Research report

Markov State Models for Protein Dynamics

Victoria Valeeva, Eugene Klyshko

August 7, 2021

Contents

1 Abstract

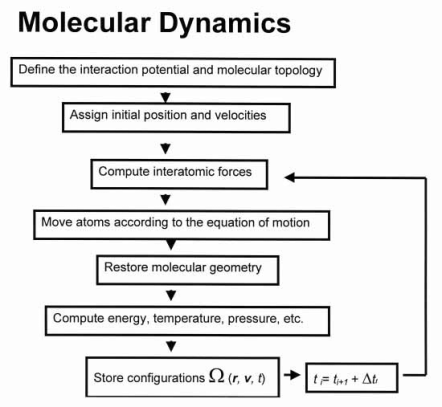
Proteins are molecular machines that perform a multitude of biological functions in living organisms. To improve existing drugs and create new therapies, we have to study the structure and dynamics of proteins using different experimental and simulation methods which complement each other, such as X-Ray Crystallography[1], Nuclear Magnetic Resonance Spectroscopy[2], and Molecular Dynamics Simulations[3]. In this project, we studied the structural and dynamical properties of the extensively studied PDZ domain of the human protein LNX2[4] in a crystal environment using MD data: a 3x3x3 supercell simulation[5], which is a representation of an X-Ray Crystallography experiment. We used a Markov State Model[6] approach to identify conformational states, specify their geometrical differences, and count the transitions between these states. Finally, we showed that the issue of equilibration time of the protein in a crystal environment can be addressed by careful consideration of its population dynamics.

2 Introduction

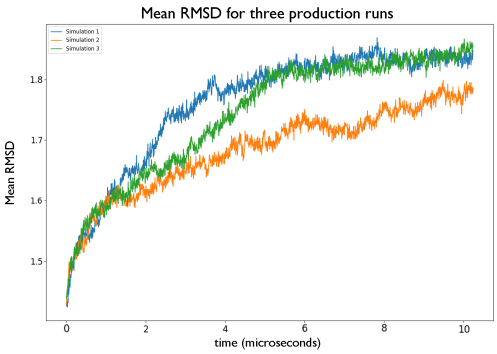
Proteins are polymerized biological molecules consisting of a sequence of monomers: amino acids. An X-Ray Crystallography experiment that applied an electric field to a protein crystal[7] was conducted to provide a new method of studying protein function. That experiment was recreated in silico[5], using molecular dynamics - a computer simulation method for analyzing the physical movements of atoms and molecules, and provided extensive sampling.

In this project I was working with the data that was generated in the equilibration simulation (electric field is turned off throughout the whole simulation): three ten-microsecond runs of a supercell system, consisting of 108 individual proteins, simulated with the forcefield CHARMM36m[8]. Force field is a computational method that is used to estimate the forces between atoms within molecules and also between molecules. More precisely, the force field refers to the functional form and parameter sets used to calculate the potential energy of a system of particles. A plot of the average RMSD (a common metric that describes the average distance between atomic positions) for that simulation suggested that the equili-

bration timescales are unusually high, as the graph for the first and the third simulations stabilizes at around six microseconds, and is unstable for the second simulation for the whole its duration.



Molecular Dynamics Pipeline



Mean RMSD for three simulation runs

To tackle this issue, I used the Markov State Model (MSM) approach that identifies stable conformational states and counts the transitions between them throughout the simulation. It specifically aided to find the equilibrium populations of these identified macrostates, which were used to answer the main question of the project. An MSMBuilder software package was used for this step of the analysis. Protein conformation's stability is determined by its free energy, where lower free energy means more stable conformation.

