BIO/CMPSC 300 Introduction to Bioinformatics
Spring 2016

Kristen Webb and Janyl Jumadinova

http://cs.allgheny.edu/sites/jjumadinova/300

Lab 9

Given: Friday April 15, 2016

Due: Friday April 22, 2016

Objectives:

- **Know how to use available tools to examine the experimentally determined structures of proteins and visualize structural an functional features.**
- **Appreciate the value and limitations of predicting 3-D structure from sequence alone**

Reading Assignment

Chapter 11 in Exploring Bioinformatics textbook.

Required Deliverables (submitted through your Bitbucket repository):

- **Answers to Web Exploration questions**
- **A properly formatted and commented Python program and an output produced by the program.**

General Guidelines for Labs

Work on the Alden Hall computers. If you want to work on a different machine, be sure to transfer your programs to the Alden machines and re-run them before submitting.

Keep all of your files! Don't delete your programs and reports after you hand them in---you might need them again later.

Back up your files regularly. Use a flash drive or Google Drive or whatever your favorite backup method is.

Review the Honor Code policy on the syllabus. Remember that you may discuss experiments and programs with others, but copying answers or programs is a violation of the Honor Code

Part I: Predicting Secondary Structure from Amino-Acid Sequence

HIV and AIDS have been a major focus of pharmaceutical discovery for more than 25 years, and indeed we have developed an unprecedented number of new antivirals, some of which resulted from the study of protein structure and rational design. In this lab, we focus on the HIV protease – an enzyme that cleaves a polyprotein product into individual functional protein units. Understanding the 3-D structure and determining the location of the active site of the HIV protease could aid in the development of an antiviral drug to combat HIV infections.

Ab initio prediction is the prediction of a protein's tertiary structure from the amino acid sequence alone. A number of approaches have been developed to tackle this problem. Today you will be using a web-based tool called PSIPRED. PSIPRED uses a neural network algorithm and integrates both a Chou-Fasman-like prediction algorithm and comparative data obtained through BLAST searches of the NCBI protein database to look for regions of the protein likely to form α-helices, β-sheets, or random coils.

1. Go to the Protein Data Bank homepage and enter 1KJF in the search bar. This is the Protein Data Bank ID (PDB ID) for the protein HIV-1 protease.

2. Click on the Download Files link and download the FASTA-formatted amino acid sequence as a text file.

3. Go to the PSIPRED homepage (http://bioinf.cs.ucl.ac.uk/psipred/)

4. From the PSIPRED page, confirm that the PSIPRED v3.3 Predict Secondary Structure method is selected.

5. Open the text file and copy the first FASTA-formatted sequences (1KJF:A). Enter the HIV-1 protease amino acid sequence, your email, and a short identifier for the submission and submit your request. Repeat with the second and third sequences. You should get an email within half an hour or less indicating the job is complete. Move on to part II of the lab while you wait.

6. When your results are ready, you can example the results in text form in the email or graphically by clicking the email link. Either way, you should see that each amino acid in the protein has been assigned a letter indicating whether it is predicted to be in an alpha (H)elix, a strand of a beta sh(E)et, or a random (C)oil.

7. Each amino acid also has a number indicating the statistical confidence of the prediction (nine is the highest).

8. In the graphical version (the PDF file provides the nicest view), the confidence value is replaced by a bar whose height shows the level of confidence, and the α-helices and β-sheets are shown graphically with cylinders and arrows, respectively.

9. Download and save your results for easy comparison.

Part II: Exploring the Structure of the HIV protease

When the structure of a protein is "solved", we know where the atoms that make up its amino acids are found in space, allowing us to generate representations that show the locations of the various amino-acid side chains and how they interact to form secondary and tertiary structures. X-ray crystallography is the current gold standard for protein structure and can under the best conditions distinguish the positions of less than 10^{-10} meters apart. Structural data are deposited in public databases, most notably the Protein Data Bank (PDB), in a standardized format that can be read by various kinds of software to visualize and work with the structures. You will use the Protein Data Bank file to explore the HIV-1 protease using the powerful visualization tool FirstGlance in Jmol.

1. Go to the FirstGlance in Jmol website (http://bioinformatics.org/firstglance/fgij/) and enter the PDB identification code for the HIV-1 protease (1KJF). When the applet loads, you should see the protease structure in a "cartoon" view where α-helices are shown as by spiral ribbons (arrows point toward the C-terminus of the protein) and β-sheets by parallel flat ribbons. Unstructured areas of the protein look like thin ropes.

2. When the program starts, the protein is rotating to show you the three-dimensional view; click on the spin button in the menu at the left to stop it. If ligands is selected, unselect that as well.

3. Notice there are thee different colors used to represent the structures of the protein (light blue, light green, and dark green). You should see that two colors represent two polypeptides with the same structure joined together. The third color shows a short peptide that represents a segment of the protein substrate in the active site of the enzyme.

