JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

COMMUNICATION

pubs.acs.org/JACS

Atomistic Folding Simulations of the Five-Helix Bundle Protein λ_{6-85}

Gregory R. Bowman,[†] Vincent A. Voelz,[†] and Vijay S. Pande*[,†,‡]

[†]Department of Chemistry and [‡]Biophysics Program, Stanford University, Stanford, California 94305, United States

S *Supporting Information*

ABSTRACT: Protein folding is a classic grand challenge that is relevant to numerous human diseases, such as protein misfolding diseases like Alzheimer's disease. Solving the folding problem will ultimately require a combination of theory, simulation, and experiment, with theory and simulation providing an atomically detailed picture of both the thermodynamics and kinetics of folding and experimental tests grounding these models in reality. However, theory and simulation generally fall orders of magnitude short of biologically relevant time scales. Here we report significant progress toward closing this gap: an atomistic model of the folding of an 80-residue fragment of the λ repressor protein with explicit solvent that captures dynamics on a 10 milliseconds time scale. In addition, we provide a number of predictions that warrant further experimental investigation. For example, our model's native state is a kinetic hub, and biexponential kinetics arises from the presence of many free-energy basins separated by barriers of different heights rather than a single low barrier along one reaction coordinate (the previously proposed incipient downhill folding scenario).

Understanding protein folding is a long-standing problem with important medical applications, such as elucidating the role of protein misfolding in disorders like Alzheimer's disease. Solving the folding problem will ultimately require a combination of theory, simulation, and experiment, with theory and simulation providing an atomically detailed picture of both the thermodynamics and kinetics of folding and experimental tests grounding these models in reality. However, modeling long-time-scale dynamics (e.g., microseconds, milliseconds, and beyond) with sufficient statistical accuracy and chemical detail to make a quantitative connection with experiments is extremely challenging. Much progress has been made with small, fast-folding proteins (less than 40 residues and 1 ms folding time scales[1]), but can the methods used be scaled to larger, slower systems? Here we report significant progress in this direction: an atomistic model of the folding of an 80-residue fragment of the λ repressor protein (λ_{6-85}) with explicit solvent that captures dynamics on a 10 milliseconds time scale.

This advance builds upon a growing body of work on describing molecular kinetics with network models called Markov state models (MSMs). MSMs are discrete-time master equation models that essentially serve as maps of a molecule's conformational space.[1−3] The states in an MSM come from kinetic clustering of atomistic simulations (i.e., grouping together conformations that can interconvert rapidly into what is called a metastable state). Thus, these models are an important advance over previous approaches, such as diffusion−collision models,[4,5] as an MSM's states are derived from dynamics in detailed simulations rather than human intuition. One can exploit the kinetic definition of states in an MSM to perform simulations efficiently[6−8] and make a direct connection to experiments.[9−11] For example, we have successfully used MSMs for all-atom ab initio structure prediction of small systems such as the villin headpiece (35 residues, microsecond folding time).[9] Noe et al.[10] predicted the relaxation kinetics of a PinWW domain (34 residues, microsecond folding time), and Voelz et al.[11] did the same for NTL9 (39 residues, millisecond folding time).

To test whether the MSM approach can be scaled to larger systems, we built MSMs for the D14A mutant of λ_{6-85} (Figure 1A).[12] D14A has the following mutations: D14A, Y22W, Q33Y, G46A, and G48A. This system was chosen because it is twice as large as the small model systems that have been studied with MSMs to date yet surprisingly is still reported to fold on a 10 μs time scale.[12] Since molecular dynamics (MD) simulations can now reach time scales of tens of microseconds on a routine basis, it should be feasible to run many folding simulations for this system. Future comparison with other large, slower-folding proteins could also help us to understand what properties of D14A allow it to fold as quickly as ultrafast folding proteins less than half its size.

MSMs for D14A. We built atomically detailed network models (MSMs) for D14A from 3265 MD simulations with explicit solvent. Each trajectory was up to 1 μs in length, for an aggregate of 1.3 ms of simulation—an enormous data set given that most simulation studies are based on only nanoseconds to microseconds of data. These simulations were started from six initial configurations drawn from replica-exchange simulations in implicit solvent.[13] One was nativelike, three were partially unfolded, and two had β-sheets. A more detailed description of our simulations is given in the Supporting Information (SI).

