SEARCHING FOR CONFORMATIONAL STATES IN THE DYNAMICS OF PROTEIN CRYSTALS

Victoria Valeeva[1], Eugene Klyshko[1,2], and Sarah Rauscher[1,2,3]

[1]Dept. of Chemical and Physical Sciences, University of Toronto Mississauga; [2]Dept. of Physics, University of Toronto; [3]Dept. of Chemistry, University of Toronto;

vicky@escape13.me

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 the careful consideration of its population dynamics.

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.

1. **FEATURIZATION**[7]: 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.
2. **DIMENSIONALITY REDUCTION**[7]: 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. A lag time of 100 nanoseconds (or 1000 frames) was selected for the analysis.
3. **CLUSTERING**[7]: 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. KMeansMiniBatch algorithm with 300 clusters was selected for this step.

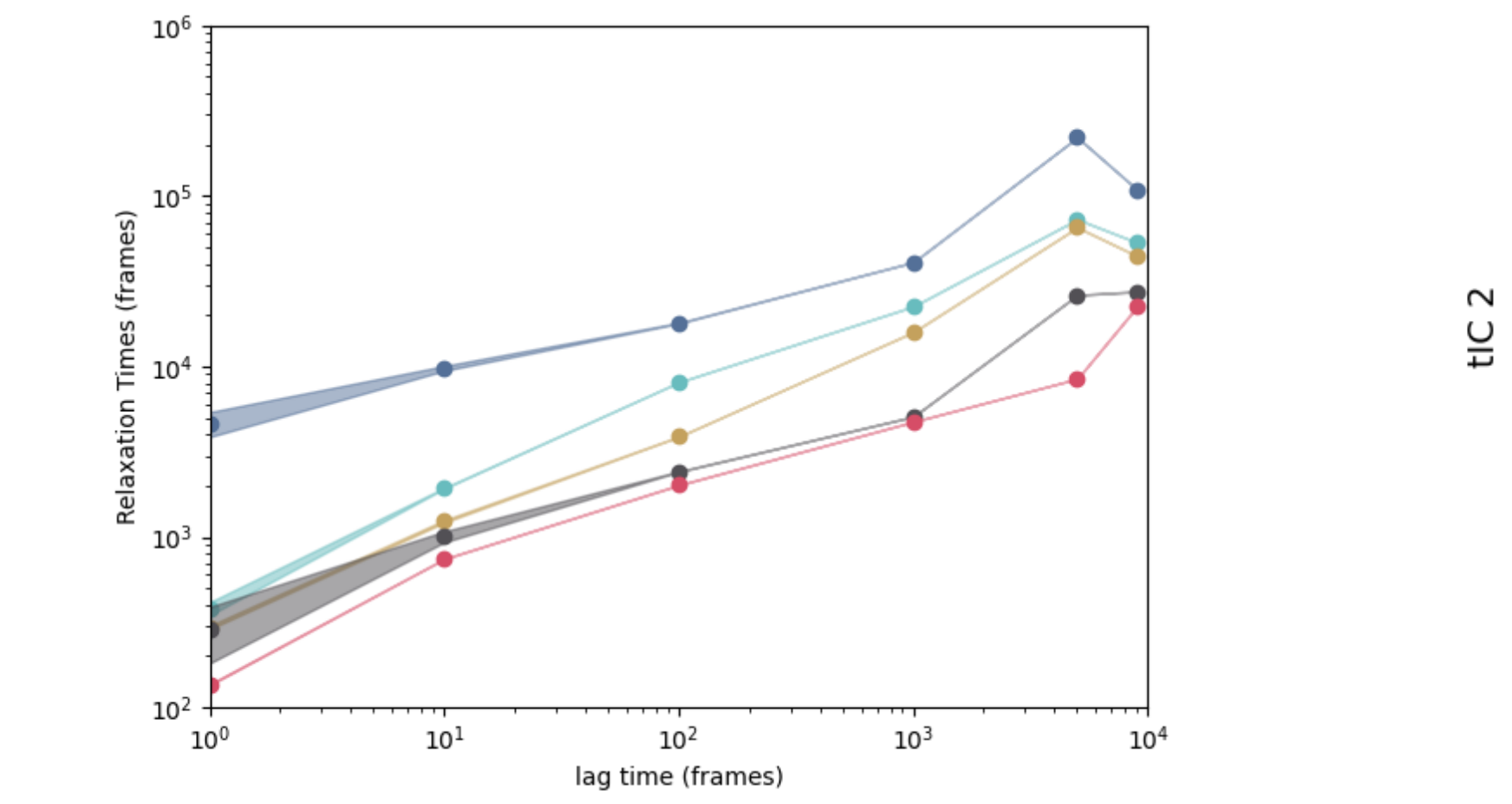


Fig. 1: Implied tICA timescales: the lag time for the model was chosen according to this plot.

