

Introduction

Over the past few decades, major advances in the field of molecular biology, coupled with advances in genomic technologies, have led to an explosive growth in the biological information generated by the scientific community. This deluge of genomic information has, in turn, led to an absolute requirement for computerized databases to store, organize and index the data, and for specialized tools to view and analyze the data.

Bioinformatics

The National Council for Biological Information (NCBI) web site defines Bioinformatics as the field of science in which biology, computer science, and information technology merge to form a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned.

Massive amounts of such data have been generated via the global-scaled Human Genome Project. This type of high-dimensional and noisy data presents special challenges in data analysis, statistical modeling and interpretation of results. Of course, there are various ways in which individual entries in sequence and structure databases can be compiled to reveal patterns and trends in biology. For example, sequence families or neighborhoods can be defined and annotated based on the similarity of each sequence to other members of the family. Common sequence features in sequence families can be identified in multiple alignments. These motifs may provide clues to the biochemical function of members of the family. Clustering of sequences into trees that reflect the degree of similarity between each sequence and all of the others in the family reveals evolutionary relationships. Finally, identification of homologs to each gene in well-characterized metabolic pathways provides information about the prevalence of that pathway in other organisms.

Statistical Learning

Vast amounts of data are being generated in many fields, and the statistician's job is to make sense of it all: to extract important patterns and trends, and understand "what the data says". This can be termed as *learning from data*.

The challenges in learning from data have led to a revolution in the statistical sciences. Since computation plays such a key role, it is not surprising that researchers in other fields such as computer science and engineering have done much of this new development.

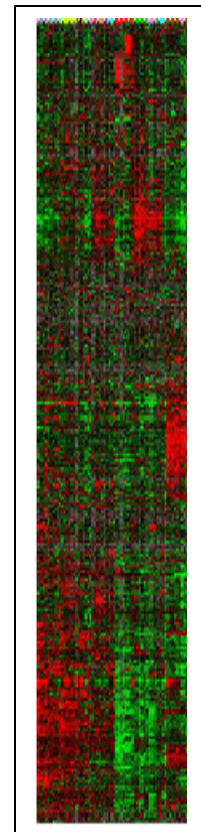
The learning problems can be roughly categorized as either *supervised* or *unsupervised*. In supervised learning, the goal is to predict the value of an outcome measure based on a number of input measures; in unsupervised learning, there is no outcome measure, and the goal is to describe the associations and patterns among a set of input measures.

DNA Expression Microarrays

(Hastie et al. (2001))

DNA microarrays measure the expression of a gene in a cell by measuring the amount of mRNA (messenger ribonucleic acid) present for that gene. Microarrays are considered a breakthrough technology in biology, facilitating the quantitative study of thousands of genes simultaneously from a single sample of cells.

A gene expression dataset collects together the expression values from a series of DNA microarray experiments, with each column representing an experiment. There are therefore several thousand rows representing individual



genes, and tens of columns representing samples, for example (see figure), which contain 6830 genes (rows) and 64 samples (columns). The figure displays the data set as a heat map, ranging from green [G] (negative) to red [R] (positive). The samples are 64 cancer tumors from different patients.

The challenge here is to understand how the genes how the genes and samples are organized. Typical questions include the following:

- a) which samples are most similar to each other, in terms of their expression profiles across genes?
- b) which genes are most similar to each other, in terms of their expression profiles across samples?
- c) do certain genes show very high (or low) expression for certain cancer samples?

We could view this task as a regression problem (*supervised learning*), with two categorical predictor variables - genes and samples, with the response variable being the level of expression. However, it is probably more useful to view it as *unsupervised learning problem*. For example, for question (a) above, we think of the samples as points in 6830 - dimensional space, which we want to *cluster* together in some way.

Graphical Presentation

The primary purpose of the image analysis step is to extract numerical foreground and background intensities for the red and green channels for each spot on the microarray. Once the numerical data is extracted, it is a good idea to use routinely a variety of exploratory graphical displays to examine the results of any microarray experiment. Graphical displays can help assess the success of the experiment, can guide the choice of analysis tools and can highlight specific problems. The most common graphical display of data from a microarray slide is a scatterplot of the two channel intensities, $\log R$ versus $\log G$. Boxplots can be useful for comparing M-values between various groups. For example, side-by-side boxplots of the normalized M-values for a series of six replicates arrays. A spatial plot of the background or M-values can often

reveal spatial trends or artifacts of various kinds.

