Exercise 10

Markov State Model Analysis of MD simulations

Deadline: Please hand in Exercise 10 in **pdf format** by **Friday, 17 July, 10.15 a.m.** to saleksic@zedat.fu-berlin.de

1 Microstate trajectory (20)

The Markov state model of WPPPPRVPR is about to be constructed on the conformational space spanned by the ϕ- and ψ-backbone torsion angles. The conformation space of WPPPPRVPR can be comprised by taking into account six flexible amino acids namely W_1,P_5,R_6,V_7,P_8,R_9. Each possible configuration of bins along the peptide chain represents a microstate.

How many microstates are in total comprising the conformational space of WPPPPRVPR peptide? (2)

Write a matlab script which loads trajectory files containing bin numbers of W_1,P_5,R_6,V_7,P_8,R_9 you produced in Ex09, and calculate the microstate number for the each frame of the simulation, which captures the dynamics of the whole peptide (see example below). (10)

Let's consider an example of 3-mer with RES1 and RES3 having two minima in their Ramachandran plains, while RES2 having three minima. The total number of possible microstates would be 2*3*2=12. Microstate number can be calculated according to the following equation:

$$MSN = b_{res1} + b_{res2}\cdot \mathsf{T}_{res1} + b_{res3}\cdot \mathsf{T}_{res1}\cdot \mathsf{T}_{res2} + ... + b_{resn}\cdot \mathsf{T}_{res1}\cdot \mathsf{T}_{res2}\cdot ...\cdot \mathsf{T}_{res(n-1)} \quad (1)$$

where MSN is a microstate number for the given frame, *b* a bin number of the respective amino acid, *T* is a total number of bins that the respective amino acid has.

frame number	RES1	RES2	RES3	microstate number
1	0	2	1	0+2*2+2*3*1=10
2	0	0	1	0+2*0+2*3*1=6
3	1	2	0	1+2*2+2*3*0=3
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
n	1	2	1	1+2*2+2*3*1=11

Table 1: Conformational space projected onto microstate numbers

Highly dimensional conformational space of 3-mer is reduced to microstate numbers between 0 and 11.

Run previously written script for both replicas and save it as *.dat files. However, not all microstates will be visited, so you need to write a matlab script (hint: use matlab command unique), which will sort the microstates numbers between 0 and N_{max}-1, where N_{max} is a maximal number of visited microstates(i.e 8 out 12 in the example above), and save them into a new microstates trajectories (*.txt files) to be used in the second exercise. How many microstates are actually visited in this MD simulation? (8)

2 Markov state model (MSM) construction (15)

Dowloaded EMMA package for the construction of Markov state model (MSM) save into your home folder (user/yourusername). Modify **MSM.sh**,by including HOMEPATH (directory for Ex09, and Ex10), EMMA (directory for EMMA software package), lag times (variable k) of 100,250,500,750,1000,2000,3000, 5000, and 10000 ps. Include modified **MSM.sh** script into your report. (2)

MSM.sh script uses the EMMA software package to:

- extract the largest connected set (EMMA command mm_connectivity)
- compute transition matrices(mm_estimate)
- compute eigenvalues-eigenvectors pairs (mm_transitionmatrixAnalysis) at different lag-times
- perform PCCA analys (mm_pcca)

2.1 Implied time scales (ITS) (13)

A way to find a lag-time for which modeling the dynamics with a Markov process is a good approximation, is to check the convergence to a plateau of the timescales.The implied timescales at each lag-time can be calculated according to the following equation:

$$ITS_i = \frac{-\tau_i}{\ln(\lambda_i(\tau))} \tag{2}$$

where λ_i, and τ_i, are respective eigenvalues, and lagtimes.