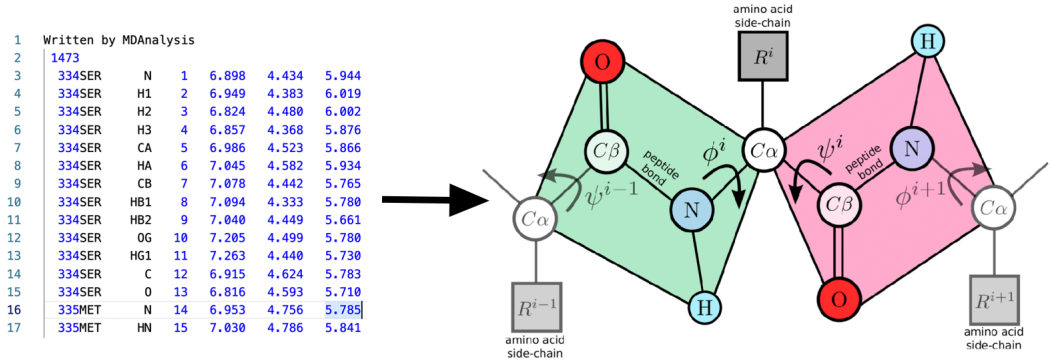
3 Methods

Markov State Models are a class of models for modelling the long-timescale dynamics of molecular systems. They model the dynamics of a system as a series of memoryless (ergodic), probabilistic jumps between a set of states. Building a Markov State Model requires a number of steps.

3.1 Featurization

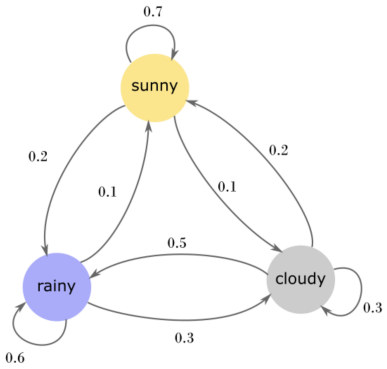
Since there's usually no special rotational or translational reference frame in a molecular dynamics simulation, it is often desirable to remove rotational and translational motion via

featurization that is insensitive to rotations and translations. The xyz-coordinates of the atoms were converted to sines and cosines of dihedral angles – the internal angles of the polypeptide chain at which two adjacent planes meet. Backbone (phi, psi) and side-chain (chi1, chi2) angles were selected, because it was observed in previous research that significant conformational changes happen in side-chains of protein residues.

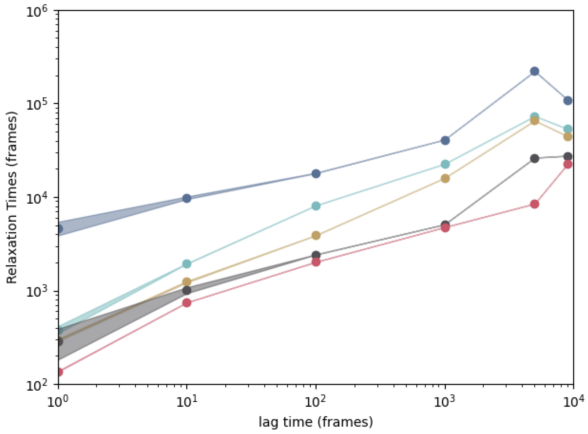


```
Written by MDAnalysis
1473
  334SER     N    1   6.898   4.434   5.944
  334SER     H1   2   6.949   4.383   6.019
  334SER     H2   3   6.824   4.480   6.002
  334SER     H3   4   6.857   4.368   5.876
  334SER     CA   5   6.986   4.523   5.866
  334SER     HA   6   7.045   4.582   5.934
  334SER     CB   7   7.078   4.442   5.765
  334SER     HB1  8   7.094   4.333   5.780
  334SER     HB2  9   7.040   4.449   5.661
  334SER     OG  10   7.205   4.499   5.780
  334SER     HG1 11   7.263   4.440   5.730
  334SER     C   12   6.915   4.624   5.783
  334SER     O   13   6.816   4.593   5.710
  335MET     N   14   6.953   4.756   5.785
  335MET     HN  15   7.030   4.786   5.841
```


