Matteo	Andriolo	2060889

Midterm test No. 2

03 / 05 / 2023

Please complete the following tasks and submit the document in **PDF format** to damiano.piovesan@unipd.it by **12:30 PM on May 17, 2022**, two weeks after the deadline. Please include your **surname** in the **file name**.

protein structure (<PDB ID>_<chain ID>) assigned =(2k5d_A).

Please answer the following questions concisely, with a maximum of **500 words** in total:

1. What is the relationship between sequence similarity and structure similarity in biological proteins?
2. What are the main steps involved in homology modelling?
3. How can you measure the quality of a structural alignment?
4. What are the differences, in terms of amino acid composition, between globular and intrinsically disordered proteins?

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).

Complete the following tasks:

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**.
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.
3. Calculate the **estimated energy** of each residue based on the weighted contribution of the **frequency of neighboring amino acids** 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.
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. Please report both the fraction and the raw count of residues with positive energy.

Theoretical questions

1. Sequence similarity refers to the extent of similarity between two protein sequences, while structure similarity refers to the resemblance of their 3D structures. High sequence similarity often implies high structure similarity, as similar sequences tend to fold into similar structures. However, proteins with low sequence similarity can still have similar structures due to convergent evolution. Any random pair of natural sequences have at least 15% sequence identity. Generally, protein with at least 30% identical residues have similar structure (shorter alignment have higher threshold).

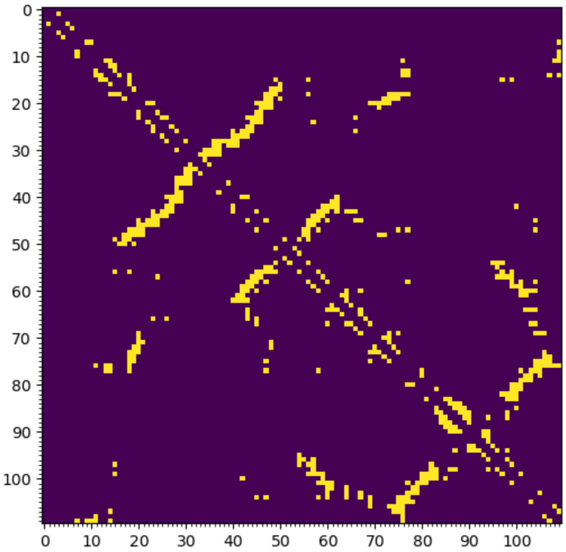
2. Homology modeling involves four main steps: (a) template identification (select protein with high sequence similarity); (b) sequence alignment (align target and template sequences) (c) model construction (the target protein's 3D structure is generated based on the template's structure); (d) model evaluation (the model's quality and accuracy are assessed).

3. The quality of a structural alignment can be measured using various metrics, including RMSD (root-mean-square deviation), which quantifies the average distance between corresponding atoms, and TM-score (template modeling score), which weights also the residue pairs at smaller distances relatively stronger than those at larger distances

4. Globular proteins have a well-defined, compact 3D structure with a hydrophobic core and hydrophilic surface, and contain a higher proportion of hydrophobic and secondary structure-forming amino acids. In contrast, intrinsically disordered proteins lack a stable 3D structure, have a higher proportion of hydrophilic and disorder-promoting amino acids, and are more flexible and dynamic in nature (have higher complexity calculated using entropy - lots of different aminoacids in the same fragment)

