J|A|C|S COMMUNICATIONS

Published on Web 01/13/2010

Molecular Simulation of *ab Initio* Protein Folding for a Millisecond Folder NTL9(1−39)

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

Departments of Chemistry and Structural Biology, Stanford University, Stanford, California 94305, and Biophysics Program, Stanford University, Stanford, California 94305

Received October 28, 2009; E-mail: pande@stanford.edu

A complete understanding of how proteins fold, i.e. self-assemble to their biologically relevant "native state," remains an unattained goal.[1] Computer simulation, validated by experiment, is a natural means to elucidate this. There is over a million-fold range in folding rates, suggesting a possible diversity in mechanisms between slow and fast folding proteins.[2] Very fast (microsecond time scale) folding proteins[3,4] appear to fold via a large number of heterogeneous, parallel paths,[5−7] potentially key for folding on such fast time scales. Does the folding of much slower proteins change this picture?

To date, the slowest-folding proteins folded *ab initio* by all-atom molecular dynamics simulations with fidelity to experimental kinetics have had folding times in the range of nanoseconds to microseconds. These include the designed mini-protein Trp-cage (∼4.1 μs),[8] the villin headpiece domain (∼10 μs),[9] a fast-folding variant of villin (<1 μs),[7] and Fip35 WW domain (∼13 μs).[10] In this communication, we report simulations of several folding trajectories, each from fully unfolded states, of the 39-residue protein NTL9(1−39), which experimentally has a folding time of ∼1.5 ms.[11]

MD Simulation. Trajectories were simulated via the Folding@Home distributed computing platform[12] at 300, 330, 370, and 450 K from native, extended, and random-coil configurations using an accelerated version of GROMACS written for GPU processors,[13] for an aggregate time of 1.52 ms. GPUs play a key role here, allowing for dramatically longer trajectories than previously possible. The AMBER ff96 force field[14] with the GBSA solvation model[15] was used, a combination previously shown to give good results folding Fip35 WW domain,[10] and shown to exhibit a good balance of native-like secondary structure for a set of small helical and β-sheet peptides studied by replica exchange.[17]

Prediction of ab Initio Folding and Folding Rates. We find that the native state (taken from the N-terminal domain of the crystal structure of ribosomal protein L9[18]) is stable in this force field at 300 K, exhibiting decreasing stability with increasing temperature (Figure 1a). Rmsd-C_α distributions after 10 μs show well-defined native and collapsed unfolded basins near 3 and 5 Å, respectively. Of the ∼3000 trajectories started from unfolded (extended and coil) states at 370 K (Figure 1b), two reach an rmsd-C_α < 3.5 Å and eight reach an rmsd-C_α < 4 Å. No productive folding trajectories were observed at lower temperatures, consistent with the enhanced forward folding rate expected by Arrhenius kinetics. Higher temperature trajectories (450 K) exceed the melting temperature of NTL9 in the force field.

The observed number of folding events n is consistent with expectations from a simple model of parallel uncoupled folding simulations[19] in which folding is modeled as a two-state Poisson process: $\langle n \rangle = \int M(t) k \exp(-M(t)\, kt)\, dt$, is the number of simulations that reach time t (Figure 1b) and k is the experimental folding rate (∼640/s).[11] This theory predicts (on average) ∼1.8 folding trajectories for the amount of sampling performed, in agreement with the two folding trajectories found in practice. Posterior distributions of folding rates given the amount of simulation time and number of folding trajectories were computed using a Bayesian approach,[16] which yield expectation values within an order of magnitude of the experimental folding rate.