Atom positions are converted into dihedral angles



Simple Markov Chain



Implied Relaxation Timescales

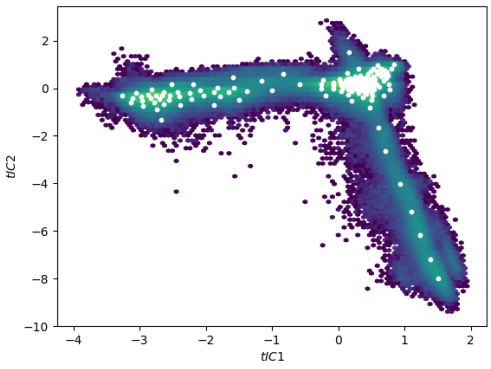
3.2 Dimensionality Reduction

As the PDZ domain was described by 329 dihedral angles, a two-dimensional representation of protein conformations was computed using Time-Structure Independent Components Analysis (tICA). It is one of the dimensionality reduction algorithms that works by finding the slowest-relaxing degrees of freedom in a time-series dataset. A lag time – an important parameter of a MSM that determines the gap between frames at which the transitions are

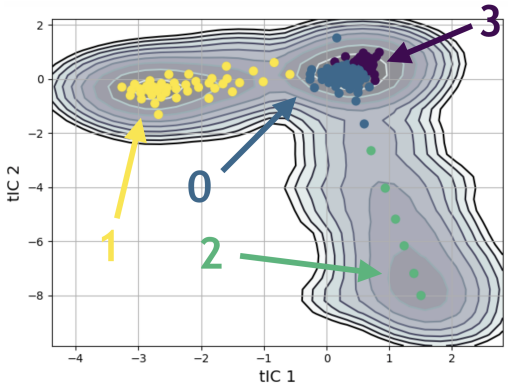
counted – of 100 nanoseconds (or 1000 frames) was selected for the analysis, because the implied relaxation timescales of the model start to converge for lag time more than 1000.

3.3 Clustering

Clustering molecular dynamics trajectories groups the data into a set of clusters such that conformations in the same cluster are structurally similar to one another, and conformations in different clusters are structurally distinct. Mini Batch K-Means algorithm with 300 clusters was selected for this step, because more complicated clustering algorithms did not work due to the large size of the data. Next, Robust Perron Cluster Analysis (PCCA+) was applied to determine four stable macrostates and classify each of the microstates that became available via Mini Batch K-Means.



Conformational Microstates

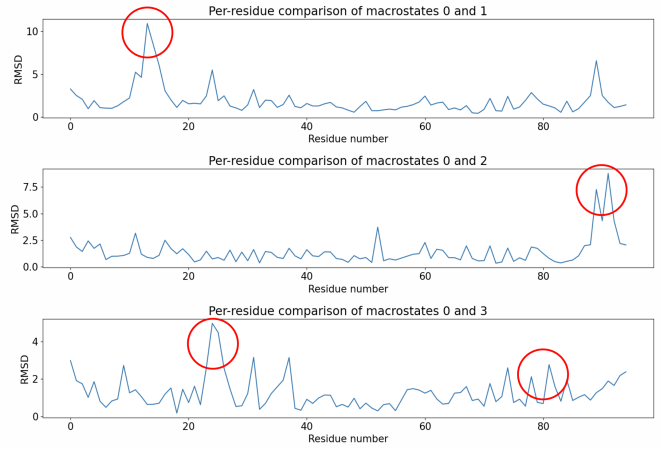


Conformational Macrostates

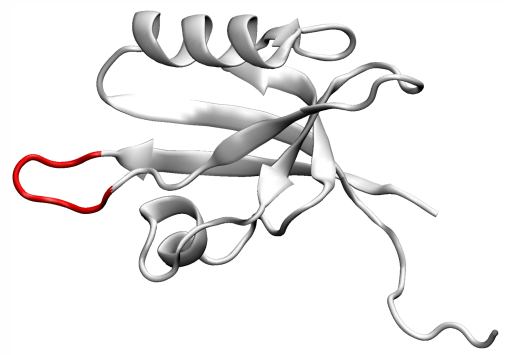
4 Results

For each of the identified macrostates, a representative with the lowest free energy was selected. Then, per-residue RMSD was calculated for each of the selected representative conformations to identify clear geometrical differences. A representative of macrostate 0, where all of the trajectories start, was taken as a reference; macrostate 1 had significant pertubations near residue 14, where a flexible loop is situated; macrostate 2 deviated the most around the C-terminus, and macrostate 3 diverged near residues 24 and 80, which correspond to a flexible loop and an alpha-helix that is getting slighly tilted in that conformation. As