The highest-resolution MSM we created for D14A has 30 000 microstates and is appropriate for making quantitative connections with experiments because of its great structural and temporal detail. A low-resolution model with 5000 macrostates was created from the high-resolution MSM to facilitate interpretation of the model. More details on our use of the MSMBuilder package[14] to construct these models are given in the SI. While no single trajectory visited every state, these MSMs were able to capture long-time-scale dynamics through the use of overlap between our simulations, which allowed them to be stitched together in a physically and statistically meaningful way (see Figure S1 in the SI for a 1 s long trajectory). Examination of the implied time

Received: August 12, 2010
Published: December 21, 2010

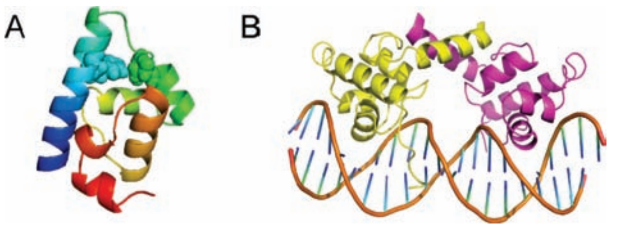

Figure 1. (A) Model of λ_{6-85} taken from (B). The Trp22−Tyr33 pair that has been monitored in *T*-jump experiments is shown as space-filled. (B) Crystal structure of the λ_{1-92} dimer bound to DNA (PDB entry 1LMB).

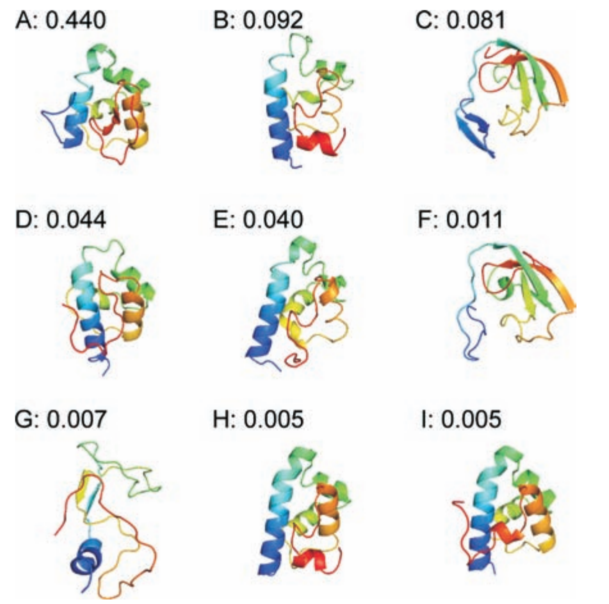

Figure 2. The nine most populated states from our coarse-grained MSM with their equilibrium probabilities.

scales of the microstate MSM shows that a 5 ns lag time yields Markovian behavior (Figures S2−S4).

The Native State. One of the first observations from our coarse-grained MSM is that our model's native state (Figure 2A) differs from the crystal structure (Figure 1A) in helix five. The crystal structure is a highly metastable state (Figure 2H), which we call the crystallographic state. However, in the native state of our model, helix five is unraveled and packed against the side of the remainder of the protein (Figure 2A). Figure 2 also shows that helix five is unstructured in many of the other highly populated states of our model.

While this difference could be due to force-field errors, we argue that helix five is actually likely to be unstructured in solution given the origins of this model system for folding. Full-length λ repressor is a 236-residue transcription factor that binds to DNA as a dimer, maintaining the λ phage in the lysogenic state. Figure 1B shows the crystal structure of a 92-residue fragment that can still dimerize and bind to DNA.[15,16] On the basis of this structure, Huang and Oas[17] selected an 80-residue fragment (λ_{6-85}) that favors the monomeric state (Figure 1A), making it appropriate for folding studies. In the 92-residue fragment, helix five is extended by seven residues and forms important packing interactions between the two members of the dimer. These extra interactions likely stabilize helix five. Truncating the sequence to favor the monomer could destabilize the fifth helix, leading to a lack of structure and a strong propensity either to fill the hydrophobic cavity normally occupied by the corresponding helix of the other member of the dimer or to adopt one of a number of the other well-populated, unstructured conformations shown in Figure 2.

