-expressed genes branch off common nodes. Similarity scores are reflected by the branch lengths of any pair of genes in the tree – the length of a branch is inversely proportional to the extent of similarity. Although construction of the tree is initiated by connecting genes that are most similar to each other, genes added later are connected to the branches that they are most closely associated with. Although each branch links two elements, the overall shape of the tree can be asymmetric. This method was pioneered in the Brown and Botstein labs at Stanford (12).

Several methods can be employed to build the tree. One method is Single–linkage clustering in which the program calculates the distance between two clusters by determining the minimum distance between two members of the cluster (nearest neighbor method). In the Complete-linkage clustering method, the farthest distance between two members of the cluster is used. This method tends to produce compact clusters of similar size. A third method is the Average-Linkage clustering method in which distances are calculated by averaging distances between members of the cluster. This method is similar to the UPGMA tree building method which calculates distances based on average expression profile for each cluster and joins the clusters separated by the smallest average distance (13). An alternate method is the Weighted-Pair group average method in which the size of the clusters is used to weight the clusters and is best used when it is expected that the cluster sizes will be uneven. The last method was used by the Botstein labs in

their pioneering study on hierarchical clustering of microarray data (12). The Pearson coefficient was used as an estimate of similarity. The dot product is calculated for each pair of correlation coefficients and used to generate the matrix. This method is advantageous because similarity scores are a reflection of the shape of the expression profiles rather than the magnitudes of the signals.

The hierarchical clustering software package introduced by the Brown and Botstein labs (TREEVIEW) (12) is one of the most widely used tools in functional genomics. From such clustering studies, fundamental relationships that exist in the cell can be discovered. For example, genes that encode functionally related proteins are often co-regulated. Potential functions for novel or uncharacterized genes can be guesstimated based on co-regulation with genes of known function. Coincident expression patterns over a range of experimental conditions increase the probability of discovering a common transcriptional regulatory program for a set of genes and computational methods that assay non-coding regulatory regions for conserved transcription factor binding sites may be informative in such cases. Crosstalk between signaling pathways may also be reflected in similar expression of genes under different conditions (7). This method is particularly advantageous in visualizing overall similarities in expression patterns observed in an experiment. The number and size of expression patterns within a dataset can be estimated quickly.

There are significant disadvantages to using this method. Hierarchical clustering ignores negative associations even when the underlying dissimilarity measure used supports them. As discussed above negative associations can be crucial to the biological process under study and may be missed completely. This method does not result in clusters that are globally optimal in that early incorrect choices in linking genes to a branch are not reversible after other branches have been added to the tree. So, this method falls under the category of greedy algorithms which provide good answers but are computationally intractable for finding the most globally optimal set of clusters (4). Such a scenario might result in poorly delimited, noisy clusters that may obscure relevant relationships in a dataset. Like all clustering methods, the number and composition of the clusters produced vary with the choice of distance metric and analysis is highly subjective. The use of a weighted average method becomes increasingly problematic as the weighted average may not accurately reflect the expression profiles of genes within the cluster. The distance metrics also assume linear relationships between the objects which is not necessarily true in biology. Methods that assume nonlinear correlations between the genes analyzed such as the use of the Spearman coefficient allow many-at-once comparisons and may be more relevant to the needs of system biologists who wish to construct a comprehensive network of specific transciptosomes. The results of clustering are very sensitive to the features used to compute the dissimilarity metric. Features are usually weighed equally and the effects of relevant features can be masked by irrelevant ones. For example, in a study of response of cancer profiles to a drug, a feature set including the entire genomic profile may not be appropriate because the response depends on a handful of target genes and inclusion of thousands of other genes might introduce a lot of noise into the analysis and make extraction of similarities of genes difficult (1).

