Structural Bioinformatics
Assignment 1 - Secondary structure assignment

Sanne Abeln, Maurits Dijkstra and Juami van Gils

February 3, 2023

Introduction

The aim of this practical is to *assign* secondary structure, given the PDB coordinates of a protein structure. Note that secondary structure assignment is something fundamentally different from secondary structure prediction - please make sure you are well aware of this difference.

Note that most students find section 1 of this assignment most difficult. If you get stuck, please continue with section 2 and ask for help in the next class.

Assignment

Make sure you run the scripts with the main folder ('Assignment1') as your working directory, to ensure that the output files are written into the 'Output' folder. Examples of how to run the scripts are provided in each of the Python files.

Since a Canvas quiz is not reliable for saving your answers, please keep them in a separate document and copy them into the quiz when you submit. Please keep your answers concise. Submit your two scripts to CodeGrade via Canvas.

1 Calculating backbone dihedral angles using coordinates

1.1 Calculation of Phi and Psi dihedral angles

Write out on paper how you can calculate the phi and psi dihedral angles, as defined by the IUPAC standards, using the coordinates of the backbone atoms from a PDB file. You will need to use the equations in the Appendix. Make sure to define each variable in each equation you state. Have a look at the lecture slides for a more detailed explanation of phi- and psi-angles and how to calculate them. You will need these calculations for your implementation of Q1.3.

1.2 Strategy for assigning secondary structure [5 points]

Describe a strategy for assigning secondary structure based on phi and psi angles. You should assign the secondary structure type for each residue, given the backbone coordinates of a PDB file. The classification should be based on the phi and psi angles of the residue from the equations above; you may use the idea of a Ramachandran plot to define your assignment criteria. Here, a classification in three groups [alpha, beta, loop] would be sufficient. Please use max 100 words.

Hint: have a look at the next question(s) before you start.

1.3 Implement your strategy for assigning secondary structure [20 points]

Implement a program that can assign the secondary structure type for each residues given the backbone coordinates of a PDB file.

You can use the attached Python script (readPDB.py). Look for the lines: "### START CODING HERE" (3x).
You need to hand in your modified code via Codegrade.

1.4 Discussion of secondary structure assignment strategy [20 points]

Have a detailed look at the assignment of secondary structure by your program. Inspect your protein through a viewer, e.g. Chimera. Specifically, consider the length of the secondary structure elements. You may also compare your secondary structure assignment to output of the DSSP. Show an example of secondary structure element assigned by your program that does not seem correct to you. Discuss the quality of your results and possible problems. Suggest improvements to your assignment strategy (to overcome these problems). Please use max 250 words.

1.5 Dihedrals versus hydrogen bonds [Discussion (ungraded)]

No currently available secondary structure assignment program (e.g. Stride or DSSP), relies solely on dihedral angles for their assignment. What are the pros and cons of using hydrogen bonds versus dihedral angles for secondary structure assignments?

2 Propensities for amino acids to be buried

2.1 Implement propensities to be buried [5 points]

Write a script that can obtain the solvent accessible area from DSSP files to calculate the propensity for each amino acid to be buried. See section 9.4 of the book and the lecture slides for an explanation of how to calculate propensities. You can define a residue as buried when the relative surface accessibility of its side chain is less than 7%.

You can use the template file provided (readDSSP.py), look for the lines '### START CODING HERE' in the functions 'decide_if_buried' and 'print_propensities'. You can ignore the 'read_dir' function for now. For quicker debugging, you can run your script on the small library at this stage.
IMPORTANT: Look at the information below before handing in your script!

2.2 'Special' amino acids [Task and Discussion question (ungraded)]

When you run the script, you will get key errors for unknown amino acids. What does every type of unknown amino acid in the small library and large library represent?. Hint: you may want to look at documentation about the DSSP file format.When you run the script, you will get key errors for unknown amino acids. Investigate what these are and how to replace them.Now also modify the 'read_dir' function in the indicated block. Make sure your script now works on the large library as well.

