Introduction to Bioinformatics
Problem Set 3: Genome Sequencing and Genome Analysis

1. Rhodopsin is serves not only to sense light in eyes but also in bacteria, including our favorites, cyanobacteria. Let's take a look at the rhodopsin-like protein from the cyanobacterium *Anabaena* PCC 7120.

1a. Get the sequence of the *Anabaena* protein p-Alr3165 and obtain a Kyte-Doolittle hydrophobicity plot. How many transmembrane regions do you think the protein has? Provide your answer and a file containing the plot (right click on the image and SaveImage).

1b. Perhaps the results are marginal, so check them by running the sequence through DAS. What do you conclude? Provide your answer and a file containing the plot.

Interlude

Suppose you'd like to know what the specific amino acid sequence is of the first transmembrane span... that's not easy to tell from the plot you get from the web, so let's do it ourselves. Fortunately, the plot is calculated in a very simple-minded way. Take a closer look at the Kyte-Doolittle plot, particularly the information in the lower left corner. Note that the "window size" is 19 (was it?) and that the effective length is 243, the full length (261) minus 18. What this means is that 19 amino acids are considered at a time, beginning with the first amino acid and ending with the last.[1] The plot shows the average hydrophobicity for each window. We can do that… if we knew the hydrophobicities.

Now back up one screen to the main Kyte-Doolittle page. You'll see a link at the bottom to Amino Acid Hydrophobicity Scores. Click on that.

1c. Do the hydrophobicity scores make sense? What are the hydrophobicity scores of the two amino acids that are negatively charged? What are the hydrophobicities of the six[2] amino acids whose side chains consist solely of carbon and hydrogen? Provide your answers, both the names of the amino acids and their scores.

1d. Now do the following in BioBIKE:

- Define phytochrome-seq as the sequence of p-Alr3165
- Write a loop that drags a 9-amino-acid window through the sequence. You can do that by considering each amino acid position, starting with the first and ending with the last minus 18.
- Within each iteration of the loop, set a variable equal to the window-sized sequence of phytochrome-seq starting from the amino acid position under consideration and ending at the same position + 18.
- Set another iteration variable equal to the separate amino acids of the window-sized sequence. You can do this with the SPLIT function.

[1] So the first window will begin at amino acid #1 and extend to amino acid #19, then it will shift to amino acid #2 and extend to amino acid #20, and so forth until the end, where it will begin at amino acid #243 and extend to amino acid #261… always 19 amino acids long.

[2] You might think the number is 5 or 6, according to how you interpret the question. Either way.

- Set another iteration variable equal to the hydrophobicity scores of each of the separate amino acids. You can do this with the HYDROPHOBICITY-OF function. You can do this with a loop, but why not let BioBIKE do the work? Recall that a function that works with one thing should work with a list as well.
- Set another iteration variable equal to window score, i.e. the sum of the hydrophobicity scores of the list made in the previous step divided by the window size
- Set another iteration variable equal to the middle amino acid in the window. You can extract any character from the middle of a string of characters like so (by example:

```
(DEFINE word AS "ABCDEFGHI")
word[3]  --> "C"
```

- Display (using DISPLAY-LINE) the middle amino acid followed by a tab (*tab*) followed by the window score

Provide the code you wrote from the above outline and your conclusion (recall that the question was what amino acids lie in the first transmembrane span).

1e. That was for show. To get results that can be ***used***, remove the DISPLAY-LINE and instead COLLECT the window scores. After running this code, save your results to disk as a tab-delimited file (one of the formats readable by Excel). To do this, use the function `WRITE-TAB-DELIMITED-FILE`. Then click on **Browse BioFiles** (at the bottom of the page), on your directory, and finally on the file you just saved. Once it appears on the screen, use the appropriate browser function to save the file to your own computer. Then bring it up in Excel (open the file, check **Delimited** (if not already checked), **Next**, **Tab** (if not already checked), and **Finish**. Plot the results by selecting column A, then clicking on the Chart Wizard (a colored bar graph), choose **Line**, and smooth lines without points. Pretty up the graph if you like, and hand it in along with the code that produced the data.

1f. Replace the middle three amino acids of the first transmembrane region with three glutamates (DDD). To do this, use a variation on:

```
(INSERT "DDD" INTO phytochrome-seq FROM 47 TO 50)
```

Rerun the code with this new sequence. What is the effect of this modification? Provide your answer and the specific reasons (i.e. specific data) that led you to it.