Write a matlab script, which loops over eigenvalue files (**evals*_maxiter.dat**) for all tested lag times τ, and extracts implied times (ITS), and produce plot ITS=f(τ). As the first eigenvalue is always 1, and represents the equilibrium distribution, take into consideration eigenvalues 2 to 4, for obtaining the ITS. In ITS plot, use ns as a time unit (values of lag time are in ps). (10)

Provide lag time τ, at which ITS representing the main three processes converged (reached plateau). What are values of ITS at that lag time? (3)

3 Long-live Conformations (60)

Once the MSM is constructed, it is possible to merge rapidly mixing microstates into larger macrostates that represent the long-lived conformations of the system. Perron Cluster Cluster Analysis (PCCA) uses the eigenspectrum of the transition matrix to assign microstates to coarser so called macrostates or long-live conformations. Starting with all microstates merged in one single macrostate, it is iteratively broken into two smaller, kinetically diverse states, based on the second eigenvector sign. The next step consist in choosing the set with greater spread in the eigenvector components, and breaking it in two using the third eigenvector sign. Further iterations of the algorithm create more macrostates.EMMA command mm_pcca performed PCCA analysis for all selected lag time τ by taking into consideration second,third,and forth right eigenvectors. Microstate number belonging to the same macrostate are written in the same row of set_X_Y.dat files, where X denots number of macrostates, and Y lag time τ.

3.1 Size of macrostates (5)

By considering set_4_Y.dat file for the lag time τ at which ITS converged, Write a matlab script which calculates the size of all four macrostates. Provide the size of each macrostates in terms of percentage of simulation length (time).

3.2 Ramachandran plots of long-lived conformations (15)

Change in the dynamics of flexible amino acids can be followed by plotting their Ramachandran plains for each detected macrostate. Write a matlab script, which produce Ramachandran plots for all six flexible amino acids in all four macrostates (hint:use matlab command find, which can extract frame numbers from microstate trajectory based on the vector containing the microstate number beloging to that specific macrostate, and that list of frames can be use for plotting of respcetive ϕ-ψ timeseries). Discuss the changes occuring in Ramachandran plots of each macrostate by comparing it with the Ramachandran plot representing equilibrium distribution (Exercise 09).

3.3 Visualization of long-lived conformations (10)

Size of provided ***.xtc** files corresponds to the size of the repsective macrostates. Open *.xtc trajectories and align frames on the top each other. See instructions below: (10)

VMD $\rightarrow$ Graphics $\rightarrow$ Representations $\rightarrow$ Trajectory $\rightarrow$ Draw Multiple Frames

In the field Draw Multiple Frames enter 1:X:N (last frame) and click ENTER. Every Xth frame will be aligned to the previous(choose suitable number, since not all frames should be aligned) . Improve trajectory smoothing by increasing Trajectory Smoothing Windows Size. In Draw Style change the drawing method to NewCartoon. Save aligned frames for all four macrostates as separated figures.

3.4 PCCA analysis (10)

Let's consider two cases. In the first case PCCA algorithm performed well in terms of spliting between macrostates (Table 2). In the second case, microstates to which an eigenvector assigns values close to zero are randomly assigned to different macrostates, when different eigenvectors are analyzied. It is known weakness of PCCA algorithm (Table 3).

eigenvector	macrostate 1	macrostate 2	macrostate 3	macrostate 4
2^{nd}	60	40		
3^{rd}	15	45	40	
4^{th}	15	20	25	40

Table 2: Spliting between macrostates - ideal case

Total number of 100 microstates were split into two macrostates of 60 microstates,and 40 microstates respectively:

$$\text{MacroS(60)} \leftrightarrow \text{MacroS(40)}$$

This is the first slowest process, occuring at the first ITS, and represents the higest energetic barrier in energy landscape. Then 3^{rd} eigenvector is considered, and MacroS(60) is further split into MacroS(15) and MacroS(45), representing the second slowest process (second ITS).

$$\text{MacroS(60)} \leftrightarrow \text{MacroS(15)} + \text{MacroS(45)}$$

Finally, when 4^{th} eigenvector was accounted, MacroS(45) was split into MacroS(20) and MacroS(45). This the third slowest process occuring at third ITS (the lowest energetic barrier).

