Biophysical*Journal*
Biophysical Letter

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

Christian R. Schwantes,[1] Diwakar Shukla,[1,2] and Vijay S. Pande[1,3,4,5,*]
[1]Department of Chemistry, [2]SIMBIOS NIH Center for Biomedical Computation, [3]Biophysics Program, [4]Structural Biology, and [5]Department of Computer Science, Stanford University, Stanford, California

ABSTRACT After reanalyzing simulations of NuG2—a designed mutant of protein G—generated by Lindorff-Larsen et al. with time structure-based independent components analysis and Markov state models as well as performing 1.5 ms of additional sampling on Folding@home, we found an intermediate with a register-shift in one of the $\beta$-sheets that was visited along a minor folding pathway. The minor folding pathway was initiated by the register-shifted sheet, which is composed of solely nonnative contacts, suggesting that for some peptides, nonnative contacts can lead to productive folding events. To confirm this experimentally, we suggest a mutational strategy for stabilizing the register shift, as well as an infrared experiment that could observe the nonnative folding nucleus.

There are many important questions surrounding the physics of protein folding that remain unanswered (1). Why do proteins fold quickly? What is the role of nonnative interactions in the folding process? Are there multiple pathways to the folded state? Molecular dynamics (MD) simulation has proven to be a powerful tool that can provide an atomic level answer to these and other biophysical questions (2–4) as well as a way for interpreting and predicting new experiments (5). Nonetheless, making sense of large and high-dimensional MD datasets can be difficult, but the analysis can be made simpler by employing Markov state models (MSMs). An MSM consists of a set of states (groups of peptide conformations) and the probabilities of interconversion between those states (6–9). There are two main steps in the MSM construction process. The first is clustering the data into some (preferably small) set of states and the second is estimating the probabilities of transitioning between states. Recent work has highlighted the importance of the state decomposition: the general features of the model depend significantly on the choice of the state space and basic structural metrics, such as root mean-square deviation (RMSD) in atom positions or dihedral angles may not be the best choice for clustering protein conformations (10). In addition, recent improvements have illustrated that distance metrics designed to ignore fast degrees of freedom can produce a superior state decomposition and provide better estimates of kinetic and thermodynamic observables (11,12). Alternative approaches, which use an energy-minimization during the clustering step, can also be useful for avoiding pathologies of strictly structural metrics (13).

In this Letter, we discuss the time structure-based independent components analysis (tICA) method applied to simulations of the peptide NuG2. This protein is a mutant of protein G, which was computationally designed to fold faster via mutations that stabilized a $\beta$-sheet between strands 1 and 2 (14). Previously, Beauchamp et al. (15) built an MSM on the dataset generated by Lindorff-Larsen et al. (16) utilizing the RMSD of atom positions as the distance metric during the clustering step. We reanalyzed the same dataset, but used tICA to build the new MSM (see MSM Construction in the Supporting Material). When we compared the RMSD-based model to one built using the tICA metric, we found that there was a new slow timescale (~180 $\mu$s) in the tICA model that was absent in the RMSD-based model (Fig. 1). This slow process corresponded to a near-native state, which had a two-residue register shift in the sheet formed between strands 1 and 2. In fact, this process was also observed in the RMSD-based model (see Fig. S1 in the Supporting Material), but the corresponding timescale was two orders-of-magnitude faster in the original model (Fig. 1). The timescale's sensitivity to the choice of state decomposition illustrated that this eigenprocess was not adequately sampled in the original simulation. In fact, this register-shifted state was only visited at the very end of a single trajectory. To improve our estimate of the register-shifted state's kinetic and thermodynamic properties, we used Folding@home to generate 1.5 ms of aggregate

*Submitted March 11, 2015, and accepted for publication March 7, 2016.*
*Correspondence: pande@stanford.edu
Editor: Nathan Baker.
http://dx.doi.org/10.1016/j.bpj.2016.03.026

