REVIEW

Analyzing microarray data using cluster analysis

William Shannon†[1,2], *Robert Culverhouse*[1] & *Jill Duncan*[1]

†*Author for correspondence*
[1]*Department of Medicine*
[2]*Division of Biostatistics, Washington Univ. School of Medicine, 660 S. Euclid Ave, Campus Box 8005, St. Louis, MO 63110, USA*
Tel: +1 314 454 8356;
Fax: +1 314 454 5113;
E-mail: shannon@ilya.wustl.edu

As pharmacogenetics researchers gather more detailed and complex data on gene polymorphisms that effect drug metabolizing enzymes, drug target receptors and drug transporters, they will need access to advanced statistical tools to mine that data. These tools include approaches from classical biostatistics, such as logistic regression or linear discriminant analysis, and supervised learning methods from computer science, such as support vector machines and artificial neural networks. In this review, we present an overview of another class of models, cluster analysis, which will likely be less familiar to pharmacogenetics researchers. Cluster analysis is used to analyze data that is not *a priori* known to contain any specific subgroups. The goal is to use the data itself to identify meaningful or informative subgroups. Specifically, we will focus on demonstrating the use of distance-based methods of hierarchical clustering to analyze gene expression data.

Keywords: consensus methods, distance calculations, heat maps, hierarchical clustering, k-means clustering, Mantel statistics, microarrays, unsupervised learning

Ashley Publications Ltd
www.ashley-pub.com

Introduction

As gene chips become more routine in basic research, it is important for biologists to understand the biostatistical methods used to analyze these data so that they can better interpret the biological meaning of the results. Strategies for analyzing gene chip data can be broadly grouped into two categories: *discrimination* (or *supervised learning*) and *clustering* (or *unsupervised learning*).

Discrimination requires that the data consist of two components. The first is the gene expression measurements from the chips run on a set of samples. The second component is the data characterizing the samples (e.g., tumor or normal tissue, time cells were harvested from a culture) or the genes (e.g., regulatory factor, oncogene). For this method, the goal is to use a mathematical model to predict a sample characteristic, say tumor subtype, from the expression values. Once this model is fit, the gene expression values of a new tumor sample can be used to make a 'prediction' of its subtype class. There are a large number of statistical and computational approaches for discrimination (i.e., supervised learning) ranging from classical statistical linear discriminant analysis [1] to modern machine learning approaches such as support vector machines [2,3] and artificial neural networks [4,5]. Microarray analysis using supervised learning methods was recently reviewed in this journal [6] and will not be discussed further in this review.

In this review, we will discuss the second group of analytical approaches for analyzing microarray data: *cluster analysis* or *unsupervised learning*. In clustering, the data consist only of the gene expression values. The analytical goal is to find clusters of samples or clusters of genes such that observations within a cluster are more similar to each other than they are to observations in different clusters. Cluster analysis can be viewed as a data reduction method in that the observations in a cluster can be represented by an 'average' of the observations in that cluster.

There are a large number of statistical and computational approaches available for clustering. These include hierarchical clustering [7,8] and k-means clustering [9] from the statistical literature and self-organizing maps [10] and artificial neural networks [4] from the machine learning literature. While these algorithms are relatively equivalent in terms of performance (i.e., one method does not dominate all others), the focus of this paper will be on hierarchical clustering. For a broad overview of the multivariate statistics used in cluster analysis the reader is referred to Timm [11]. For a broad overview of both unsupervised and supervised learning methods from both the statistics and machine learning literature, the reader is referred to Hastie *et al.* [12]. For a broad overview of the application of these methods to biological data the reader is referred to Legendre and Legendre [13]. Each of these references cover hierarchical and other clustering methods in more mathematical detail than presented here and show their application to data for illustration.

2003 © Ashley Publications Ltd ISSN 1462-2416

Raw data

Gene expression data measured by gene chips (microarrays) are preprocessed using image analysis techniques to extract expression values from images and scaling algorithms to make expression values comparable across chips. These preprocessing steps are generally done with a microarray-platform vendor's software or through software developed by researchers interested in improving the estimates of the expression data [14,15]. While these steps can have a significant impact on the quality of the data and are an area of active research, our review will start with the assumption that these preprocessing steps have already been performed and the estimates of the expression level are as good as can be obtained.

Expression data are typically analyzed in matrix form with each row representing a gene and each column representing a chip or sample. For a study with 20 samples run on Affymetrix GeneChips™, the dimensions of the data matrix would be (approximately) 12,000 rows (one for each gene) by 20 columns. Newer chips have even more genes on them. Often there will be one additional column giving the gene label for identification. However, this column is excluded from analysis and only the chip columns containing expression values are used.

We represent the data matrix by the symbol X and denote the data as follows:

$$X = \begin{array}{c|cccc} \textit{Gene} & \textit{Chip}1 & \textit{Chip}2 & \cdots & \textit{Chip}20 \\ \hline 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,20} \\ 2 & x_{2,1} & x_{2,2} & \cdots & x_{2,20} \\ 3 & x_{3,1} & x_{3,2} & \cdots & x_{3,20} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 12{,}000 & x_{12000,1} & x_{12000,2} & \cdots & x_{12000,20} \end{array}$$

The matrix entries correspond to the expression value of a gene (row) and chip (column). For example, $x_{1,1}$ is the expression value of gene 1 in sample 1, $x_{3,20}$ is the expression value of gene 3 in sample 20, etc. In general the notation $x_{i,j}$ corresponds to the expression level of gene *i* in sample *j*. While this notation may seem clumsy at first, it is important to understand the 'structure' of the data to learn how the analysis is done and how the results should be interpreted.