$$\text{MacroS(45)} \leftrightarrow \text{MacroS(20)} + \text{MacroS(25)}$$

The three slowest kinetic processes can be sumirized in the following way (macrostates can be also noted as clusters):

Process 1: MacroS(60) =C1+C2+C3 ↔ C4=MacroS(40)
Process 2: MacroS(45) =C2+C3 ↔ C1=MacroS(15)
Process 3: MacroS(20) C2 ↔ C3=MacroS(25)

eigenvector	macrostate 1	macrostate 2	macrostate 3	macrostate 4
2^{nd}	60	40		
3^{rd}	15	43	42	
4^{th}	15	21	24	40

Table 3: Spliting between macrostates - non-ideal case

Since PCCA analysis was not capable of uniform assignment of microstates to respective macrostates, the only solution is to treat it, as it was an ideal case explained above.

By considering sets_X_Y.dat files (X=2-4) for the lag time of ITS convergence , deduce the three main kinetic processes ocurring in WPPPPRVPR peptide by filling the following table, and summerize the splitting process like shown in example above. (10)

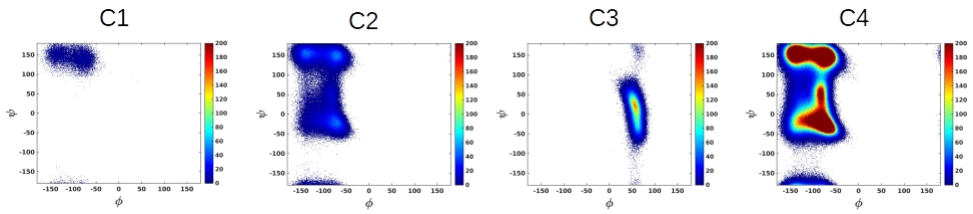
eigenvector	macrostate 1	macrostate 2	macrostate 3	macrostate 4
2^{nd}				
3^{rd}				
4^{th}				

Table 4: Spliting between macrostates

3.5 Kinetic processes visualized via Ramachandran plots (10)

Ramachandran plots produced in Exercise 3.2 interpret in terms of kinetics processes determined in Exercise 3.4 (see example below).

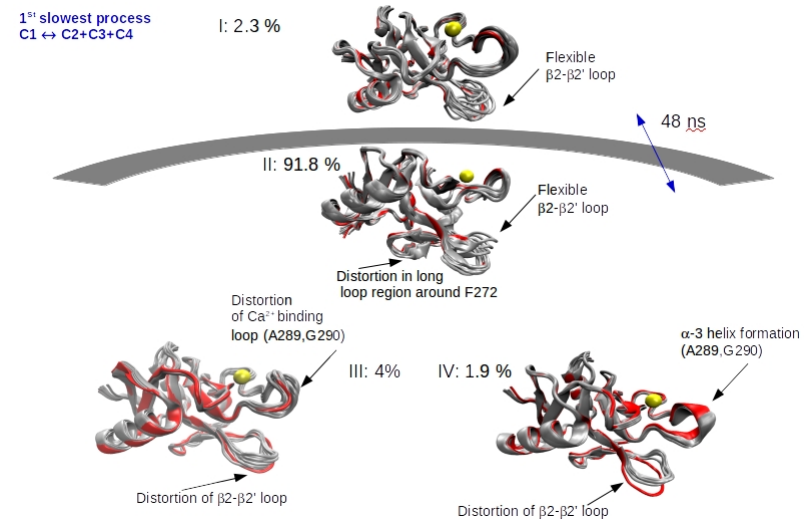
Let's consider again the previously shown three kinetic processes:
Process 1: C1+C2+C3 ↔ C4
Process 2: C2+C3 ↔ C1
Process 3: C2 ↔ C3



