Chapter 20
EXPLORATORY GENOMIC DATA ANALYSIS

Larry Smith

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health,Bethesda, MD 20894

Chapter Overview

In this chapter, an introductory description of the exploration of genomic data is given. Rather than attempt an exhaustive overview of the types of genomic data and methods of analysis, the chapter focuses on one type of data, gene expression profiling by microarray technology, and one method of analysis, cluster analysis for discovering and sorting mixed populations. This type of data and method of analysis is very common in bioinformatics. It illustrates recurring problems and solutions. And a major portion of bioinformatics dealing with exploratory genomic data analysis can be viewed as a refinement and extension of this basic analysis.

Keywords

gene expression profiling; microarray technology; cluster analysis; mixed populations

1. INTRODUCTION

Exploratory genomic data analysis draws on mathematical, statistical, and computational methods to discover meaningful genetic relationships from large-scale measurements of genes. It is a continuously growing area that is constantly being seeded with new approaches and interpretations. Most of this new material is easily accessible given a familiarity with basic genetics and multivariate statistics. It is my hope that this introduction to prototypical methods and case studies will be enough to get started.

Genomics researchers are interested in the genetics of specific, controllable, cellular phenomena. Microarray assay technology makes it possible for them to measure genomic variables, such as relative mRNA expression level, associated with thousands of genes in multiple controlled situations. Normally, the usefulness of experimental data ends with the experiment, when the results are evaluated in relation to a stated hypothesis. But with its vast scope, microarray data provides a unique view of the complex and largely unknown genetic organization of cells, a view into a world beyond all preconceived hypotheses. It is not surprising that large-scale microarray assays are undertaken for the purpose of scrutinizing the results and searching for potentially new genetic knowledge. This is where exploratory data analysis enters.

Data mining, also called knowledge discovery and descriptive statistics, is a process of reformulating a large quantity of unintelligible data with a view to suggest hypothetical relationships. With luck, some of these hypotheses, when validated experimentally, will lead to real discoveries. But experimentation is usually expensive and time consuming, and it is not wise to pursue arbitrarily generated hypotheses. Therefore, it is not merely enough to generate hypotheses; they must be evaluated and selected according to their significance and plausibility. Significance means that a generated hypothesis is not likely to be based on experimental errors. To establish significance requires a statistical test based on a model of the experimental errors, also called a null hypothesis. Plausibility means that a generated hypothesis is reasonable in light of current scientific research and knowledge. To establish plausibility requires a theoretical justification and a connection of the new hypothesis to previous research. This is where literature mining proves to be useful.

In this chapter we give an introduction to data mining of genomic data with an intuitive explanation of the fundamental technique of cluster analysis of mixed population data. Three published studies are described, each with a different analysis of the same gene expression profile data set.

The material found here on data mining and genomic data is by no means complete. Rather, the goal of this chapter is to introduce the field to new students, who might afterwards follow their own interests and inclinations into any of the numerous directions of current research involving genomic data analysis.

2. OVERVIEW

Gene expression profiling is any laboratory procedure that measures the expression of genes in differing controlled cellular conditions. The controlled conditions might be selected phenotypes (e.g. cancer types), stages of development or differentiation, or responses to external stimuli (e.g. drug dosages). Gene expression is the quantitative level of messenger RNA found in cells. Expression levels of genes are highly regulated in a cell and depend on many factors. The up- or down-regulation of a particular gene in a particular situation is an important clue to its function. Genes whose expressions are correlated may actually be co-regulated, and so they are potentially participating in coordinated functions. Therefore, knowledge about one gene might be applicable to the other. And genes whose regulation correlates with an externally controlled condition are potentially involved somehow in that condition. This is a modest motivation for applying data mining to expression profile data.

2.1 Gene Expression Data

Gene expression profile data is collected in large quantities using microarray technology. The basic science behind this technology is not difficult to understand. In brief, a cDNA (complementary DNA) microarray is a glass slide containing thousands of robotically miniaturized hybridization experiments. To conduct a microarray assay a sample is prepared and exposed to the slide under controlled hybridization conditions. A sample is actually a solution containing DNA that has been collected from prepared cellular samples, concentrated and purified, and infused with radioactive or fluorescent nucleotide markers. After the hybridization, each spot on the slide emits a photon signal whose strength is determined by the amount (or amounts) of sample DNA binding at that spot. The photon signals from the slide are measured with an optical scanner, and the pixelized image is loaded into a computer for data analysis.

Microarray manufacturers typically provide slides and scanning equipment as well as computer software to analyze image data. Image analysis software must assess the quality of the slide globally, identify anomalies such as dust specs, identify the precise areas of each spot and its image quality, and ultimately reduce each spot to a single numeric quantity.

