Homework Assignment

Scientific Python Course

Orsolya Réka Molnár (MS1YRR)

István Szepesi-Nagy (K45SFS)

2022. 05. 11.

Orsolya Réka Molnár (MS1YRR)
István Szepesi-Nagy (K45SFS)

1 Introduction

The aim of the project is to identify the structural domains of a given protein. The protein PDB file can be imported from locally or from the PDB database.

The program calculates the least connected parts in the protein and plots the results on the GUI. The workload is distributed evenly, but Orsolya will be mainly responsible for the GUI implementation, while István will be responsible for the domain finder algorithm implementation. For version control and cooperative work, GitHub will be used.

2 DOMAK algorithm

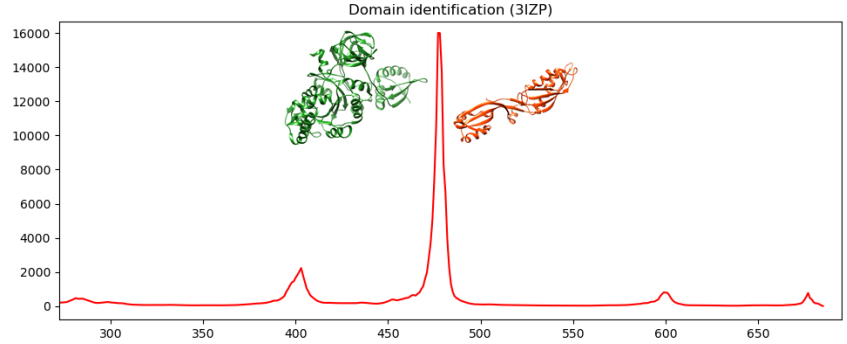
The DOMAK algorithm is used to create the 3D domain database. It calculates a 'split value' from the number of each type of contact when the protein is divided arbitrarily into two parts. This split value is large when the two parts of the structure are distinct. The split value can be any number, but in this task it is fixed to 8 Ångstroms distance.

The algorithm iterates through the PDB file and stores the coordinates of all Cα atoms. Than calculates the distances and only under the split value are kept. Lastly the inter - and intradomain contact parts are calculated within the matrix.

$$\frac{Intra_A}{Inter_{AB}} \cdot \frac{Intra_B}{Inter_{AB}}$$

1. Equation - DOMAK formula.

The result is a plot which represents the upper mentioned value for residue in the protein. High peaks represent good quality separation place between protein domains.



1. Figure - Example plot result for protein 3IZP.

3 The function of the program

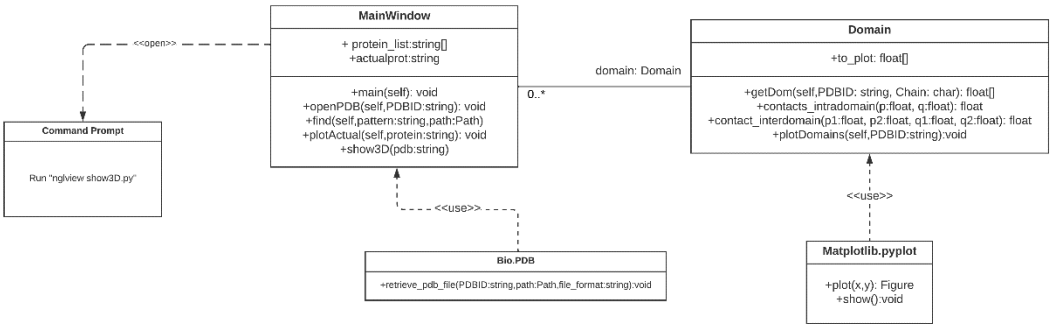
The Python program has three main functions implemented:

- PDB file download from webserver
- Domain calculation and plotting
- 3D representation of protein with jupyter-notebook

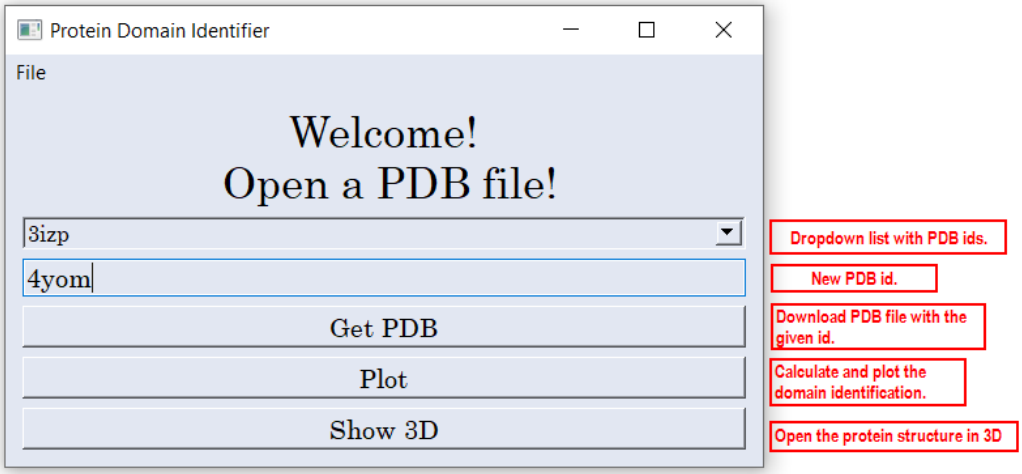
PDB file collection is used with BioPython's PBDList class, where the desired file is downloaded in *.ent format.