Systat offers more scientific and technical graphing options than any other desktop statistics package.

Classification

(Smyth et al. (2003))

Two very important uses for microarray data are to generate gene expression profiles which can (i) discriminate between different known cell types or conditions, e.g. between tumor and normal tissue or between tumors of different types or (ii) identify different and previously unknown cell types or conditions, e.g. new subclasses of an existing class of tumors. The same problems arise when it is genes that are being classified: one might wish to assign an unknown cDNA sequence to one of a set of known gene classes, or one might wish to partition a set of genes into new functional classes on the basis of their expression patterns across a number of samples.

These dual tasks have been described as class prediction and class discovery in the influential paper by Golub et al. (1999). In the machine learning literature they are known as supervised and unsupervised learning, the learning in question being of the combinations of measurements - here gene expression values - which assign units to classes. In the statistical literature they are known as discrimination and clustering. The distinction is important. Clustering or unsupervised methods are likely to be appropriate if classes do not exist in advance. If the classes are preexisting, then discriminant analysis or supervised learning methods are more appropriate and more efficient than clustering methods.

Cluster methods tend to be over-used in microarray data analysis relative to discrimination methods. A common practice for example is to suppress existing class assignments, use an unsupervised learning technique to define new classes and assign the units to these classes, and then see how well the existing class assignments are reflected in the new classes. A more direct and efficient approach would be to use a supervised method to discriminate the classes in conjunction with a method such

as cross validation to evaluate the repeatability of the results on new data. The efficiency of direct discrimination over clustering becomes increasingly important, as the prediction problem becomes more challenging.

Discrimination methods include linear discriminant analysis in various forms, nearest neighbor classifiers, classification trees, aggregating classifiers, neural networks and support vector machines. The first three methods are simple to apply once the genes have been filtered. The other methods are more sophisticated and require considerable skill in their application.

Other techniques, such as Principal Components Analysis and Correspondence Analysis (Fellenberg et al., 2001) have also been used to analyze gene expression data.

Systat's Cluster provides three procedures for clustering: Hierarchical Clustering, K-means, and Additive Trees. The Hierarchical Clustering procedure comprises hierarchical linkage methods. The K-means Clustering procedure splits a set of objects into a selected number of groups by maximizing between-cluster variation and minimizing within-cluster variation. The Additive Trees Clustering procedure produces a Sattath-Tversky additive tree clustering.

Discriminant Analysis performs linear and quadratic discriminant analysis, providing linear or quadratic functions of the variables that "best" separate cases into two or more predefined groups. Systat can select the variables in the linear function in a forward or backward stepwise manner, either interactively by the user or automatically.

The Systat's TREES module computes classification and regression trees. Classification trees include those models in which the dependent variable (the predicted variable) is categorical. Regression trees include those in which it is continuous. Within these types of trees, the TREES module can use categorical or continuous predictors, depending on whether a CATEGORY statement includes some or all of the predictors.

Factor analysis provides principal components analysis and common factor

analysis (maximum likelihood and iterated principal axis). Systat has options to rotate, sort, plot, and save factor loadings. With the principal components method, you can also save the scores and coefficients. Orthogonal methods of rotation include varimax, equamax, quartimax, and orthomax. A direct oblimin method is also available for oblique rotation.

Systat's Correspondence analysis allows you to examine the relationship between categorical variables graphically. It computes simple and multiple correspondence analysis for two-way and multiway tables of categorical variables, respectively. Tables are decomposed into row and column coordinates, which are displayed in a graph. Categories that are similar to each other appear close to each other in the graphs.

References (in order of appearance)

Trevor Hastie et al. (2001). *The Elements of Statistical Learning*, Springer-Verlag, New York.

Gordon K. Smyth et al. (2003). 'Statistical Issues in cDNA Microarray Data Analysis', In: *Functional Genomics: Methods and Protocols*, M. J. Brownstein and A. B. Khodursky (eds.), *Methods in Molecular Biology* Volume 224, Humana Press, Totowa, NJ, 2003, pages 111-136.

T. R. Golub et al. (1999). 'Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring', *SCIENCE*, VOL 286, 531-537.

Kurt Fellenberg et al. (2001). 'Correspondence analysis applied to microarray data', *PNAS*, vol. 98, no. 19, 10781-10786.