There are also additional data manipulations that can be performed after image analysis, such as quality filtering, background subtraction, and channel equalization.

In addition to data quality depending on the image, quality also depends on the manufacture of the microarray slide itself. Manufacturing flaws are possible, for example. And because so many different kinds of cDNA are used to make the slide, it is not rare for several spots on a slide to contain cDNA that is not properly identified. In any exploratory data analysis, it is important to keep control of data quality and, somehow, to prevent poor quality data from influencing the results of the analysis.

It is possible to give a general and sufficient description of expression profiling data without referring to the details of the technology. Several assays are conducted using the same microarray layout. Each assay is associated with controlled experimental conditions, such as cell phenotype, stage of development or differentiation, exposure to drug or radiation, *etc*. Each microarray spot has an associated DNA sequence, which is related to one or more genes in the genome map of the species. The data can therefore be pictured as a matrix of numeric data whose columns correspond to experimental conditions and whose rows correspond to DNA sequences. Each value in the matrix is a relative measure of the expression of that row's DNA in that column's experimental condition. Alternatively, each row of the data matrix is a vector of numbers representing the expression of a particular DNA sequence in all of the experimental conditions. Simultaneously, each column is a vector of numbers representing the gene expression in a particular condition.

The concepts of data mining are independent of the categories of the underlying data. Any matrix of data, or any list of values associated to members of some population, can be used with data mining. For this reason, in most of this section I will refer to data points, individuals, or members of a population instead of genes.

2.2 Mixed Populations

One of the fundamental goals of data mining is to determine if the data arose from a mixture of several distinct populations. If it did, then each population might have a distinctive distribution of data points. If data were two or three dimensional, we could visually inspect a graph of the data for different populations. Figure 20-1 shows some examples of how two populations might appear in two dimensions. In 1a, the two populations appear to have identical distributions. In this case, there is no amount of mathematical or statistical manipulation able to find a meaningful difference. In 1b and 1c, the populations are located in non-overlapping regions. In these examples we say there are mean differences. In 1b the distribution of the two populations has the same shape around their respective means. But in 1c, the shape of the distributions are different, we would say that these have variance differences. Not shown is an example where the subpopulations have the same mean but different variance (which is a case, similar to 1a, which is difficult to detect). Of course, distributions in reality need not be shaped into spheres or ellipses, as shown in 1d they might assume arbitrary shapes while still falling into separate regions.

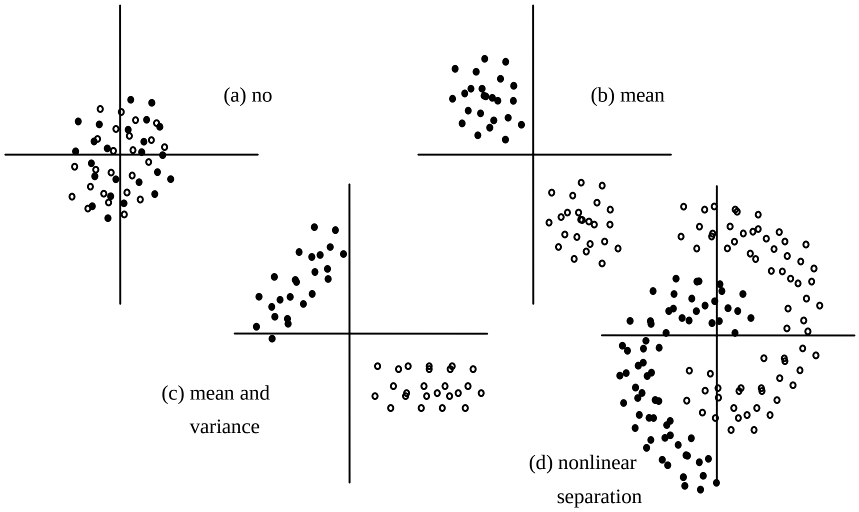


Figure 20-1. Different ways that a mixture of populations might be reflected in 2 dimensional data

In higher dimensions, visual inspection is impossible. There are well known techniques able to reduce high dimensional data to 2 or 3 dimensions for visual inspection. These include principal components analysis (PCA), multidimensional scaling, and factor analysis. These methods should preserve the appearance of subpopulations with significant mean differences. The K-means algorithm is able to detect mixed populations with less significant mean differences, or with significant variance differences. And nonlinear differences, like that shown in 1d, can sometimes be detected using (connected) clustering. The number of dimensions not only confounds our ability to visualize, but a phenomenon called the "curse of dimensionality" makes differences in the data less obvious. It is difficult to give a brief explanation of why this happens, but it has to do with the fact that small random variations in each of many dimensions tend to make points appear to be more uniformly spaced – a mean difference “signal” is drowned out by the high dimensional “noise.”

