**CMPSC 300**
**Introduction to Bioinformatics**
**Fall 2017**

**Lab 8: Investigating Protein Structural Qualities and Details**
**Save this lab assignment to:** `labs/lab8`

## Objectives

- Know how to use available tools (both, from web-sites and stand-alone) to examine the experimentally determined structures of proteins and to visualize their 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.

## 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 its 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.

An *ab initio* prediction is the prediction of a protein's tertiary structure using only its amino acid sequence. A number of approaches have been developed to tackle this problem. Today you will be using a web-based tool called PSIPRED which uses a neural network algorithms 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 that are likely to form $\alpha$-helices, $\beta$-sheets, or random coils.

## Steps to Take

1. Go to the *Protein Data Bank* homepage `https://www.rcsb.org/` 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. Note: You are likely to have more than one protein in your FASTA file. Make sure that you are entering one at a time into the *PSIPRED* online tool in the next steps.
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. **Please be patient; 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 -helices and -sheets are shown graphically with cylinders and arrows, respectively.

9. Download and save your results for easy comparison.

## 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.

## Steps to Take

1. Go to the *FirstGlance in Jmol* website (`http://bioinformatics.org/firstglance/fgij/`) and enter of the PDB identification code (as from above: 1KJF ) for the HIV-1 protease. When the applet loads, you should see the protease structure in a "cartoon" view where $\alpha$-helices are shown as by spiral ribbons (arrows point toward the C-terminus of the protein) and $\beta$-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 three 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 *Views* tab, click *Secondary Structure*. Now the $\alpha$-helices, $\beta$-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 are finished by clicking on the Reset link, stopping the spin and un-selecting 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

1. What do these the molecules of your work look like, according to each tool? Place these screen shots into your LibreOffice file document (Please do not use MS Word for your submitted document as the instructor does not have this software to open and read your file.)

2. 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).

3. *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?

4. 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.

5. 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?

### Required Deliverables

All of the deliverables specified below should be placed into a new folder named 'lab08' in your Bitbucket repository (`cs300f2017-bbill`) and shared with the instructor by correctly using appropriate Git commands, such as `git add -A`, `git commit -m ''your message''` and `git push` to send your documents to the Bitbucket's server. When you have finished, please ensure that you have sent your files correctly to the Bitbucket Web site by checking the `source` files. This will show you your recently pushed files on their web site. Please ask questions, if necessary.

1. A document responding to the questions above, including screenshots of the molecules of your work. Please include these screen shots of the molecule structures of your work. Be sure to places these graphics directly into your document. Do not leave them as scattered files in your repository.

2. Make sure to submit a LibreOffice file, not a text file or MS Word file, with proper formatting and **your name included at the top of the document**.

Please see the instructor, if you have questions about assignment submission.