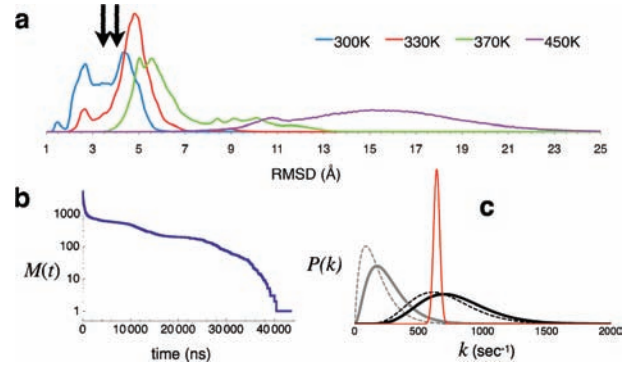

Figure 1. (a) Distributions of rmsd-C_α for native-state simulations of NTL9(1−39) after 10 μs. The arrows indicate thresholds defined for the native basin at 3.5 and 4 Å. (b) The number of parallel simulations $M(t)$ started from unfolded states at 370 K that reach time t. (c) Posterior predictions of the folding rate given the amount of simulation time and observed folding events for 3.5 Å (dashed) and 4 Å (solid) thresholds, using uniform (black) and Jeffrey's (gray) priors, using methods from ref 16. In red is a Gaussian distribution representing the experimental rate mean and standard deviation.

In addition to native-like conformations, we see near-native configurations, which show heterogeneity in hydrophobic packing, most notably in alternative side chain arrangements in the β-sheet structure (Figure 2). Most common of these is a non-native hydrophobic core involving residues I4, I18, and I37 (which normally contact the C-terminal helix in the full-length protein) with F5 solvent-exposed.

Insight into Folding Mechanisms. To describe the kinetics and mechanistic aspects of folding, we employ a new paradigm for sampling the global free energy landscape of folding, using Markov State Models (MSMs). MSM approaches, by automatically identifying a set of kinetically metastable states (such as foldons[20]) and efficiently sampling transitions between these states, can model long-time scale kinetics from much shorter trajectories.[21−24]

Our strategy for simulating slow-folding proteins is first to generate an initial series of kinetically connected states from both the folding and unfolding directions and then to use adaptive resampling techniques[25] to produce statistically converged estimates of metastable basins and the transition rates between them. In the

[†] Department of Chemistry.
[‡] Department of Structural Biology.
[§] Biophysics Program.

10.1021/ja9090353

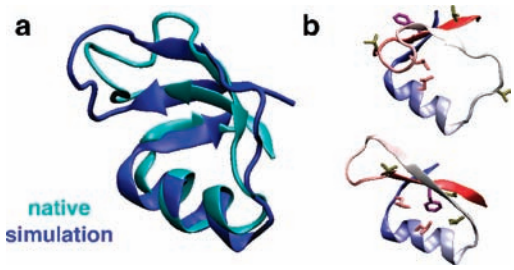

Figure 2. (a) A snapshot from a folding trajectory (dark blue) achieves an rmsd-C_α of 3.1 Å compared to the native state (cyan). (b) Non-native (top) and native-like (bottom) hydrophobic core arrangements observed in low-rmsd conformations of folding trajectories. Highlighted are side chains of residues F5 (magenta), V3,V9,V21 (tan), and L30,L35 (pink).

remainder of this communication, we report progress toward the first goal, by constructing an MSM from the entire set of 370 K trajectory data,[26,27] which we will use to seed future rounds of transition sampling. While additional rounds of adaptive sampling could likely aid in increasing the quantitative power of this model, there are several notable observations which can be made with the current data set.

Key to accurately identifying metastable states is the clustering of trajectory conformations into *microstates* fine-grained enough to be used for lumping into groups of maximally metastable *macrostates*.[26] 100 000 microstate clusters were calculated using an approximate *k*-centers algorithm,[28] each with an average radius of 4.5 Å rmsd-backbone. Lag times ranging from 1 to 32 ns were used to build a series of MSMs. The implied time scales predicted by these models (obtained by diagonalizing the rate matrix) show a clear spectral gap separating the slowest relaxation time scale from the rest, indicative of single-exponential kinetics (see Figure S1). The implied time scale of the model levels off beyond a lag time of ~10 ns to an implied time scale of ~1 ms, close to the experimental folding time.