Practical questions

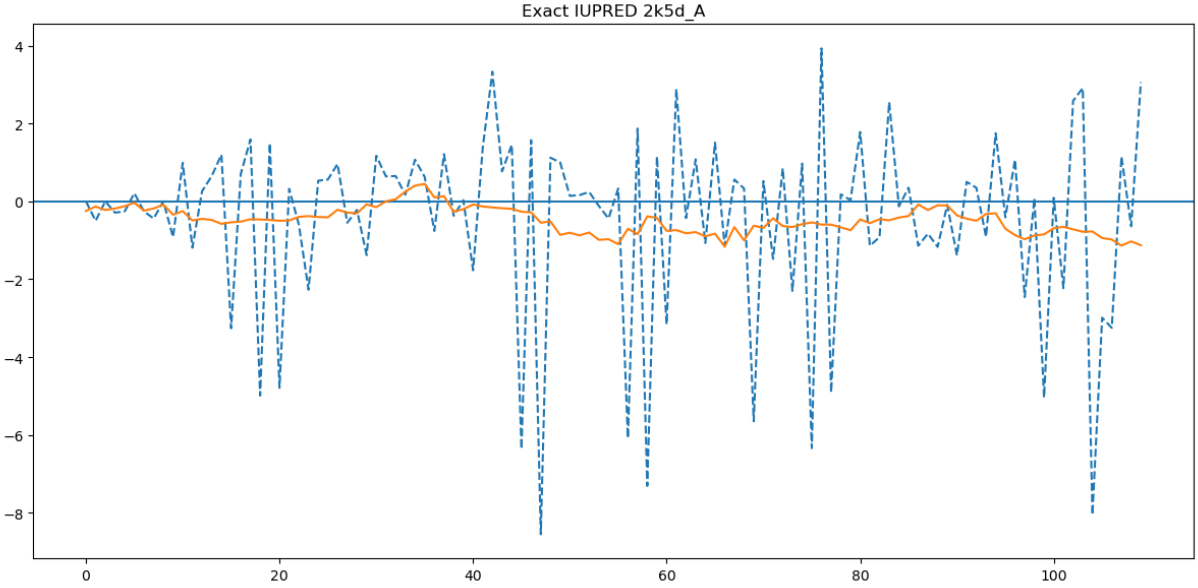
1. Number of contacts is 286



2. The contact matrix is (as above) calculated considering sequence separation of 2 and minimum distance threshold of 3.5. The **exact energy** of each residue is calculated using the formula

$$E_i^p = \sum_{j=1}^{20} M_{ij} c_j^p$$

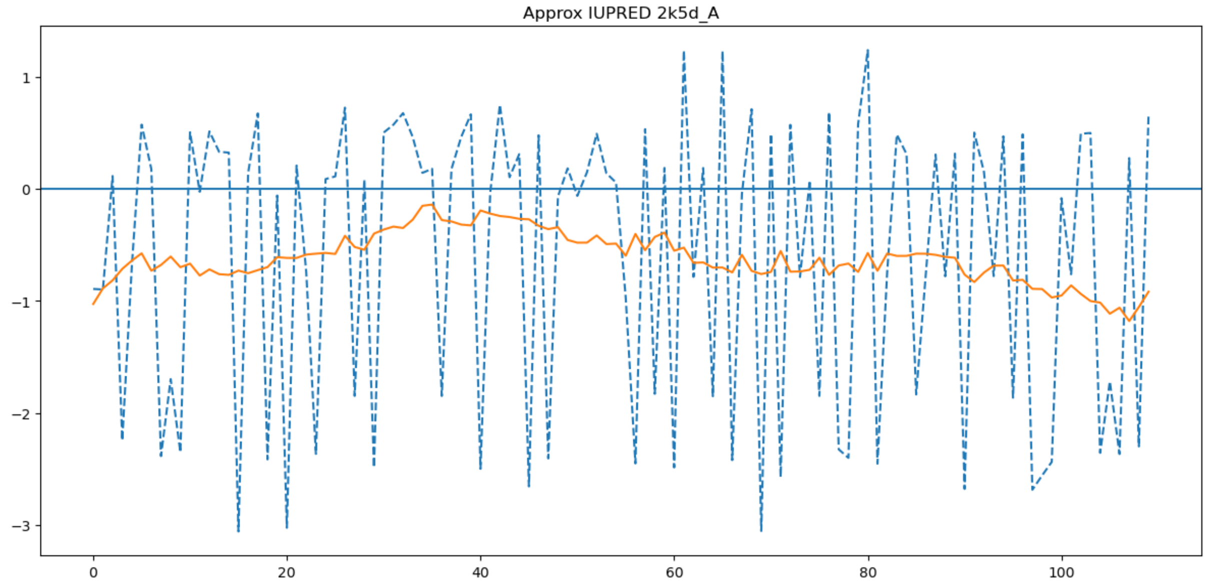
Where matrix M is given and c_i^p is the vector (dimension 20) containing number of residues (of type j) in contact with amino acid in position p of type i



3. The estimated energy is calculated, for each amino acid p of type i, with the formula

$$e_i^p = \sum_{j=1}^{20} P_{ij} n_j^p$$

by considering all amino acids in a window of size $\pm 2 - 100$ using the matrix P_{ij} given.



4.

	count ≥ 0	fraction ≥ 0
Energy	60	0.545455
energy_smoothed	7	0.063636
est_energy	56	0.509091
est_energy_smooth	0	0.000000

Formulas used in point 2 and 3 are taken from the IUPRED paper

GitHub repository