

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

Christian R. Schwantes,<sup>1</sup> Diwakar Shukla,<sup>1,2</sup> and Vijay S. Pande<sup>1,3,4,5,\*</sup>

<sup>1</sup>Department of Chemistry, <sup>2</sup>SIMBIOS NIH Center for Biomedical Computation, <sup>3</sup>Biophysics Program, <sup>4</sup>Structural Biology, and <sup>5</sup>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 ( $\sim 180 \mu\text{s}$ ) 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

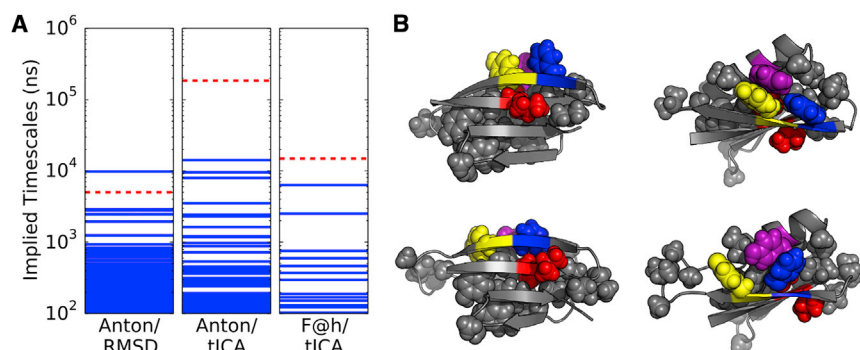
Submitted March 11, 2015, and accepted for publication March 7, 2016.

\*Correspondence: [pande@stanford.edu](mailto:pande@stanford.edu)

Editor: Nathan Baker.

<http://dx.doi.org/10.1016/j.bpj.2016.03.026>

© 2016 Biophysical Society



**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  $\sim 15 \mu\text{s}$ . (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<sup>16</sup> (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.

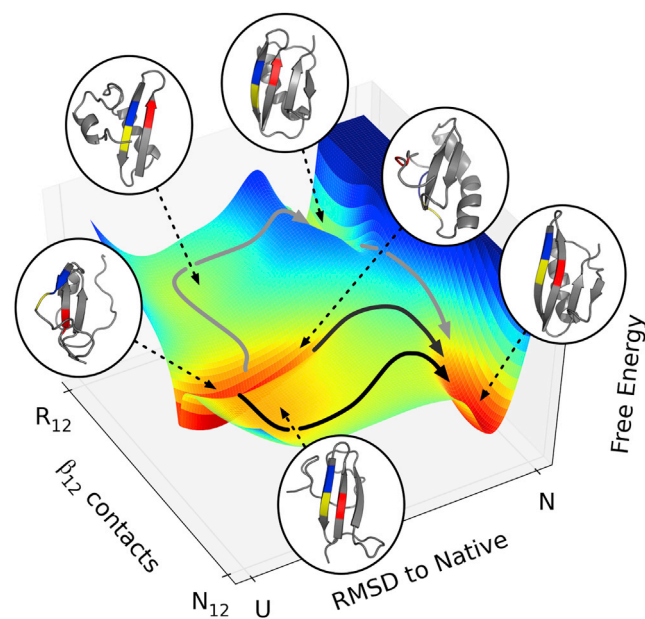
sampling (17) (see Folding@home Sampling in the [Supporting Material](#)). The folding timescale of the new MSM was slightly faster ( $\sim 6 \mu\text{s}$ ) than previous models, likely due to using one order-of-magnitude fewer states. However, recent results have illustrated that models built with greater 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,  $\sim 1\%$  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  $\sim 2$  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<sub>1-39</sub> (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 muta-

tion (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 ( $\sim 55\%$ ) proceeded through the native sheet between strands 1 and 2 ( $N_{12}$ ) and then to the folded state. The second largest pathway ( $\sim 10\%$ ) proceeded by first forming the sheet between strands 3 and 4, and then folding to the native state. Additionally, a low-flux ( $\sim 1\%$ ) 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  $^{13}\text{C}=^{18}\text{O}$  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](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