An improvement to hierarchical clustering is k-means clustering when the exact number of clusters to be created is known. Initially cluster centers are selected randomly. Since k is specified randomly, primary component analysis is first performed to estimate the number of regulations on the data set. This technique reduces the dimensionality of the dataset by projecting the data in n-dimensional space onto a Cartesian coordinate system. This provides a visually intuitive interface for making general assumptions about the diversity and content of the dataset. Different k values can also be tested initially. Often times, k values are over estimated since the success of the analysis rests on the quality of the clusters produced (7). For every iteration, all the profiles are assigned to a cluster that they are nearest to and then the cluster center is recalculated based on the profiles within the cluster. A recursive algorithm is used to choose expression vectors, calculate the inter and intra distances and move the vector only if the selected cluster is more similar to the point than the original one. K-means clustering allow biological knowledge to be

integrated into the clustering method but biological significance needs to be judged manually. This method has been successfully used the Altman lab to distinguish between two types of lymphomas (1). This approach may produce clusters that are more stable than those produced by solely binary comparisons. It is possible to address hypothesis driven questions by seeding clusters with expression profiles of interest such as a molecular signature that is characteristic of a particular cancer type prior to running the k-means clustering method. However this process is still a subjective process that is sensitive to the initial assumptions about the expression profile diversity in the matrix.

Self Organizing Maps

Self Organizing Maps (SOM) is a neural network based clustering system that is better designed for exploratory analysis. Similar to the k-means clustering, the total number of clusters needs to be estimated initially. The genes are first represented as points in multidimensional space. Each biological sample is considered a separate axis of this space and the expression levels are coordinates. This is easily visualized using 3 or fewer microarrays but can be extended to n dimensions. Euclidean distance is the most common measure of similarity used but the other metrics are also applicable. A map is set with the centers of each cluster to be arranged in an initial arbitrary configuration. As the method iterates, the centroids move towards randomly chosen genes at a decreasing rate. This is continued until there is no further movement of the centroids. At each iteration, a data point P is randomly selected and the node closest to the data point Np is moved most while the other nodes are adjusted proportionately depending upon their distance to Np in the initial configuration. This algorithm has been implemented in the software GENECLUSTER (13). A weighing factor learned from the test set ensures that the closest nodes are moved more than the distant nodes. Thus, each node comes to define a cluster of similar gene expression profiles and adjacent clusters are likely to contain genes that have related expression patterns or kinetics.

The primary advantages of the SOMs include easy visualization of expression patterns and reduced computational requirements compared with methods that require comprehensive pair wise comparisons such as Hierarchical clustering (4). The nodes are fit according to a learned weight function and the positions of the nodes reflect the distribution of objects in the expression space and thus nodal organization in the SOMs is less arbitrary than those derived from pair-wise comparisons of similarity scores. Adjacent nodes in a SOM are more closely related than distantly located nodes. Although, use of this method does not guarantee discrete clusters, changing the starting configuration provides a starting point for addressing these issues. The use of SOMs in conjunction with Hierarchical clustering or Principal Component analysis gives a reasonable estimate of the number of nodes required to efficiently describe the dataset (7).

The disadvantages of this method include an arbitrary initial configuration of the SOM and hence random movement of the centroids which makes the final configuration not always reproducible. Similar to Hierarchical clustering method, crucial negative associations are missed. Even after the centroids have stopped moving, further techniques are required to separate the boundaries of each cluster. Finally, genes can belong to only one cluster at a time (4).

Relevance Networks

Relevance networks perform expression analysis at the genome or network level. They allow networks of features to be built, whether it is of genes or phenotypes or clinical measurements. This method first compares all features of the genes in a pair-wise manner similar to the clustering methods. Two genes are typically compared with each other by plotting all samples on a scatter plot using expression levels as coordinates. A correlation coefficient (or any other dissimilarity metric) is calculated. A threshold is chosen and only those features above those are kept, while the others are discarded. These are displayed in a graph with genes as nodes and relationships as edges. The closer the association, the thicker is the edge. This method involves permuting the entire original dataset to preserve the distribution of gene

expression values but breaking the link between expression value and a particular condition or tissue. The pair-wise association strengths are recalculated for each permutation and the largest value of association obtained is recorded. After a large number of permutations, this maximum number becomes the minimum threshold value for any association in the unpermuted datasets. The threshold is chosen using permutation analysis but can act as a dial, increasing or decreasing the number of associations (4).

