William Schlough

For the BME 160 Final Project, I was unsure of what I wanted to create and I didn't have any immediate ideas but I eventually landed on trying to create a generalized hydropathy plot (that is, one that could be used with any standard nucleotide in-sequence). A hydropathy plot represents the hydrophobic and hydrophilic tendencies of a specified amino acid sequence, and can be used to partially determine the shape of the protein coded by said amino acid sequence. The plot's *y* values come from averaging a window of *z* values at a specified *x* point, and then moving to the next point and repeating the average, storing it as a new *y* value. The end result is a graph that has *x* values spanning the length of the amino acid chain, and for every *x* value there exists a *y* value that is equal to the average of the sum of all values within the frame of *z* length(reference graph here) . From the graph, we can make predictions based on shifts from negative to positive and positive to negative, along with making predictions based off of the sections that span a long range and sit in either a positive or negative region. The predictions being made are based on if a transmembrane domain (protein) sits in a certain region along the amino acid chain, and if it is of alpha-helix type. Since the program created is of generalized type, the result is a functional program that creates a hydropathy plot for every open reading frame in an input nucleotide sequence, and then prints those plots to a single PDF file in the program directory.

The problem at hand has multiple parts - how to create a hydropathy plot, how to print out the plot, and what can I do with the plot - which can be dealt with individually. The hydropathy plot requires a hydropathy scale / index to function, and along with that there is also the issue of mapping the points to a graph, so a way to average the values at each point is also necessary. For printing, I don't want the user to have to view each graph within whatever IDE they are working with, so printing each graph to an outfile would be a great addition. The last step of the problem is what can be done with the plot, which is where the program being generalized is a slight issue - I chose not to focus on any specific sequence, so while the program might be functional, it hasn't been applied to anything in particular yet. That is to say, solving this problem has more to do with producing working code rather than using said code to research.

For the hydropathy plot, it is necessary to use a hydropathy scale / index, which is a list of values that describe each amino acid's relative hydrophobicity (positive values) or

hydrophilicity (negative values). I used a predetermined hydropathy scale from Kyte J., and Doolittle R.F., translated the nucleotide in-sequence to an amino acid sequence, and matched those AA's to their predetermined values from the hydropathy scale. Once I had this list, I tried calculating a moving weighted average by hand to form the hydropathy plot with, but the program was running slow and I couldn't take large averages. Instead I found a numpy function called convolve(), which aligns two lists and slides the second list down the first, cross multiplying and adding all the values up (it does this until the second list no longer matches up with the first list). I spoofed the function by creating a list of ones that would be the length of the window for my moving weight average, and then calling convolve() with list1 (matching AA hydropathy scale values) and the list of ones, and then dividing by the number of objects in the list of ones.

After evaluating the accuracy of my program, I believe that while it does what it is intended to do with good success, it could use some additions that would greatly improve what can be drawn from the output. I would like to first clean up the graphs and how they print in the file for easier viewing, but in terms of taking the program further I want to try and create my own hydropathy scale from scratch based on solubility scores and bonding angles for each chemical in each amino acid. If I were able to do this, I think my program would have a lot more viability in terms of use and accuracy.

Overall, this project was much more tedious with regards to the coding, but I gained some great insight on working with graphing in single and multiple files and using other modules like numpy to execute equations that would be much more difficult by hand.

REFERENCES

"A simple method for displaying the hydropathic character of a protein" - Kyte J., and Doolittle R.F., 1982
https://www.sciencedirect.com/science/article/pii/0022283682905150?via%3Dihub