- Dill, K. A., and J. L. MacCallum. 2012. The protein-folding problem, 50 years on. *Science*. 338:1042–1046.
- 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.
- 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.
- 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.
- 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.
- 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.
- Prinz, J.-H., H. Wu, ..., F. Noé. 2011. Markov models of molecular kinetics: generation and validation. *J. Chem. Phys.* 134:174105.
- Schütte, C., F. Noé, ..., E. Vanden-Eijnden. 2011. Markov state models based on milestoning. *J. Chem. Phys.* 134:204105.
- 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.
- 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.
- 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.
- Pérez-Hernández, G., F. Paul, ..., F. Noé. 2013. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* 139:015102.
- Wales, D. J. 2010. Energy landscapes: some new horizons. *Curr. Opin. Struct. Biol.* 20:3–10.
- Nauli, S., B. Kuhlman, and D. Baker. 2001. Computer-based redesign of a protein folding pathway. *Nat. Struct. Biol.* 8:602–605.
- 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.
- Lindorff-Larsen, K., S. Piana, ..., D. E. Shaw. 2011. How fast-folding proteins fold. *Science*. 334:517–520.
- Shirts, M., and V. S. Pande. 2000. COMPUTING: screen savers of the world unite! *Science*. 290:1903–1904.
- 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.
- 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.
- 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.
- Huang, R., L. Wu, ..., T. A. Keiderling. 2009. Cross-strand coupling and site-specific unfolding thermodynamics of a trpzip  $\beta$ -hairpin peptide using  $^{13}\text{C}$  isotopic labeling and IR spectroscopy. *J. Phys. Chem. B*. 113:5661–5674.
- 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.

**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,<sup>†</sup> Diwakar Shukla,<sup>†‡</sup> Vijay S. Pande<sup>†+§¶</sup>

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

## Contents

|          |   |          |
|----------|---|----------|
| <b>1</b> | <b>Methods</b>  | <b>2</b> |
| 1.1      | Simulation Details . . . . .                            | 2        |
| 1.1.1    | Folding@home MD Simulations . . . . .                   | 2        |
| 1.1.2    | Starting Structures for Folding@home Sampling . . . . . | 2        |
| 1.2      | MSM Construction . . . . .                              | 3        |
| 1.3      | MSM Analysis . . . . .                                  | 4        |
| 1.3.1    | Register Shift Eigenprocesses . . . . .                 | 4        |
| 1.3.2    | Transition Path Theory . . . . .                        | 5        |
| 1.3.3    | “Free Energy” Surface Definition . . . . .              | 6        |
| <b>2</b> | <b>Miscellaneous Results</b>                            | <b>7</b> |
| 2.1      | Misassignment in the RMSD MSM . . . . .                 | 7        |
| 2.2      | Convergence of Folding@home Dataset . . . . .           | 8        |
| 2.3      | Suggested Mutations . . . . .                           | 9        |
| 2.4      | Experimental Observables . . . . .                      | 10       |

# 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<sup>+</sup> and Cl<sup>-</sup> 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  $\mu$ s. 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  $\mu$ s 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  $\mu$ s 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<sup>th</sup> of the data (because of the limits of the clustering method) with the RMSD metric on atom positions. This model used  $\sim 3500$  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 quantitative calculation; it was only used to guide our Folding@home simulations.



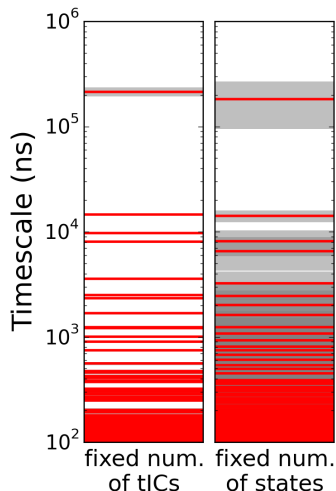


Figure 1: Many MSMs were built on the Anton dataset using tICA. Three models built with 15,000, 20,000, and 25,000 states with six tICs are compared on the left. The means of the timescales across these models are drawn in red, and the gray areas represent one standard deviation in either direction. 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

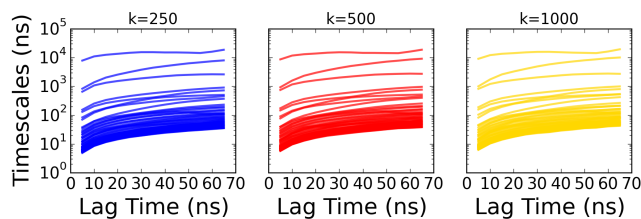


Figure 2: The implied timescales converged within 50 ns for the full Folding@home dataset. Other models, however, did not have well-behaved timescales and were not used for the final analysis.

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.

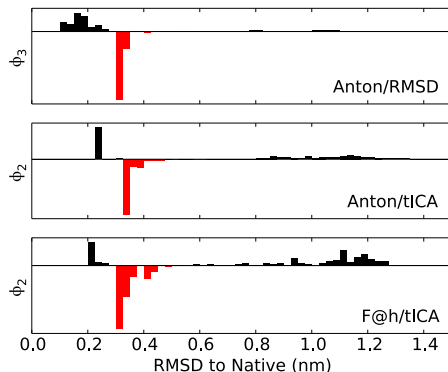


Figure 3: The eigenvectors of all three MSMs were projected onto each state’s average RMSD to the native conformation. In all cases, the eigenprocesses corresponded to the exchange between near-native states (red) and other states in the simulation. These near-native states correspond to the register-shifted 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  $\beta_{12}$  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,

