Problem Set #3 BCH394P/364C Systems Biology/Bioinformatics Marcotte Spring 2025
Due by 10 PM Monday, March 24, 2025

1. Download the human protein sequences from the course web site. These represent roughly 1,800 of the ~20K human proteins. This subset is enriched for fairly highly expressed, well-conserved, and reasonably well-annotated proteins. Many of the corresponding genes were known prior to the genome sequence, but a substantial fraction were new and a few remain uncharacterized. The file is in .fasta format, so each entry begins with ">", a protein's common name, then a UniProt database identifier (from https://www.uniprot.org/), then a description of the protein. The name, id, and description are separated by vertical bars so you can keep track of them together but still separate them if you need to.

2. Calculate the frequencies of each amino acid for each of the proteins. I recommend you use the BioPython library function SeqIO.parse() for this, but if you write a custom parser, be sure to skip lines starting with ">" for your calculations but keep track of the protein name/location so that you can generate a feature vector of amino acid frequencies for each protein. Some entries contain non-amino acid characters, so skip these characters as well and just keep track of the 20 standard amino acids. Write the amino acid frequencies of each protein as a vector, separated by tabs. (In Python, you can print a tab character using "`\t`", just as you might print a newline character with "`\n`".)
You should write out a vector for each protein sequence in the file, in the following format:
ProteinName|ID|Description 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|ID|description of the protein, followed by a tab, the fraction of Alanine (A) in the protein, a tab, the fraction of Cysteine (C), etc. Be sure to write the amino acid frequencies in the identical order for each protein!

3. We're going to explore these data using clustering. There are many tools for clustering, many of which you can download and install locally on your computer, but we'll take advantage of a nice web-based tool called Morpheus, available from the Broad at https://software.broadinstitute.org/morpheus/

4. The program should be fairly self-explanatory. Modify your amino acid frequency vectors from question 2 above to work with the Morpheus program by adding a line to the beginning of your data file that consists of the word PROTEIN, followed by a tab, then each amino acid in order separated by tabs, e.g.:
PROTEIN A C D E F G H I K L M N P Q R S T V W Y
Open your amino acid frequency vectors into the Morpheus program. Perform hierarchical clustering on your data by clicking the wrench button and selecting "hierarchical clustering", selecting "rows" (or "rows and columns", if you prefer), a similarity measure of your choice, and "average linkage clustering". After hierarchical clustering, you should see a clustered set of feature vectors. You can interactively navigate the proteins to see how they cluster together.

5. Do the clusters group proteins in a fashion consistent with their functions? Are there any functions clearly clustered better than others? Support your answer with 2 examples captured as screen shots of the clusters.

If you think about it a bit, clustering proteins by their amino acid frequency vectors is mostly going to group proteins with similar sequences, especially duplicate genes and close homologs. So, let's look at some other features that will start to group proteins by their functions independently of their sequence similarities.

6. Download the file of human protein phylogenetic profiles from the course web site. Each entry in this file is the phylogenetic profile of one human protein, which captures the phylogenetic distribution of that gene family across species. Following the name are 100 numbers, each indicating whether or not that protein has an ortholog in that particular organism. The names of the organisms are given in the first line.

7. In Morpheus, load in the phylogenetic profile data you downloaded in step 6. Cluster all of the proteins with hierarchical clustering. Do proteins with related functions cluster? Several cellular functions show better clustering than others, not least because these organisms include both ciliated and non-ciliated species, so you might check out some ciliary proteins (like the intraflagellar transport proteins) to see how they behave. Print out 2 screen shots of clusters where several of the proteins function together.

8. Likewise, try clustering the proteins based on their biochemical co-purification patterns (also downloadable from the course web site). This is a larger file, so you will probably only want to cluster rows, not columns, and it may take a bit of time. To measure these data, various human cell lines were lysed and native protein extracts were made. The proteins were separated chromatographically, and the proteins in each chromatographic fraction identified by mass spectrometry. Each row indicates the chromatographic elution profile for a given protein, with the numbers indicating protein abundances. (The results from several different separations have been concatenated together; the rather cryptic column headers denote the specific biochemical fractions analyzed in each separation.) The key idea here is that proteins that are bound to each other will tend to co-elute, which you should see especially for large protein complexes like the proteasome. Again, print out 2 clusters where many of the proteins are known to interact or function together.

9. Which types of features, co-purification, amino acid frequencies, or phylogenetic profiles, seem to be working the best to organize the proteins in a manner consistent with their functions?

10. You may have noticed that some of these features have only limited power (especially for the relatively small datasets we're considering) to group co-functioning proteins together. Why should you expect these features to inform us at all about protein functions? Explain your logic for each of the 3 data types.

11. Even though not all the proteins with the same function cluster together, the data can still be used to predict protein function. For example, one such strategy exploits the fact that small clusters of co-inherited (or co-expressed or similar composition) proteins have similar functions. You can explore this a bit in Morpheus by selecting a set of proteins using the search bar function on the top left of the screen. (The arrow button to the left of the search bar allows you to "Match all search terms" to get more specific searches.) If you press the small vertical arrow to the right of the search bar, you can redraw the clustergram to move your "matches to top" so you can see them. It is then possible to search for the "Nearest neighbors" (in the pulldown tools menu) of proteins you select. For a dataset of your choice, select a set of proteins of your choice (specified e.g. by name or function; examples include "proteasome alpha", "flagellar", "tRNA ligase", etc.) & search for their nearest neighbors using Morpheus. You might test a few examples to find one that seems to have a bit of predictive power. Turn in a screen shot of your results.