Introduction to Bioinformatics – AS 250.265
Assignment 7

At the start of the course, we discussed three different aspects of bioinformatics that we will be learning about: First, we wanted to learn how to use the tools. Then, we wanted to learn how some of those tools work in detail. Finally, we wanted to learn about the biology (and physics) behind the tools.

In the lab four, you learned quite a bit about how to use several popular proteomics web servers. This assignment, then, will help to round out your exposure to proteomics, both from a biological and an technological perspective.

This assignment is worth 100 points toward the homework portion of the class. The laboratory questions you do in class constitute 40 of those points, and this handout makes up the remainder. Both parts are due on April 7 at the start of class. Electronic portions of the assignment should be submitted through the course website with appropriate filenames (e.g. `hw07-q3b.py`). The written portions of the assignment may be submitted by hand or uploaded as a Word document or PDF file.

Question 1: Laboratory Number 4 (40 points)

Submit your solutions the laboratory questions on proteomics that you completed in class.

Question 2: Unknown Residues (15 points)

Sometimes in bioinformatics application, the complete primary structure (that is, the sequence) of a protein is unknown. There are times, for example, when the sequencing fails and you cannot identify the residue at all. Other times, it may be possible to eliminate all except the most closely related amino acids, i.e. asparagine and aspartate or glutamine and glutamate.

Because this problem happens occasionally, some bioinformatics servers will allow you to include alternate inputs: X represents a totally unknown residue, B represents D or N, and Z represents E or Q. In this question, we will investigate these residues from a molecular weight perspective using the pI/MW server at `www.expasy.org`.

a. Determine the molecular weight used for B, Z, and X at the pI/MW server. To do this, you will need to calculate the molecular weights for B, BB, Z, ZZ, etc. and calculate the slope between the two points. The slope between the two points is the molecular weight for B, Z, or X. (5 points)

b. Calculate the complete molecular weight for Asp (neutral), Asn, Glu (neutral), and Gln. Is the molecular weight you calculate in part (a) what you would predict from averaging the appropriate molecular weights? (Hint: what atoms are added when extending a peptide chain?) (5 points)

c. Why do you think the molecular weight for one residue, B, Z, or X is different from the actual value of molecular weight used incrementally? In other words, why does the slope from part (a) have a nonzero intercept? (5 points)

Question 3: GTP Binding Proteins (15 points)

One of the useful features of the EcoCyc website is that it allows you to search for enzyme substrates as well as enzyme functions. Thus, if you are interested in a protein that breaks down or changes a particular chemical, you can identify several proteins without performing an exhaustive literature search. In this question, we'll explore the though process that might go into designing an actual bioinformatics experiment.

Visit the EcoCyc website and search the *E. coli* genome for a GTP binding protein. The protein you identify will be the Era G-protein. G proteins are essential in all cells for performing various reactions, most of which are cyclic: when GTP is bound, the protein is ready to bind a particular substrate. Once the substrate is bound, GTP is hydrolyzed to GDP, and the substrate is often transported or transferred to another protein. Finally, with the substrate gone, the G-protein is ready to bind GTP once more and start the process again.

Using BLASTP, search the human proteome for homologues to the Era GTPase. In your BLAST results, you will find several real known homologues of this protein, but you will also find several significant matches that are unnamed or have unknown function.

a. In your BLAST search, what matrix did you use? What protein sequence accession number did you choose as the most relevant unknown protein? (3 points)

b. Is the specific function of Era known? If so, what does it do? Where did you look to find out? (5 points)

c. Has the crystal structure of Era been determined? If so, what is the relevant PDB ID? (2 points)

d. Design an experiment that will test whether your protein of unknown function will bind GTP. Be thorough, describing both how you will test for binding as well has how you will confirm that the protein bound is your indeed the protein of interest. (5 points)

Question 4: Hydropathy Profiles (30 points)

In this last question, we will implement a simple yet powerful proteomics tool that will output what is known as a hydropathy plot. These plots are useful for predicting transmembrane helices in proteins that are known to be incorporate helices as their strategy for passing through the lipid bilayer in cells. The plots are not useful for non-membrane proteins or membrane proteins that use other methods for traversing the membrane.