Most software programs use the data matrix *X* in this form to cluster genes. There is no reason that clustering cannot also be done directly on columns. However, to simplify discussion in this paper and to be consistent with many statistical packages, to cluster samples we will use the *transposition* of *X*. This is obtained by flipping the matrix across the diagonal so that the columns become the rows and the rows become the columns. This changes the dimensions from the original 12,000 rows by 20 columns to a matrix of dimension 20 rows by 12,000 columns. In this format the samples are the rows and the genes are the columns. We denote the transposition of *X* by X^T:

$$X^T = \begin{array}{c|ccccc} \textit{Chip} & \textit{Gene}\ 1 & \textit{Gene}\ 2 & \textit{Gene}\ 3 & \cdots & \textit{Gene}\ 12000 \\ \hline 1 & y_{1,1} & y_{1,2} & y_{1,3} & \cdots & y_{1,12000} \\ 2 & y_{2,1} & y_{2,2} & y_{2,3} & \cdots & y_{2,12000} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 20 & y_{20,1} & y_{20,2} & y_{20,3} & \cdots & y_{20,12000} \end{array}$$

The matrix entries of X^T, coded as $y_{i,j}$, correspond to the expression value in a chip (row) for a given gene (column). For example, $y_{3,20}$ is the expression value in sample 3 for gene 20. In general, the notation $y_{i,j}$ corresponds to the expression level in sample *i* for gene *j*, so $y_{i,j} = x_{j,i}$.

Filtering

The first step in analyzing microarray data is to filter out genes that are not expressed or do not show variation across sample types. In our experience, this usually reduces the data set by 3000–5000 genes. The Affymetrix GeneChips contains a variable for each gene that declares whether the gene was expressed, not expressed or indeterminate. We always remove from the analyses the rows corresponding to genes that were not expressed on any of the chips. Other strategies for gene filtering include filtering at a threshold of the variance of the gene across chips or if two or more tissue types are represented in the experiments, filtering at a threshold of a test statistic. For example, if gene chips are used to analyze tumor and normal tissues, the two groups can be compared using t-statistics calculated for each gene. An arbitrary threshold based on a value for the t-statistic or to filter out a certain percentage of the genes can be used.

These methods of filtering genes are arbitrary (except, perhaps, for filtering out genes based on the expression/no expression call by Affymetrix software). However, if used conservatively to filter out only the least differentially expressed genes, the analyst should be protected from eliminating any important genes.

Standardized data

Although clustering methods can be applied to the raw data, it is often more useful to precede the analysis by standardizing the expression values. Standardization in statistics is a commonly used tool to transform data into a format needed for meaningful statistical analysis [16]. For example, *variance stabilization* is needed to fit a regression model to data where the variance for some values of the outcome Y may be large, say for those values of Y corresponding to large values of the predictor variable X, while the variance of Y is small for those values corresponding to small values of X. Another use of standardization is to *normalize* the data so a simple statistical test (e.g., t-test) can be used. Transformations specifically designed to allow standard statistical tests to be applied to microarray data are currently being proposed [17,18].

Transformation of microarray data for cluster analysis has a different purpose than transformations used to meet assumptions of statistical tests as described above. Cluster analysis depends on a distance measure (discussed in the next section). Since distance measures are sensitive to differences in the absolute values of the expression values (scale), microarray data for clustering often needs to be transformed to adjust for different scales. To illustrate this, consider three hypothetical genes A, B and C, whose expression levels have been measured in four normal tissue samples and four diseased tissue samples. The results of these measurements are displayed in **Figure 1A**. Genes A and B are tightly coregulated and differentially expressed across tissue types (i.e., higher in diseased tissue relative to normal tissue) but gene A is expressed at a much higher level than gene B. Gene C is not differentially expressed across tissue types but happens to have average expression levels similar to that of gene A. We typically want to find clusters that place genes A and B together because they appear to be coregulated (low in normal tissue, high in diseased tissue) but would not cluster them with gene C which is constant across all tissue samples. Clustering using the raw expression profiles would separate genes A and B and cluster genes A and C. **Figure 1B** shows the expression profiles for the same three genes after normalization (see below) across samples. In this transformed data, we see the expression values for genes A and B are closely aligned. In contrast, the values for gene C fluctuate randomly. This transformation results in the representations for genes A and B being near each other and thus increases the likelihood that they are clustered together.

Normalizing a gene across samples is accomplished by subtracting from each expression level the mean of the expression levels for that gene and then dividing by the standard deviation of that gene. Our matrix notation in the last section can now be used to clarify how the normalization is done. The data matrix X consists of rows of genes we want to normalize. Consider the first gene at row 1 consisting of the expression levels $x_{1,1}$, $x_{1,2}$, ... , $x_{1,20}$ corresponding to gene 1 in sample 1, gene 1 in sample 2 etc. We calculate the mean of gene 1 by

$$\bar{x}_1 = \frac{x_{1,1} + x_{1,2} + \ldots + x_{1,20}}{20}$$

and the standard deviation of gene 1 by

$$s_1 = \sqrt{\frac{(x_{1,1} - \bar{x}_1)^2 + (x_{1,2} - \bar{x}_1)^2 + \ldots + (x_{1,20} - \bar{x}_1)^2}{20 - 1}}$$

The notation '+ ... +' indicates to add all the terms between $x_{1,2}$ and $x_{1,20}$. With these terms, the normalized expression values are:

