BIN 515 – Structural Bioinformatics

Assignment 2

Due date: April 5, 2020

Late policy: For each assignment, 20 points deduction will be applied for one day late, and 10 points additional deduction for each extra day.

Use your preferred programming language to implement the required methods to solve the problems below. Please, submit a complete report with the program codes. Discuss your results in your report. The instructor has the right to request a demo of the codes any time after the submission of the assignment through online tools and can determine the final score based on the demo performance.

Select a **single chain human** protein from PDB with **at least 150 and at max 250 residues**. Use the advanced search option at PDB to select the protein. Proceed with that protein to perform the following actions.

Q1. Write a script and run it to prepare the contact matrix of this protein. A residue is defined to be contacting with another residue if the distance between C^{α} atoms of these two residues is less than 7 Å.

Q2. Draw this contact matrix to create the contact map. Label the secondary structures on the contact map (a-helices, parallel or anti-parallel beta sheets).

Q3. Coordination number (CN) of a residue in a protein is defined as the total number of **non-bonded contacts** of a residue. Find the coordination number of each residue (Hint: You may use the contact matrix in Q1 to find the coordination numbers).

- **a.** Which residue is the most connected one?
- **b.** Find the most connected top 10 residues. Draw the protein in VMD in new cartoon representation. Label those 10 residues in vdW representation in VMD. Discuss at which region of the protein these residues are located (surface or core region) just by checking visually.
- **c.** Check if there is any correlation between the coordination number and B-factors of the residues. Draw an x-y scatter plot where the x-axis is B-factors (temperature factor) and y-axis is the coordination number. Discuss your results in your report.

Q4. Write a script to find the centroid of the coordinates of the protein you have selected.

- **a.** Calculate the new coordinates when the centroid is shifted to the origin and write the new coordinated into a file.
- **b.** Using your script show that the new coordinates satisfy the condition: $\sum_{i=1}^{N} x^{(i)} = 0$ (Note: This total may not exactly equal to zero because of rounding in arithmetic calculations. So, check out $0 \leq \sum_{i=1}^{N} x^{(i)} \leq 10^{-10}$

Q5. Find an NMR structure in PDB. Take a look at the PDB file. How many models are available for that protein in the PDB file? Write a script to parse the coordinates of Cα atoms of each models. Compare each model with the first model in terms of RMSD. (Hint: Models in PDB file are seperated by the keyword ENDMDL or ENDMODEL keyword.) Plot the models versus RMSD values and discuss the resulting plot in your report.

Q6. Accessible surface area (ASA) of a residue is defined as the solvent exposed surface area. Maximum ASA is the maximum possible solvent accessible surface area for the residue in an isolated state. Relative accessible surface area (RSA) is used to assess how much surface area is lost by the residue in its protein state and calculated by the formula;

$$RSAi = \frac{ASAi}{MaxASAi}$$

where *i* is the residue number.

Lower RSA value implies more buried residues. There are many tools and online servers to calculate ASA. For this assignment use getarea server (http://curie.utmb.edu/getarea.html) with its default parameters.

In the output “Ratio” column is the RSA value.

- **a.** Use your selected protein and calculate the RSA value for each residue.
- **b.** Check if there is any correlation between the coordination number (calculated in Q3) and RSA of the residues (Hint: You may plot a x-y scatter plot where the x axis is the coordination number and y axis is the RSA value).
- **c.** A residue is labelled as located on the surface if the RSA value is greater than 5%, otherwise it is labelled as core residue. Count the number of surface and core residues. Is there any difference between core and surface in terms of the physicochemical properties of the aminoacids (small, large, hydrophobic, polar, charged etc.)