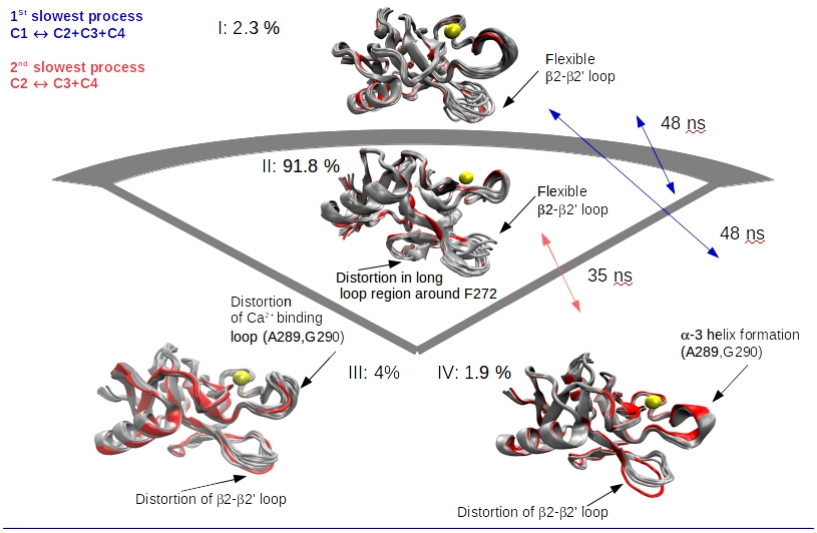
In the first kinetic process, there is a exchange between equilibrium distribution (visualy merge plots for C1,C2,C3, all three regions visited) and α+β minima (C4). In the second kinetic process, there is exchange between β-region and equilibrium distribution (visualy merge plots for C2,C3, all three regions visited). Finally in the third kinetic process, there is exchange between α+β minima (C2) and L-α minimum.

3.6 Hiearchy of the free energy surface landscape (15)

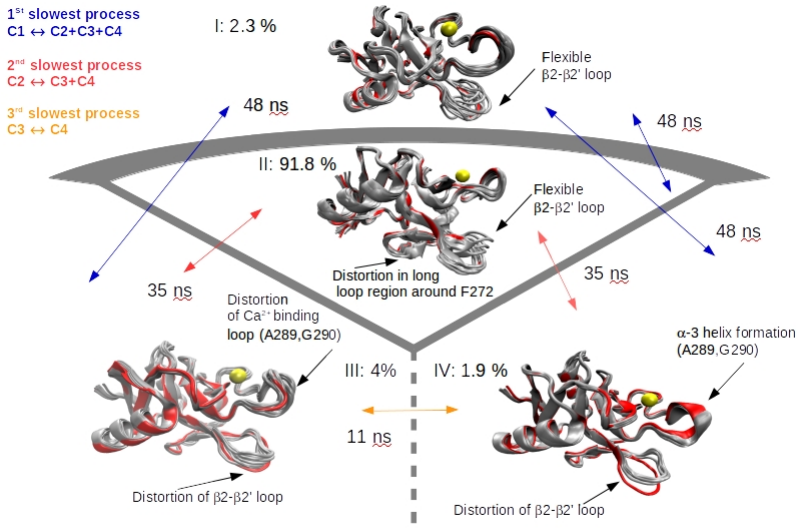
In the last step, based on the detected kinetic process, provide an energy surface landscape, as shown in the figures below. The thickest energetic barrier should separate participating macrostates in the first kinetic process, the medium thick energetic barrier should separate the macrostates splited in the second kinetic process, and the tinnies energetic barrier should separate two macrostates separated in the third kinetic process. Also provide ITS for all three kinetic processes. (15)



1^{st} kinetic process, separated by the highest energetic barrier



2^{nd} kinetic process, separated by the medium energetic barrier



3^{rd} kinetic process, separated by the lowest energetic barrier

4 Files

EMMA software package can be downloaded under:

https://www.dropbox.com/s/i6rvoosrjri9jzj/EMMA.zip?dl=0

Script can be downloaded under:

https://www.dropbox.com/s/5ofqnoqeyctxuyd/MSM1.sh?dl=0

Macrostates trajectories can be downloaded under:
https://www.dropbox.com/s/ri2n6jd923avj74/clusters.zip?dl=0