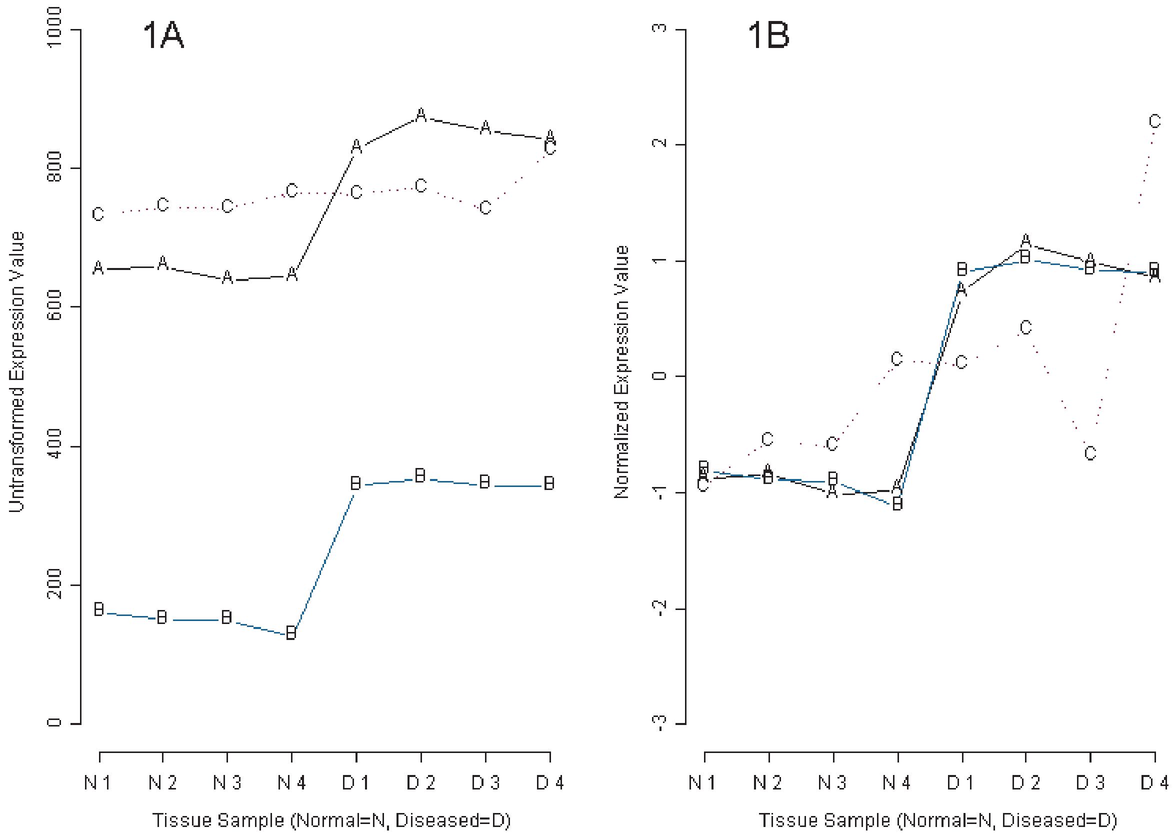
$$\frac{x_{1,1} - \bar{x}_1}{s_1}, \frac{x_{1,2} - \bar{x}_1}{s_1}, \ldots, \frac{x_{1,20} - \bar{x}_1}{s_1}$$

Most statistical programs can apply this normalization to each row of the matrix X. While conceptually this normalization might also be applied to each row of the transposed matrix X^T, we have found this not useful for uncovering structure in cluster analysis. We recommend that normalization be applied to genes across samples only.

Distance measures

Many cluster analysis methods, including hierarchical clustering, use distances measured between rows of the data matrices X or X^T. Measuring distances can be thought of as placing a ruler between two points and recording how far apart they are. To make this idea more clear before we present formulae for calculating distances, consider a simple example where the data matrix X consists of gene expression values (rows) measured on only two samples (columns).

Figure 1. Gene expression profiles before and after normalization.



With just two samples (chips) we can plot each gene as a point on a two-dimensional scatter plot where the X-axis corresponds to the first chip and the Y-axis corresponds to the second chip. Consider three genes (A, B and C) whose expression levels are measured in the two samples:

Gene	*Chip1*	*Chip2*
A	-2.0	1.0
B	-1.5	-0.5
C	1.0	0.25

These three genes can be plotted on a standard scatter plot as shown in **Figure 2**.

In addition to the gene labels A, B and C, this graph shows the calculated distances between each of these genes where the distances are calculated using the Euclidean distance formula. Specifically the distance between genes A and B is calculated by the formula