An important strength of MSMs is their ability to gain insight at coarser scales by "lumping" the kinetic transitions into a simpler model with fewer states. To gain a mesoscopic view of the folding free energy landscape, we lumped our 100 000- microstate MSM into a 2000-macrostate model. From this view, we find that the metastable states are diffuse collections of conformations over which multiple possible folding pathways can occur, indicating a vast heterogeneity of folding substates that need to be understood in greater detail. At the same time, we can identify highly populated "native" (state *n*) and "unfolded" (state *a*) macrostates that dominate the observed relaxation rates (Figures 3 and S2).

The 10 pathways with the highest folding flux from macrostate *a* to *n* were calculated by a greedy backtracking algorithm (see Supporting Information (SI)) from the macrostate transition matrix using transition path theory[29,30] (TPT). The diversity of pathways demonstrates the power of the MSM approach: although we observe only a few folding trajectories directly, a network of many possible pathways can be inferred from the overlapping sampling of local transitions.

While NTL9(1−39) folds quickly for a two-state folder, it is similar in size to many ultrafast (submillisecond) folders that appear to exhibit so-called "downhill" folding. Hence, we would like to understand the structural features that limit the overall folding rate. As in a macroscopic two-state model, the highest-flux pathways in our mesoscopic model are $a \rightarrow m \rightarrow n$ and $a \rightarrow l \rightarrow n$ direct routes from disordered to structured macrostates, reminiscent of nucleation−condensation. These pathways by themselves, however, account for only ~10% of the total flux, and the structural diversity seen in all pathways is reminiscent of more hierarchical folding models

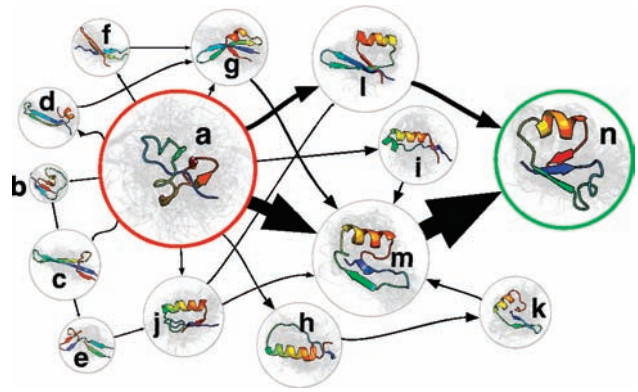

Figure 3. A 2000-state Markov State Model (MSM) was built using a lag time of 12 ns. Shown is the superposition of the top 10 folding fluxes, calculated by a greedy backtracking algorithm (see Supporting Information). These pathways account for only ~25% of the total flux and transit only 14 of the 2000 macrostates (shown labeled *a*−*n*, for convenient discussion). The visual size of each state is proportional to its free energy, and arrow size is proportional to the interstate flux.

such as diffusion−collision. Thus, we sought to more fully study the 14 macrostates transited by the top 10 folding pathways.

To examine structural changes along the folding reaction, we considered three main native structural elements: the central helix (α), the pairing of strands 1 and 2 (β_{12}), and the pairing of strands 1 and 3 (β_{13}). To quantify the extent of native-like structuring for each of these elements we calculated Q_α, $Q_{\beta12}$, and $Q_{\beta13}$, respectively (see SI for details). The *Q*-value is a number between 0 and 1 that quantifies the extent of native-like contacts. We then examined, for each macrostate, the *Q*-values in relation to the p_{fold} value (committor), a kinetic reaction coordinate. The p_{fold} value is computed from the macrostate transition matrix.[24,29,30]