There is also a reasonable amount of experimental data corroborating our hypothesis that helix five is unstructured in solution. First, the stability of this system seems to be insensitive to mutations in helix five.[13] A crystal structure for λ_{6-85} also has high *B* factors in helix five.[18] Therefore, it is plausible that helix five is stabilized by packing interactions in this crystal but is still intrinsically unstable and likely to be more unstructured in solution.

Further support for our hypothesis comes from theoretical studies. For example, helix five has negligible helical propensity according to Agadir[19] (Figure S5). Similar results were also found in a Go model study, where helix five tended to undock from the rest of the protein.[20] However, those models did not include non-native interactions, so helix five was not found to unravel or pack against the protein in that work.

***β*-Sheet States.** Figure 2 also shows that a number of the most populated states in our model have significant β-sheet content. The prediction of β-sheet states in the unfolded ensemble is somewhat surprising for a helical protein; however, experiments have shown that the unfolded and denatured states of many systems can have significant populations of compact, β-sheet structures yet still display the random coil statistics characteristic of expanded conformations.[21,22] Thus, our prediction of compact β-sheet structures is not unreasonable.

Folding Kinetics. While the experimentally reported folding time for D14A is 10 μs, analysis of our high-resolution MSM revealed the presence of microscopic transitions on time scales up to 10 ms. These time scales were preserved in subsamples of the data set and an independent data set run at a lower temperature (Figures S3 and S4), indicating that they are a robust feature of the simulated system.

Analysis of our coarse-grained MSM revealed that this long time scale corresponds to exchange between the compact β-sheet structures in the unfolded ensemble and the crystallographic state through multiple parallel pathways (Figure 3 and Figure S6). A more detailed view of one of these pathways and portions thereof are shown in Figure 4 and Figure S7. In this particular pathway, the compact β-sheet structure first expands (A−E), breaking apart the β-sheets. Next, helices 1 and 4 begin to form, and this is followed by collapse into a nativelike topology (F and G). Finally, the remaining helices form (G and H). The ability to extract these detailed pathways highlights one of the advantages of MSMs over conventional analysis techniques such as projections of the free-energy surface, which tend to oversimplify folding and paint different pictures depending on the order parameters chosen (Figures S8 and S9 and ref 20). However, one must take care in interpreting these pathway diagrams because they show the net flux from one state to another, leaving out the backward steps and excursions that molecules in solution make as they stochastically explore the conformational space.

One possible explanation for the difference between our simulation results and experiment is that the experimental probe used to monitor folding is not sensitive to the slow transition from compact β-sheet structures to the crystallographic state. To test this hypothesis, we used our MSM to calculate the macroscopic rate measured experimentally by modeling the relaxation

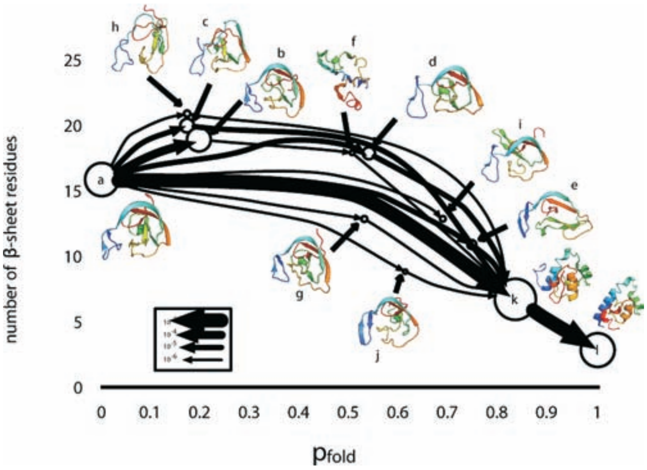

Figure 3. Coarse-grained view of the 10 ms time scale transition in which the size of the circle representing each state is proportional to the logarithm of that state's equilibrium probability and each arrow width is proportional to the logarithm of the flux along the corresponding edge (see the key in the figure). The states are laid out in terms of the average number of β-sheet residues (calculated from 100 random conformations from each state) and the value of p_{fold} (the probability of reaching state l before state a).

