Bioinformatics – Lecture Notes

Announcements

I need to collect project proposals, copies of the articles, and group meeting times.

Reference: Microarray Data Analysis and Visualization by Arun Jagota

Class 24 – April 11, 2002

I. Introduction

A. Microarray Data Analysis

Gene chips allow the simultaneous monitoring of the expression level of thousands of genes. Many statistical and computational methods are used to analyze this data. These include:

- statistical hypothesis tests for differential expression analysis
- principal component analysis and other methods for visualizing high-dimensional microarray data
- cluster analysis for grouping together genes or samples with similar expression patterns
- hidden Markov models, neural networks and other classifiers for predictively classifying sample expression patters as one of several types (diseased, ie. cancerous, vs. normal)

B. What is micorarray data?

In spite of the ability to allow us to simultaneously monitor the expression of thousands of genes, there are some liabilities with micorarray data. Each micorarray is very expensive, the statistical reproducibility of the data is relatively poor, and there are a lot of genes and complex interactions in the genome.

Microarray data is often arranged in an n x m matrix $\mathbf{M}$ with rows for the n genes and columns for the m biological samples in which gene expression has been monitored. Hence, m_{ij} is the expression level of gene i in sample j. A row $\mathbf{e}_i$ is the *gene expression pattern* of gene i over all the samples. A column $\mathbf{s}_j$ is the expression level of all genes in a sample j and is called the *sample expression pattern*.

There are several popular microarray technologies, each with data in a slightly different form:

cDNA microarray – The expression level e_{ij} of a gene i in sample j is expressed as a log ratio, $\log(r_{ij}/g_i)$, of the log of its actual expression level r_{ij} in this sample over its expression level g_i in a control. When this data is visualized e_{ij} is color coded to a mixture of red ($r_{ij} >> g_i$) and green ($r_{ij} << g_i$) and a mixture in between.

Nylon membrane and plastic arrays (by Clontech) – A raw intensity and a background value are measured for each gene. The analyst is free to choose the raw intensity or can adjust it by subtracting the background intensity.

Oligonucleotide silicon chips (by Affymetrix) – These arrays produce a variety of numbers derived from 16-20 pairs of perfect match (PM) and mismatch (MM) probes. There are several statistics related to gene expression that can be derived from this data. The most commonly used one is the *average difference* (AVD), which is derived from the differences of PM-MM in the 16-20 probe pairs. The next most commonly used method is the *log absolute value* (LAV), which comes from the rations PM/MM in the probe pairs. Note: The Affymetrix gene-chip software has a absent/present call for each gene on a chip. According to Jagota, the method is complex and arbitrary so they usually ignore it.

Note: Each new version of a microarray chip is at least slightly different from the previous version. This means that the measures are likely to change. This has to be taken into account when analyzing data.

C. For what do we used micorarray data?

Genes with similar expression patterns over all samples – We can compare the expression patterns e_i and $e_{i'}$ of two genes i and i' over all samples. If we use cluster analysis, we can separate the genes into groups of genes with similar expression patterns (trees). This will allow us to find what unknown genes have altered expression in a particular disease by comparing the pattern to genes know to be affiliated with a disease. It can also find genes that fit a certain pattern such as a particular pattern of change with time. It can also characterize broad functional classes of new genes from the known classes of genes with similar expression.

Genes with unusual expression levels in a sample – In contrast to standard statistical methods where we ignore outliers, here outliers might have particular importance. Hence, we look for genes whose expression levels are very different from the others.

Genes whose expression levels vary across samples – We can compare gene expression levels of a particular gene or set of genes in different samples. This can be used to look compare normal and diseased tissues or diseased tissue before and after treatment.

Samples that have similar expression patterns – We might want to compare the expression patters of all genes between two samples. We might cluster the genes into gene with similar expression patterns to help with the comparison. This can be used to look compare normal and diseased tissues or diseased tissue before and after treatment.

Tissues that might be cancerous (diseased) – We can take the gene expression pattern of sample and compare it to library expression patterns that indicate diseased or not diseased tissue.

D. Statistical methods can help

Experimental Design – Since using microarrays is costly and time consuming, we want to design experiments to use the minimal number of micorarrays that will give a statistically significant result.

Data Pre-processing – It is sometimes useful to preprocess the data prior to visualization. An example of this is the log ratio mentioned earlier. It is often necessary to rescale data from different microarrays so that they can be compared. This is due to variation in chip to chip intensity. Another type of preprocessing is subtracting the mean and dividing by the variance.