This analysis yields several key insights into the folding mechanism of NTL9(1−39) on the mesoscale. We find the "unfolded" state *a* is compact and contains a baseline level of residual native-like structure, with Q_α near 0.5, and $Q_{\beta12}$ and $Q_{\beta13}$ near 0.2. In general, across the 14 macrostates studied, *Q*-values increase as p_{fold} values increase, although the relative balance of Q_α, $Q_{\beta12}$, and $Q_{\beta13}$ varies, indicating pathway heterogeneity: i.e., native-like structures can form in different orders (Figures 4, S4, and S5). An exception to this, however, is observed for β_{12} strand

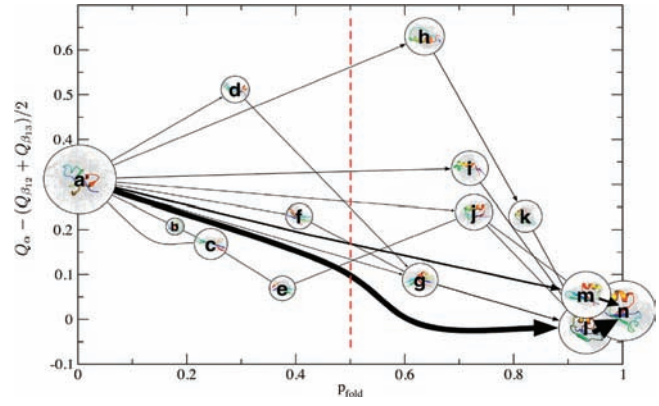

Figure 4. The 14 macrostates involved in the top 10 folding pathways, plotted along structural and kinetic reaction coordinates. The balance between native-like helix and sheet structure is quantified by $Q_\alpha - (Q_{\beta12} + Q_{\beta13})/2$ (vertical axis), and progress along the folding reaction is quantified by the p_{fold} (committor) value (horizontal axis). It can be seen that the "unfolded" state (a) contains residual native-like helical propensity, and that pathways involving various ordering of native-like helix and sheet formation are possible.

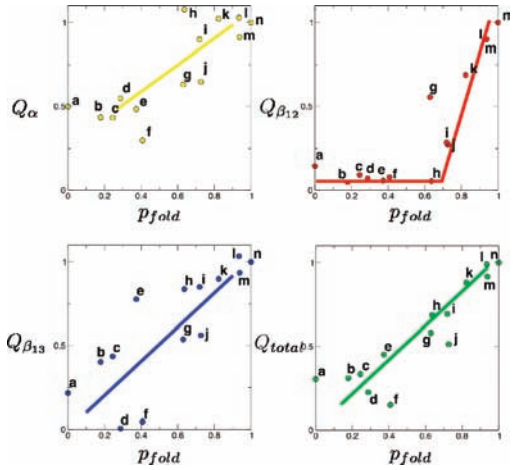

Figure 5. *Q*-values, which capture the extent of native-like structures, plotted versus p_{fold} (committor) values. The lines are to guide the eye.

pairing. Only for macrostates with $p_{fold} > 0.5$ (states $g-n$) does appreciable β_{12} strand pairing occur (Figure 5). This suggests that the formation of a *local* strand pair (β_{12}), rather than a *nonlocal* strand pair (β_{13}), is rate-limiting. This effect is not predicted by strictly topological models of folding in which loop closure entropy loss dominates[31] but instead may result from sequence-specific details. Unlike the β_{13} strand pair, which has a small interaction surface stabilized by hydrophobic contacts, the β_{12} hairpin contains seven of the protein's eight lysine residues and three of its five glycine residues in a flexible loop region, features which may imbue β_{12} with larger barriers to folding. This proposed role of β_{12} is also consistent with the large changes in kinetics and stability seen experimentally for mutations in the β_{12} hairpin.[11]

It is natural to compare our results with previous unfolding simulations of NTL9(1−39) K12 M by Snow et al.[32] In that work, a detailed characterization of the transition state ensemble required the definition of strand-pairing reaction coordinates corresponding to β_{12} and β_{13} formation. In our MSM analysis, no such predefinition is required. Snow et al. also note the difficulty in resolving kinetic intermediates not captured by the chosen order parameters. Indeed, our structural analysis can resolve subtle kinetic intermediates within the native basin, corresponding to alternative rearrangements of the β_{12} hairpin loop (Figure S6).