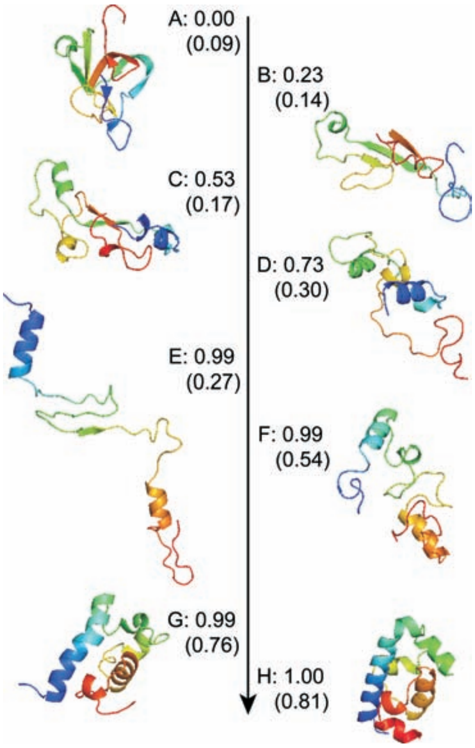

Figure 4. Representative high-resolution pathway that occurs on a 10 ms time scale, with p_{fold} values (probabilities of reaching state H before state A) shown. The proportion of native contacts is also given in parentheses as an estimate of how nativelike the topology is. Relative contact orders for each state are given in Table S1.

of a surrogate for the Trp22−Tyr33 quenching interaction measured in temperature-jump (T-jump) experiments (Figure S10). We also calculated the relaxation of the C$_\alpha$ root-mean-square deviation (rmsd) from the crystal structure to test whether a more global metric could capture longer time scales than the experimental probe, which could capture only local relaxations (Figure S10). Both exhibit biexponential relaxation (a characteristic of D14A that has been used to argue that it is a downhill folder) and have similar time scales, but their slow phase is ∼2 orders of magnitude slower than in experiment (1 ms vs 10 μs).

This result suggests that this slow transition is not present in solution because the experimental probe would have captured it if it were. Further support comes from the fact that ignoring simulations started from β-sheet structures yielded better agreement between the simulations and experiment (Figure S11). First, the Trp22−Tyr33 surrogate has a 1 μs fast phase and a 4.3 μs slow phase, in reasonable agreement with the experimental values of 2 and 10 μs.[12] Second, the rmsd relaxes on different time scales in this case, consistent with the observed probe-dependent kinetics.[23,24]

While it is natural to consider the potential flaws in a force field when confronted with a discrepancy between simulation and experiment, we suggest that there are alternative possibilities as well. The folding rate of λ_{6-85} is known to be highly sensitive to solvent viscosity.[25,26] For example, one variant of λ_{6-85} folds on a 210 μs time scale in the absence of denaturant but a 5 ms time scale in the presence of only 0.5 M GuHCL.[26] Force-field errors are known to destabilize proteins, so it is possible that our simulated system is more like D14A in mild denaturant than it is like D14A in aqueous solution. It is also still possible that future experiments will reveal the presence of a 10 ms time scale for D14A. Indeed, one might expect D14A, with its sizable hydrophobic core, to fold on longer time scales since the wild-type villin headpiece (which is less than half the size of D14A and barely has a hydrophobic core) is also reported to fold in just under 10 μs.[27]

Fully resolving this issue will likely require more experiments and simulations to yield more points of comparison between simulation and experiment. Regardless of the outcome, our work shows that MSMs built from atomistic simulations can now sample 10 ms time scales, reproduce qualitative phenomena such as biexponential relaxation, and possibly even provide quantitative agreement with experiment. Moreover, the ability to make such direct comparisons on long time scales opens the door to further improvements of atomistic models used in MD simulations.