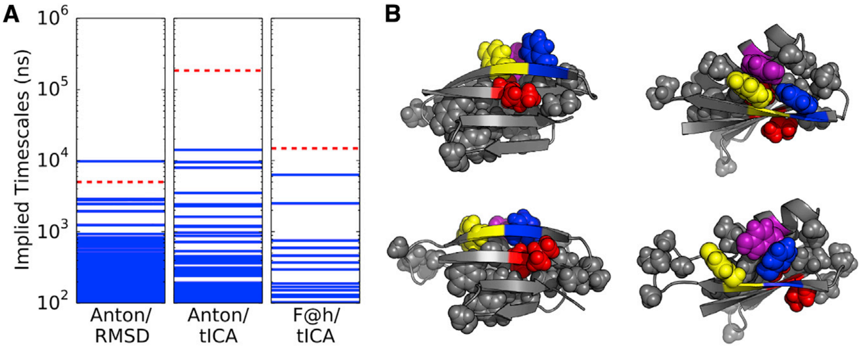

**FIGURE 1** (*A*) A slow eigenprocess (*red, dashed line*) was found in the both the tICA and RMSD models built with the Anton dataset, but with two very different timescales indicating a large uncertainty in the model. After producing a new dataset on Folding@home, we found that the timescale was best estimated to be ~15 $\mu$s. (*B*) The slow timescale corresponded to a two-residue register shift in strand two. The native state is shown in the top two images, and the register shift in the bottom two. There is a corresponding two-residue shift in the hydrophobic core, with Tyr$^{16}$ (*yellow*) contacting Tyr33 (*purple*) in the native state, while in the register-shifted state, Phe14 (*blue*) stacks with Tyr33 (*purple*). To see this figure in color, go online.

sampling (17) (see Folding@home Sampling in the Supporting Material). The folding timescale of the new MSM was slightly faster (~6 $\mu$s) than previous models, likely due to using one order-of-magnitude fewer states. However, recent results have illustrated that models built with greater than a few hundred states may be overfit to the observed data (18), and we are therefore more confident in this model than the previous ones.

Using transition path theory (19), we found that most of the reactive flux went through a pathway that first forms the sheet between strands 1 and 2, and then forms the remainder of the native secondary structure (Fig. 2). However, ~1% of the flux flowed through the register-shifted state along a pathway that again first forms the sheet between strands 1 and 2, but with a two-residue register shift. After the formation of this sheet, the remainder of the secondary structure forms as it is in the native state and then in a final step, strand 2 shifts to the native register. Because the register shift contacts are nonnative, this pathway is nucleated by a state that is made up of entirely nonnative secondary structure.

The register-shifted state is fairly stable, having an equilibrium free energy ($\Delta G$) of ~2 kcal/mol above the native state. Interestingly, the two-residue shift in secondary structure is also accompanied by a two-residue shift in the hydrophobic contacts in the core (Fig. 1). A similar shift in the hydrophobic contacts was observed for the register-shifted states found in NTL9$_{1-39}$ (11), suggesting a larger trend that register-shifted states can be stabilized by favorable hydrophobic packing. In fact, these results suggest that register shifts are more probable when the shift in register does not significantly disrupt the hydrophobic core.

By comparing the hydrophobic contacts in the register-shifted and native folds, Tyr16 and Phe14 appear to be in competition for forming a contact with Tyr33 (Fig. 1). Therefore, by mutating one of these residues we believe it is possible to stabilize (or destabilize) the register-shifted state relative to the native fold. For instance, the Y16T mutation (which reverses one of the mutations made by Nauli et al. (14)) would remove the Tyr16-Tyr33 contact formed in the native fold and possibly force the protein to adopt a register-shifted conformation with Phe14 contacting Tyr33. The stacking of two benzenes has been estimated to be between 1 and 2 kcal/mol (20), and so, this mutation could shift the population significantly. However, we note that other mutations, which disrupt the contacts with Tyr33 or the hydrophobic packing, could also be useful. Because the