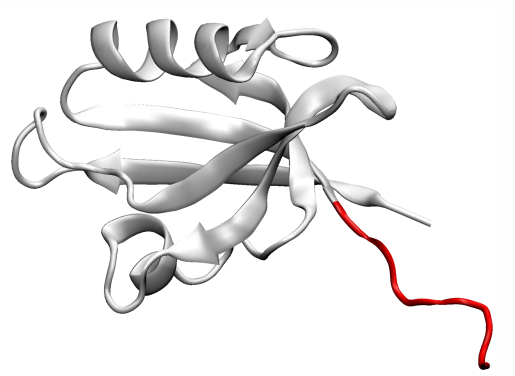
observed, it is the most disordered regions of the protein that deviate the most.



Per-residue RMSD



Macrostate 1

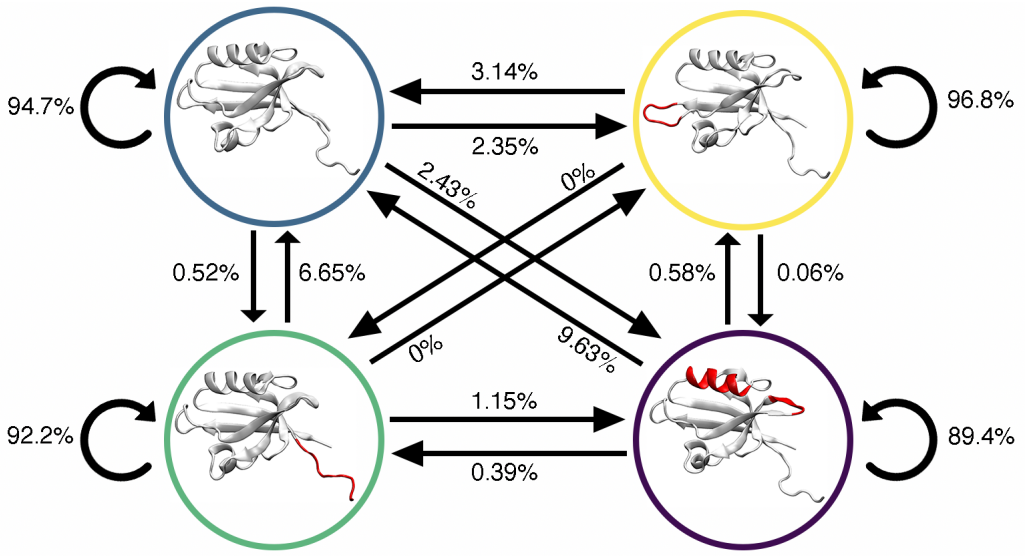


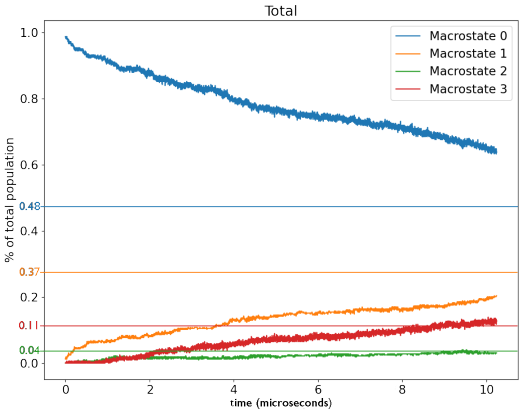
Macrostate 2

Macrostate 3

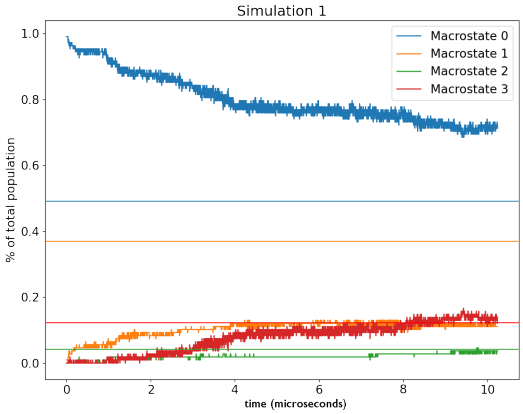
Next, the transitions between the identified macrostates were counted with a lag time of 100 nanoseconds. Then, a transition matrix describing probabilities of one state transitioning to another (or staying in the same one) was computed.

Equilibrium Population is a hypothetical distribution of states to which the system should converge at time $\to \infty$. The closest macrostate was assigned to each structure of each trajectory, and a percentage of the total population was computed for each of the identified macrostates. On the figures below, equilibrium populations of each macrostate are displayed as straight lines.



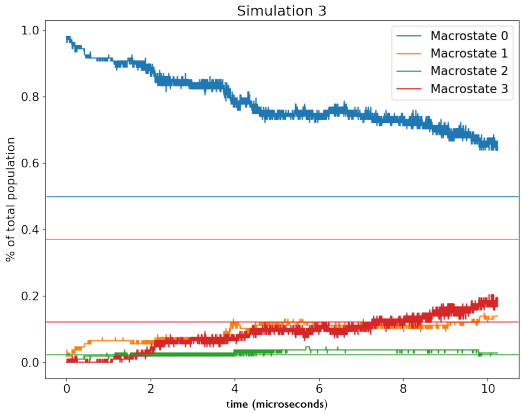


Total Population Dynamics



Simulation 1 Population Dynamics

Simulation 2 Population Dynamics



Simulation 3 Population Dynamics

As can be seen from the figures, the population of energetically-distant macrostate 1 significantly grows throughout the simulation, which explains the unusually high equilibration times.

5 Discussion

Markov state models prove themselves to be a solid method to identify conformational macrostates, which also allows for an accurate approximation of overall population dynamics. It also helped to identify the problem of the long equilibration process in CHARMM36m, which occurs because some of the populous macrostates are energetically distant from the conformations at the start of