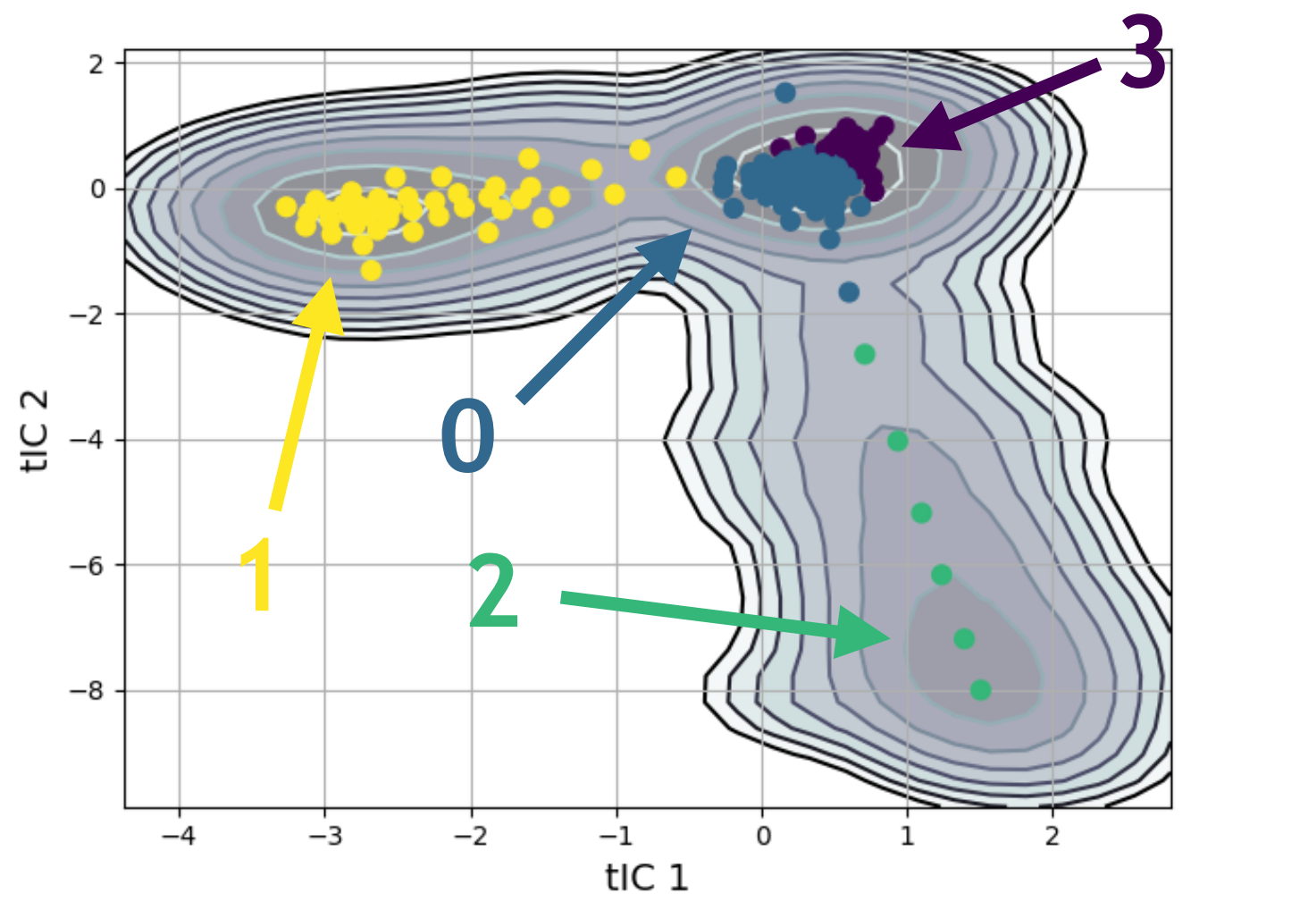


Fig. 2: PDZ Macrostates: the darker areas correspond to lower free energy.