Relevance networks can be used for analysis of gene expression dynamics (16). The primary motivation for studying gene expression dynamics is that static expression levels may not provide all the information required for identifying important relationships. For instance, consider the hypothetical example in which gene A codes for an enhancer protein that upregulates the expression levels of gene B. Since the expression level of B can be high or low, simply examining the correlation between the statical patterns of gene expression does not enunciate such relationships. Linear correlation coefficients are used as the measure of association. For dynamic analyses, slopes are calculated between each adjacent pair of expression data points yielding a dataset where each row is the time series of a particular gene's expression dynamics. Pair-wise Pearson's correlation coefficients are then calculated between all possible combinations of rows. They are then squared after which the original sign is appended to indicate negative or positive correlation (16).

Correlation coefficients are sensitive to outliers which can bias downstream data analysis. Two symmetric outliers can artificially raise the correlation coefficient of an otherwise nonlinear distribution. An entropy based filter is used to remove genes with outlying values by ranking the genes based on calculated entropies and discarding the bottom 5% from the analysis. Other issues pertinent to dynamic analysis include the issue of stasis. Most of the genes do not exhibit a change in their expression values over time, and this may lead to seriously misleading analyses as genes that remain stationary together can lead to an artificially high degree of association. To address this issue, the stationary points are filtered out by setting an exclusion range around the zero slope range. Since many data points are removed, the remaining data can become very sparse making spurious associations possible (16). This method has been used to study the effect of anti cancer agents on genes in cancer cell lines measured using microarray analysis (17). Specific clusters were found through analysis of RNA expression and anticancer agent susceptibility. A putative link was also found between a single gene and anti cancer agent susceptibility.

There are many advantages to using relevance networks. Features of more than one data type can be represented together. For example, if a strong association exists, a link can be visualized between systolic blood pressure and the expression level of a particular gene. Features can have a variable number of associations. It is possible for a transcription factor to be associated with more genes than some downstream factor. This can be contrasted with clustering methods discussed earlier which can only link each feature to one other feature, typically the one that it is most strongly correlated to but not to other links. This algorithm allows visualization of negative as well of positive associations, lack of which was one of the strongest drawbacks of the other methods (4).

Disadvantages include the degree of complexity observed at lower thresholds at which many links are possible between genes, linking them all together in a single network. Completely connected components of these graphs known as cliques cannot be found easily. Further, there is no modeling of noise making correlation coefficients calculated from low intensity expressions similar to high intensity expressions (17). Another serious limitation of using relevance networks in dynamic analysis is the exclusion of outliers. Outliers may sometimes be a true indicator of biological function as for instance when a gene acts as a step function. They may also be present when a gene or a pharmaceutical agent acts only on a single cell line. Therefore some of the valid hypotheses may be missed in order not to have a high false positive rate.

Nearest Neighbors

This method can be used in an unsupervised manner; however it is usually used in a supervised manner to determine genes that have functional properties similar to designated queries. For example, an ideal gene pattern might be one that is highly expressed in one condition and expressed at low levels in another condition. All the genes that are being analyzed can be compared to the ideal pattern and ranked based on their similarities. This method was used to distinguish acute lymphatic leukemia from acute myelogenous leukemia in one the first publications that showed how microarray analysis can assist in difficult clinical diagnosis (15).

Although this technique results in a set of genes that splits the data into two distinct sets, it does not give the smallest set of genes that most accurately makes the splits. For example, the expression levels of two genes might split the two conditions perfectly, but these two genes may not be the ones closest to the idealized query (4). This technique also finds use in toxicogenomics. Tissue exposed to various compounds that are known to induce toxicity at different time points as well as normal tissue is subjected to microarray analysis and makes up the training set that creates an implicit model of toxicity. Newer compounds can be tested on these tissues (on a high throughput basis) and the distance of the expression patterns from the training set can be calculated to make decisions on the similarities of mechanisms of toxicity.

