Exercise 09

Advance Analysis of Long MD Simulation

Deadline: Please hand in Exercise 09 in **pdf format** by **Friday, 10 July, 10.15 a.m.** to saleksic@zedat.fu-berlin.de

1 Peptide sequence (5)

In order to mimick the behavior of a peptide as if it was part of a longer protein sequence, it is a common practice to cap peptides with acetyl group at the N-terminus, and with methyl amino group at the C-terminus. For this exercise, you are provided with two independent simulation runs of a length of 1 μs (trajectory was saved every 1 ps) performed in NVT ensemble at 300K. For each replica, ϕ-ψ dihedral timeseries were extracted using the Gromacs command g_rama. Based on the provided files, determine the sequence of a capped nona-peptid Ace-X_9-NHMe, where wild card X can be any of 20 proteogenic amino acids. Provide sequence in one letter code. (5)

2 Backbone flexibility (50)

2.1 Ramachandran plots (10)

In a folder dedicated to this exercise, create two subfolders called rep1 and rep2. The provided bash script called **grep.sh** extracts ϕ-ψ dihedral timeseries for each residue from rama*.xvg files, and then it merges them into a single file representing the sampling of a equilibrium distribution. Modify the following line of grep.sh, so it represents peptide structure:

```
for k in RES-2 ... RES-10
```

Bash scripts are exectuted:

```
sh bash_script.sh
```

Include modifed bash script in your report. (2)

Modify the provided matlab script called **Ramachandran_plots.m** in such fashion, that it produces only Ramachandran plots for the amino acids of the 9-mer.Include the modifed matlab script and produced plots into your report. How many amino acids do have multiple minima in their Ramachandran plots? Name those amino acids(one letter code) and discuss visited regions of respective Ramachandran plains. (8)

2.2 MD trajcetory discretization (40)

In odrer to construct a Markov state model (MSM), one should select a proper reaction coordinate onto which dynamics is projected. It is well documented in literature that discretization based on the backbone diheadrals can be used to determine the conformational dynamics of peptides. For each of the amino acids, marked as flexible in section 2.1, create a matlab script, which bins allowed regions into two to three bins depending on the number of minima occuring in Ramachandran plain of the repsective amino acid. In order to perform the task,you can use script **Ramachandran_plots.m** ,and specially function script **rama.m** as a template. Additional lines to be introduced to Ramachandran plots represent boarders separating energy minima (examples of binned Ramachandran plots are presented in Fig. 1). Include in report binned Ramachandran plots of flexible amino acids, and state the values of ϕ-ψ used as boarders. (20)

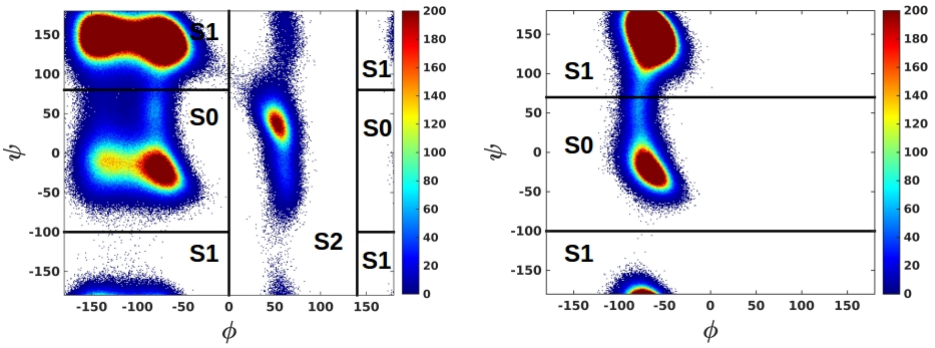

Fig.1 Examples of binned Ramachandran plots (S0 represents α-helical region, S1 represents β-sheet or PPII region, and S2 represents L-α-helical region)

Based on ϕ-ψ values used to define boarders, write a script in which ϕ-ψ trajectory is projected onto the disretization:

frame number	ϕ	ψ	region	bin number
1	-20	-20	α-helical	0
2	20	20	L-α-helical	2
3	-20	-150	β-sheet	1
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
n	-20	-150	β-sheet	1