Data

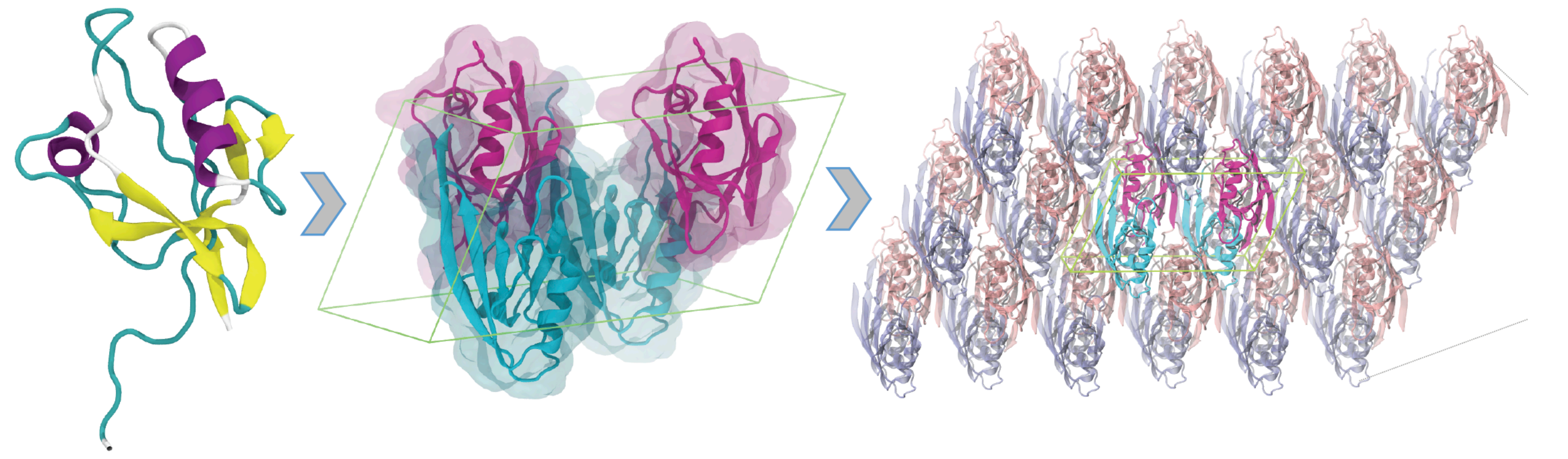


Fig. 3: Data representation: 3 10-microsecond runs of the 3x3x3 supercell of the PDZ domain of protein LNX2 (108 x 3 = 324 trajectory files), simulated with the CHARMM36m[8] forcefield (different forcefields have different ways of estimating forces for atomic or molecular interactions).