If a scientific theory predicts subpopulations, there might be some expectation as to what they are, their number, and even how the distributions of the data points may differ. This would be an enormous help in analyzing the data. But without prior knowledge, unanalyzed data is a territory in complete darkness, and the various analysis techniques may seem to provide little more than a dim light. It is impossible to predict in advance whether it will be easy, difficult, or impossible to identify subpopulations. As easy as it might be to define the objective and the approach, in practice, data mining is a challenge to be respected.

And even if subpopulations are found, it is not enough to report the fact. Each subpopulation must be identified and described. The simplest approach, when it is logically justified, is to find exemplars in each subpopulation, or members (*e.g.* genes) that are well known, then to generalize what is known about them to the entire subpopulation. If there are no exemplars, or if the exemplars do not generalize, the distribution of data might be correlated in distinctive ways with the experimental conditions (this is the implication of significant mean differences), and this might suggest the function of members of the population. If no clear identification of any subpopulation can be made, the validity of the analysis must be questioned.

2.3 Methods for Mixed Populations

Encyclopedic knowledge may be available for some genes, but for most of them at the present time, little is known beyond some portions of their nucleotide sequence. Even less is known, relatively, of the ways in which genes are related. Microarray experiments generate data for thousands of genes, some known and some not. A fundamental goal in analyzing microarray data is to find new relationships between genes.

Scientists begin to understand diverse populations of things by categorizing individuals based on their similarities and differences. Data mining is able to uncover categories in a large population provided that they are reflected in the data.

K-means analysis is a statistical modeling technique (and several variations) that is able to find subpopulations of data that are distributed differently. To apply it, however, requires an estimate of the number, K, of subpopulations. It works well when data conforms to the model, as the number K can be found by trial and error. But in less ideal situations, the results are difficult to interpret.

Principal components analysis (PCA) and factor analysis, on the other hand, apply simpler linear models to the data. An ideal data set consists of points (in some high dimension) that are compactly distributed in an ellipsoidal shape. For this type of data, the first principal component is the measurement along the longest diameter of the ellipse, the second component along the next longest diameter, perpendicular to the first, *etc*. Factor analysis is different from PCA in that it allows more flexibility in the components (after the first one), but this requires subjective choices. The usual way of using PCA, or factor analysis, is to represent each data point with the first 2 or 3 components and graph them for visual examination. Data sets reflecting distinct subpopulations are not ideal, but one can imagine nevertheless that the data is distributed within an ellipsoid with gaps. Whatever the cause, PCA, or factor analysis, is frequently effective with mixed population and non-ideal data, and subpopulations can be seen in the component graph in lower dimensions. Sometimes this dimension reduction is carried out to make subsequent analysis (for example, cluster analysis to be discussed next) simpler or more robust.

Some analysis methods, however, are not based on the distribution of data but rather on the "distances" between all pairs of points. Multidimensional scaling (MDS), for instance, finds a configuration of points in lower dimension whose pairwise distances are the "closest" possible to the distances in the original higher dimensional space (variations of MDA depend on how closeness is measured). The graph that results from MDA is similar to the graph obtained from PCA, and usually there is no discernable advantage to either one, though the PCA algorithm is simpler (in fact, there is a variant of MDS that produces graphs identical to PCA).

Cluster analysis is very popular because it is both simple and intuitive. It produces a tree structure suggesting an intuitive organization of the data. Think of the cluster tree as a collection of data points, forming the leaves, and many branches that join the leaves together. Each branch has a number of leaves and may also have many branches below it. The "height" of a branch is a composite measure of the distance (sometimes called the diameter) between the leaves of that branch. The root branch is the unique branch containing all of the data. The graphical representation of a cluster tree is called a dendrogram, an example is shown in Figure 20-2. The vertical scale represents the height, and the horizontal axis lists all of the data points in the order dictated by the tree. A horizontal segment is used to represent each branch, located at the height of the branch. Vertical segments connect a branch with the data or branches that are immediately below it. The data and sub-branches of a branch can be put in any order but, by convention, they are usually sorted left to right by increasing height of the sub-branches.