Support Vector Machines

Support Vector Machines (SVM) addresses the problem of finding combinations of genes that better split the dataset into distinct groups. Using previous information about gene expression, the SVM learns expression features for a specific class and then classifies the genes based on their expression levels as either included in the class or excluded from it. As the SVM learns to distinguish between class members and outliers, an optimal hyper plane is drawn to divide these points. Even if it is not possible to use the genes alone to create the separation, combination of features of the genes may be used to make the delineation possible. Each biological sample is considered as a point in multidimensional space in which each dimension is a gene and the coordinate of each point is the expression level of the gene. Using SVMs this multidimensional space acquires more dimensions based on mathematical combinations of the gene expressions. The goal of this method is to find a plane that perfectly splits two or more sets of samples. Using this method, the delineating plane has the largest possible margin from samples in the two conditions and therefore avoids data over-fitting which has been discussed previously as one of the problems associated with supervised methods of microarray analysis. The SVM software is included in a package called GIST and can be downloaded from Columbia University.

Since the number of genes being analyzed in microarray experiments is very large, it might not always be possible to split them up into two separate groups. The total number of features available in the dataset is expanded by this method by combining genes using mathematical operations called kernel functions. For example, in addition to using the expression levels of two genes A and B, the combination features A x B, A/B, $(A \times B)^2$ etc can also be used to effectively separate the dataset into two biologically distinct groups. The dot product of two normalized vectors if often seen to be the most effective way to measure similarity between genes as can be seen in Hierarchical clustering (12). Other higher powers of kernel functions have also been explored previously (14). Features for all d-fold interactions that take place in the dataset (where d is any positive power used in the kernel function) are taken into account and multiple iterations are performed before an optimal kernel function is derived. Although SVMs with higher power kernel functions are more effective than simpler ones, none of them correctly identified all the genes analyzed. Another concern in classification schemes arises from the relative imbalance of true positives versus negatives in the dataset. In such cases where the magnitude of noise in the negative data set exceeds the magnitude of the positive signal, incorrect false negative classifications confound the data analysis. In case, it is not possible to generate the delineating plane, a soft margin is implemented which allows some members of the training set to be misclassified. The soft margin along with the modified

kernel function, which includes a diagonal element to correct for the imbalance between the number of positive and negative objects in the dataset, enable more optimal delineations. One problem when a higher kernel function is used to split to data set is that even though the function may be mathematically sensible, its biological interpretation is extremely difficult.

Conclusion

The field of microarray expression analysis is a booming one and the literature is filled with novel algorithms as well as potential uses of this data in drug discovery and clinical diagnosis. The goal of this short review is to provide an introduction to the different aspects of data analysis and mining to yield useful information. I have discussed in detail the more popular methods while briefly alluding to the numerous other techniques in existence. This review also seeks to provide a framework for evaluating novel methods that may be proposed. Algorithms may be supervised or unsupervised depending on the extent of external information used to drive the analysis. The algorithms can be used for static or dynamic analysis, at the level of single or multiple genes or at the genomic level. The primary challenge is to apply these techniques in a manner than will provide reliable answers to questions of biological interest.

Uses for supervised method are easy to imagine. One example to illustrate this is the use of supervised methods for hypotheses in toxicogenomics. For instance, hypotheses such as 'some genes in the genome influence liver metabolism of a particular compound' can be answered with a technique that finds genes whose expression levels are markedly different between samples with drug and without. This question is a critical one in pharmacokinetic analysis of compounds that go through the clinical pipeline of most pharmaceutical companies. A second use is in the development of diagnostic tests to determine biomarkers for specific diseases such as 'a combination of gene expression measurements that accurately distinguish malignant from non-malignant tumors'. The use for unsupervised methods is less intuitive because the questions answered by such algorithms are less direct. For example questions about the number and type of responses in a period of time after treatment with a pharmacological agent cannot be found in a supervised manner. These techniques survey all genes and cluster them together based on expression patterns. True genetic regulatory networks can be found by methods such as Bayesian networks which have not been discussed here. Since there are no ideal answers being sought, it is not clear which method needs to be used. However, unsupervised methods can be instrumental in the early discovery process.

