Statistical Analysis of Genetic Data
Math 155, Spring 2008
Jo Hardin
Project 9: PAM clustering

Due Friday April 25

1 Instructions

1. The goal for this part of the project is to cluster genes from your data using Partitioning Around Medoids (PAM). You will also compare the results of the clusterings (among themselves and with previous clusterings).

2. You can use either of the subsets of genes from last week or a new subset of genes (you might try using more genes).

3. Run PAM for a variety of k (# of groups) values.

2 Things to put on your web site (next)

- Again, the idea here is that you, as the statistician, are trying to get some insight into the co-expression patterns of your genes. You want to find what you think of as the best clustering, and then explain to the biologist (i.e., the reader of your website) why it is the best clustering.

- You won't have nice plots here, so your decisions will be made based on measures like average silhouette width, average dissimilarity (objective function), or maximum dissimilarity.

- Try a variety of k values. For average silhouette width, you'll want the maximum. For average or maximum dissimilarity, you'll want the bend of an "L".

- After you have a few PAM clustering vectors (remember, you also different distances to use), compare some of your results using the `table` and `adjrand` functions.

- Give a discussion / conclusion about your results. Do you think there really are any clusters? Is it all just noise? Are there genes which truly seem co-expressed? (Feel free to look at scatterplots of genes to assess the last question).