For the plotting Matplotlib's Pyplot library was used. It gets the parameters from an integer list containing the DOMAK values.

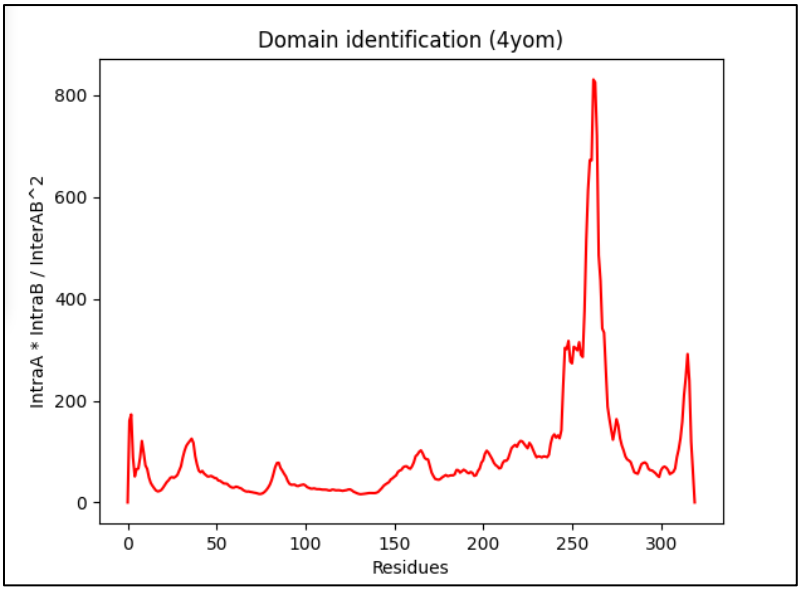
For the 3D representation, a jupyter-notebook widget was used (nglview library), which can open and represent the PDB structure in 3D. The notebook is started from the GUi, which starts a command prompt to start the jupyter-notebook. Unfortunately, automated start of the notebook is not available, so human interaction is necessary.



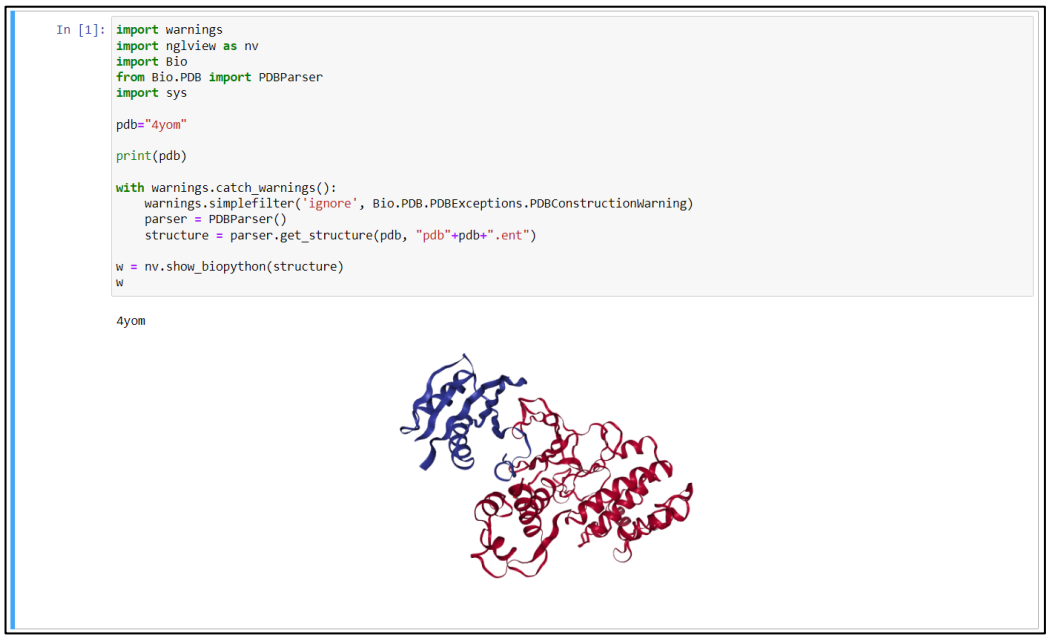
2. Figure - UML diagram of the program with basic functions also represented.



3. Figure - Graphical user interface with described functions.



4. Figure - Example plot results of 4YOM protein.



5. Figure - 3D representation in jupyter-notebook of 4YOM opened from the GUI.