Identification of Conformational Differences

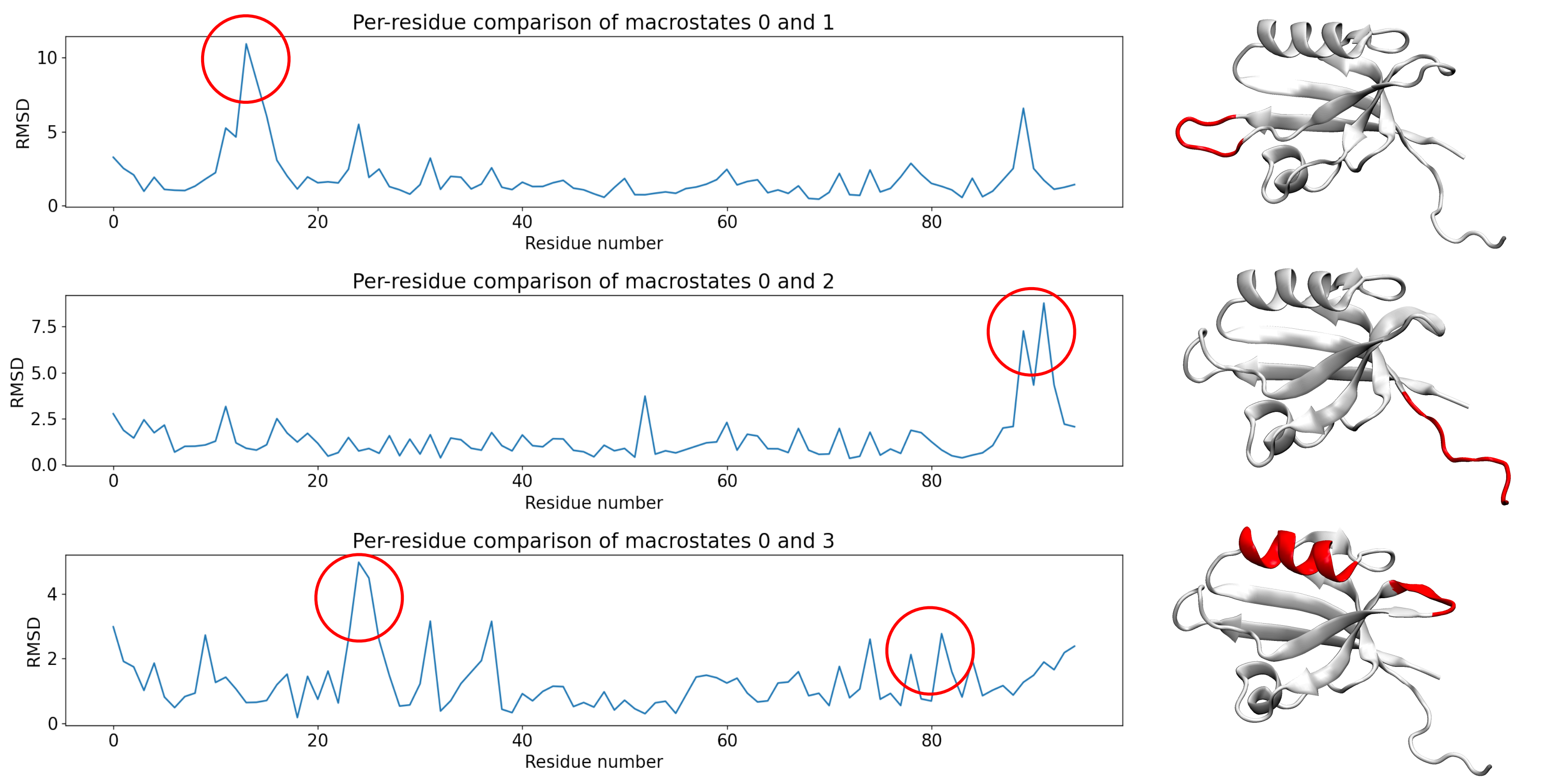


Fig. 4: Per-residue comparison of identified macrostates: a representative with the lowest free energy was selected from each macrostate, and root mean square deviation (a metric that describes the average distance between atomic positions) was calculated for pairs of macrostate representatives. The figure suggests significant differences near residue 14 (a flexible turn), residue 24 (a flexible loop), residue 81 (a tilting helix), and near the C-terminus.

Timescales and Transition Matrix

Implied Relaxation Timescales for the dataset have shown a tolerable level of consistency with population dynamics. The relaxation timescales are:

