Capstone Project Milestone

Problem: Find a faster, more efficient and more cost-effective method for identifying protein structure and function using sequenced genomic data.

Proteins are the building blocks of all cells and are responsible for all of the essential processes of life. One of the biggest distinguishers of function amongst proteins is whether they are located within the membrane of the cell, or on the inside of the cell. Accurately identifying the location of a protein is thus an important task and improving upon existing laboratory methods of identification will save huge amounts of both time and money.

Because of the essential role that proteins play throughout the various cells and tissues of the human body, they are a major target of research efforts into human disease and drug discovery. Being able to identify protein function and location is an important step for scientists conducting research in both the public and private sectors. For decades, in order to study a protein, it had to be isolated and precipitated individually from a cell, usually at an incredibly high cost, both in terms of time and money. With the advancement of genomic sequencing technologies, we are now capable of generating huge amounts of biological data at ever decreasing costs. Unlocking the insights from this data is critical to current research endeavors and accurate protein prediction has been a major emphasis in the field.

The Data Set:

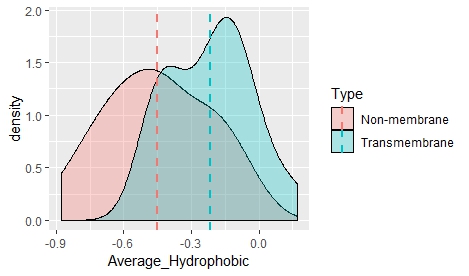
The data I will be using will be mRNA sequence data acquired from the NCBI (National Center for Biotechnology Information) database. The data will come in a FASTA format and will simply be the entire base code, represented by the letters A, C, G, and T, for my chosen proteins. I will choose 40 different proteins; 20 transmembrane and 20 non-membrane. Descriptions of the proteins can be found in the same database and I will pick proteins whose locations in the cell, i.e. transmembrane vs non-membrane, have been previously verified in the laboratory.

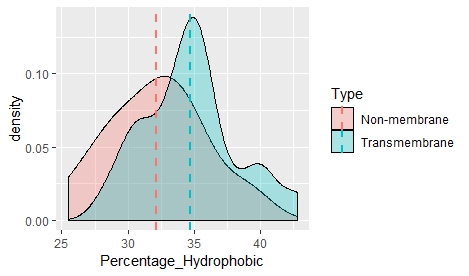
In order to solve the problem, a fair amount of data wrangling and cleaning is required. First, I will need to convert the base code into the protein's amino acid profile. Each amino acid corresponds to a triplet codon of base pairs and every amino acid has specific chemical properties. In my project, I will begin by breaking the code into three base triplets (aka codons) and then create a function to convert these into their corresponding amino acid. I will then assign every amino acid a specific hydrophobic value which matches data produced by Kyte and Doolittle and their work on identifying transmembrane regions within proteins. Biological membranes are extremely hydrophobic and any protein that passes thru or resides within a membrane will, according to my hypothesis, also have a high percentage of hydrophobic amino acids within the section that interacts directly with the membrane. Once I have the amino acid sequence, I can translate that into a plot of hydrophobic values. Part of my code will produce a hydropathic plot for each protein analyzed. These plots have traditionally been used to identify transmembrane regions within proteins. Additionally, I will calculate averages for various characteristics of the proteins based off their amino acid profile to hopefully find identifiable characteristics that can be used to distinguish membrane proteins versus non-membrane proteins. Finally, I will incorporate these findings into a model that will be able to identify protein type based off genomic data.

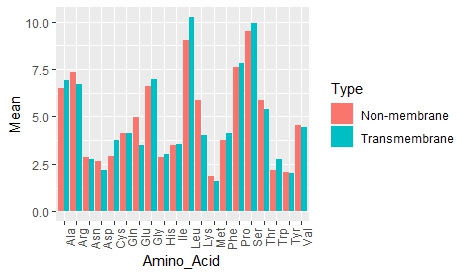
Initial Findings:

After analyzing each of the 40 proteins and compiling their features into a single data frame I started my exploratory data analysis. The first step I took was to calculate summary statistics on the various features of my data frame. Knowing the chemical nature of biological membranes and their affinity towards hydrophobic molecules, I have come up with the hypothesis that transmembrane proteins should contain more amino acid residues of a hydrophobic nature than their counterpart non-membrane proteins. Based off this hypothesis, I have focused my statistical analysis on three key measurements. First, I calculated the average hydrophobic value for both types of proteins. If my hypothesis is correct, transmembrane proteins will contain more amino acids with higher hydrophobic values and thus have a higher overall average hydrophobic value than non-membrane proteins. Next, I found the percentage of all hydrophobic amino acids within a protein compared to the total number of amino acids in the protein. As with the average hydrophobic value, I’m predicting transmembrane proteins will have a larger amount of hydrophobic amino acids and in turn a higher percentage of hydrophobic amino acids. Finally, I will examine the percentage of each individual amino acid within a protein and try to ascertain any significance between the two types of proteins. I made plots of each of these features to get an idea on any significance there might be.

Density Plots







Based on the above plots there does appear to be significance between transmembrane and non-membrane proteins in both the average hydrophobic values and the percentage of hydrophobic amino acids. Additionally, the amino acids Glu, Leu, Lys and Trp may also have significance in determining the type of protein. To ascertain any statistical significance of these features, I ran t-tests on each. For the amino acids, Glu had a p-value of 0.015, Leu had a p-value of 0.033, Lys had a p-value of 0.003, and Trp had a p-value of 0.146. This shows that Lys seems to be the most significant in determining protein type, while Glu and Leu also show significance. Trp is above the 0.05 p-value threshold and may not be as significant. The p-value for the average hydrophobic value was 0.001 suggesting strong significance while the p-value for percentage of hydrophobic amino acids was 0.033, indicating a smaller significance. After running the t-tests there appears to be enough features showing significance to develop a model that can accurately and efficiently predict protein type from genetic code data. I will attempt to develop two separate models, a logistic regression model and a decision tree model, in order to accomplish this task.