These challenges in data analysis at the level of data normalization, determining ideal analytical techniques is a short term one and after the functional genomics pipeline has been established, the rate limiting step shifts to post analytical challenges. The findings from microarray analysis need to be linked to other aspect of the discovery pipeline. Finding a 'list of genes' from the microarray analysis should not be the end in itself and should be validated by conducting relevant biological experiments such as PCRs, northern blots etc, numerical certification of the results using alternate techniques and by linking the data to other sources of information such as Entrez, Locuslink, GO, Genbank in order to improve the range and conclusions that can be drawn. Techniques are also emerging to reconstruct networks of genetic interactions in order to create integrated and systematic models of biological systems (18).

Figures describing techniques discussed in the paper

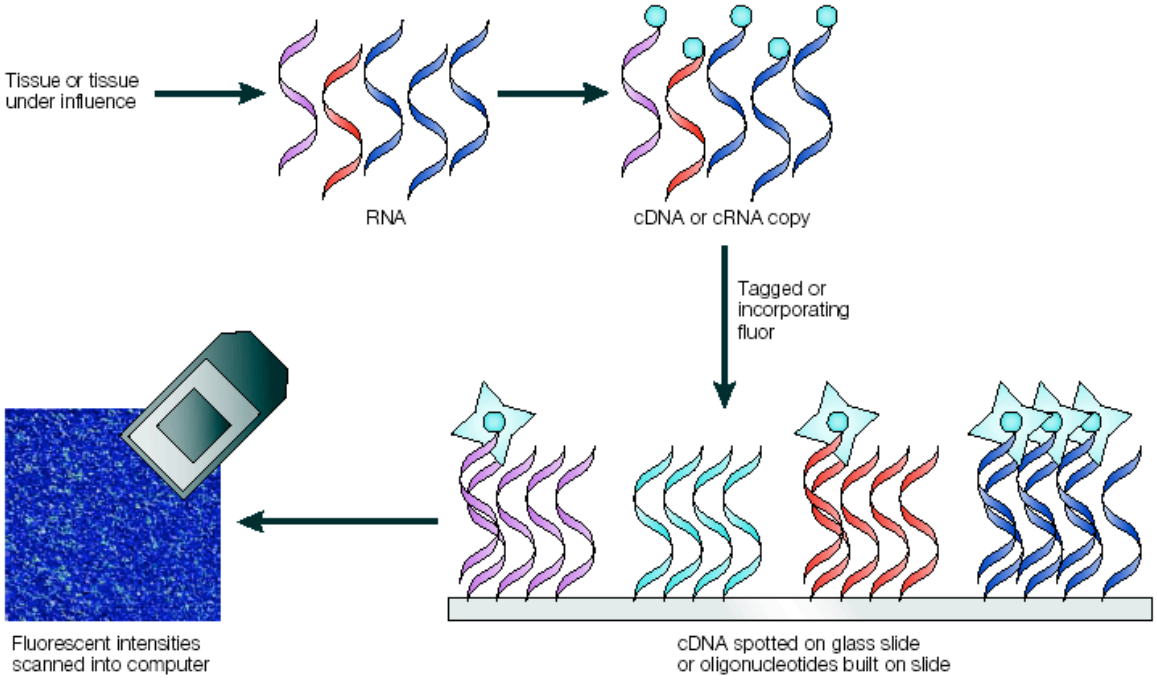

Fig 1: Experimental protocol for microarray analysis. Slight differences exist for oligonucleotide microarrays and cDNA microarrays (4)

