Problem Set 1. Clustering

Due Date: March 15th, 2016 via email to ml4bio@gmail.com. Please include any R code you used for answering the questions, as well as a short description of what you did to answer those questions and any figures that you generated. We can read PDF or Word documents.

1. Use R to calculate the objective function for K-means for the Quaid Data Box. Show that there really are two identical optima.

2. Use a "held out data"/"cross-validation" approach (described in Lecture 1) to find the optimal number of clusters for the CD4/CD8 data with K-means and pam. (hint: you can get both kmeans and pam in the cluster package in R and use plot to make the plots. hint: analyze the data in log space for better behavior)
   a) Make a plot of the objective function on the held out data for various choices of k for both clustering methods
   b) Make plots that shows the clusters and the data for the optimal values of k
   c) Does this analysis resolve the disagreement that I showed in class? Explain briefly.

3. Look at the manual for Gene Cluster 3.0. Explain the weighting function used to downweight dimensions.

4. Download blosum 62 and calculate the distances between these sequences (you can get blosum62 in the peplib package, and use the Biostrings package to read sequences in R) :

```
>ENSP00000386200_Hsap/300-357
GVHSMEDNGIKHGGLDLTTNNSSSTTSSNTSKASPPITHHSIVNGQSSVLSARRDSSS
>ENSPTRP00000033573_Ptro/276-333
GVHSMEDNGIKHGGLDLTTNNSSSTTSSTTSKASPPITHHSIVNGQSSVLNARRDSSS
>ENSMMUP00000010501_Mmul/299-356
GVHSMEDNGIKHGGLDLTTNNSSSTTSSTTSKASPPITHHSIVNGQSSVLNARRDSSS
>ENSMUSP00000111137_Mmus/274-331
GVHSMEDNGIKHGGLDLTTNNSSSTTSSTTSKASPPITHHSIVNGQSSVLNARRDSSS
>ENSRNOP00000069190_Rnor/270-327
GVHSQEDNGIKHGGLDLTTNNSSSTTSSTTSKASPPITHHSIVNGQSSVLNARRDSSS
```

   When this was discovered, researchers speculated that there might be positive selection on the human sequence. Why?