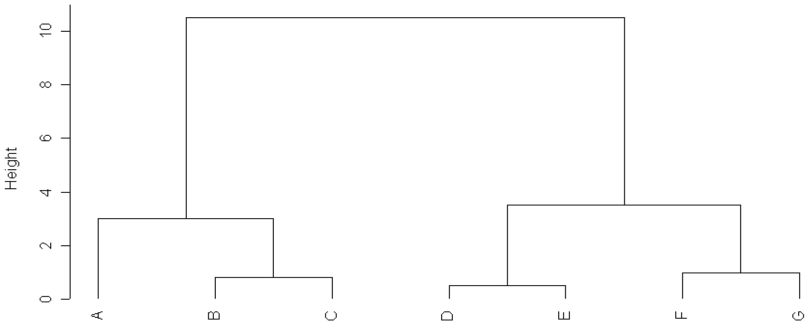


Figure 20-2. Example of a cluster tree dendrogram

There are several different clustering algorithms, and most of them are fast and efficient and use only the distances between points to be given. If the data is a mixture of well-separated subpopulations (such as the illustrations in Figure 20-1b and 20-1c), then virtually any cluster tree will have two or more distinct branches joined by a root branch at a significantly greater height. For separated but intertwined populations (such as the illustration in Figure 20-1d) a kind of clustering called *connected clustering* is sometimes able to show them, but only if it is generally the case that for each point there is at least one point in the same subpopulation that is closer to that point than any point in the other subpopulations.

One problem with cluster trees is that they must be interpreted. Cluster analysis always produces a result, even when there are no clearly delineated subpopulations. A branch that is contained in a branch at a significantly greater height is a candidate for a subpopulation. This conclusion is not automatic because it depends on interpreting the height scale, which can be arbitrary. Even when there are no obvious subpopulations, it may happen that the population members with known similarities cluster together, making it possible to postulate an “organizing principal” for the cluster tree. More often, however, cluster trees will appear to have no obvious “organization.” The decision to search for meaning in a random cluster tree or to give up and try a different approach can be a vexing one, compounded by psychological, social, and even political pressures. But there are a great many variations to try; the key is in the calculation of distances, which I will discuss in the next section.

Finally, if a cluster analysis reveals subpopulations, they cannot merely be reported, they must be explained. At least some of the subpopulations must be identified by some common characteristics, as mentioned in Section 2.2. I will have more to say about this in Section 2.5.

2.4 Distance

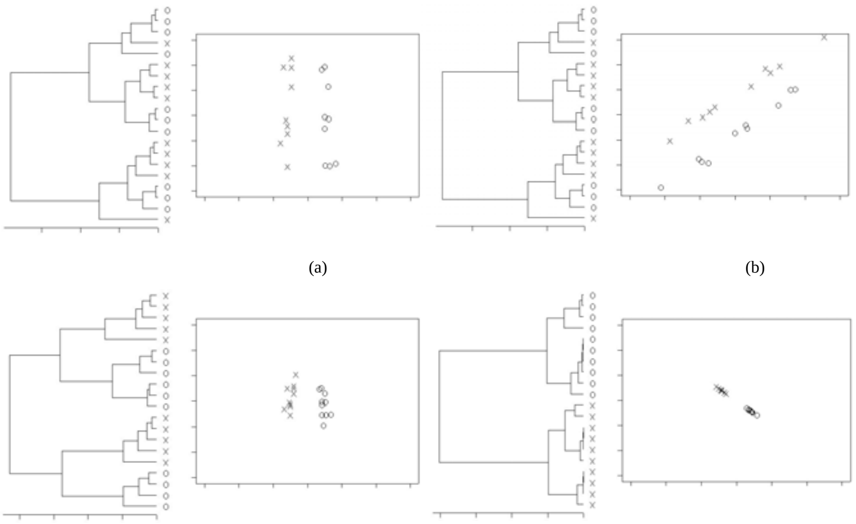
The result of cluster analysis, regardless of the algorithm used, depends primarily on the calculated distances between individuals. After defining and collecting the raw data, deciding how to calculate distances is the next crucial step towards successful analysis, and there are many options. In this section, I will describe and illustrate some basic distance calculation methods.

The distance between two vectors of data is given by a formula that combines the differences in each component into a single value. There are many choices to make when calculating the distance between points. Each distance formula is sensitive to particular kinds of differences, and clustering results reflect those differences.

The obvious distance formula is Euclidean distance (the sum of squared differences of each component). With the Euclidean distance, each coordinate is considered equally important, and this might not be desired. But it is important to be aware of some undesirable features that are both subtle and common.

Since the components of a data vector come from separate processes (*e.g.* different experimental conditions), the distribution of component values can be very different from one another. Consider, for example, what would happen if one experiment has a high variance compared with the others. This is illustrated in Figure 20-3a for hypothetical data with two distinct subpopulations. The distances between two points in the one population are sometimes greater than the distances between points in two different populations. The problem here is that the variance of the second component is very large compared with the variance of the first. It can be corrected by "normalizing" each component. That is, each data value is modified by dividing by the standard deviation (square root of variance) computed over all values of that component. An alternative solution in this example is to note that the populations are distinguished by their first component, so the distance could be computed by disregarding the second component (this would be an example of subjective factor analysis).

