Homework 9

Instructions: This homework is due at the beginning of lab, Wednesday, November 15th. You can edit this file and copy/paste into it by opening it in Libreoffice on Linux, PDFfiller Google App, Acrobat, MS Word, and others. Bioinformaticians are especially agile at using the internet to gather and share information, so if you do not remember or see the answer in class materials, feel free to use the web, but do not plagiarize it. Also, like anything you can obtain from the internet, information about bioinformatics may be wrong or only partially correct. Be sure to check the source of the information to help judge its quality. This particular homework will require you to write three Python scripts. I would like you to submit a digital copy of these scripts on Blackboard, along with your answers to questions and output files, all compressed into one .zip file. Make sure to run each script to ensure they work. I also encourage you to write and test each script one line at a time and test every variable you define. For all scripts please use argparse!

1. Answer a few quick questions about protein structure
    - a. What is the typical size of a protein (circle one)?
        - i. About the size of a virus
        - ii. About 1/100th the size of a virus
        - iii. About 1/100th the size of a human cell
        - iv. About 1/1,000th the size of a human cell
        - v. About 1/1,000,000th the size of a human cell
    - b. If amino acid backbones are all the same, what part of the amino acid makes the difference between hydrophobic and hydrophilic amino acids? What does hydrophobic and hydrophilic mean?
    - c. What are the only 2 degrees of freedom in the flexibility of an amino acid's backbone?
    - d. List the three kinds of secondary structures discussed in lecture.
    - e. Are parallel or antiparallel beta sheets more stable (in most environments)?
2. Download the PBD file for "2GB1" from www.RCSB.org. Write a Python script that uses the PBD file, extracts x,y, and z coordinates for all backbone atoms (N,CA,C) and outputs the information in a tab delimited file with the format:

    aaNumber atomNumber atomName x y z
3. Write a Python script that calculates the mean and standard deviation (if you cannot get numpy or scipy modules to work for standard deviation, you may use my function in stDev.py on Blackboard) of the following features and outputs the information into a tabular file: (tip: if you missed the lab there is help on this question in the slides)
    - a. All bond lengths: N_i-CA_i, CA_i-C_i, C_i-N_{i+1}
    - b. All bond angles (in degrees): N_i-CA_i-C_i, CA_i-C_i-N_{i+1}, C_i-N_{i+1}-CA_{i+1}
    - c. Distance between adjacent CAs (CA_i and CA_{i+1})

    For the output use the tab-delimited format:

    Category mean standardDev

4. Make a copy of your script from question 3 and modify it such that all the same statistics are calculated separately for each secondary structure described in the "HELIX" and "SHEET" lines in the PDB file. Based on the results, what are the main differences you see between the different kinds of secondary structures? How could you use this information to predict protein structure? For the script the output format must be tab delimited and as follows:

   kindOfSecondaryStructure	strand	sheetID	category	mean	standardDev

5. Write a script that calculates the torsional angles (in degrees) psi and phi for all amino acids in "2GB1" and outputs the information in a tab delimited file with the format:

   aaNumber	aaName	psiAngle	phiAngle

   For help calculating torsional angles see my example script, torsionalAngles.py, and the websites referenced in it on Blackboard.