Biophysical Journal, Volume *110*

Supplemental Information

Markov State Models and tICA Reveal a Nonnative Folding Nucleus in Simulations of NuG2

Christian R. Schwantes, Diwakar Shukla, and Vijay S. Pande

Supplementary Material for: "Markov State Models and tICA Reveal Non-Native Folding Nucleus in Simulations of NuG2"

Christian R. Schwantes,[†] Diwakar Shukla,[†‡] Vijay S. Pande[†+§¶]

[†] *Department of Chemistry,* [‡] *SIMBIOS NIH Center for Biomedical Computation,* [+] *Biophysics Program,* [¶] *Structural Biology, and* [§] *Department of Computer Science, Stanford University, Stanford, CA 94305.*

Contents

1 Methods

1.1 Simulation Details

1.1.1 Folding@home MD Simulations

Distributed molecular dynamics simulations were performed using GROMACS (1) on the Folding@home (2) computing platform. The CHARMM22* (3) forcefield was used for the protein along with the TIP3P (4) water model. The all-atom starting structures were solvated in a 60 Å cubic solvent box with TIP3P water molecules such that water extended at least 10 Å away from the surface of the protein. Na^+ and Cl^- ions were added to the system to neutralize the charge, corresponding to a salt concentration of approximately 100 mM. Covalent bonds involving hydrogen atoms were constrained with LINCS (5) and Particle Mesh Ewald (PME) (6) was used to treat long-range electrostatic interactions. The structures obtained after an initial equilibration for 1 ns at constant temperature and pressure and with constraints on the heavy atom positions were used as the starting conformation for the distributed molecular dynamics simulations. Production MD simulations were carried out at constant temperature and pressure of 350 K and 1 atm, respectively, with a time step of 2 fs. Trajectory snapshots were recorded every 100 ps.

1.1.2 Starting Structures for Folding@home Sampling

Three rounds of simulations were started using the initial structures generated from the molecular dynamics simulations of Nug2 reported by Lindorff-Larsen et al. (7) They performed simulations of N37A/A46D/D47A triple mutant of the redesigned protein G variant NuG2 (8) for a total simulation time of 1154 μs. The first set of $\sim$13000 simulations were started from the 100 states with minimum population obtained from the tICA MSM of Nug2 (see "MSM Construction" below) for an aggregate simulation time of 450 μs on the Folding@home platform. Another set of 10000 simulations were performed starting from only the register-shifted states observed in the MSM. The aggregate simulation time of 650 μs was obtained for this set using the Folding@home platform. In a third round of simulations, we started a total 3,217 simulations from the states in the original 15,000 state MSM which had not been adequately visited by the initial rounds of

simulation. These states were selected by:

- Assign the Folding@home simulations from the first two rounds to the 15,000 state MSM, using the original six tICs.
- Select states in the MSM which had 10% or less assignments from the Folding@home simulations relative to the original dataset from Lindorff-Larsen et al. (7).
- Select states in the MSM which had an average distance to their assigned points in the Folding@home dataset greater than one. Although this isn't an interpretable distance, almost all states have an average distance to the generator less than 0.6.

1.2 MSM Construction

All MSMs were constructed and analyzed with the MSMBuilder software package (9). Three MSMs are discussed in this article. The first MSM was presented in Beauchamp et al. (10) and was built on 250^{th} of the data (because of the limits of the clustering method) with the RMSD metric on atom positions. This model used ~ 3500 states and a lag time of 50 ns.

On the full Anton dataset, we first featurized each conformation into a vector corresponding to all residue-residue distances in the protein. For a given pair of residues, the distance was defined as the shortest distance between heavy atoms in the residues. In this feature space, we computed the slowest linear projections using tICA with a correlation lag time of 200 ns. Using the k-centers clustering algorithm and a lag time of 100 ns, we built many MSMs varying both the number of tICs and number of states. We built three models with 15,000, 20,000, and 25,000 states using six tICs. These models had essentially identical timescales (Fig. 1). We also built seven models using 15,000 states but varying the number of tICs between three and nine. The timescales were far less robust to the addition of more tICs. As more tICs were used, the slower timescales tended to become faster, while the faster timescales became slower. Nonetheless, the variance in the slowest timescales led to error bars within an order of magnitude (Fig. 1). For the MSM presented in the main text, we used the model built on 15,000 states and six tICs. This model, however, was not used for any quantitive calculation; it was only used to guide our Folding@home simulations.

