Statistical Analysis of Genetic Data
Math 155, Spring 2008
Jo Hardin
Project 8: hierarchical clustering

Due Friday April 18

1 Instructions

1. The goal for this part of the project is to cluster genes from your data.

2. You are going to come up with clusterings for two sets of genes: (1) a set of 20 random genes, (2) a set of 20 genes that were "significant" in your previous analyses. For example, for the second set of genes you could find 20 from the same 2 group comparison, but 10 of the genes are up regulated and 10 are down regulated. Another example might be to find 10 genes that are significant in comparing groups 1 and 2 where a different 10 are significant when comparing groups 3 and 4 (assuming you have at least 4 groups). Email me if you aren't sure how to find the 20 genes from last week's assignment.

3. I want you to play around with the distance options and the linking method. You should try out a bunch of combinations when you are working on your analysis. However, you only need to report 3 different combinations for each set of genes.

4. For the R code, you should use the functions `dist, cor, as.dist, hclust, plot, identify.hclust, cutree`. Remember you can use "?" to find out more about any of the functions.

5. `cor` has a method argument where two of the options are Pearson and Spearman.

6. `hclust` uses only things that have a mode of "distance" function, so you need to transform the correlation matrix into something with mode distance by using the `as.dist` function.

7. Remember that if you have an M matrix which is 100 genes by 20 arrays, your correlation (or distance) matrix should be 100 x 100 (not 20 x 20). Though if it is an object of mode "distance", it'll be a vector (not a matrix) with length $\binom{n}{2}$, here that would be $\binom{100}{2} = 4950$.

8. `cutree` tells you which groups each of your items is in.

2 Things to put on your web site (next)

- At least 3 dendrograms for each of the sets of genes (that is, you should have at least 6 dendrograms). Within the 3 dendrograms, you should have at least 2 different distances and at least 2 different linking methods.

- Ideally your plots will be on one page of the website (use `par(mfrow=c(2,2))` or `par(mfrow=c(2,3))` to get all the plots on one sheet). You want the user to be able to compare all the plots simultaneously without clicking back and forth.

- A discussion of the plots and whether you see any patterns. Address the difference between the random genes and the genes selected from the lists of differentially expressed genes.

  - For a given set of 20 genes, are there differences in the distance metrics you chose?
  - For a given set of 20 genes, are there differences in the linking method you chose?
  - In comparing the two sets of genes, do you see differences in distances? That is, at what distance (on the y-axis) do the objects link up?
  - Do you the differentially expressed genes show more patterns?
  - Give me some insight into the images (think of me as a naive biologist who doesn't know how to interpret the results!)