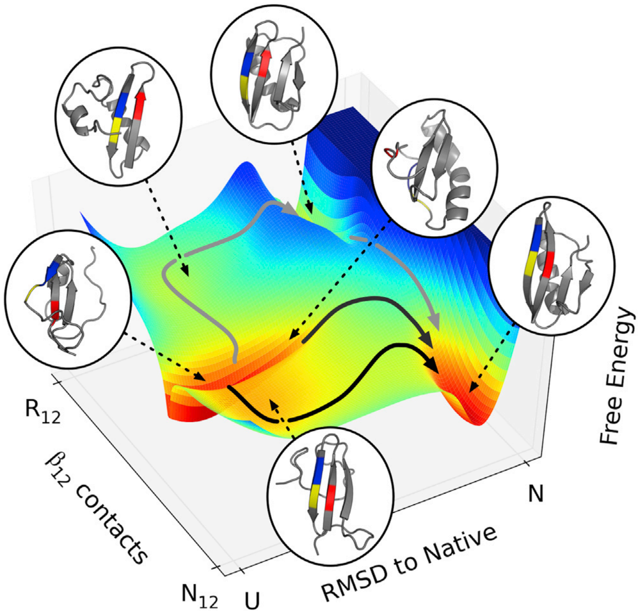

**FIGURE 2** Folding occurred along three main pathways. The largest flux pathway (~55%) proceeded through the native sheet between strands 1 and 2 ($N_{12}$) and then to the folded state. The second largest pathway (~10%) proceeded by first forming the sheet between strands 3 and 4, and then folding to the native state. Additionally, a low-flux (~1%) pathway proceeded through the register-shifted state ($R_{12}$). To see this figure in color, go online.

register-shifted conformation has a significantly different backbone hydrogen-bonding network, infrared (IR) spectroscopy with carefully placed $^{13}C{=}^{18}O$ probes provides a powerful tool to confirm our observations. The stability of the register-shifted state puts observing this conformation at the cusp of what is possible with a conventional IR experiment. Therefore, we suggest leveraging the between-strand coupling of two heavy carbonyl labels to observe the state. Briefly, when two labeled carbonyls are on adjacent strands and within one or two residues, an anomalously large IR absorption occurs (21). We suggest labeling Leu5 and Thr12, which will be adjacent in the register-shifted conformation but separated by two residues in the native fold (see Fig. S6).

Our results provide compelling evidence for the necessity of kinetically informed distance metrics, and shed light on the limits of simple structural metrics such as RMSD. In addition, our new dataset indicates that the folding of NuG2 can proceed through a minor pathway that is nucleated entirely by nonnative secondary structure. Many have assumed that nonnative contacts can only give rise to so-called glassy free-energy landscapes that will slow the folding process (22). However, the mutations made by Nauli et al. (14) that introduced these stable nonnative contacts actually increased the rate of folding by several orders of magnitude. These observations are consistent with previous work from Clementi and Plotkin (23), who found that favorable nonnative interactions can speed up the folding reaction in simplified models, but because we studied a real protein, we can go further to suggest specific experiments that can verify (or refute) our conclusions. In fact, register shifts have been observed in many other MD studies (10,24), and these nonnative folding nuclei may be quite prevalent. So long as the nonnative structure can correct itself without completely unfolding, then nonnative contacts may lead to productive folding events.

## SUPPORTING MATERIAL

Supporting Materials and Methods, Supporting Results, and eight figures are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(16)30107-2.

## AUTHOR CONTRIBUTIONS

C.R.S designed and performed the research, analyzed the data, and wrote the article; D.S. ran simulations and wrote the article; and V.S.P. designed the research and wrote the article.

## ACKNOWLEDGMENTS

We thank Robert T. McGibbon, Thomas J. Lane, and Kyle Beauchamp for discussions regarding MSMs as well as Caitlin Davis, Yu-Shan Lin, and Stephen Fried for useful discussions regarding IR spectroscopy.

The V.S.P. group gratefully acknowledges support from grants No. NIH-R01-GM062868 and No. NSF-MCB-0954714 and the SIMBIOS NIH Center for Biomedical Computation through the National Institutes of Health Roadmap for Medical Research Grant (No. U54-GM07297).

## SUPPORTING CITATIONS

References (25–29) appear in the Supporting Material.