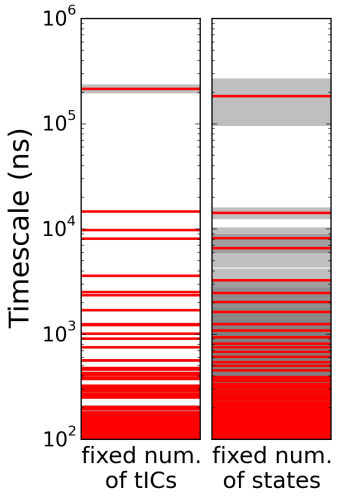

Figure 1: Many MSMs were built on the Anton dataset using tICA. Three models built with 15,000, 20,000, and 25,000 states with six tICs are compared on the left. The means of the timescales across these models are drawn in red, and the gray areas represent one standard deviation in either direction. The timescales were essentially the same for the models built with different numbers of states. On the right, seven models built with 15,000 states but with the number of tICs varying between three and nine are compared. The timescales were less robust to the choice of tICs, however the errors are within an order of magnitude for the slower timescales.

For the Folding@home dataset we used the same featurization, but recalculated the slowest tICs without the Anton dataset. The implied timescales were not converged for models built with greater than the top two tICs, and so we did not use these models. The choice of number of states was somewhat flexible as the timescales for 250-, 500-, and 1000-state models were essentially the same (Fig. 2). For the analysis presented in the main text, we used a 250 state model built with the hybrid k-medoids clustering algorithm(9) using two tICs and a lag time of 50 ns.

1.3 MSM Analysis

1.3.1 Register Shift Eigenprocesses

The timescales referred to by a red dash in Fig. 1 of the main text all corresponded to roughly the same eigenprocess (Fig. 3). By viewing the second or third eigenvectors of each MSM projected

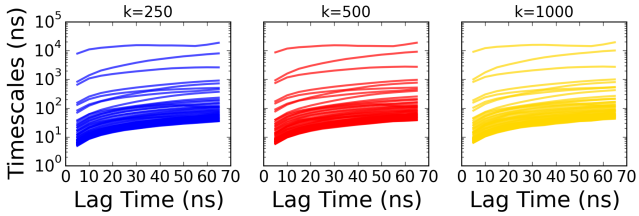

Figure 2: The implied timescales converged within 50 ns for the full Folding@home dataset. Other models, however, did not have well-behaved timescales and were not used for the final analysis.

onto a single order parameter (RMSD to the native state), we can see that for all models, the process corresponded to exchange between a near-native intermediate and the other structures in the simulation.

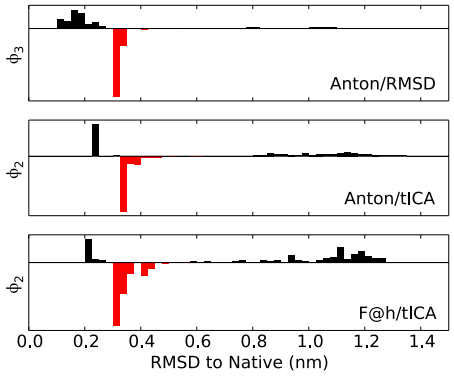

Figure 3: The eigenvectors of all three MSMs were projected onto each state's average RMSD to the native conformation. In all cases, the eigenprocesses corresponded to the exchange between near-native states (red) and other states in the simulation. These near-native states correspond to the register-shifted intermediate shown in Fig. 1 of the main text.

1.3.2 Transition Path Theory

We used residue-residue contacts to distinguish between the intermediates in the simulation. For each conformation in the dataset we computed contacts in the sheet between strands one and two. A contact was considered formed if any two heavy atoms were closer than 6 Å. For both the register shift and native folds, there were a total of four contacts made in the sheet. A state in the MSM was considered register shifted if (on average) there were two more register-shifted

contacts than native register contacts formed in the sheet. Additionally, unfolded states were defined as states whose average RMSD to the native structure was greater than 6 Å.

