Bioinformatics: Fall 2019
Homework Assignment 1

Milestone report due Sept. 20, 2017
Final product due Sept. 27, 2017

For this assignment, your overall task (working in teams) is to write a program that will (i) translate RNA sequences into amino acid sequences and (ii) *annotate* the resulting sequences: specifically, to identify putative transmembrane domains within each translated protein.

1. Write a function that imports RNA sequences in FASTA format and determines the corresponding amino acid sequence.

2. Thoroughly review the ***primary research literature*** (not Wikipedia or textbooks!) to identify specific features that characterize proteins' transmembrane domains. These features may include the presence of specific 2° structures, regions rich in specific biochemical classes of amino acids, etc.

3. Choose a set of features from your list above, and modify your code from part 1 to detect those features. Add appropriate front-end **AND** in-line documentation sufficient for a new user to understand exactly what each part of the code is doing. The program's output should clearly summarize the specific features for which you scanned, and the results of each such scan.

4. Use your modified program to scan the file *Assignment1Sequences.txt* for transmembrane domains, and print the output. For each of the three sequences, determine whether that sequence is likely to encode a membrane protein, and clearly explain your reasoning.

5. Write a short scientific report (7 pages maximum) summarizing your work. This report should clearly describe:
    a) the specific features you used to identify putative transmembrane domains,
    b) specific reasons why you chose those features rather than others,
    c) the results of scanning each sequence for the chosen features,
    d) an interpretation of these results, and
    e) a list of works cited.

    On a separate page, clearly indicate the specific contributions of each group member to this assignment. **Each group member must sign this statement to indicate agreement regarding his or her individual contribution.**

Milestone report (due at start of class on Sept. 20)

This report will consist of the following items, submitted as a **hard copy**:

1. A brief description of the specific methods your group plans to use to identify transmembrane domains (including your choices of key parameter values such as window size, hydrophobicity scale, etc.).
2. A list of your references, demonstrating a thorough review of the relevant literature and an understanding of what sources are appropriate to cite in professional work.
3. Your Python code for translating the given RNA sequences into amino acid sequences, including appropriate front-end and in-line documentation.
4. The output of your translation program for the three given RNA sequences.

Final write-up (due at start of class on Sept. 27)

This write-up will consist of the following items:

1. Your Python program, e-mailed to Drs. Weisstein and Beck as a file named *Bioinfo_Assignment1_Name1_Name2_Name3.py*, where each *Name* is the last name of a group member.
2. Your scientific report, attached to the same e-mail as a file with the same name but the suffix **.docx** .
3. A hard copy of your group's contribution statement, signed by each group member.

Expectations for all assignments and projects

Literature Review

1. Your literature review should demonstrate that you have read extensively on a particular research topic and are familiar with multiple approaches for conducting such research. See: *http://www.library.arizona.edu/help/tutorials/litreviews/index.html* .

2. All factual claims must be supported with appropriate in-line citations. **Exception:** factual claims that can reasonably be considered common knowledge. See: *http://homeworktips.about.com/od/researchandreference/qt/whentocite.htm* . Sources should represent a broad range of major studies relevant to the topic: be careful not to just repeatedly cite the same few sources. We will discuss citation format in class.

3. All references must be from ***peer-reviewed*** publications (e.g., articles from scientific journals). Textbooks, professional society websites, Wikipedia, etc. are NOT appropriate, although they may help lead you to more appropriate sources.

Main Body of Paper

4. Methods should be described in enough detail that the reader could repeat the analysis without having access to the Python code. For example, if a weighted sliding window approach is used, this section should explicitly state the window's length and the weight of each position within the window; if a hydrophobicity index is used, either tabulate the hydrophobicity of each amino acid or include a reference to such a table.

Python Code and Comments

5. Code should be as efficient as reasonably possible: for example, to run a long series of commands multiple times, use a loop rather than copy-and-pasting the command block repeatedly within the code.

6. Your program should also be flexible. Avoid hard-coding any parameters that might reasonably be expected to vary from one problem to another, but do hard-code any fixed biological parameters.

7. Comments should be sufficient for a naïve user to understand each code block and the entire program. Also note any essential points that might easily be overlooked or misunderstood. All code and comments should consistently use correct biological terminology.

Collaboration

8. Because this assignment is a collaborative project, **all** components should show clear evidence of substantive contribution from **each** group member. For example, even if a CS major writes most of the Python code, the biology major(s) should still read it over to suggest clarification and check for biological accuracy. Likewise, even if the biology major(s) write most of the paper, the CS major should read it over to ensure that the code's operation is described correctly.