Another problem is illustrated in Figure 20-3b. The data of 3a has been rotated by 45 degrees, and it is still the case that the distances between populations are sometimes less than the distances within populations. In fact, the cluster trees are identical. One approach would be to rotate the data so that it looks like 3a and then normalize (this too would be an example of subjective factor analysis). But in 3b the problem can be understood more



(a)

(b)

Figure 20-3. Two examples of mixed populations with unequal variances before (above) and after (below) modification of the data. In (a) each component is divided by the standard deviation, and in (b) each point is mean-centered.

simply. Each point in the sample has a random constant added to all components. It is as if each point has randomly reset the “zero-point” for its measurements. The solution is to “mean center” the data. That is, each data value is modified by subtracting the average of all of the components of that data point. This solution has the same effect as the approach to 3a of calculating the distances using only the first coordinate.

One way to account for unequal importance of different components is to multiply each component by a constant weight reflecting its relative importance. The decision to do this cannot be motivated by the data, but must be based on knowledgeable assessment of the different experimental conditions. For example, if there are 10 experimental conditions, but 5 of them are duplicates of the same condition, it would be sensible to weight each of the 5 duplicate components with a weight of 1/5.

Other distance measures use different formulas but are similar in spirit to Euclidean distance. The Manhattan distance, for example, is the sum of the absolute values (instead of square) of the differences in each component. Correlation distance is one minus the Pearson correlation coefficient of the paired components of two data points. Interestingly, correlation distance is equivalent to Euclidean distance if the data is first mean-centered and normalized.

Finally, a problem that frequently arises is missing data. In any experiment, it is to be expected that some data is usable and some is not. In the case of microarray data, this depends on quality assessments coming from the manufacture of the array, the certainty of identification of the cDNA in each spot, and the quality of the scanned image. There is no obvious way of computing distances between points that have some values missing. If the components are disregarded, the calculated distance will be grossly distorted. A common approach is to fill in the missing values with their averages or to use some estimate of the difference in each component with missing values.

2.5 Hypothesis Selection

Let us summarize the data mining process we have discussed up to now. First, raw data is obtained as vectors of data for many members of some population (*i.e.* genes). Pairwise distances are computed between members and this is used to cluster the data. The distance calculation is refined and the data is re-clustered a few times until it looks as if some interesting subpopulations have been found. The next thing to do is to shore up the analysis with evidence that the discovery is not a mirage and that it is worth the time to consider for further experimentation.

Gene expression profile data reduced to pairwise distances and visualized by clustering suggest relationships or subpopulations defined by similarities between genes. Clustering analysis does not, however, say what those similarities are based on. For a potential relationship to be worthy of further experimentation, it must satisfy three conditions: 1) it must be statistically significant, 2) it must be strong, and 3) it must be theoretically plausible.

Statistical significance. The idea of statistical significance of a conclusion is to show that it could not have arisen by accident. To do this requires a probabilistic model of how the conclusion might have been made by accident, which is then used to compute the probability that it could have happened that way. If that probability, or p-value, is sufficiently small (usually less than 0.05), the conclusion is said to be statistically significant. A detailed model for the “null hypothesis” is usually impractical to construct. And simplified models are likely to be doubted by the scientific community anyway. An alternative to this is to apply the idea of bootstrapping.

Suppose we observe a distance between two population members that appears to be very small compared to other distances, and we want to conclude from this that the members are somehow related, perhaps even duplicates. The statistical significance of the fact is the probability of a certain “random” model producing such a small distance. A reasonable alternative to constructing a model, called bootstrapping, is to assume a model in which the data values arise randomly. To estimate significance using this model, the data is shuffled a large number of times, each time calculating the distance between the populations in question and noting the percentage of times that the distance is less than or equal to the distance in question. This percentage is the p-value. Data is shuffled by randomly permuting the values in the matrix. The model can be refined by allowing only certain types of permutations. For example, depending on the experiment, it might be logical to assume that data within a particular row or column arise randomly, and to satisfy this model the random permutations used in shuffling the data could be chosen to satisfy such constraints.

A similar idea could be applied to a cluster analysis. For example, suppose that a number of genes are found in a branch of the cluster tree, and the difference in heights between that branch and the branch immediately containing it is large enough to suggest that they are in a distinct subpopulation. By shuffling the data many times, and clustering each time, the percentage of times that this height difference (or larger) is observed for the same set of genes can be used as a p-value.