For describing the main folding pathways, we used transition path theory as described by Noé et al. (11). We defined the folded ensemble as the set of states in the MSM with the lowest average RMSD to the native structure such that the total population of the set of states was greater than 10%. The unfolded ensemble was selected in the same way but by collecting states with the highest average RMSD to the native structure. Then three intermediates were distinguished:

- R_{12}: Register shift in strand two; these states had at least two of the four register shifted contacts formed, but fewer than two contacts formed in the sheet between strands three and four
- N_{12}: Native register in strand two; these states had at least two of the four native contacts formed, but fewer than two contacts formed in the sheet between strands three and four.
- N_{34}: Native sheet formed between strands three and four; these states had at least two of the four native contacts formed between strands three and four, but fewer than two contacts formed in either the register or native conformation of strands one and two.

We calculated all paths that contributed to 99.9999% of the total flux through the network and assigned each to one of the three intermediates or "undetermined." Paths were labeled as undetermined if they did not visit any of the intermediate states or visited more than one of them. The result was that 55% of the flux visits the N_{12} intermediate, 10% visits the N_{34} intermediate and 1% visits R_{12}.

1.3.3 "Free Energy" Surface Definition

Fig. 2 in the main text was generated from the populations of the states in the MSM. The β_{12} axis was calculated by subtracting the number of register shifted contacts from the native contacts in the sheet between strands one and two. For each state in the MSM, we computed the average and standard deviation in this quantity as well as the RMSD to the native conformation. We calculated a free energy surface by placing a two dimensional Gaussian at the mean of each

state's order parameters with a covariance matrix given by the states' standard deviations. These gaussians were weighted by the equilibrium probabilities computed from the MSM. Finally, the free energy was calculated by taking the natural log of this probability distribution. We note that this is only a useful illustration, and the barriers in this surface do not correspond to the rates of transferring between states.

2 Miscellaneous Results

2.1 Misassignment in the RMSD MSM

The discrepancy in the estimation of the register shift relaxation timescale between the RMSD and tICA MSMs on the initial dataset can be explained by not having enough sampling. As such,

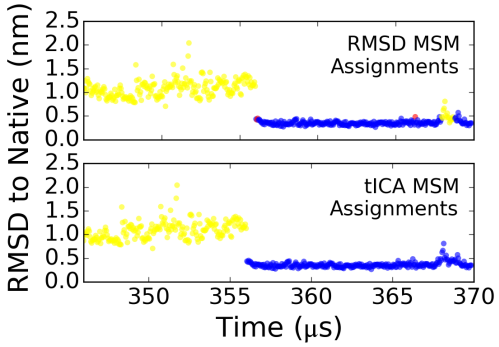

Figure 4: Above, the color indicates the macrostate that a point is assigned to according to the RMSD or tICA model: unfolded (yellow), folded (red), register-shifted (blue). Because the RMSD MSM misassigns a few register-shifted states to the folded and unfolded states, the timescale is much faster than the corresponding process in the tICA model.

a few misassignments can change the timescales drastically. In the initial dataset, the register-shifted state is only visited at the end of a single trajectory, but the RMSD MSM incorrectly assigns a few frames to the native state, which results in a relaxation timescale that is faster than it should be (Fig. 4). The tICA MSM, however, correctly assigns all of these frames separate from the native state, and so the relaxation timescale is estimated to be much slower. (In either the RMSD or tICA model, there are still a few misassignments to this state, which means the state does not get trimmed during the MSM construction process.)

2.2 Convergence of Folding@home Dataset

To track the convergence of the simulation, as data came in we built three MSMs using tICA and k-medoids clustering on the entire dataset. Each model used 250, 500, or 1000 states with the top two tICs (we did not use more tICs because these models did not have well-behaved implied timescales). We were most interested in the register shifted intermediate so we tracked