A Native Hub. The biexponential relaxation of D14A and other variants of λ_{6-85} has previously been attributed to incipient downhill folding. The incipient downhill folding model is similar to the more conventional two-state model often used to describe folding but has a lower barrier (on the order of k_BT) separating the folded and unfolded states (Figure S12A). As a result, there is believed to be a moderate population of proteins on top of the barrier that can slide downhill into the native state, giving rise to a fast phase, in addition to an unfolded population that must cross the barrier before folding, giving rise to the slow phase.

Projections of the free energy onto a kinetically meaningful order parameter (p_{fold}, the probability of folding before unfolding[28]) are consistent with incipient downhill folding. When the full data set was used, such projections appeared to be two-state, but when simulations started from the compact β-sheet conformations were removed (thereby yielding better agreement with experiment), the barrier between the folded and unfolded states was greatly reduced, consistent with incipient downhill folding (Figure S13).

Further analysis of our MSM, however, indicated that the biexponential relaxation of D14A may be due to metastability and a hublike native state rather than incipient downhill folding. When a single non-native state was chosen as the starting point for p_{fold} calculations, the other non-native states actually appeared to have p_{fold} values near 1 (i.e., they appeared on the native side of the projection), indicating that the folding is more complex than the incipient downhill folding scenario. The MSM revealed that there are many metastable states separated by barriers of different heights (e.g., there is reasonable variability in the transition times between states), and the convolution of these dynamics gives rise to biexponential relaxation and fast folding. These states are arranged in such a way that the native state acts as a kinetic hub, as has been observed for a number of smaller systems.[29]

A first hint that D14A may also have a native hub comes from the large number of connections to our native state (Figure S12). The native state in our model makes direct connections to 98% of the non-native states, while non-native states connect to only 0.1% of the other states on average. Moreover, the mean first-passage times (MFPTs) to the native state were typically found to be ~10 times shorter than the MFPTs between non-native states, as shown in Figure S14, and this held regardless of whether the β-sheet simulations were included in the analysis. Therefore, molecules in non-native states can generally fold faster than they can transition to other non-native states. The fastest way to transition between two randomly selected non-native states is then to fold and unfold. The large number of folding pathways that result from this topology is hidden by projections of the free energy onto p_{fold}.

Conclusions. The combination of simulations and MSMs can now access ~10 milliseconds time scales for moderately large (~80 residue) systems with explicit solvent, greatly increasing the common ground between simulation and experiment (the previous state of the art was 1 ms time scales for ~40 residue proteins in implicit solvent). The ability of our MSMs to capture biexponential kinetics also indicates that proteins previously designated as incipient downhill folders actually have many barriers of differing heights. In addition, our model leads to a number of predictions for D14A: (1) helix five unfolds and fills a hydrophobic pocket in the native state and lacks structure in other well-populated states; (2) there is significant β-sheet structure in the unfolded ensemble; (3) there are structural rearrangements on 10 ms time scales that were not detected in past experiments, or alternatively, the simulated system reflects dynamics in mild denaturant; and (4) the native state acts as a kinetic hub. Our ability to reconcile these observations with existing experiments suggests that more experimental data are necessary to provide a detailed description of how D14A and other variants of λ_{6-85} fold. We suggest that MSMs could be used to help design such experiments and lead to important new insights into folding or, at the very least, provide more data for refining existing force fields and improving the agreement between simulation and experiment.

ASSOCIATED CONTENT

Supporting Information. Methods, Figures S1−S14, and Tables S1 and S2. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

pande@stanford.edu

ACKNOWLEDGMENT

We thank B. Swope and J. Pitera for starting structures, D. L. Ensign for simulations, K. A. Beauchamp for figures and comments, M. Gruebele for comments, the Folding@home users and CNS-0619926 for computing resources, and the NIH (R01-GM062868) for funding. G.R.B. was funded by the Berry Foundation.