Data Visualization – Principle component analysis and multidimensional scaling are two useful techniques for reducing multidimensional data to two and three dimensions. This allows us to visualize it.

Cluster Analysis – By associating genes with similar expression patterns, we might be able to draw conclusions about their functional expression.

Probability Theory – We can use statistical modeling and inference to analyze our data. Probability theory is the basis for these.

Statistical Inference – This is the formulation and statistical testing of a hypothesis and alternative hypothesis.

Classifiers for the Data – We can construct classes from data, such a diseased vs. non-diseased tissue. We can build a model (such as a hidden Markov model) that fits know data for the different classes. This can then be used to classify previously unclassified data.

II. Preprocessing Microarray Data

Before microarray data can be analyzed or stored, a number of procedures or transformations must be applied to it. In order to analyze the data correctly, it is important to understand what the transformations might be doing to the data.

A. Ratioing the data

This is the most popular transformation. The expression level e_{ij} of a gene i in sample j is expressed as a ratio, (r_{ij}/g_i), of its actual expression level r_{ij} in this sample over its expression level g_i in a control. This tells us the level of under- or over- expression of a gene i in the sample j. If the control value g_i is very small, it can make the ratio very big. This can skew results incorrectly.

B. Log-tranforming rationed data

This is also a popular transformation. The expression level e_{ij} of a gene i in sample j is expressed as a log ratio, log(r_{ij}/g_i), of the log of its actual expression level r_{ij} in this sample over its expression level g_i in a control. This will suppress outliers caused when the control value g_i is very small. However, it creates a new outlier when r_{ij} is very small.

C. Alternative to ratioing the data

An alternative that eliminated both of the outlier problems above is

$$\frac{r_{ij}}{r_{ij} + g_i}$$

This gives a value in [0,1] and can be interpreted at the probability of gene i is higher in sample j than in control.

D. Differencing the data

Another transformation is to difference the data ie. $r_{ij} - g_i$. This is not really appropriate in our previous context. However, this is used by Affymetrix in a different context. In their data r_{ij} is the strength of the match of the target i to a specific probe j and g_i is he strength of the match of the target i to a control for this probe.

E. Scaling data across chips to account for chip-to-chip difference

As mentioned previously, different chips might display different intensities. When comparing different chips the data might need to be scaled so that they are on the same scale. Alternatively, they can be normalized so that they are between [0,1] and compared.

F. Zero-centering a gene on a sample expression pattern
This in effect the same as subtracting the mean expression pattern.

Suppose that **x** is an expression pattern for a particular gene g_i whose components are log-ratios. Let $x' = x - \bar{x}$ where $\bar{x}$ is the 'average expression pattern' or control. Then **x** indicates whether the gene g_i is induced or repressed relative to control. (Remember the **x**'s are vectors).

Subtracting the mean expression pattern and dividing by the standard deviation can accomplish this.

$$x' = \frac{x - \bar{x}}{s}$$

(Remember the **x**'s are vectors).

G. Weighting the components of a gene or sample expression pattern differently

If we have a matrix of weights **W**=diag($w_1,\ldots,w_n$), we can weight the expression patterns by

$$\mathbf{x}_w = \mathbf{Wx}$$

In this way, we can weight the contributions from different genes differently.

H. Handling missing data

Sometimes components of an expression pattern **x** are missing. To fix this, the missing values can be replaced by the mean over the non-missing values in **x**.

I. Variation filtering expression patterns

When we are performing cluster analysis on gene or sample expression patterns, patterns with low variance will all seem sufficiently similar to each other and might form a cluster. This cluster will probably not reflect any interesting result.

J. Discretizing expression data

Sometimes we might want to convert gene or sample expression pattern into discrete values. For example, if we have log-ratio, we may want to simply look at whether something is up- or down-regulated. To do this, we can do the following:

$$x_b = (x_{b,i}) \quad where \quad x_{b,i} = \begin{cases} +1 & when\ x_i > 0 \\ -1 & when\ x_i < 0 \\ 0 & when\ x_i = 0 \end{cases}$$

In this case +1 would indicate up-regulation, 0 would indicate no change and –1 would indicate down-regulation.

NEXT TIME

III. Measuring Dissimilarity of Expression Data
IV. Visualizing Micorarray Data
V. Cluster Analysis of Microarray Data