Strength of finding. Independent of p-value, that is to say, even if the p-value is very small, a discovery is probably not worth discovering if it is not very strong. This happens frequently, for example, with correlation coefficients. When many independent measurements of co-occurring phenomena are taken, a very small non-zero correlation coefficient can be very significant, *i.e.* the p-value can be very small. It is correct to conclude from a small p-value that something "real" is causing the two phenomena to be related. The problem is, because the effect is so weak, it will be difficult to find the common cause. Or once it is found it will be difficult to prove that it is indeed the cause. Or even if the cause is discovered and proven, it may be so slight that it is of little value to science.

The same is true with data mining of genomic data. If the height difference of a subpopulation is determined to be statistically significant, yet it is apparently a very small difference, the distance formula is so sensitively refined that it is able to reveal similarities between members that are probably very remote. Granted, if the subpopulation is known to exist, then it can be automatically concluded that the analysis has detected it. But this is hardly a breakthrough. Data mining is looking for relationships that were not previously known.

Plausibility. Now suppose that a subpopulation has been found, it is statistically significant and strongly delineated. It is not interesting if the subpopulation is already known in the scientific literature, except as a demonstration of the approach, so suppose also that it is new. To be taken seriously, it is necessary to give some plausible explanation of what it is that characterizes this subpopulation. It is the job of the scientist to know the field and to know how to make plausible arguments. But in data mining, the field covers a very large collection of genes, and the literature bearing on all of these genes is too massive for any human scientist to grasp. In this situation, scientific research means library research: searching for and assessing research articles, exploiting electronic and print resources to expand the search, and then taking the time to carefully read and understand the articles that seem to be relevant.

Search tools are being developed and refined today to help scientists find articles focused to highly technical needs. The need is so great that some research teams include computational linguists who can develop new search tools tailored to their specific research. The ideal search tool should not only be able to retrieve articles, it should be able to summarize and rank the relevance of each article to the stated goal. Future NLP applications should be able to take a given relationship, say between two or more genes, and determine if there is a common thread in the literature that might connect them. The retrieved articles should be summarized and grouped according to the underlying explanation that is being suggested. What is more, they should be able to take as input many different possible relationships and search all of them, returning the relationships with supporting literature. All of this has been done manually in the past and is done increasingly with the help of computers.

It is important to understand that a strong relationship that is statistically significant might have an uninteresting cause. But the cause may still be worth identifying, even if it is not scientifically important. An uninteresting cause of relationships in the data may reveal a deficiency in the design of the experiment or ways in which the data could be modified to correct for the uninteresting effects.

3. CASE STUDIES

In this section, I will discuss two papers that introduced the gene expression profiles of 60 previously studied cancer cell lines. The two groups that collaborated to produce this data published separate analyses. The paper “Systematic variation in gene expression patterns in human cancer cell lines” (Ross, et al., 2000) describes the cluster analysis of the genes and cells. And “A gene expression database for the molecular pharmacology of cancer” (Scherf, et al., 2000) correlates this data with anti-cancer drug response data for the same 60 cell lines that has been collected at the NCI since the 1960s. I will also describe a third study that was done independently, some time later, for a data mining competition based on this data.

NCI60. The 60 cell lines of NCI60 have been extensively studied for many years. They represent a variety of human tissues, organs, and cancer types. The gene expression profile study measures the expression of 9,703 genes for each cell line which is a significant addition to the body of scientific data, making it possible to explore genetic correlations with virtually any of the previously studied phenomena. The data for each gene and cell line is the ratio of the expression of the gene in the cell line to the average expression of the gene in a pooled mixture of the cell lines. For example, a value of 2 for a gene and cell line indicates that the gene is expressed in the cell line twice as much as it is on average in the pool. The NCI60 data set is publicly available and might still be used to discover new relationships in the genetics of cancer.

Ross. The study of Ross, et al. explored the gene expression profiles with regard to the cell types. For this analysis, each cell line is represented by the vector of data formed from the expression values of all of the genes. Distances between cell lines were calculated using the correlation distance (this is equivalent to using Euclidean distance after mean-centering and normalizing the data). They observed that cells with the same histologic origin clustered together, indicating that they had similar gene expression patterns. Also a larger organization of cell lines was seen in the cluster tree, grouping epithelial cells and stromal cells. To explain why this occurred, they turned to the cluster tree of the genes. There they noted several large branches of genes including genes expressed characteristically in specific cell types. For example, one branch of genes included many genes related to melanin synthesis and were expressed mostly in the melanoma cell lines. These cell-specific genes probably drive the organization of the cluster tree of the cell lines along histological type. Similarly, the presence of cluster branches dominated by genes characteristically expressed in epithelial cells or stromal cells could explain the large-scale organization of the cell line cluster tree.

