Introduction to Bioinformatics – AS 250.265
Assignment 6

In the field of bioinformatics, it is important both to understand the biology of the sequences you are working with as well as the physics and chemistry. Understanding the biology is necessary for relating the sequence to the context in which it is working in the cell. Understanding the chemistry and physics is necessary for making inferences into how the protein is functioning in that context.

In this assignment, we will explore the biophysics of protein structure. Part of the assignment will test your ability to visualize proteins using the skills you learned in laboratory number three. The rest of the assignment will give you practice analyzing protein structures using some of the tools you have already learned. As you work through this assignment, you should keep in mind that, while our treatment of protein structures is largely geometrical, there are physical principles underlying why the protein adopts a specific folded conformation. One a base level, the principles of quantum mechanics describe the geometry of the bonds, and on other levels, hydrogen bonds, hydrophobicicty, and sterics explain why the three dimensional structure is as it is.

This assignment is worth 100 points toward the homework portion of the class. The laboratory questions you do in class constitute 20 of those points, and this handout makes up the remainder. Both parts are due on March 31 at the start of class. Electronic portions of the assignment should be submitted through the course website with appropriate filenames (e.g. `hw06-q1e.png`). The written portions of the assignment may be submitted by hand or uploaded as a Word document or PDF file.

Question 1: Laboratory Number 3 (20 points)

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

Question 2: Practice Using PyMol (30 points)

For the following questions, you will construct PyMol PNG files that display the scenarios described. You should upload these to the course web site, and answer the questions associated with each scenario. Your written answers may be uploaded to the web site or submitted by hand.

a. As discussed in class, the red dots that appear around the 1RBP molecule when you first load PyMol are crystallographic water molecules. What is the residue name used to describe these waters?

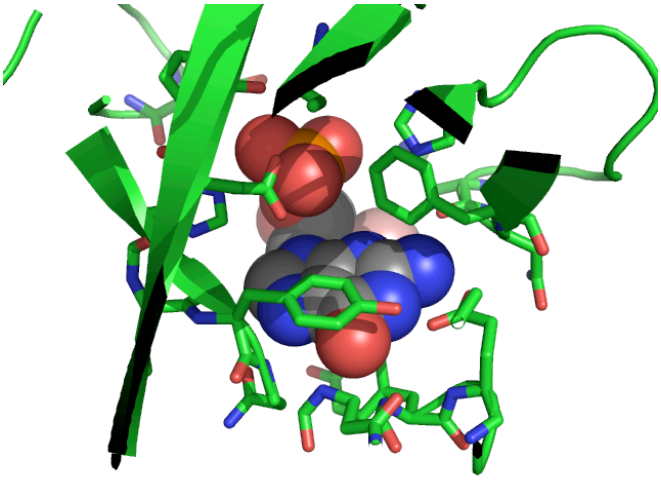
   Construct a representation of 1RBP where the protein is displayed as sticks and the water molecules are displayed as spheres. Color the water molecules yellow. The retinol in your display should be hidden. Save an image of your display and submit it to the course web server.

   Is most of the water in the interior or on the exterior of the protein? Give an example of a water molecule that appears on the interior of the protein. (10 points)

b. Construct an image of the retinol binding pocket in 1RBP. Display all non-hetero atoms within 5 angstroms of the retinol as spheres, hiding all other protein atoms. Display the retinol as a yellow molecule in the sticks representation. Make sure your final image is zoomed in appropriately, and submit the ray traced PNG image.

What is the distance of the closest protein atom to the retinol? (Hint: adjust the "within" selection you used to create the image until just one or two atoms appear, then find the closest using the `dist` command with an appropriate distance cutoff. (10 points)

c. Download the structure for ribonuclease T1 (1RNT) from the PDB. Ribonuclease (rnase) T1 is an enzyme that cleaves RNA at single stranded G bases. The following image highlights the active site region of the enzyme as viewed from inside the protein.



As closely as possible, reconstruct this view of the molecule. In this image, the view was centered on the G base, and the slab was adjusted until the active site was visible through the bulk of the protein. Sticks were drawn for all atoms within 7 angstroms of the G base, and the carbon atoms (element C) in the base were colored gray for clarity. Other protein atoms were rendered as cartoons.

Examining the structure, what two basic (positively-charged) residues compensate the negatively charged oxygen atoms in the phosphate group (PO_4^{-2})? Give both the residue names and indices in your answer. (10 points)

Question 3: Homology Modeling (30 points)

By now you should have received your results from SWISS-MODEL and 3D-JIGSAW. You should also at this time revisit the PHD results from lab three. You should save the contents of the SWISS-MODEL PDB file to your computer, and copy the contents of the 3D-JIGSAW email into a blank text file called "3djigsaw.pdb." You can view both of these files in PyMol and compare them with the original PDB entry, 1WM3.