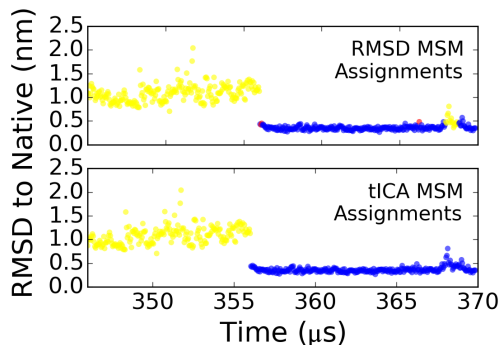


Figure 4: Above, the color indicates the macrostate that a point is assigned to according to the RMSD or tICA model: unfolded (yellow), folded (red), register-shifted (blue). Because the RMSD MSM misassigns a few register-shifted states to the folded and unfolded states, the timescale is much faster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

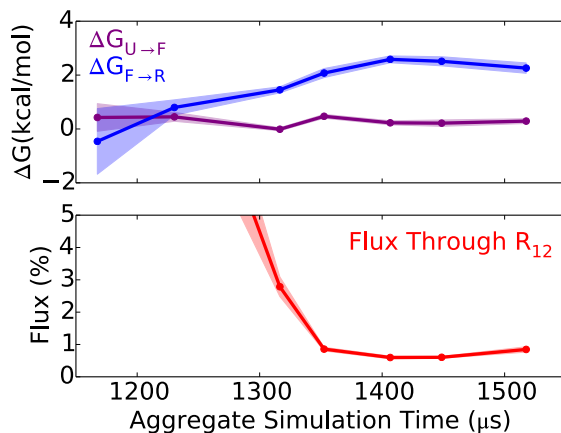


Figure 5: By building a range of MSMs on the same dataset we were able to judge the convergence of the simulations along a few observables corresponding to the register shifted state. After a total of 1.5 ms, the MSMs’ estimates of the register shifted intermediate’s stability have approximately converged.

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

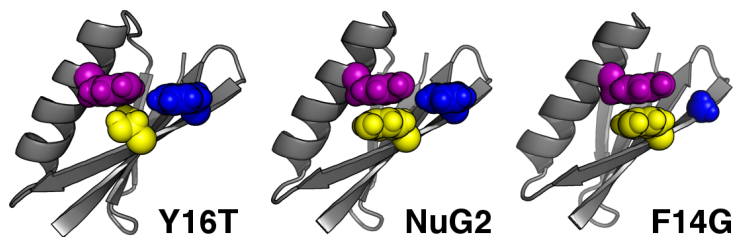


Figure 6: The Y16T (yellow residue) mutation may stabilize the register shifted intermediate relative to the native state by removing the Tyr16-Tyr33 interaction formed in the native state. Conversely, F14G (blue residue) should force the peptide to prefer the native fold by removing the Tyr33-Phe14 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 ad-

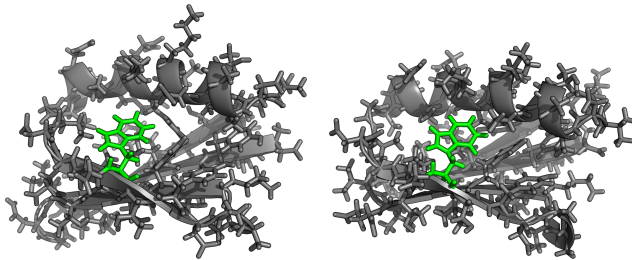


Figure 7: Trp42 is in strand three of the folded state in NuG2. As a result, it exists in largely the same environment regardless of whether there is a register shift in strand two or not. This means that tryptophan fluorescence experiments would be unable to discern between the native (left) and register-shifted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

ditional tryptophan fluorescence would likely be the same in either the register-shifted or native states.

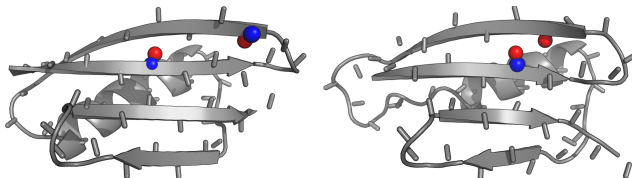


Figure 8: Two  $^{13}\text{C}=\text{}^{18}\text{O}$  labels can be used to detect the presence of the register shifted state in NuG2. These probes will be coupled via the beta sheet when in the register shifted conformation (right), while uncoupled in the native register (left). This coupling gives rise 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  $^{13}\text{C}=\text{}^{18}\text{O}$  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  $\beta$ -hairpin peptide using  $^{13}\text{C}$  isotopic labeling and IR spectroscopy. *J. Phys. Chem. B* 113:5661–5674.