Capstone Project: Data Wrangling Zach Young

Summary:

The data wrangling for this project will consist of two essential processes. First, I must convert each protein from its base genetic code into the corresponding amino acids and then assign a hydrophobic value to each amino acid. In this step, I will create a separate data frame for each of the 40 proteins with columns for codons, amino acids, hydrophobic value, amino acid position and finally, a column for hydropathic index. Once this is done I will find summarize the data from each of the 40 proteins and combine the summarized data from each protein into one final data frame for analysis.

Step 1: Base code to amino acids and hydrophobic values

The raw data that I will be starting out with will simply be the base genetic code of each protein. This consists of a string of four letters, A, T, C and G, each corresponding to a nucleotide in a strand of DNA. The first step in making sense of this data is splitting the letters into groups of three. Each group of three letters is termed a codon and each codon corresponds to a specific amino acid. To accomplish this, I first will create a long list of nucleotides and then use the strsplit function to break the list into groups of three. Next, after converting the list of codons into a data frame so that each codon is a separate row, I will create custom functions to convert each codon to its corresponding amino acid and assign each amino acid a hydrophobic value. The custom functions will utilize the gsub function and then I will use sapply to apply the functions to each row. Finally, I will use a rolling average function from the zoo library on the hydrophobic value column to create a hydropathic index column. This will be used to create hydropathic plots for every protein.

Step 2: Combining summarized protein data into one data frame

Once I have a data frame for each individual protein I will calculate summary statistics for the protein including the average hydrophobic value, the percentage of all hydrophobic amino acids in the protein along with the percentage of each of the hydrophobic amino acids in the protein. Once I have these values I will put each into its own column of a data frame and create another column identifying whether the protein is membrane or non-membrane. Finally, I will join all 40 of these individual data frames into one final data frame using the full join function. Once this is done I will have one data frame that I will be able to use to compare membrane vs. non-membrane proteins and perform analysis on.