2.07 μs 1.124 μs 0.683 μs

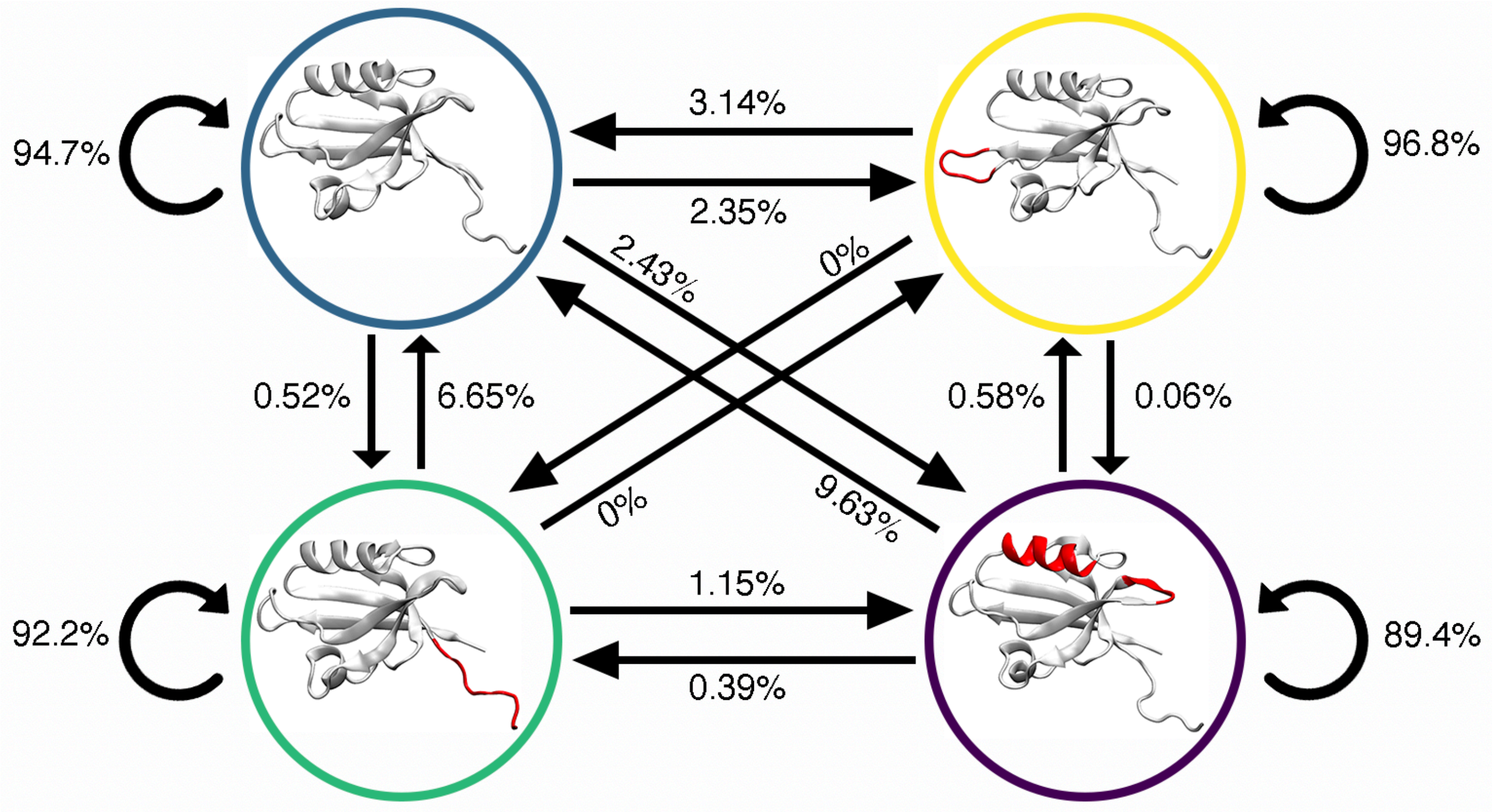
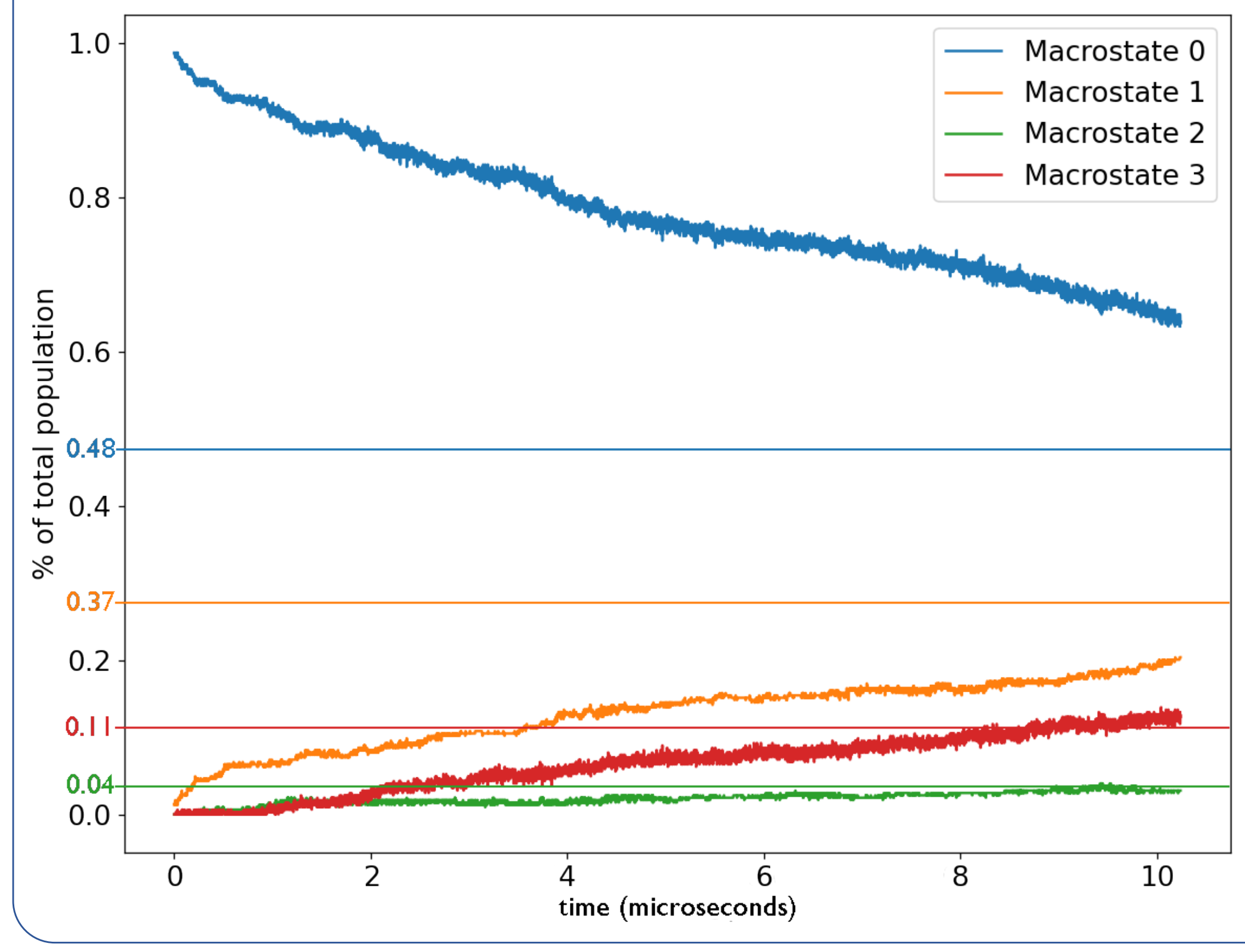