They explored the cluster of genes further and found several branches containing genes with common functions. For example, there was a cluster of cell-cycle-related genes, RNA processing genes, genes regulated by interferons, and genes with possible roles in drug metabolism. The expression pattern of genes in a branch containing ribosomal genes correlated with the cell doubling time of the cell lines, supporting the interpretation that the branch contained genes supporting cell proliferation rate. It is interesting to consider what the result would be if the cell lines had been clustered using only the genes from one of these functional branches. A picture would emerge, not necessarily organized around histological type, reflecting the status of the particular function in the cell lines.

They also used cluster analysis to test the feasibility of using this same microarray assay to identify the tissue of origin of actual tumor biopsies. They obtained biopsies from two breast cancer patients and performed microarray assays on samples of cell extracts. Cluster analysis was used with the new data together with some of the breast cancer and leukemia cell lines and it was found that the biopsy samples appeared in a branch independent of those cell lines. They hypothesized that the biopsy samples contained cells of several different types, unlike the pure origin of each of the 60 cell lines. To support the hypothesis, they found that genes characteristic of different cell lines, including leukocytes, epithelial cells, and stromal cells, were expressed in the biopsy samples. They also confirmed the hypothesis by staining the biopsy specimens and demonstrating different cell types.

Scherf. The study of Scherf, et al. used the same NCI60 microarray data together with drug response data collected at the Developmental Therapeutics Program of the National Cancer Institute. This drug response data is the IC50, or the dosage at which cell growth is inhibited by 50% in culture, measured for thousands of drugs that have been considered candidates for anticancer treatment. The cluster tree of cell lines obtained using the gene expression profiles was compared with the cluster tree obtained using the drug response data. The differences reiterate the conclusion that the predominant gene expression pattern is tied to the histological origin of the cell. But the cluster tree of cells based on drug response has a different pattern, reflecting different modes of drug susceptibility that is not necessarily related to tissue of origin. For example, the ovarian and colon cancer cell lines were scattered into disparate branches in the cluster tree based on drug response. The two sets of pairwise differences between cell lines, obtained from the gene expression data and from drug response data, were compared directly with the Pearson correlation. The correlation of the distances was 0.21, which is statistically significant. This number is an indicator of the positive but small degree to which gene expression could theoretically be used to predict drug response in a cell line.

They also looked for possible relationships between specific genes and specific drugs. They did this by computing the correlation between the 60 expression values for 1,376 genes and the 60 drug response values for 118 drugs with established modes of action. This resulted in 162,368 correlations, which had to be searched for significant, strong, and meaningful correlations. Three examples were chosen for illustration having strong, statistically significant correlations (all negative). Supporting arguments were made for the plausibility of these associations and references were given for the required background. It was not feasible to consider all of the significant and strong correlations carefully, because each correlation required a literature search, assessment of relevance, and a creative leap to a possible underlying mechanism. A key search tool was developed, called MedMiner, which performed MEDLINE literature searches for user selected gene-drug pairs and presented summarizing sentences for evaluation. The lead scientist estimated that MedMiner decreased the time involved in literature search by a factor of 10. But many correlations, including some potential breakthroughs and other unexpectedly low correlations, could not be pursued because of the infeasibility of finding literature to suggest or support an underlying commonality.

Coombes. The NCI60 data set is publicly available, and has been the object of further analysis since the twin studies of Ross and Scherf. The data analysis competition, “Critical Assessment of Microarray Data Analysis” or CAMDA (http://www.camda.duke.edu), is an annual open competition where new approaches to analyzing a given data set are judged. In 2001 the NCI60 data set was used as the data set for the competition, and the winner that year was K.R. Coombes, et al. for “Biology-Driven Clustering of Microarray Data: Applications to the NCI60 Data Set.” By performing cluster analysis of cell lines, similar to Ross, they showed that some cancer types were easily distinguished by gene expression, *e.g.* colon or leukemia, but not breast and lung. Then by restricting the expression data to various subsets of genes, *i.e.* by calculating the distance between cells based on specific subsets of genes, they were able to show that some meaningful subsets were able to distinguish cancer types while others were not. For example, genes on chromosome 2 were, but genes on chromosome 16 were not. Also, genes associated with certain functions, such as signal transduction and cell proliferation were, but genes associated with apoptosis and energy metabolism were not able to distinguish cancer types. They reached these conclusions by examining the cluster tree in each case to see if the established cell type branches appeared.

4. CONCLUSIONS

This chapter described approaches to exploratory genomic data analysis, stressing cluster analysis. It is not necessary to have a deep understanding of genomic data in order to carry the analysis out, but a deep understanding of some facet of molecular biology is required to complete the analysis. It is likely that a large-scale microarray study will be carried out in one or more labs and with several collaborators. In this circumstance, specialists can carry out data mining, producing a list of possible conclusions (chosen for statistical significance and strength) that can be handed over to others trained in molecular biology to make the final selections and provide supporting arguments. But even with this division of labor, a team of molecular biologists does not have the expertise to judge a very large list of possible hypotheses. Sophisticated computer tools for retrieving relevant scientific literature are indispensable. Future advances in the ability of computers to make human-like judgments may relieve the human bottleneck.

