Processing of Biological Data

Prof. Dr. Volkhard Helms
Winter Semester 2018-2019

Saarland University
Chair for Computational Biology

Exercise Sheet 5

Due: January 21, 2019 23:59

Please feel free to contact Pratiti (pratiti.bhadra@gmail.com) for any clarifications.

Submit your solutions (in single pdf file) and code to pratiti.bhadra@gmail.com. Subject of the email should be in the following format : Assignment5-"your name".

100 points

The Protein Data Bank (PDB) (45 points)

Q-1 From the PDB homepage https://www.rcsb.org/ download the file 2an6.pdb (protein: PDBID 2AN6). Familiarize yourself with a molecular graphics software of your choice, e.g. VMD or Pymol

- Identify all protein-peptide interaction chain pairs (5 points)
- Write a program to calculate the probability of the 20 amino acid types in the interface of chain A and E, by interface of two chains we understand all residue pairs that have atoms within 0.5 nm from each other. You are encouraged to use Python. Your program should output a text file containing amino acid 3-letter codes and the corresponding probabilities. Plot the probability vs residues and interpret it. (30+5+5 = 40 points)

Molecular Dynamics (MD) Simulation (55 points)

As an illustration for **Q-2** - **Q-4** watch the MD simulation video of protein-ligand complex when the protein is adsorbed on self-assembled-monolayer surface (given at the course website *Supp:Vedio.avi*).

Q-2 From the course website download the file *Supp-Q-2::COMdist-Protein-Ligand.xls*. The file contains the center of mass distance (*COMdist*) between protein and ligand at 0.1 ns intervals during a 100 ns long MD simulation in two different conditions, namely A and B. (25 points)

- Plot the data over time (both condition A and condition B data on the same graph). What do you get from the graph? Which condition is suitable for the protein-ligand complex and why? (5+2+1+2 = 10 points)
- Only for condition A (15 points)
  - Compute block averages of *COMdist*. Block size = 10 time points. Average of the data points for each block is known as block average. To calculate the block averages the data points should be divided into blocks of equal length (*block size*). Here divide the data points into blocks of 10 time points. Calculate and report the average on each block (*block averaging*) (5 points)

- If we define 3 states of protein-ligand interaction by time slot (a) 20 ns to 40 ns (b) 50 ns to 70 ns (c) 80 ns to 100 ns. What should be the appropriate names of the three states (a), (b) and (c). (6 points)
- Calculate the *standard error* of the block average values of state (b).(4 points)

Q-3 File *Supp-Q-3::P(x).xls* provides the experimental probability distribution of x ($P_{exp}(x)$) along with two observed probability distribution of x ($P_{obs1}(x)$ *and* $P_{obs2}(x)$) obtained from two different simulations. Quantify the divergence between experimental distribution and observed distributions, e.g. Kullback-Leibler divergence or any correlation coefficient. Find which simulation is able to produce a result closer to the experiment. (15 points)

Q-4 Plot the distribution of the cosine of the angle between the dipole moment vector of the protein molecule and the surface normal in the last 25 ns of simulation data. The surface normal vector is $\{0, 0, 1\}$ and the dipole moment vectors are given in the file *Supp-Q-4::dipole-moment-vector-Protein* with the corresponding simulation time. (15 points)