ϕ-ψ timeseries is (columns 2 and 3 in the table above) loaded, and combination of ϕ-ψ angles determines to which region of Ramachandran plain, amino acid is placed in i-th frame. Print output files containing the trajectory consisting of bin numbers instead of the ϕ-ψ values for all flexible amino acids separetly. Determine the probability of finding amino acid in each of visited regions respectively. Script should contain code for all flexible amino acids. (20)

3 Sidechain flexibility (45)

Download provided trajectory and compute χ_1 angles for all amino acids of 9-mer by using Gromacs command **g_chi**. Delete the first 12 lines in output files, so *.xvg files can be importated in matlab. (5)

Write a matlab script, which loops over all 9 amino acids of this peptide and plots χ-1 values as a histrogram with 360 bins.Label x- and y-axis, so x-axis represent χ-1 values, and y-axis represents occurence of χ-1 valus in each of 360 bins.Include a matlab script and χ-1 plots into your report. (5)

3.1 Normalized Euclidean distance (10)

The Euclidean distance or Euclidean metric is the "ordinary" (i.e straight line) distance between two points in Euclidean space. With this distance, Euclidean space becomes a metric space. Euclidean distance can be used to measure the difference between the histograms of χ-1 angles of a pair of amino acids.

Write a matlab script, which calculates the normalized Euclian distance between the χ-1 histograms of R_6 and V_7, according to the following equation:

$$\delta_{RV} = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(R_i - V_i)^2} \quad (1)$$

where δ_{RV} is a normalized Euclian distance for the given pair of amino acids, N=360 (number of bins), R_i probability of finding χ-1 of R_6 in i-th bin, and V_i probability of finding χ-1 of V_7 in i-th bin.

3.2 Kullback-Leibler divergence (10)

In probability theory, and information theory, the Kullback-Leibler divergence is a non-symmetric measure of the difference between two probability distributions P and Q. Similarly to Euclidian distance,Kullback-Leibler divergence can be used to estimate the difference between the histograms of χ-1 angles of a pair of amino acids. For discrete probability distributions P and Q, the Kullback-Leibler divergence of Q from P is defined to be:

$$\delta_{PQ} = \sum_{i=1}^{N} P(i) ln \frac{P(i)}{Q(i)} \tag{2}$$

By implementing equation (2), write a matlab script which computes Kullback-Leibler divergence of discretized χ-1 histograms for R_6 and V_7 (N=360).

3.3 Mutual information (25)

In probability theory and information theory, the mutual information (MI) of two random variables is a measure of the variables' mutual dependence. The mutual information of two discrete random variables X and Y can be defined as:

$$MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \tag{3}$$

where p(x,y) is the joint probability distribution function of X and Y, and p(x) and p(y) are the marginal probability distribution functions of X and Y respectively. The values of MI are not confined to a certain interval. Therefore, in practice, one uses the normalized mutual information NMI, which is confined to [0,1], with NMI=0 corresponding to absence of mutual dependence.The normalized mutual information is given as:

$$NMI(X,Y) = \frac{MI(X,Y)}{min(H(X), H(Y))} \tag{4}$$

where H(X) is the informational entropy of the marginal probability distribution of variable X, given as:

$$H(X) = -\sum_{x \in X} p(X) log(p(X)) \tag{5}$$

Based on exations 3 to 5, write a matlab script, which calculates informational entropies, mutual information, and normalized mutual information of χ-1 angles for R_6 and V_7.

4 Files

Backbone diheadrals can be downloaded under:

https://www.dropbox.com/s/awzfmw61x90aph0/rama.zip?dl=0

Trajectory for computing the sidechain diheadrals can be downloaded under:

https://www.dropbox.com/s/9cfj4p7uo5smc6i/rep1.zip?dl=0

Scripts can be downloaded under:

https://www.dropbox.com/s/bzqo8myjg4yi33a/scripts.zip?dl=0