$$d(A,B) = \sqrt{(-2.0-(-1.5))^2+(1.0-(-0.5))^2} = 1.58$$

between genes A and C by

$$d(A,C) = \sqrt{(-2.0-1.0)^2+(1.0-0.25)^2} = 3.09$$

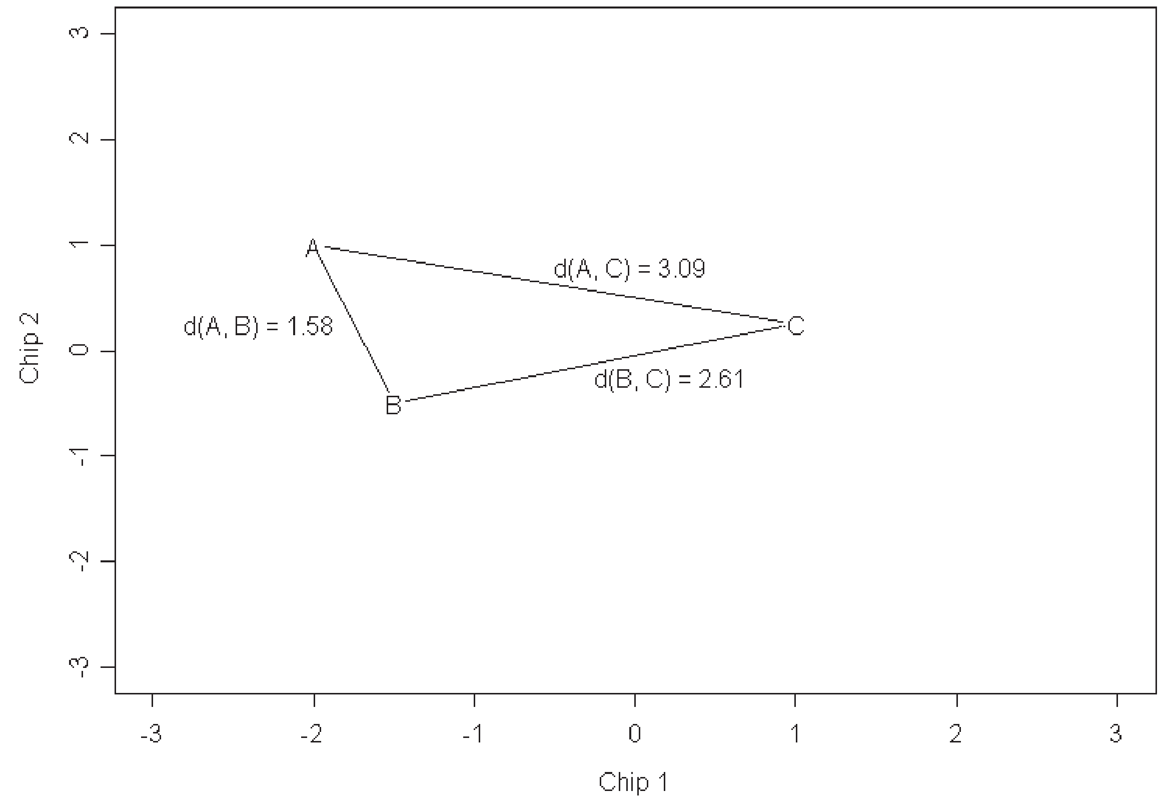
and between genes B and C by

$$d(B,C) = \sqrt{(-1.5-1.0)^2+(-0.5-0.25)^2} = 2.61$$

For convenience we record distances in a distance matrix

$$D = \begin{array}{cc} & \begin{array}{ccc} A & B & C \end{array} \\ \begin{array}{c} A \\ B \\ C \end{array} & \begin{bmatrix} 0.00 & 1.58 & 3.09 \\ 1.58 & 0.00 & 2.61 \\ 3.09 & 2.61 & 0.00 \end{bmatrix} \end{array}$$

Figure 2. Scatterplot of three genes from two samples.



The entries correspond to the distances between the genes denoted on the row and column (e.g., d(A,B) = 1.58). Note that the distances on the diagonal are all 0, the distances are all non-negative and the matrix is symmetric (e.g., d(A,B) = d(B,A)).

We now want to generalize the idea of the Euclidean distance matrix for any microarray data. Specifically, recall the data matrix is

	Gene	*Chip*1	*Chip*2	…	*Chip*20
	1	$x_{1,1}$	$x_{1,2}$	…	$x_{1,20}$
X =	2	$x_{2,1}$	$x_{2,2}$	…	$x_{2,20}$
	3	$x_{3,1}$	$x_{3,2}$	…	$x_{3,20}$
	⋮	⋮	⋮	⋮	⋮
	12,000	$x_{12000,1}$	$x_{12000,2}$	…	$x_{12000,20}$

where each row represents a gene and each column represents a chip. We will assume the data has been normalized. Distances are calculated between each pair of rows in X: d(1,2) is the distance from row 1 to 2, d(1,3) is the distance from row 1 to 3, d(1,12000) the distance from 1 to 12000, d(2,3) the distance from row 2 to 3, etc. The Euclidean distance for any two rows, say rows *i* and *j*, is calculated using the expression values for all the chips in those two rows as follows:

$$d(i,j) = \sqrt{(x_{i,1} - x_{j,1})^2 + (x_{i,2} - x_{j,2})^2 + \ldots + (x_{i,20} - x_{j,20})^2}$$

Notice that the subscripts on the$x$'s change for the column (chip) number. In words, this calculation is performed by subtracting the expression level of gene *j* from gene *i* and squaring it in each chip from 1 to 20, adding these values together and then taking the square root of the sum. No matter how many chips there are, for 12,000 genes this produces a 12,000 by 12,000 distance matrix containing 144,000,000 numbers indicating the computational complexity involved in cluster analysis.

There are many other distance measures that could be used (i.e., Manhattan distance) though we believe the Euclidean distance is generally appropriate for normalized microarray data.

Hierarchical clustering

Several different algorithms will produce a hierarchical clustering from a pair-wise distance matrix. Each of these algorithms follows the same general strategy. Suppose we are clustering genes. The algorithms begin with each gene by itself in a separate cluster. These clusters correspond to the tips of the *clustering tree* (dendrogram). The algorithms search the distance matrix for the pair of genes that have the smallest distance between them and merge these two genes into a cluster. The distance matrix is recalculated to now include the distance between genes not clustered and the new cluster formed by the two genes. For simplicity, we will assume that only two genes are merged at each step, though more could be merged at any step.

Many algorithms follow this series of steps to produce hierarchical clustering of data. Variations between the algorithms can lead to different dendrograms and hence different clusters. We will consider an *average linkage* algorithm called *unweighted centroid clustering* for illustration and then compare it to other hierarchical clustering algorithms. It should be noted that different authors define average clustering in different ways. For example, others refer to the definition of average clustering used by Hastie *et al.* [12] as *unweighted arithmetic average clustering*. Readers interested in more technical descriptions of four average clustering algorithms should refer to Legendre and Legendre [13].