4. On the View tab, click Secondary Structure. Now the α-helices, β-sheets, and random coils have distinct colors. Mousing over the other links will provide a brief description. Explore the other viewing options in the Views tab. Reset the view when you're finished by clicking on the Reset link, stopping the spin and unselecting the ligands.

5. In addition to these preset views, there are additional viewing options that can be accessed by right-clicking on the molecule. Right-click on the structure window and choose Select | All then Style | Structures | Backbone. You should now be able to see the peptide backbone of the molecule.

6. To better distinguish between the chains, right-click and choose Select | All then Style | Scheme | CPK Spacefill to show the space-filling model and Select | All then Color | Atoms | By Scheme | Chain to highlight the individual chains. Now click on molecule and watch the display at the bottom to see which amino acids you have chosen and where they are on the chain.

7. These are just a few examples of how the 3-D structure of the protein can be viewed and analyzed. You will need to continue to explore the viewing options to answer the Web Exploration Questions below.

Web Exploration Questions

a. Compare the 3-D structure of the HIV-1 protease as displayed using FirstGlace in Jmol with the prediction results from PSIPRED (if you have not yet received your results, .pdf files can be downloaded from the shared repository). How well did PSIPRED predict the secondary structures of the HIV protease? Provide specific examples of structures predicted accurately by PSIPRED, predicted structures not found in the actual structure, and actual structures not predicted (one example of each, as

applicable).

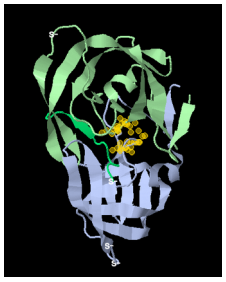
The PSIPRED prediction for this protein is really very accurate. As shown in the table below, nearly all of the structures found by PSIPRED were actually present in the crystal structure—even some unlikely-seeming short β-strands. There were some small differences in where these structures begin and end, however. The major difference is that PSIPRED predicted an α-helix at amino acids 20-25 that does not exist in the crystal structure and is in fact part of an adjacent β-strand. PSIPRED also split a long β-strand from 51-67 into two pieces. On the whole, however, one would have to say that this *de novo* prediction is quite accurate.

structure	PSIPRED	crystal structure
b-strand	3-4	2-5
b-strand	10-15	10-16
b-strand	18-19	17-25
a-helix	20-25	
b-strand	32-35	32-35
b-strand	45-48	42-49
b-strand	52-59	51-67
b-strand	62-67	
b-strand	70-77	69-78
b-strand	84-85	84-85
a-helix	86-93	86-94
b-strand	96-97	96-99

b. PSIPRED uses a prediction algorithm not unlike the Chou-Fasman algorithm we discussed in class. However, instead of applying the algorithm directly to your input sequence, it first does a BLAST search to get a collection of sequences related to your input. It then applies its prediction algorithm to the results. Why might this method be advantageous in improving the program's ability to identify genuine secondary structure?

If a given amino acid is part of an actual secondary structure, such as an α-helix, then we would expect that mutations that change this to an amino acid with poor α-helix potential would be selected against and amino acids with good α-helix potential would tend to be conserved at this position. By using a sequence comparison to look at conservation prior to predicting structures, PSIPRED can “filter out” amino acids that don’t genuinely contribute to structures and thus improve its accuracy.

c. The HIV protease is a member of the aspartyl protease family that can be recognized by the three-amino-acid motif Asp-Thr-Gly. Normally, the HIV protease contains this motif, but in order to obtain a crystal structure with a peptide in the active site, a mutation changing the Asp to Asn (structurally similar) was used for the 1KJF structure. Using FirstGlance in Jmol, locate the chains Asn-Thr-Gly protease motif in 1KJF. Copy and paste a screen shot of the 1KJF molecule where the chains are clearly displayed. (hint: using the FirstGlace in Jmol “Find” feature may be helpful here).



d. What are the numbers of the amino acids, and thus the location of the active site, on each chain that form the Asn-Thr-Gly protease motif in 1KJF?

The protease motif is Asn 25, Thr 26 and Gly 27 on each subunit.

Part III: Exploring the Structure of the HIV protease through BioPython

1. Write a Python program that calculates the alpha-carbon (CA) to nitrogen (N) distances of HIV protease studied in parts I and II. You can calculate the distance between atoms in BioPython by using a simple subtraction, as `atom1 – atom2`.

2. Plot the distribution of the calculated distances using Python. To review graphing functionality of Python, you can look back at the program BioPythonBlast4 from March3 class, Python Syntax Guide handout for Chapter 4, and/or Python documentation for matplotlib module in Python.

3. In your opinion, would you expect this distribution to be consistent across all proteins. Provide 2-3 sentence rationale for your answer.