## REFERENCES

1. Dill, K. A., and J. L. MacCallum. 2012. The protein-folding problem, 50 years on. *Science.* 338:1042–1046.
2. Lane, T. J., D. Shukla, …, V. S. Pande. 2013. To milliseconds and beyond: challenges in the simulation of protein folding. *Curr. Opin. Struct. Biol.* 23:58–65.
3. Buch, I., T. Giorgino, and G. De Fabritiis. 2011. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA.* 108:10184–10189.
4. Kelley, N. W., V. Vishal, …, V. S. Pande. 2008. Simulating oligomerization at experimental concentrations and long timescales: a Markov state model approach. *J. Chem. Phys.* 129:214707.
5. Voelz, V. A., M. Jäger, …, V. S. Pande. 2012. Slow unfolded-state structuring in Acyl-CoA binding protein folding revealed by simulation and experiment. *J. Am. Chem. Soc.* 134:12565–12577.
6. Senne, M., B. Trendelkamp-Schroer, …, F. Noé. 2012. EMMA: a software package for Markov model building and analysis. *J. Chem. Theory Comput.* 8:2223–2238.
7. Prinz, J.-H., H. Wu, …, F. Noé. 2011. Markov models of molecular kinetics: generation and validation. *J. Chem. Phys.* 134:174105.
8. Schütte, C., F. Noé, …, E. Vanden-Eijnden. 2011. Markov state models based on milestoning. *J. Chem. Phys.* 134:204105.
9. Bowman, G. R., X. Huang, and V. S. Pande. 2009. Using generalized ensemble simulations and Markov state models to identify conformational states. *Methods.* 49:197–201.
10. Kellogg, E. H., O. F. Lange, and D. Baker. 2012. Evaluation and optimization of discrete state models of protein folding. *J. Phys. Chem. B.* 116:11405–11413.
11. Schwantes, C. R., and V. S. Pande. 2013. Improvements in Markov state model construction reveal many non-native interactions in the folding of NTL9. *J. Chem. Theory Comput.* 9:2000–2009.
12. Pérez-Hernández, G., F. Paul, …, F. Noé. 2013. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* 139:015102.
13. Wales, D. J. 2010. Energy landscapes: some new horizons. *Curr. Opin. Struct. Biol.* 20:3–10.
14. Nauli, S., B. Kuhlman, and D. Baker. 2001. Computer-based redesign of a protein folding pathway. *Nat. Struct. Biol.* 8:602–605.
15. Beauchamp, K. A., R. McGibbon, …, V. S. Pande. 2012. Simple few-state models reveal hidden complexity in protein folding. *Proc. Natl. Acad. Sci. USA.* 109:17807–17813.
16. Lindorff-Larsen, K., S. Piana, …, D. E. Shaw. 2011. How fast-folding proteins fold. *Science.* 334:517–520.
17. Shirts, M., and V. S. Pande. 2000. COMPUTING: screen savers of the world unite! *Science.* 290:1903–1904.
18. McGibbon, R. T., C. R. Schwantes, and V. S. Pande. 2014. Statistical model selection for Markov models of biomolecular dynamics. *J. Phys. Chem. B.* 118:6475–6481.
19. Noé, F., C. Schütte, …, T. R. Weikl. 2009. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. USA.* 106:19011–19016.
20. Jorgensen, W. L., and D. L. Severance. 1990. Aromatic-aromatic interactions: free energy profiles for the benzene dimer in water, chloroform, and liquid benzene. *J. Am. Chem. Soc.* 112:4768–4774.
21. Huang, R., L. Wu, …, T. A. Keiderling. 2009. Cross-strand coupling and site-specific unfolding thermodynamics of a trpzip $\beta$-hairpin peptide using 13C isotopic labeling and IR spectroscopy. *J. Phys. Chem. B.* 113:5661–5674.
22. Onuchic, J. N., and P. G. Wolynes. 2004. Theory of protein folding. *Curr. Opin. Struct. Biol.* 14:70–75.

23. Clementi, C., and S. S. Plotkin. 2004. The effects of nonnative interactions on protein folding rates: theory and simulation. *Protein Sci.* 13:1750–1766.

24. Baiz, C. R., Y.-S. Lin, …, A. Tokmakoff. 2014. A molecular interpretation of 2D IR protein folding experiments with Markov state models. *Biophys. J.* 106:1359–1370.

25. Hess, B. 2008. P-LINCS: a parallel linear constraint solver for molecular simulation. *J. Chem. Theory Comput.* 4:116–122.

26. Hess, B., H. Bekker, …, J. G. E. M. Fraaije. 1997. LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.* 18:1463–1472.

27. 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.

28. Jorgensen, W. L., J. Chandrasekhar, …, M. L. Klein. 1983. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79:926–935.

29. 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.