To illustrate our average linkage algorithm, recall the distance matrix calculated above for three genes A, B and C. Suppose we have added a fourth gene D and recalculated the distance matrix *D*,

$$D = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0.00 & 1.58 & 3.09 & 4.74 \\ 1.58 & 0.00 & 2.61 & 5.00 \\ 3.09 & 2.61 & 0.00 & 2.70 \\ 4.74 & 5.00 & 2.70 & 0.00 \end{bmatrix} \end{array}$$

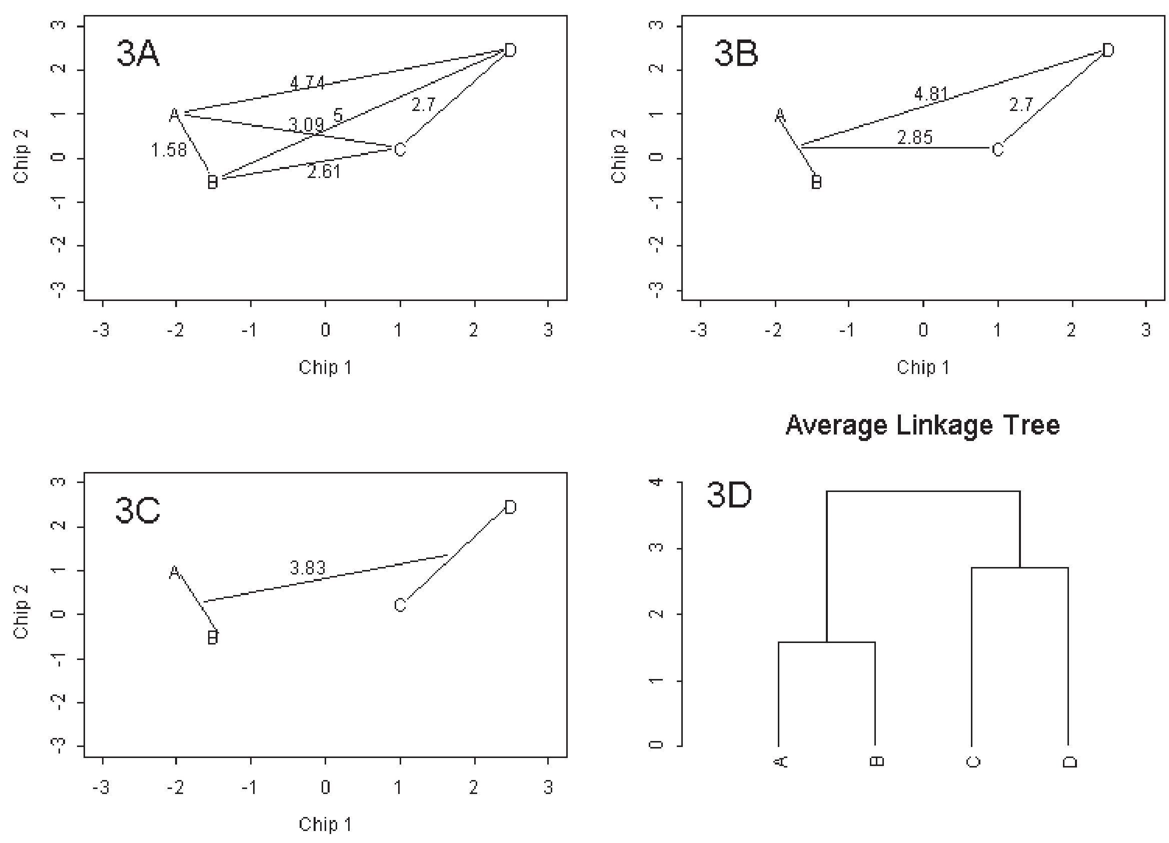
Figure 3A–D shows the steps of the average linkage clustering and the dendrogram obtained. In Figure 3A the four genes are plotted and the distance between each pair is indicated on the line connecting them. Initially, the algorithm finds the pair of genes closest to each other and merges them into a cluster. For this example, the first step merges genes A and B whose distance is 1.58. The distances are updated as follows: Replace the two genes A and B by the midpoint (AB) between them and recalculate the distance of gene C to this midpoint (d(AB, C) = 2.85) and gene D to this midpoint (d(AB, D) = 4.81). Note that d(C, D) = 2.7 is unchanged. The updated distances are shown in Figure 3B. The algorithm then repeats by finding the genes (or clusters) that have the smallest distance between them. In this iteration, genes C and D are clustered and replaced by their midpoint. The distance to all other gene clusters (such as AB) from this midpoint is calculated and the algorithm is repeated. Figure 3C shows the final distance d(AB, CD) = 3.83. Gene clusters AB and CD are merged in the last step of the algorithm.

Figure 3D summarizes the results of applying the average linkage algorithm to this data in a single graph known as a *dendrogram*. Initially, the four genes A, B, C and D are represented as single clusters along the bottom of the plot. Genes A and B are merged first at the level 1.58, followed by genes C and D being merged at the level 2.7, followed by AB and CD being merged at level 3.83. The dendrogram was fit and displayed using S-Plus (Seattle, Washington) software.

We presented the average linkage algorithm in Figure 3 in two ways to emphasize the relationship between a dendrogram and the pair-wise spatial distances of the genes. In this example, there were two chips so that the genes could be plotted in a scatter plot but the concept is the same for experiments with more than two chips. In general, the genes are points in a space whose number of dimensions equals the number of chips in the experiment. Regardless of the dimension of the problem, the Euclidean distance between genes or gene clusters can be calculated and each iteration of the algorithm merges the genes or gene clusters that have the smallest distance.

The final clustering of the genes is determined by where the dendrogram is cut. For example, cutting the dendrogram at level 3 (on the y-axis) results in the two clusters AB and CD, while cutting the dendrogram at the level 2 produces the three clusters AB, C and D. This dendrogram

Figure 3. Iterations of a hierarchical clustering and the resulting dendrogram.



can produce four distinct cluster results: ABCD when the dendrogram is not split; AB and CD when split into two groups; AB, C and D when split into three groups; and A, B, C and D when split into four groups.