Fig. 5: Markov Chain: four conformational macrostates were identified, and transitions between them were counted with a lag time of 100 nanoseconds. Then, a transition matrix describing probabilities of one state transitioning to another (or staying in the same one) was computed.

Population Dynamics



Equilibrium Population is a hypothetical distribution of states to which the system should converge at time → ∞.

Fig. 6: Population of each macrostate over time: equilibrium populations are shown as straight lines. 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.

Conclusion

Markov state models are 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.

Acknowledgements

Victoria Valeeva is an incoming undergraduate student at the University of Toronto Mississauga who hopes to major in Molecular Biology. She enjoys applying computational techniques to problems in natural sciences.

Eugene Klyshko is an enthusiastic researcher in computational biophysics that uses the powers of machine learning, statistics, and high-performance computing. He is working towards his PhD degree at the University of Toronto.

Dr Sarah Rauscher is a professor of physics and a principal investigator of the lab that works on problems in computational biophysics with a focus on intrinsically disordered proteins at the University of Toronto Mississauga.

I would like to thank Eugene and Professor Sarah Rauscher for allowing me to make such a huge step in my university career and helping me to love the field of research in computational sciences much more than before.

References

[1] Smyth, M. S., and J. H. J. Martin. "x Ray crystallography." *Molecular Pathology* 53.1 (2000): 8.
[2] Levitan, Abraham. "NMR Spectroscopy."
[3] Karplus, Martin, and Gregory A. Petsko. "Molecular dynamics simulations in biology." *Nature* 347.6294 (1990): 631-639.
[4] Young, Paul W. "LNX1/LNX2 proteins: Functions in neuronal signalling and beyond." *Neuronal signaling* 2.2 (2018): NS20170191.
[5] Klyshko, Eugene, et al. "EF-X in Silico-Modeling Protein Dynamics in an Electric Field." *Biophysical Journal* 118.3 (2020): 504a.
[6] Husic, Brooke E., and Vijay S. Pande. "Markov state models: From an art to a science." *Journal of the American Chemical Society* 140.7 (2018): 2386-2396.
[7] Harrigan, Matthew P., et al. "MSMBuilder: statistical models for biomolecular dynamics." *Biophysical journal* 112.1 (2017): 10-15.
[8] Huang, Jing, et al. "CHARMM36m: an improved force field for folded and intrinsically disordered proteins." *Nature methods* 14.1 (2017): 71-73.

computecanada