REFERENCES

Efron, B. and Tibshirani, R.J. (1993*) An Introduction to the Bootstrap*. Chapman & Hall, New York.

Manly, B. (1986). *Multivariate Statistical Methods, A Primer*. Chapman & Hall, London.

Quackenbush, J. (2001). “Computational Analysis of Microarray Data,” *Nature*, 2, 418-427.

Ross, et al. (2000). “Systematic Variation in Gene Expression Patterns in Human Cancer Cell Lines,” *Nature Genetics*, 24, 227-235.

Scherf, et al. (2000). “A Gene Expression Database for the Molecular Pharmacology of Cancer,.” *Nature Genetics*, 24, 236-244.

Tanabe, L., et al. (1999) “MedMiner: An Internet Text-mining Tool for Biomedical Information, with Application to Gene Expression Profiling,” *Biotechniques*, 27(6), 1210-4, 1216-7.

Voorhees, E.M. and L.P. Buckland, eds. (2003). *Text Retrieval Conference (TREC 2003).* National Institute of Standards and Technology, Gaithersburg.

Weinstein, et al. (2002). “The Bioinformatics of Microarray Gene Expression Profiling,” *Cytometry*, 47, 46-49.

Weinstein, et al. (1997). “An Information-Intensive Approach to the Molecular Pharmacology of Cancer,” *Science*, 275, 343-349.

SUGGESTED READINGS

There are many books on bioinformatics that deal specifically with the analysis of microarray data, and more appear every year. There are too many to name or single out. Most provide a good introduction to the subject with programming guidelines and exercises.

For a description of microarray technology, see the January, 1999 issue of Nature Genetics (volume 21) “The Chipping Forecast.”

A comprehensive overview of data mining techniques for microarray data can be found in F. Valfar “Pattern Recognition Techniques in Microarray Data Analysis.” Ann. N. Y. Acad. Sci. 980, 41-64, which can be obtained in full text through PUBMED.

ONLINE RESOURCES

The R statistical package, available for free download under the GNU public license. http://www.gnu.org/software/r/R.html.

Data and analysis tools for the NCI60 data set (and others) can be found at http://discover.nci.nih.gov/.

The NCBI web site, http://www.ncbi.nlm.nih.gov, is the entry point to the PUBMED database and sequence databases. You can also find tools for data analysis of sequence data (under

"Tools") and a bioinformatics primer (in "About NCBI"). In addition there is a database of microarray assay data, called GEO, at http://www.ncbi.nlm.nih.gov/geo/.

The annual competition Critical Assessment of Microarray Data Analysis (CAMDA) can be found at http://www.camda.duke.edu.

QUESTIONS FOR DISCUSSION

The following questions are open ended.

1. When exploring data, testing all of your ideas with simulated data is an invaluable habit. The statistical package R is ideal for this. Obtain the R package and write a script that generates random microarray data and perform a cluster analysis. Here is a simplified R script to do this:

```
data <- rnorm(100, 0, 1)
data.dist <- dist(data)
data.clust <- hclust(data.dist)
plclust(data.clust)
```

   Generate random data for mixed populations and cluster them.

2. If you are familiar with any field of compiled data, scientific or other, try to find an application of cluster analysis. For example if you were interested in baseball, what data would be useful to explore different kinds of batters and pitchers? Can you draw an analogy between your approach and exploratory genomic data analysis?

3. Microarray studies generate data for thousands of genes. How many data points are you able to cluster using the R package before running out of memory? One approach to cluster analysis of huge data sets is to select a subset, either randomly or by careful selection. What kind of conclusions could be drawn from cluster analysis of random subsets of genes? How could you improve on this?

4. In the Scherf, et al. study, both gene expression data and drug response data was available for all 60 cell lines. An approach called "co-clustering" treats the drug response data as if it were gene expression data and analyzes it together with the genes. What new kind of result is possible with co-clustering? Most often, the results are disappointing. What goes wrong?

5. Suppose some of the data in a microarray study is missing or flagged as unreliable. How could you calculate distance when there are missing values? Predict the circumstances where your method would lead to

misleading results. When should genes be discarded for having too many missing values?

6. Suppose a data mining system were able to generate a list of possible hypotheses and provide each one with a p-value, a measure of strength, and an indication of the amount of accompanying literature. Could these three factors be combined in a way to rank the hypotheses from “best to “worst”?

7. Formulate some features of an ideal literature search tool to aid in data mining.