Average linkage is one of many hierarchical clustering algorithms that operate by iteratively merging the genes or gene clusters with the smallest distance between them followed by an updating of the distance matrix. Many of these differ only in how the distance matrix is updated. In average linkage, as shown above, when two genes are clustered, the distances of the other genes and gene clusters to this new cluster is based on the midpoint of the new cluster. In contrast, single linkage calculates the distances between each gene in the new cluster to each of the genes in another cluster and takes the smallest distance. Complete linkage uses the largest distance of all these distances as the distance between the clusters. For example, in Figure 3A the first merging clustered genes A and B and the distance of this new cluster to gene D was d(AB, D) = 4.81. For single linkage, the distance would be d(AB, D) = 4.74 and for complete linkage the distance would be d(AB, D) = 5.

In practice, we have found the average linkage algorithm generally works well with standardized microarray data and single linkage generally performs poorly.

Difficulties and pitfalls of cluster analysis

Unlike standard statistical methods, such as the t-test and analysis-of-variance, hierarchical clustering does not have a probabilistic foundation. Because of this, hierarchical clustering has no statistical test to guide the decision of where to cut the dendrogram. While it is possible to compute a formal test statistic, such as an F-test statistic, the assumptions of the statistical test are not met. Thus, the p-value listed in a statistics table would not represent the probability of the test-statistic value arising under the null hypothesis. In other words, the p-value has no meaning and is not a measure of the statistical significance of the clusters being different.

In the absence of formal statistical tests, external criteria are typically used to choose the number of clusters. One such criterion is that if splitting a tree at a particular point produces clusters of genes or samples that are nearly homogeneous with regard to an important property, the split would be deemed appropriate. For example, if splitting a tree at a particular height resulted in mostly tumor samples in one cluster and mostly normal samples in the other, the split would likely be considered interesting. Such a split is considered to be evidence that some of the genes used to generate the tree may be involved with the biology of the tumor and hence the genes warrant further scrutiny. The obvious problem with this approach is the subjective nature of deciding which external criteria to use.

A second difficulty with cluster analysis is that the algorithms are guaranteed to produce clusters from any data and there is currently no generally accepted way to test a null hypothesis of no clusters (e.g., data are distributed uniformly). For this reason, caution is required in interpreting the results of a cluster analysis method. The results always need to be examined to see if it is plausible that they are indeed natural clusters and not just artifacts of the algorithm.

In spite of these two problems, cluster analysis is a powerful tool for data reduction. One must remember that data reduction is the chief purpose of a cluster analysis. Since microarrays present the researcher with thousands of gene expression values, the data must be reduced before a human can tell an explanatory story about the relationship between genes and the phenotypes. Putative relationships between clusters of genes and phenotypes need to be recognized as nothing more than hypotheses generated by clustering methods. The clustering process has not statistically validated the relationships and they must be formally validated through additional experiments.

Methods are currently being developed to address the weaknesses of cluster analysis. We believe that the current interest in applying cluster analysis to genomics will generate enough research effort to successfully meet this challenge. For example, one method to determine the number of clusters without resorting to external criteria is to use the number that optimizes the Gap statistic, a statistic comparing within-cluster dispersion (spread of data points) to dispersion under the null hypothesis [12,19]. Another approach uses a perturbation method (sensitivity analysis). It introduces a small amount of random noise to the expression data, reclusters the data and then compares the results to the original clustering [20]. Our lab has begun investigating a formal statistical approach based on graph theory and a probability distribution on graphical objects [21] and another approach based on Mantel statistics which are briefly discussed below [22]. In spite of these efforts, the problem of selecting the correct number of clusters remains open after fifty years of study.

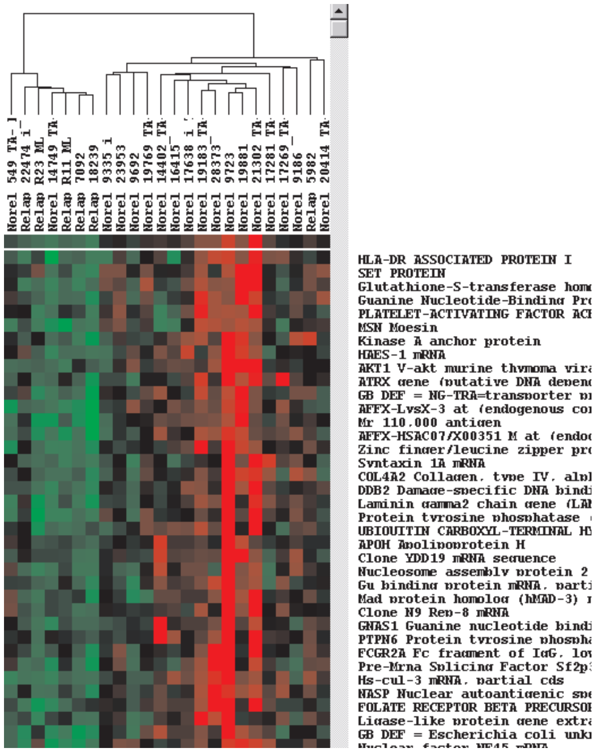
In summary, in spite of the danger of misusing or misinterpreting the results of cluster analysis, as long as one keeps in mind that cluster analysis is only appropriate for data reduction and hypothesis generation, the pitfalls can be reduced or avoided.

Heat maps

Hierarchical clustering is used to produce what have been called ‘heat maps’ in papers reporting on microarray data analyses. The heat map presents a grid of colored points where each color represents a gene expression value in the sample. Figure 4 is an example taken from a recent paper using expression levels for cancer classification [23]. The grid coordinates correspond to the sample by gene combinations. In this case, the columns (samples) are tumors, some from patients who have relapsed and some from patients who have not relapsed. The rows represent 348 genes found to distinguish the patients according to their relapse status. In the heat map colors at a particular point (i.e., row by column coordinate) are assigned to represent the level of expression for that gene (row) in the sample (column) with red corresponding to high expression, green corresponding to low expression and black corresponding to an intermediate level of expression.