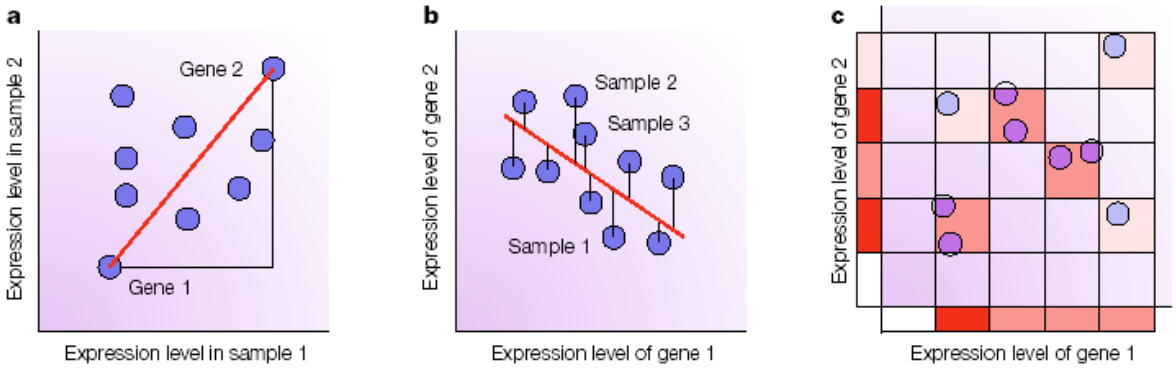

Fig 2: Dissimilarity measures used for clustering analysis. 2a represents Euclidean distance 2b represents Pearson's coefficient and 2c represents mutual information (4)

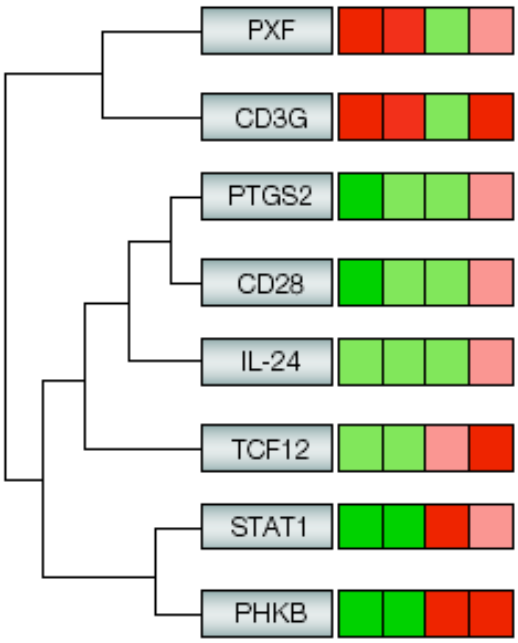

Fig 3: Hierarchical Clustering analysis that separates genes into clusters based on similarity. Red indicated upregulation while green indicated downregulation of the gene (4).

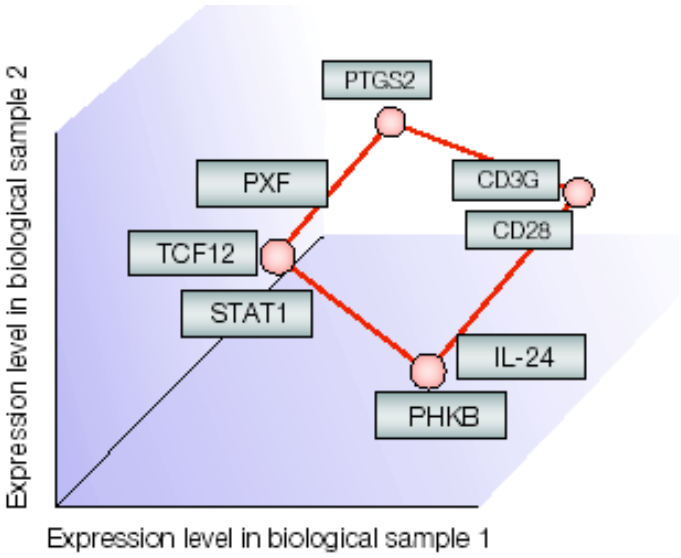

Fig 4: Self organizing networks find variable sized clusters of genes that are similar to each other, given the number of clusters that need to be found (4)

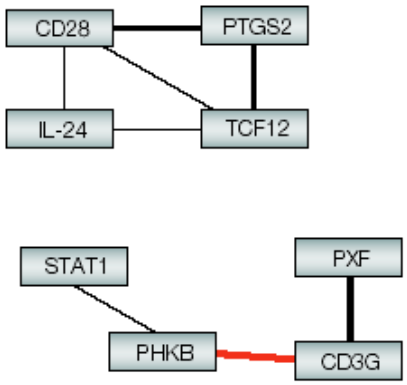

Fig 5: Relevance networks find and display pairs of genes with strong positive or negative correlations and then construct a network of these gene pairs. The strength of the correlation is proportional to the thickness of the lines between the genes, with red indicating a negative correlation (4).