Recall from class the transfer free energy of a residue from water to lipid:

$$X_{membrane} \leftrightarrow X_{H_2O}$$

This reaction, where X is a specific amino acid residue, has a transfer free energy, $\Delta\overline{G}_{tr}^{o}$, that can be determined experimentally through various thermodynamic cycles and with various approximations for the lipid environment. One of these "hydrophobicity scales" was calculated by Stephen White and William Wimley at UC Irvine. A table for all 20 amino acids, adapted from their 1996 *Nature Structural Biology* paper, is below. You can find the reprint of this paper available on the assignment website. Recall that negative free energies represent favorable transfer from lipid to water.

Amino Acid	**$\Delta\overline{G}_{tr}^{o}$ (kcal/mol)**
A	-0.17
R	-0.81
N	-0.42
D	-1.23
C	0.24
Q	-0.58
E	-2.02
G	-0.01
H	-0.17
I	0.31
L	0.56
K	-0.99
M	0.23
F	1.13
P	-0.45
S	-0.13
T	-0.14
W	1.85
Y	0.94
V	-0.07

On the website, you will find a template program that you will use in implementing your own hydropathy plot program. Your program will take a FASTA file as input from the command line and will output on each line the residue number and its hydropathy value, separated by a space. This output can then be copied into Microsoft Excel and printed as an actual graph.

In calculating your hydropathy plot, we will use a window of five residues. That is, the hydrophobicity of a given residue in the plot will be calculated as the average hydrophobicity of that residue plus the four neighboring residues. In the terms of an equation:

$$H_{i,plot} = \frac{H_{i-2} + H_{i-1} + H_{i} + H_{i+1} + H_{i+2}}{5}$$

Using a window will allow the plot to be smoothed and it will compensate for the fact that even highly polar residues can sometimes be inserted into the membrane. Of course, using a window will also limit the size of the proteins we can examine, and it will be your responsibility to ensure that FASTA files containing five residues or less are not accepted by your program.

As you look at the template, you will see that the interface has been written for you already. You will simply have to implement the actual hydropathy calculation.

a. Write the hydropathy program using the template given on the assignment website. You may test it using the sample FASTA files given on the assignment website. (15 points)

Now that you have written your program, you should be able to visualize hydropathy profiles for an arbitrary FASTA file. Large stretches of hydrophobic residues will be positive on your plot—these will tend to form hydrophobic alpha helices that will insert into the membrane (or be inserted by the translation machinery). The problem we now face is: how many hydrophobic residues will it take to span the membrane?

b. Using the Internet, determine the translation per residue of an alpha helix, in angstroms. This is sometimes called the helix "rise." You may find the NCBI books database useful for finding this information. Given that the hydrophobic part of the lipid bilayer is typically 25-30 Å How many residues must have positive hydropathy in order to span the lipid bilayer? (3 points)

Now that we've finished with the preliminaries, we can apply our program to some real proteins. First, let's look at a *bona fide* transmembrane protein. The protein we will examine bovine rhodopsin, one of the few membrane proteins whose structure has been determined. On the course website, you can download the FASTA file (`1f88.fasta`) containing the sequence, as well as a PyMol session file (`1f88.pse`) containing the structure. The rainbow coloring was generated using the PyMol command `util.chainbow`.

c. Construct a hydropathy profile for this protein in Excel and submit it to the web server. What segments would you predict to be transmembrane helices from your hydropathy plot? How does your prediction compare to the actual structure? (6 points)

Now we will look at a second protein, myoglobin. This protein is entirely alpha-helical, but it is not a membrane protein. The files you should examine are `1mbo.fasta` and `1mbo.pse`.

d. As before, construct a hydropathy profile for myoglobin and submit it to the web server. Comment on the predicted transmembrane segments and the actual helices in myoglobin. If a protein's localization to the membrane were not known for certain, how useful do you think hydropathy profiles would be? (6 points)