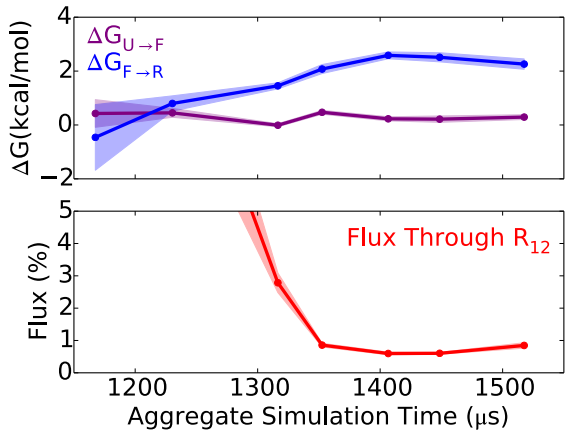

Figure 5: By building a range of MSMs on the same dataset we were able to judge the convergence of the simulations along a few observables corresponding to the register shifted state. After a total of 1.5 ms, the MSMs' estimates of the register shifted intermediate's stability have approximately converged.

three quantities as new data was added:

- The free energy between unfolded and folded states. A state was labeled folded if the average RMSD to the crystal structure was less than 6 Å and unfolded otherwise.
- The free energy between the folded and the near-native register-shifted states. A state was considered register-shifted if at least two of the register-shifted contacts were formed and the RMSD to the crystal was less than 6 Å.
- The flux through the register-shifted intermediate.

We stopped sampling once these observables had converged (Fig. 5). It is important to note that we do not have a good way to compute or even estimate the statistical uncertainty in these

quantities. The error bars shown in Fig. 5 simply corresponds to the standard deviation in the estimates for several models built with different numbers of states. Since the space of possible models is extremely large, it is difficult to sample enough models for the standard deviation reported in Fig. 5 to be a good estimate for the actual uncertainty in our estimates. Guided by intuition, we believe our models to be accurate to approximately one kCal / mol and one percentage point of flux, which would mean that these models have converged. Admittedly, this analysis would greatly benefit from a method for computing the statistical uncertainties more reliably, but such a method – to our knowledge – does not currently exist.

2.3 Suggested Mutations

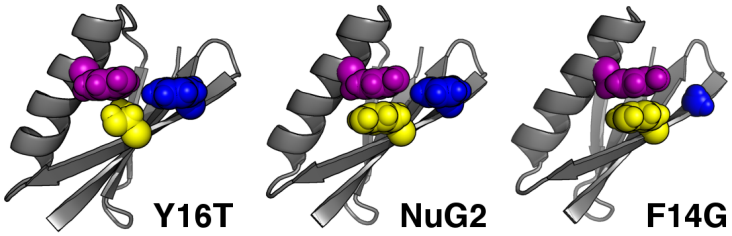

Figure 6: The Y16T (yellow residue) mutation may stabilize the register shifted intermediate relative to the native state by removing the Tyr16-Tyr33 interaction formed in the native state. Conversely, F14G (blue residue) should force the peptide to prefer the native fold by removing the Tyr33-Phe14 interaction formed in the register-shifted state. These mutations correspond to reversing the mutations made at those positions by Nauli et al. (8). Above, each mutation is depicted in the native conformation and Tyr33 is colored purple.

As discussed in the main text, Phe14 and Tyr16 appear to be competing for a contact with Tyr33 in the register shifted state and native state, respectively. Therefore, by mutating Tyr16 away, we may be able to destabilize the native state relative to the native state by removing the Tyr16-Tyr33 contact formed in the native fold. Conversely, we can attempt to destabilize the register-shifted state by mutating Phe14 and removing the Phe14-Tyr33 contact formed in the register shift Fig. 6. We emphasize that this is only a hypothesis, and without doing additional simulations we can't say for sure whether this interaction is a major contributor to the stability of the register shift. However, together with previous results, we believe that stable, register-shifted states likely only occur when the hydrophobic core is not disrupted. Therefore, the hydrophobic contacts represent one way to mutate the peptide and disrupt the native fold relative to the

register shift.

2.4 Experimental Observables

The original experiments performed by Nauli et al. (8) to monitor folding were based on tryptophan fluorescence. However, the tryptophan in NuG2 is on strand three and its environment in the native and register-shifted states is essentially the same (Fig. 7). This means that the additional tryptophan fluorescence would likely be the same in either the register-shifted or native states.

