Alberto	*Crivellari*	*2061934*

Midterm test No. 2

09 / 05 / 2022

Please answer all questions below and submit this document in **PDF format** by **23 May 2022, 12:30 PM** (two weeks after) to **damiano.piovesan@unipd.it**. Each student is assigned a different **protein structure (<PDB ID>_<chain ID>)**. The first part of the test is open questions, the second part is based on the analysis of the assigned structure. **Assignment file here.**

Answer the following questions concisely (**max 500 words in total**).

1. What is the relationship between the sequence similarity and structure similarity in biological proteins?

   - Proteins with at least 30% of sequence similarity implies that they have similar structure, so high sequence similarity usually means high structure similarity.

2. What are the main steps in homology modelling?

   - Identify related structures;
   - select templates;
   - align target sequence with template structures;
   - build model;
   - evaluate model

3. How can you measure the quality of a structural alignment?

   - calculating RMSD and kabsch algorithm

4. What are the differences between globular and intrinsically disordered proteins in terms of amino acid composition?

   - Globular ones are mostly hydrophobic while intrinsically disordered ones have low hydrophobicity

Download the assigned PDB structure and consider only **standard (non-hetero) residues** of the specified chain (<PDB ID>_<chain ID>). Calculate the contact map (question 1) and the conformational energy (questions 2 and 3) as described in the IUPRED paper. The *M* and *P* matrices are available from the ***iupred_data.py***. The smoothed energy is the moving average of the raw energy over a window of 21 residues (±10 residues around the current position).

1. Calculate and plot the contact map of your chain. Use the ***NeighborSearch*** module and the ***search_all(3.5, level="R")*** method. Consider only contacts between positions with a **sequence separation ≥ 2**.

   - Retrieve the structure and the chain, 2k5d model 0 chain A
   - Use neighborSearch and search_all with seqsep 2 to get all pairs of contacting atoms
   - Fill a matrix 110x110 (number of standard residues) and i put 1 for contact, 10 for others
   - Plot the contact map

2. Calculate the **exact energy** of each residue based on the weighted contribution of its **contacts** (as calculated above) and plot the raw and smoothed energy for each residue on the same figure. Use the ***M matrix*** to calculate the contact energy.

   - Start from the function iupred in the practical lesson, and modify it by basing the weight on contacts of the contact map

   - Use m matrix of iupred data to calculate the energy and plot both raw and smoothed energy

3. Calculate the **estimated energy** of each residue based on the weighted contribution of the **frequency of neighboring residues** in the sequence and plot the raw and smoothed energy for each residue on the same figure. Use the ***P matrix*** to calculate the estimated energy. Neighboring residues are those 2-100 residues apart from the current position.

   - Simpler one, can just use the function iupred from the notebook of the practical lesson

4. Report the **disorder content** for the two different calculations. Disorder content can be calculated as the fraction of **residues with positive energy** (≥ 0) over the length of the sequence. Report both the fraction and the raw count of residues with positive energy.

   - We calculate disorder content for both predicted energy and exact energy of the structure
   - Exact energy: 63 raw count, 0.5727272727272728 disorder content
   - Predicted energy: 57 raw count, 0.5181818181818182 disorder content