Fig 6: The nearest neighbor method first involve construction of hypothetical genes that fit the individual patterns, and then finds genes that are the most similar to the hypothetical genes.

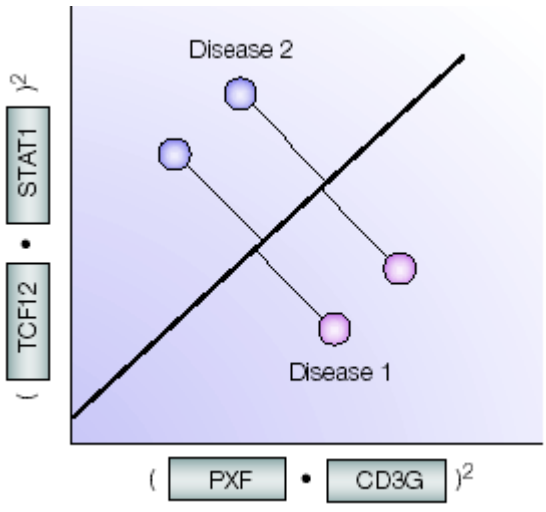

Fig 7: Support Vector machines try mathematical combination of genes to find the line that best separates biological samples

References

1. Raychoudhuri S, Sutphin P. D., Chang JT, Altman RB, Basic microarray analysis: grouping and feature reduction. *Trends Biotechnol*. 2001 May; 19(5):189-93.
2. Bassett DE Jr., Eisen MB, Boguski MS, Gene expression informatics--it's all in your mine. *Nat Genet.* 1999 Jan; 21(1 Suppl):51-5.
3. Debouck C, Goodfellow PN, DNA microarrays in drug discovery and development. *Nat Genet.* 1999 Jan; 21(1 Suppl):48-50.
4. Butte A, The use and analysis of microarray data. *Nat Rev Drug Discov.* 2002 Dec; 1(12):951-60.
5. Tan EC, A critical review of statistical methods for differential analysis of 2-sample microarrays, BIOC218 Project Spring '03 project
6. Brown PO, Botstein D, Exploring the new world of the genome with DNA microarrays. *Nat Genet.* 1999 Jan; 21(1 Suppl):33-7.
7. Davies E, A critical review of computational methods used to manage microarray data sets, BIOC218 Winter'03 project
8. Nadon R, Shoemaker J, Statistical issues with microarrays: processing and analysis. *Trends Genet.* 2002 May; 18(5):265-71.
9. Kerr KM, Churchill GA, Statistical design and the analysis of gene expression microarray data. *Genet Res*. 2001 Apr; 77(2):123-8.
10. Bolstad BM, Irizarry RA, Astrand M, Speed TP, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics.* 2003 Jan 22; 19(2):185-93.
11. Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB, Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*. 2002 Nov; 18(11):1454-61.
12. Eisen MB, Spellman PT, Brown PO, Botstein D, Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci* U S A. 1998 Dec 8; 95(25):14863-8.
13. Smetana CR, Overview of microarray data analysis, BIOC218 Spring'03 project
14. Brown MP, Grundy WN, Lin D, Christianini N, Sugnet CW, Furey TS, Ares M Jr, Haussler D, Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci* U S A. 2000 Jan 4; 97(1):262-7.
15. Golub TR, Slonim TK, Tamayo P, Huard C, Gaseenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caliguiri MA, Bloomfield CD, Lander ES, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*. 1999 Oct 15; 286(5439):531-7.
16. Reis BY, Butte AS, Kohane IS, Extracting knowledge from dynamics in gene expression. *J Biomed Inform*. 2001 Feb; 34(1):15-27.
17. Butte AJ, Tamayo, Slonin D, Golub TR, Kohane IS, Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc Natl Acad Sci* U S A. 2000 Oct 24; 97(22):12182-6.
18. Altman RB, Raychaudhuri S, Whole-genome expression analysis: challenges beyond clustering. *Curr Opin Struct Biol.* 2001 Jun; 11(3):340-7.