REFERENCES

(1) Bowman, G. R.; Huang, X.; Pande, V. S. *Cell Res.* **2010,** *20*, 622–630.

(2) Noe, F.; Fischer, S. *Curr. Opin. Struct. Biol.* **2008,** *18*, 154–162.

(3) Schütte, C.; Fischer, A.; Huisinga, W.; Deuflhard, P. *J. Comput. Phys.* **1999,** *151*, 146–168.

(4) Karplus, M.; Weaver, D. L. *Nature* **1976,** *260*, 404–406.

(5) Burton, R. E.; Myers, J. K.; Oas, T. G. *Biochemistry* **1998,** *37*, 5337–5343.

(6) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Dill, K. A. *Multiscale Model. Simul.* **2006,** *5*, 1214–1226.

(7) Hinrichs, N. S.; Pande, V. S. *J. Chem. Phys.* **2007,** *126*, No. 244101.

(8) Bowman, G. R.; Ensign, D. L.; Pande, V. S. *J. Chem. Theory Comput.* **2010,** *6*, 787–794.

(9) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. *J. Chem. Phys.* **2009,** *131*, No. 124101.

(10) Noe, F.; Schutte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. *Proc. Natl. Acad. Sci. U.S.A.* **2009,** *106*, 19011–19016.

(11) Voelz, V. A.; Bowman, G. R.; Beauchamp, K. A.; Pande, V. S. *J. Am. Chem. Soc.* **2010,** *132*, 1526–1528.

(12) Yang, W. Y.; Gruebele, M. *Nature* **2003,** *423*, 193–197.

(13) Larios, E.; Pitera, J. W.; Swope, W.; Gruebele, M. *Chem. Phys.* **2006,** *323*, 45–53.

(14) Bowman, G. R.; Huang, X.; Pande, V. S. *Methods* **2009,** *49*, 197–201.

(15) Pabo, C. O.; Lewis, M. *Nature* **1982,** *298*, 443–447.

(16) Clarke, N. D.; Beamer, L. J.; Goldberg, H. R.; Berkower, C.; Pabo, C. O. *Science* **1991,** *254*, 267–270.

(17) Huang, G. S.; Oas, T. G. *Proc. Natl. Acad. Sci. U.S.A.* **1995,** *92*, 6878–6882.

(18) Liu, F.; Gao, Y. G.; Gruebele, M. *J. Mol. Biol.* **2010,** *397*, 789–798.

(19) Munoz, V.; Serrano, L. *Nat. Struct. Biol.* **1994,** *1*, 399–409.

(20) Allen, L. R.; Krivov, S. V.; Paci, E. *PLoS Comput. Biol.* **2009,** *5*, No. e1000428.

(21) Yang, W. Y.; Larios, E.; Gruebele, M. *J. Am. Chem. Soc.* **2003,** *125*, 16220–16227.

(22) Hoffmann, A.; Kane, A.; Nettels, D.; Hertzog, D. E.; Baumgartel, P.; Lengefeld, J.; Reichardt, G.; Horsley, D. A.; Seckler, R.; Bakajin, O.; Schuler, B. *Proc. Natl. Acad. Sci. U.S.A.* **2007,** *104*, 105–110.

(23) DeCamp, S. J.; Naganathan, A. N.; Waldauer, S. A.; Bakajin, O.; Lapidus, L. J. *Biophys. J.* **2009,** *97*, 1772–1777.

(24) Ma, H.; Gruebele, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005,** *102*, 2283–2287.

(25) Yang, W. Y.; Gruebele, M. *Biophys. J.* **2004,** *87*, 596–608.

(26) Yang, W. Y.; Gruebele, M. *Philos. Trans. R. Soc., A* **2005,** 363.

(27) Kubelka, J.; Hofrichter, J.; Eaton, W. A. *Curr. Opin. Struct. Biol.* **2004,** *14*, 76–88.

(28) Du, R.; Pande, V. S.; Grosberg, A. Y.; Tanaka, T.; Shakhnovich, E. S. *J. Chem. Phys.* **1998,** *108*, 334–350.

(29) Bowman, G. R.; Pande, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2010,** *107*, 10890–10895.