The questions for this part of the assignment are rather open-ended, but are designed to prepare you for answering similar questions in your final project. For each question, please write no more than a paragraph. Content and clarity is what counts here (and will count in your project), not simply quantity.

a. How does the PHD prediction compare to the secondary structures predicted by the homology-modeling servers? Do the strands and helices correspond or not? Are there any systematic patterns? Given that the original test sequence was generated using a PAM model, can you explain the quality of the PHD prediction? (10 points)

b. How do the SWISS-PROT and 3D-JIGSAW structures compare to the original structure on a gross topological level? (Hint: compare the cartoon diagrams.) Are there any differences? Can you quantify your result in some way? (10 points)

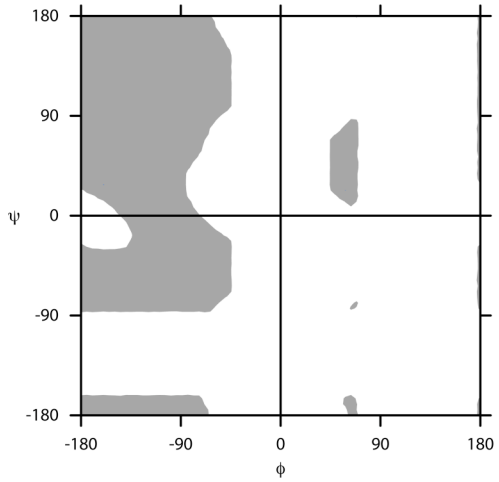
c. How do the SWISS-PROT and 3D-JIGSAW structures compare to the original structure on a microscopic level? (Hint: where residues are identical, compare the side-chain conformations between the structures.) Do you think that homology modeling can predict the chemistry of enzymes, which requires accurate positioning of active-site side chain atoms? (10 points)

Question 4: Ramachandran Plot (10 points)

a. Given that the backbone torsion ϕ is determined by the atoms C, N, C_α, and C, and that the backbone torsion angle ψ is determined by the atoms N, C_α, C, and O, what limitations are there in calculating ϕ and ψ for the ends of a protein? (3 points)

b. On the assignment six web page, you will find a file `rbp_torsions.txt` that contains the backbone torsions ϕ and ψ for retinol binding protein. Using Microsoft Excel, construct a Ramachandran scatter plot of this protein. Your plot should be on its own worksheet, with a title and labeled axes. Submit your Microsoft Excel workbook to the course website.

The "canonical" Ramachandran plot for alanine is shown below. Recall that this plot displays the sterically allowable ϕ and ψ combinations assuming all atoms are hard spheres and that no two atoms are allowed to be in the same place at the same time.



Comparing your graph to the graph of "allowable" conformations above, how well are proteins approximated by simple hard spheres? In other words, how often do you see ϕ and ψ combinations that fall outside of allowed Ramachandran space? What are some reasons you can think of why proteins might not strictly obey the Ramachandran plot? (7 points)

Question 5: Alpha Carbon Distances (10 points)

On the course web server, you will find a file called `cadist.py` that is intended to print the average C_α - C_α distance in a PDB file. Several parts of this program have already been written, including `read_pdb`, a function that reads a PDB file and returns a list of atoms. As an example of the usage of this function, consider the following snippet of code:

```
pdb = read_pdb('hw06_q5.pdb')

for atom in pdb:
     res_nam = atom[0]
     res_idx = atom[1]
     atm_nam = atom[2]
     x = atom[3]
     y = atom[4]
     z = atom[5]

     print 'Atom', atm_nam, 'is at', x, y, z
```

This code would print the name and location of each atom in the PDB file stored in `hw06_q5.pdb`.

In addition to `read_pdb`, the program also contains functions for calculating the average and variance of a list of numbers as well as calculating the distance between two atoms.

For this part of the assignment, you must complete the function `main` so that it calculates and displays the average distance and variance between two consecutive alpha carbons in a protein stored in a PDB file.

As you examine the `cadist.py` file, you'll notice that it is structured differently than other programs we have written in class. Specifically, it is able to take a filename as a command line argument. Thus, to run this program, you simply need to type the following at the command line:

```
python cadist.py hw06_q5.pdb
```

You do not have to understand exactly how this aspect of the program works; you should simply assume that the single argument to the `main` function is the filename you type on the command line. Other than that, you need only concern yourself with the functions `average`, `variance`, and `distance`, which have been included for your use.

a. [program] Fill in the `main` function to calculate average distance between consecutive alpha carbon (CA) atoms in the protein chain. You may assume that the only input your program will be given is the file `hw06_q5.pdb`, which is included on the website.

   Hint: Implement your solution so that it can store both the current alpha carbon as well as the previously identified carbon. Then, each time you find a new alpha carbon, append the distance between the two carbons to your list, which can be averaged at the end of your function. You may assume that CA atoms will be listed in the file in consecutive order and that no atoms will be missing. (8 points)

b. Given your answer to part (a), what is the ratio of the variance and the average value? Can you explain the size of the variance between alpha carbons? (Hint: what torsion angle is defined by C_α, C, N, C_α?) (2 points)