The ordering of the rows and columns was determined using hierarchical clustering and the associated dendrogram for the samples shown at the top of the figure. In this example, six relapsing patients were clustered together to the left and the non-relapsed patients clustered to the right. The heat map gives an overall view that the 348 genes have low expression in the relapse patients as indicated by the green color in the left-hand columns under the relapse patients. In

Figure 4. An example of a heat map. Reproduced from [23].



contrast, the non-relapse patients have higher expression levels for these genes as indicated by the red and black colors in their columns.

Other methods

Our group has developed and used several other methods based on clustering. We will provide a brief description of three of these methods and provide references for detailed descriptions. We have found these methods useful with data from various studies including psoriasis [24], oncology [25] and pharmacogenetics [26].

K-means clustering and self-organizing maps

Hierarchical clustering assumes a hierarchical structure in the data wherein all the genes start separately in their own cluster at the bottom of the dendrogram and iteratively merge into larger clusters as one goes up the tree. K-means and self-organizing maps (SOMs) cluster genes without assuming a hierarchical structure. Instead K-means starts with k genes sampled randomly from the data. Each of these genes is used as the starting center of one of the k clusters. Distances are calculated from each gene in the data to each of these *k* centers. Genes are then assigned to the closest center. Each center is then replaced by the average of the genes assigned to it. The algorithm repeats by recalculating the distance from each gene in the data to these new centers and reassigning genes to the closest center. This repeats until no genes are reassigned. For example, consider two gene clusters involved in the regulation of non-overlapping metabolic pathways. It may not be reasonable ever to merge these two groups of genes, as would be required by the hierarchical structure of a dendrogram. SOMs are similar to k-means but with slightly different iteration and update steps. The details of this technical distinction are not pertinent to this overview article [11,12,24,27,28].

Mantel statistic

Mantel statistics provide a method to assess the correlation between two distance measures on the same data [29,30]. We applied this method to microarray data measured on brain tumors to statistically correlate the expression patterns with clinical covariates [22,25]. Since we wanted the clustering done on the samples, we used the transposed data, X^T. Two distance matrices, one based on the microarray expression values and the other using the clinical information, represent the pair-wise distances between the same samples in terms of two different factors. If samples that are far apart in the distance based on the microarray data are also far apart in the distance based on the clinical data and samples that are close in the microarray data are close in their clinical data, the pair-wise distances are positively correlated. This can indicate that clinical differences are related to gene expression differences. The Mantel statistic provides a formal statistical framework for quantifying these relationships and permutation tests can provide accurate p-values for testing significance.

Consensus methods

This is a mathematical framework to combine the results of multiple cluster analyses into a final cluster result [22,31,32]. Conceptually, if two genes are very similar, they will be clustered together by most hierarchical clustering algorithms, distance measures and reasonable stopping rules. A consensus method will put those two genes into the same cluster in the final analysis. Similarly, if two genes seldom appear together, the consensus method will not put them in the same cluster. We are currently investigating how consensus methods might automate the choice of where to cut a dendrogram using bootstrapping to generate multiple cluster results and have applied it with encouraging results.

Conclusion

We have focused on presenting an overview of hierarchical clustering of microarray data as a tutorial, emphasizing the relationship between a dendrogram and spatial representations of genes. We believe this relationship provides an intuitive understanding of how to analyze microarray data and can make it easier to interpret the results of a cluster analysis in a biological framework. The fact that the ‘heat maps’ found in the majority of microarray publications are based on hierarchical clustering indicates that an understanding of this general method is valuable to those who are just beginning to read the microarray literature and even to those who are using supervised methods.

We have avoided a discussion of implementation since most major statistical packages provide methods for cluster analysis and visualization and the choice of the package will depend on the level of computational and statistical expertise available in the particular lab. In our case, as professional statisticians, mathematicians and computer scientists, we use two advanced statistics packages: SAS (Cary, North Carolina) and S-Plus (Seattle, Washington). These packages contain many standard-clustering approaches used with microarray data and can be programmed to perform novel methods such as Mantel statistics or consensus methods. However, these packages require a high level of programming skill and most research groups will want to look for a statistical package that is easier to use.

We presented a brief description of hierarchical and some non-hierarchical clustering methods based on distance measures that our lab has used with success. There are many non-distance-based methods available, including principal component analysis, gene shaving, Bayesian methods and mixed-models approaches. We cannot present all of them in this review article and have not yet had the need to use them in our own work. Our view is that microarray data should be analyzed using distance-based methods instead of parametric model approaches because the assumptions for parametric models are currently hard to justify. We realize, of course, that as researchers gain more experience analyzing microarray data using parametric models and develop a solid probabilistic foundation for these approaches, some of these non-clustering methods may later become the *de facto* analytical framework of choice.

We emphasize the complexity and technical difficulty of performing cluster analysis. We do not see these methods as trivial to implement and would encourage researchers to begin building long-term collaborations with statisticians. However, cluster analysis and microarray data present novel problems with which many statisticians will have had no experience. Therefore, the collaboration will require a significant investment to introduce the statistician to these fields. One attraction to this field for a statistician is the opportunity for novel statistical methods research. This should be emphasized and supported as these collaborations develop.

Finally, we mention the Classification Society of North America [101] as an excellent cluster analysis resource. This organization supports the development of clustering and classification methods and the application of these methods to many academic fields. The society also publishes the prestigious *Journal of Classification* [102], which publishes fundamental papers on cluster analysis. In addition, the society maintains the *class-l* list server, an excellent forum for raising questions about cluster analysis. It is accessible through their web page [102].

