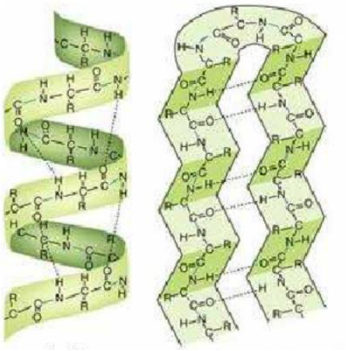
CSCI 474/597K
Lab 4a : Protein Structure, PDB Files
Spring 2019

WESTERN WASHINGTON UNIVERSITY
Computer Science



In an earlier lab we took a look at a PDB file. In this lab, we'll explore PDB structure files 1ody, 1hhp, and 1bve more in-depth than the PDB file you explore in the first lab. Specifically, we'll look at the "raw" protein structure data. This lab will serve as a good motivation for the need of visualization tools which we'll learn about in the next portion of the class.

As you've done in the past, please compose answers to the questions throughout this lab, and submit your answers via Canvas as proof of completing this lab. Most questions ask you to provide an answer for each of the three structure files -- 1ody, 1hhp, and 1bve.

The DPB is a rich repository of protein (and other molecules) structure data. It contains a wealth of (excellent) educational materials, too. PDB-101 is an especially good resource:

- https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/
- https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/dealing-with-coordinates
- https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/biological-assemblies

I. Preliminaries

Questions 1-3 can be answered without looking at plain-text structure data files. It is enough to look on the "front page" of each PDB entry.

Most questions have multiple parts – be sure to answer EACH part of each question

Q1 : What are the proteins in the files 1ody, 1hhp, and 1bve? Be concise, but provide more than 1- or 2-word answers. You may look elsewhere (online web search) to learn about the protein, but don't over do it :)

Q2 : What experimental method was used to collect the structure data/information contained in the three PDB files. Provide both the acronym, and full name, and a brief description of the experimental method (for the non-bio and non-chem students, this will require a web search). Brief means a sentence or two or three.

Q3 : What is the resolution for each of the three files? What is the resolution unit, and how many of those units are equivalent to a single meter? Who is the resolution unit named after? Briefly explain what it means for a file to have a specific resolution.

II. FASTA sequence

Look at the FASTA sequence of each of the three files.

Q4 : How many chains and residues/amino acids are in the FASTA sequence of each file?

Q5 : The three files are of the "same" protein, but they have different counts of chains. Explain this (briefly). Hints : Ask Filip for help, do a web search, or buddy-up with a chem/bio student who has biochemistry experience. Or, look at the "Biological assemblies" portion of PDB-101.

III. PDB File Format

There is a great deal of data in each PDB file. Not only data about the atoms in a protein, but data about the experimental methods used. The formatting of a DPB file -- in all its gory details -- is described at the following URL:

- http://www.wwpdb.org/documentation/file-format

There is MUCH more information there (in many formats) that is digestible in a reasonable amount of time. However, to help you answer **most** (but not all) of the remaining questions in this lab, pay attention to the following sections:

- http://www.wwpdb.org/documentation/file-format-content/format33/sect9.html#ATOM
- http://www.wwpdb.org/documentation/file-format-content/format33/sect9.html#HETATM
- http://www.wwpdb.org/documentation/file-format-content/format33/remarks.html
- http://www.wwpdb.org/documentation/file-format-content/format33/remarks2.html
- http://www.wwpdb.org/documentation/file-format-content/format33/sect5.html#SHEET
- http://www.wwpdb.org/documentation/file-format-content/format33/sect5.html#HELIX

Look closely at those resources. What does "column" refer to?

IV. PDB Files

Inspect the three PDB files (view Download -> PDB Format).

Q6 : How many total "structures" are in each file? For example, is there a single chain, multiple chains, multiple models, ligands/drugs? Provide both a list of "structures" in each file, as well as their designations (for example *chain A* and *chain B*). List them all for each of the three PDB files (except water molecules). **Hint : look at the front-page as well as data file for each PDB entry to find these details.**

Experimental methods are not perfect, and/or sometimes the experimental method conditions are such that the data gathered is inaccurate or imprecise. In other cases, the experimental method -- due to a variety of factors -- is not able to fully resolve individual atoms, and rather than guessing the coordinates of individual atoms, the person/lab submitting to the PDB designates that this-and-that atom or residues is/was not resolved.

Q7 : Do any of the three DPB files contain missing residues or atoms? If so, list them. Be sure to identify the PDB file(s) which have missing the atom(s) and/or residue(s).

In lecture we've briefly discussed the alpha helix and beta sheet, the two secondary structures found among proteins.

Q8 : How many alpha helices are in each PDB file? Specify the numerical count. For example, 3 alpha helices. If a file has more than 1 model, answer this question for the first model only. Hint : look for the HELIX keywords.

Q9 : How many residues is the longest alpha helix for each of the three PDB files? What are the residues that make up the longest alpha helix?

Q10 : How many sheets are in each PDB file? Specify the numerical count. For example, 3 beta sheets. If a file has more than 1 model, answer this question for the first model only. Hint : look for the SHEET keywords.

Q11 : How many residues is the largest sheet for each of the three PDB files?

V. ATOM and HETATM data

"The" atomic structure data in the PDB file is most often contained in lines that begin with ATOM or HETATM.

Q12 : What is the difference between a line that starts with ATOM, versus a line that starts with HETATM.

Q13 : What are the first three residues (amino acids) in each DPB file? Provide the single letter abbreviation, three letter abbreviation, as well as full name.

Atom names are either stand-alone names (such as N, C, O), or letter/letter-word combinations such as CG, OE1, and CG2.

Q14 : What columns in a line that begins with ATOM contain the atom name data?

Q15 : What do single-letter atoms (such as C and N) refer to? What makes them different from other atom names that are made up for 2 and 3 letters/characters? Don't answer "they have a single letter."

Q16 : If an atom name is made up of 2 or 3 characters, what does the second character of the atom name represent?

Q17 : For atom names that are made up for 2 or 3 characters, what are all of the possible letters that are in the second character position of the atom name? What are they abbreviations of?

Q18 : For atom 50 of each of the three PDB files, what is :

- the residue (provide the three letter abbreviation and full name) to which the atom belongs
- the x, y, and z coordinates of the atom
- the chain identifier of the residue that the atom belongs to
- the residue number that the atom belongs to
- the name of the atom and its position name (epislon 1, for example)

V. Submission and Rubric

Submit your answers to the 10 questions to Canvas.

Component of Lab	Points
Answers to questions 1-18	36
Total	36 points