2.3 Script to determine amino acid propensities to be buried [10 points]

Hand in your script on Codegrade via the Canvas Quiz.

2.4 Discussion on propensities [10 points]

Look at the propensities you find for the group of hydrophobic amino acids. What was your expectation for this group? Discuss the amino acids propensities that do not match your expectation based on the chemical properties of this group and the biological environment of the protein [max 100].

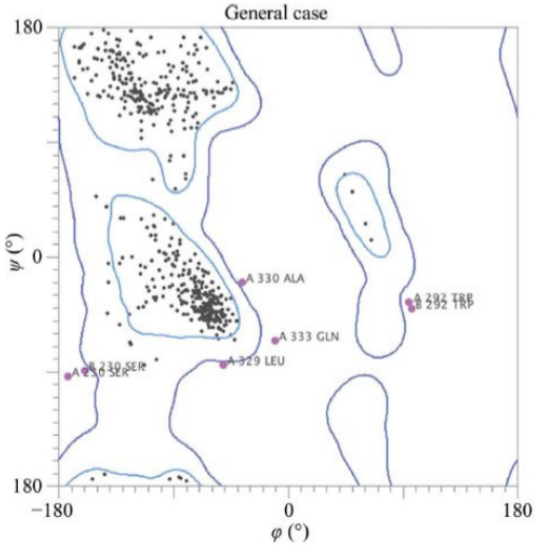
2.5 Discussion on database used [10 points]

What possible effect could the database, which you have used to calculate propensities, have on your results? What kind of biases may there be in the database, and what effect would they have on your propensities? Please do not exceed 100 words in your explanation.

3 Ramachandran plots

3.1 Contour lines [10 points]

Have a look at the Ramachandran plot shown below. For each of the clusters there are so-called contour lines that indicate when a point can be considered part of the cluster. How are these contour lines drawn? What do the datapoints in the plot represent? Explain why the data is distributed like this. Please do not exceed 100 words in your explanation.



3.2 A different dataset [10 points]

Consider a case in which you base your Ramachandran plot on a reference set of structures which score favourably according to MolProbity. How do you expect the contour lines to change? Why do you expect this change? Please do not exceed 100 words in your explanation.

Appendix

Helpful equations

A dihedral angle is the angle between two intersecting planes. To calculate this angle, here are some equations you can use:

Normal vector

A normal vector to a plane is a vector that makes a 90 degree angle with the plane. The normal vector, $\vec{n}$, to a plane can be calculated as follows:

$$\vec{n} = \vec{b_1} \times \vec{b_2} \tag{1}$$

where $\vec{b_1}$ and $\vec{b_2}$ are two vectors that determine the plane, i.e. two vectors that lie in the plane and are not parallel to each other.

Unit vector

A unit vector is a vector of length one. We can obtain a unit vector, $\hat{u}$ from a vector $\vec{v}$, by normalising it by its own length:

$$\hat{u} = \frac{\vec{v}}{\|\vec{v}\|} \tag{2}$$

Angle between vectors

An angle ϕ between two vectors $\vec{v_1}$ and $\vec{v_2}$ can be calculated as follows:

$$\cos(\phi) = \frac{\vec{v_1} \cdot \vec{v_2}}{\|\vec{v_1}\|\|\vec{v_2}\|} \tag{3}$$

$$\sin(\phi)\hat{u} = \frac{\vec{v_1} \times \vec{v_2}}{\|\vec{v_1}\|\|\vec{v_2}\|} \tag{4}$$

where $\hat{u}$ is the unit normal vector of the plane defined by $\vec{v_1}$ and $\vec{v_2}$

Atan2 function

The atan2 function, implemented in python through math.atan2() is also a useful function when determining an angle:

$$\phi = \text{atan2}(\sin(\phi), \cos(\phi)) \tag{5}$$