Outlook

Current methods of analyzing microarray data based on hierarchical clustering use 'off-the-shelf' algorithms developed over the last 50 years. Little work to date has been done to modify these methods for microarray data taking into account biological knowledge such as expected clusterings based on genes involved in metabolic pathways or genes sharing regulatory sites. Incorporating this type of knowledge will require a significant investment of time and support for statistical methodologists, but the added value of this research investment to pharmacogenomic studies should be huge.

Highlights

- Supervised learning methods can predict membership in predetermined groups and identify genes important for classification. They require training data with known group assignment for each data point.
- Cluster analyses attempt to detect natural groups in data and identify genes important for classification. No *a priori* group assignments are required.
- Cluster analysis consists of a collection of distance-based unsupervised learning methods including hierarchical clustering, k-means clustering, self-organizing maps, principal components analysis, and Mantel statistics.
- Gene expression microarray data is typically filtered and normalized before using cluster analysis.
- Cluster analysis results should be used for data reduction and hypothesis generation.
- The heat map, a useful data visualization and summary tool, is a product of hierarchical clustering.
- Drawbacks of cluster analysis include lack of statistical tests for determining the number of clusters or the strength of cluster membership.

Bibliography

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Fisher R: The use of multiple measurements in taxonomic problems. *Ann. Eugenics* 7, 179-188 (1936).
2. Brown MP, Grundy WN, Lin D *et al.*: Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA* 97(1), 262-267 (2000).
3. Cristianini N, Shawe-Taylor J: *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods.* Cambridge University Press, Cambridge (2000).
4. Bishop C: *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford (1996).
5. Khan J, Wei J, Ringner M *et al.*: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* 7, 673-679 (2001).
6. Ringner M, Peterson C, Khan J *et al.*: Analyzing array data using supervised methods. *Pharmacogenomics* 3(3), 403-415 (2002).
7. Everitt B, Rabe-Hesketh S: *The Analysis of Proximity Data*. John Wiley, New York City (1997).
8. Eisen M, Spellman P, Brown PO *et al.*: Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* 95, 14863-14868 (1998).

• **This paper presents the first application of hierarchical clustering to microarray data and is a landmark publication.**

9. Hartigan J, Wong M: A k-means clustering algorithm. *Applied Statistics* 28, 100-108 (1979).
10. Kohonen T: The self-organizing map. *Proc. IEEE* 78, 1464-1479 (1990).
11. Timm N: *Applied Multivariate Analysis.* Springer, New York City (2002).
12. Hastie T, Tibshirani R *et al.*: *The Elements of Statistical Learning*. Springer, New York City (2001).
13. Legendre P, Legendre L: *Numerical Ecology.* Elsevier, New York City (1998).
14. Li C, Hung Wong W: Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol.* 2(8), RESEARCH0032 (2001).

• **This paper takes a more traditional statistical modeling approach to improve the estimate of the genes expression and identify genes differentially expressed across sample groups. The model-based approach reduces the variability of low expression estimates, and provides a natural method of calculating expression values. The standard errors attached to expression values can be used to assess the reliability of downstream analysis.**

15. Wolfinger RD, Gibson G, Wolfinger ED *et al.*: Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.* 8(6), 625-637 (2001).
16. Steele, Torrie: *Principles and Procedures of Statistics*: *a Biometrical Approach*. McGraw-Hill, New York City (1980).
17. Yang Y, Dudoit S *et al.*: Normalization for cDNA microarray data. Dept of Statistics Technical Report, University of California Berkeley (2001).
18. Durbin B, Hardin J *et al.*: A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics* 18, S105-S110 (2002).
19. Hastie T, Tibshirani R, Eisen MB *et al.*: 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.* 1(2), 1-21 (2000).

• **This paper presents a novel approach for clustering microarray data which combines unsupervised and supervised learning methods. The method presented here also allows genes to belong to more than one cluster.**

20. Bittner M, Meltzer P, Chen Y *et al.*: Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 406(6795), 536-540 (2000).
21. Shannon W, Banks D: Combining classification trees using maximum likelihood estimation. *Stat. Med.* 18(6), 727-740 (1999).
22. Shannon WD, Watson MA, Perry A, Rich K: Mantel statistics to correlate gene expression levels from microarrays with clinical covariates. *Genet. Epidemiol.* 23(1), 87-96 (2002).

• **This paper is one of the first to focus on the problem of statistically correlating expression profiles with clinical covariates. The method presented here uses distance-based calculations like in hierarchical clustering and thus avoids the problem of distribution assumptions.**

23. Slonim DK: Transcriptional profiling in cancer: the path to clinical

pharmacogenomics. *Pharmacogenomics* 2(2), 123-136 (2001).

24. Bowcock AM, Shannon W, Du F *et al.*: Insights into psoriasis and other inflammatory diseases from large-scale gene expression studies. *Hum. Mol. Genet.* 10(17), 1793-1805 (2001).
25. Watson MA, Perry A, Budhjara V *et al.*: Gene expression profiling with oligonucleotide microarrays distinguishes World Health Organization grade of oligodendrogliomas. *Cancer Res.* 61(5), 1825-1829 (2001).
26. Zhang W, Shannon W *et al.*: Protein expression profiling to define CPT-11 therapy strategies in common cancers. (2002) (In Preparation).
27. Tamayo P, Slonim D, Mesirow J *et al.*: Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* 96(6), 2907-2912 (1999).
28. Tavazoie S, Hughes JD, Campbell MJ *et al.*: Systematic determination of genetic network architecture. *Nat. Genet.* 22(3), 281-285 (1999).
29. Mantel N: The detection of disease clustering and a generalized regression approach. *Cancer Res.* 27(2), 209-220 (1967).
30. Smouse P, Long J *et al.*: Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Systematic Zoology* 35(4), 627-632 (1986).

Websites

101. www.cs-na.org
Classification Society of North America.
102. http://link.springer-ny.com/link/service/journals/00357/
Journal of Classification.