Problem Set #4 CH391L Bioinformatics Marcotte Spring 2011
Due Tuesday, April 12, 2011

1. Download the yeast protein sequences from the course web site. These represent the complete sets of proteins (the "proteome") from *Saccharomyces cerevisiae*. The protein sequences were translated from predicted genes found when the genome was sequenced, and many were later verified by other means. Many of these genes were known prior to the genome sequence, but about ~1/3 of the genes were new. Each entry begins with a protein name (a common name or a unique id code from the Entrez genome database: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome). The name is followed immediately by the known location that the protein occupies inside a yeast cell. (Actually, the name and the location are fused together by an underscore character so that you can keep track of both of them together.)

2. Calculate the frequencies of each amino acid for each of the proteins. The file contains lines starting with ">" that separate the protein sequences—be sure to skip these lines for your calculation, but keep track of the protein name/location so that you can generate a feature vector of amino acid frequencies for each protein. Also, some entries contain non-amino acid characters (e.g., when protein sequences are ambiguous), so skip these characters as well, just keeping track of the 20 amino acids. Write the amino acid frequencies of each protein as a vector, separated by tabs. (In Perl, you can print a tab character using "\t", just as you might print a newline character with "\n".)
So, you should have a vector for each protein sequence encoded by the genome, in the form:
ProteinName_location 0.069 0.011 0.048 0.069 0.054 0.058 0.021 0.072 0.088 0.112 0.023 0.059 0.033 0.037 0.035 0.068 0.044 0.057 0.007 0.037
where the first word is the name_location of the protein, followed by a tab, the fraction of Alanine (A) in the proteome, a tab, the fraction of Cysteine (C), etc. Be sure to write the amino acid frequencies in the same order for each protein!

3. Download and install the free clustering and display software named "Cluster" and "Treeview" from http://rana.lbl.gov/EisenSoftware.htm
These are 2 programs written by Michael Eisen for automating hierarchical and k-means clustering and analysis of gene expression vectors. The programs can be installed on any PC running the Windows operating system (there is also a Linux/Mac version if you prefer).

4. The programs should be fairly self-explanatory. Modify your amino acid frequency vectors from question 2 above to work with the Cluster program by adding a line to the beginning of your data file that consists of the word UNIQID, followed by a tab, then each amino acid in order separated by tabs, e.g.:
UNIQID A C D E F G H I K L M N P Q R S T V W Y
"Load" your amino acid frequency vectors into the Cluster program. Perform the *k*-means clustering algorithm choosing some *k* of intermediate size (say, *k*=100) to cluster the data vectors. (Note: the program is intended to be used with gene expression data, which this is not, so where it refers to "genes", this means your proteins, and where the program refers to "arrays", this is whatever features your vectors are made of. In this case, these would be amino acid frequencies.) The program will automatically write out the results of the clustering in your program or data directory (the file with a name ending .kgg. The clusters in the file are numbered sequentially.) Report several of the final clusters. You can play with the value of *k* to see the effects on the clustering.

Note: Both Cluster and TreeView come with sample data so you can make sure the programs are installed & working properly with these data sets. Also, you are welcome to program any of these clustering methods yourself if you would rather do that than use the Cluster and TreeView programs.

5. Do the clusters group proteins in a fashion consistent with their sub-cellular locations? Are there some locations clearly clustered better than others?

6. Perform hierarchical clustering on your data using Cluster by clicking the "hierarchical clustering" tab, pressing the "arrays" button to turn off clustering on the columns of data, and pressing the "average linkage clustering" button.
Run the TreeView program to visualize the results. In TreeView, after loading the results of your hierarchical clustering, you should see a clustered set of feature vectors on the left side of the screen. Clicking on different levels of the tree should bring up the corresponding vectors on the right side of the screen. TreeView is set up by default to display gene expression data, which has positive and negative values. You can change the display colors to show your data better with the "Settings" tab ("Options"-> "colors", then set a color for zero.").

7. How do the hierarchically clustered data agree with your k-means clustering results from question 4 above?

8. Download the file of yeast protein phylogenetic profiles from the course web site. Each entry in this file is the phylogenetic profile of a yeast protein. Following the name_location are 149 numbers, each indicating the similarity of the protein to the best matching protein in one of 149 genomes. The numbers correspond to the genome names listed in the first line of the file.

9. In Eisen's Cluster program, load in the phylogenetic profile data you downloaded in step 8. Cluster all of the genes with hierarchical clustering. Do genes with similar subcellular locations cluster? Several subcellular compartments show better clustering than others---print out 3 clusters where many of the proteins come from the same location.

10. Likewise, try clustering the genes (hierarchical clustering will be fine) based on their microarray-measured mRNA co-expression patterns (also downloadable from the course web site). Print out 3 clusters where many of the proteins come from the same location.

11. Which types of features, mRNA expression, amino acid frequencies, or phylogenetic profiles, seems to be working the best to organize the proteins in a manner consistent with their subcellular locations? Why should you expect each of these features to inform us about subcellular locations of proteins?

12. Even though not all of the genes from the same location cluster together, the data can still be used to predict the subcellular location of proteins. Suggest a strategy for predicting the subcellular locations of yeast genes that exploits the fact that very small clusters of co-inherited (or co-expressed or similar composition) proteins have similar subcellular locations.

13. How might you predict the subcellular locations of the proteins based on combining the information from the different types of features associated with each gene?

14. Finally, some next-gen sequencing!

As I'm sure you all remember, on the first day of class, we collected 5 environmental samples for analysis. Three came from the UT turtle pond; the remaining two I collected.
They were, in no particular order:

(A) Green water from a clear, airtight, plastic bottle which I filled with pond water & sealed off more than a year ago, then set in the sun on my office windowsill. The sides of the bottle had collapsed inwards, suggesting the air in the bottle had mostly been consumed.
(B) A soil sample from the edge of the upper turtle pond
(C) Scrapings from the green, mucky ooze between the upper and lower turtle ponds
(D) Clear water collected from the middle of the lower turtle pond
(E) A surface soil sample from my compost heap

From each of these, we isolated DNA (using the MoBio PowerSoil DNA extraction kit), then sequenced it using our SoLID next-generation DNA sequencer, collecting roughly half a million reads from each sample. Two sets of reads were collected per sample, a set of 35mers (called "F5") and a set of 50mers (called "F3"). Taejoon experimented with a few different ways to analyze this data, and summary files from one of these approaches are posted on the course web page. In this approach, Taejoon used the mapping program SHRiMP to map the sequenced reads to collections of DNA sequences drawn from a wide variety of microbes, including representative species of bacteria and of archae (whole genomes from the NCBI web site), and a set of mitochondrial and chloroplast genomic DNA spanning ~3,000 eukaryotic organisms (from the European Bioinformatics Institute, http://www.ebi.ac.uk/genomes/organelle.html). Note that only one or a few species of bacteria/archae from each genus were included—a sequence hit to an organism thus gives evidence for that genus being present, but might not be due to that specific species. I mention this before you get too concerned about some of the particular organisms apparently present in these samples! The raw mapping data are available in the linked folder from the course web page. We've also put together a summary file for each sample listing the number of reads from each organism in these databases.

Based on these data, tell me which set of sequences (obscurely named V3BC21-25) corresponds to which sample (A-E above).

15. Is it really likely that there are Neanderthals (or shrews or gorillas) in the UT turtle pond? We can probably assume these to be false mappings. What's going on here? How might you better detect or flag such false mappings from these data?