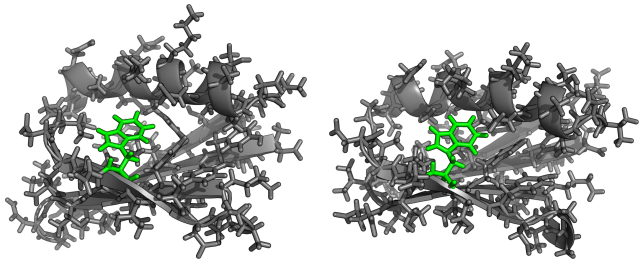

Figure 7: Trp42 is in strand three of the folded state in NuG2. As a result, it exists in largely the same environment regardless of whether there is a register shift in strand two or not. This means that tryptophan fluorescence experiments would be unable to discern between the native (left) and register-shifted (right) folds.

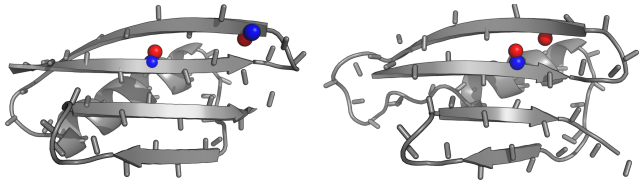

Figure 8: Two $^{13}C{=}^{18}O$ labels can be used to detect the presence of the register shifted state in NuG2. These probes will be coupled via the beta sheet when in the register shifted conformation (right), while uncoupled in the native register (left). This coupling gives rise to a large absorption in the IR spectrum (12).

We believe that a more targeted experiment can be used to verify the presence of this register-shifted state. Since the backbone's hydrogen bonding network is different in the register-shifted state, we suggest adding $^{13}C{=}^{18}O$ labels to Leu5 and Thr12, which are on strands one and two, respectively (Fig. 8). In the native fold, these residues are separated, whereas they are highly

coupled in the register shifted state. When two heavy carbonyls are coupled, they give rise to an anomalously large IR absorption (12), which should be easily detected at equilibrium.

References

1. Hess, B., 2008. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput.* 4:116–122.

2. Shirts, M., and V. S. Pande, 2000. Screen savers of the world unite! *Science* 290:1903–1904.

3. Piana, S., K. Lindorff-Larsen, and D. E. Shaw, 2011. How robust are protein folding simulations with respect to force field parameterization? *Biophys. J.* 100:L47–L49.

4. Jorgensen, W. L., J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, 1983. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* 79:926–935.

5. Hess, B., H. Bekker, H. J. Berendsen, J. G. Fraaije, et al., 1997. LINCS: a linear constraint solver for molecular simulations. *J. Comp. Chem.* 18:1463–1472.

6. Darden, T., D. York, and L. Pedersen, 1993. Particle mesh Ewald: an N log (N) method for Ewald sums in large systems. *J. Chem. Phys.* 98:10089–10092.

7. Lindorff-Larsen, K., S. Piana, R. O. Dror, and D. E. Shaw, 2011. How fast-folding proteins fold. *Science* 334:517–520.

8. Nauli, S., B. Kuhlman, and D. Baker, 2001. Computer-based redesign of a protein folding pathway - Nature Structural & Molecular Biology. *Nat. Struct. Biol.* 8:602–605.

9. Beauchamp, K. A., G. R. Bowman, T. J. Lane, L. Maibaum, I. S. Haque, and V. S. Pande, 2011. MSMBuilder2: Modeling Conformational Dynamics on the Picosecond to Millisecond Scale. *J. Chem. Theory Comput.* 7:3412–3419.

10. Beauchamp, K. A., R. T. McGibbon, Y.-S. Lin, and V. S. Pande, 2012. Simple few-state models reveal hidden complexity in protein folding. *Proc. Natl. Acad. Sci. U.S.A.* 109:17807–17813.

11. Noé, F., C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, 2009. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. U.S.A.* 106:19011–19016.

12. Huang, R., L. Wu, D. McElheny, P. Bouř, A. Roy, and T. A. Keiderling, 2009. Cross-strand coupling and site-specific unfolding thermodynamics of a trpzip β-hairpin peptide using 13C isotopic labeling and IR spectroscopy. *J. Phys. Chem. B* 113:5661–5674.