the simulation. However, while some general equilibration trends can be observed, provided enough data sampling, strong specific conjectures cannot be drawn, as proteins are highly susceptible to initial conditions. While the first and the third simulations showed similar population dynamics, in the second simulation, the population of macrostate one was growing more rapidly, although the initialization conditions were the same. Further studies to determine the behaviour of the PDZ supercell model simulated with different forcefields and external stimuli applied (ex. electric fields) are to be conducted.

References

[1] Smyth, M., 2000. *x Ray crystallography*. Molecular Pathology, 53(1), pp.8-14.

[2] Biological NMR Spectroscopy. *In NMR Spectroscopy Explained*; John Wiley and Sons, Inc.: Hoboken, NJ, USA, 2007; pp 551–626.

[3] Karplus, M.; Petsko, G. A. Molecular Dynamics Simulations in Biology. *Nature* **1990**, 347 (6294), 631–639.

[4] Young, P. W. LNX1/LNX2 Proteins: Functions in Neuronal Signalling and Beyond. *Neuronal Signal.* **2018**, 2 (2), NS20170191.

[5] Klyshko, E.; McGough, L.; Kim, J. S.; Ranganathan, R.; Rauscher, S. EF-X in Silico - Modeling Protein Dynamics in an Electric Field. *Biophys. J.* **2020**, 118 (3), 504a.

[6] Husic, B. E.; Pande, V. S. Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* **2018**, 140 (7), 2386–2396.

[7] Hekstra, D. R.; White, K. I.; Socolich, M. A.; Henning, R. W.; Šrajer, V.; Ranganathan, R. Electric-Field-Stimulated Protein Mechanics. *Nature* **2016**, 540 (7633), 400–405.

[8] Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D., Jr. CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods* **2017**, 14 (1), 71–73.