GS01 0163 Assignment 5 Fall 2007
Due Date: 15 November 2006

1. Download the follicular lymphoma microarray data set (`FL.Rda`) from the class web site. This file is an R data object; you can use the `load` command to read it into R. Describe the two objects that are contained in the data set. (Hint: you probably need to load one of the BioConductor libraries in order to interpret the class of one of the objects.) How many samples are there? How many genes? Where did the data set come from? How was it processed?

2. Use the `esApply` function to compute the median expression of each gene across this sample set. Summarize the median expression values. How many genes have a median expression greater than 9?

3. In this problem, we will work with the subset of genes whose median expression is greater than 9. Using these genes and Pearson correlation to define distance, perform a hierarchical clustering of the samples:

   (a) Using single linkage

   (b) Using complete linkage

   (c) Using average linkage

   Which linkage rule gives the clearest separation into different subtypes of follicular lymphoma?

4. We will continue working with clusters based on Pearson correlation using genes whose median expression is greater than 9. Using the linkage rule you chose in the previous problem, perform a bootstrap cluster test to determine if the main clusters are reproducible in this data set. Plot an image of the resulting bootstrap object.

   (Hints: You may need to try several different values for the number of clusters. You should also use the `dendrogram` option of the `image` command to produce a sensible image of the results.)

5. Assuming that you found at least two large and somewhat reproducible clusters in the previous problem, can you tell if these clusters match any of the clinical parameters that accompany the data?

6. We continue working with the subset of genes whose median expression is greater than 9. Perform a principal components analysis (PCA) on the samples and plot the result. Do you see clear signs of different groups? Using the factors in the accompanying clinical data, replot the data coloring different groups to see if any of them appear to separate in the plot.

7. Use the partitioning around medoids (PAM) algorithm to cluster the same data set. Use plots of the silhouette widths to decide which value between 4 and 10 clusters best describes the data.