Conclusion. The above results suggest that existing force field models using implicit solvent are indeed accurate enough to fold proteins *ab initio* at long time scales (milliseconds), opening the door to simulating more structurally complex proteins. Moreover, our work demonstrates that there need not be a single pathway or single, dominant mechanism for the folding of a given protein: since the theories proposed for how proteins fold are based on broadly relevant physical principles, it is natural to imagine that multiple mechanisms could be *simultaneously* present but that the sequence of the protein, coupled with the chemical environment, would control the balance to which each mechanistic pathway is seen.

Acknowledgment. We thank the NSF for support through FIBR Grant EF-0623664, the NIH through R01-GM062868 and Simbios U54-GM072970, and NSF Award CNS-0619926 for computer resources.

Supporting Information Available: Detailed description of simulation methods, results, analysis, and Supporting Figures S1−S7. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Dill, K. A.; Ozkan, S. B.; Weikl, T. R.; Chodera, J. D.; Voelz, V. A. *Curr. Opin. Struct. Biol.* **2007**, *17*, 342–346.
(2) Plaxco, K. W.; Simons, K. T.; Baker, D. *J. Mol. Biol.* **1998**, *277*, 985–994.
(3) Yang, W. Y.; Gruebele, M. *Nature* **2003**, *423*, 193–197.
(4) Kubelka, J.; Chiu, T. K.; Davies, D. R.; Eaton, W. A.; Hofrichter, J. *J. Mol. Biol.* **2006**, *359*, 546–553.
(5) Kubelka, J.; Hofrichter, J.; Eaton, W. A. *Curr. Opin. Struct. Biol.* **2004**, *14*, 76–88.
(6) Udgaonkar, J. B. *Annu. Rev. Biophys.* **2008**, *37*, 489–510.
(7) Ensign, D. L.; Kasson, P. M.; Pande, V. S. *J. Mol. Biol.* **2007**, *374*, 806–816.
(8) Pitera, J. W.; Swope, W. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 7587–7592.
(9) Zagrovic, B.; Snow, C. D.; Shirts, M. R.; Pande, V. S. *J. Mol. Biol.* **2002**, *323*, 927–937.
(10) Ensign, D. L.; Pande, V. S. *Biophys. J.* **2009**, *96*, L53–L55.
(11) Horng, J.-C.; Moroz, V.; Raleigh, D. P. *J. Mol. Biol.* **2003**, *326*, 1261–1270.
(12) Shirts, M.; Pande, V. *Science* **2000**, *290*, 1903–1904.
(13) Friedrichs, M. S.; Eastman, P.; Vaidyanathan, V.; Houston, M.; Legrand, S.; Beberg, A. L.; Ensign, D. L.; Bruns, C. M.; Pande, V. S. *J. Comput. Chem.* **2009**, *30*, 864–872.
(14) Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049–1074.
(15) Onufriev, A.; Bashford, D.; Case, D. *Proteins* **2004**, *55*, 383–394.
(16) Ensign, D. L.; Pande, V. S. *J. Phys. Chem. B* **2009**, *113*, 12410–12423.
(17) Shell, M. S.; Ritterson, R.; A, K. *J. Phys. Chem. B* **2008**, *112*, 6878–6886.
(18) Hoffman, D. W.; Davies, C.; Gerchman, S. E.; Kycia, J. H.; Porter, S. J.; White, S. W.; Ramakrishnan, V. *EMBO J.* **1994**, *13*, 205–212.
(19) Shirts, M. R.; Pande, V. S. *Phys. Rev. Lett.* **2001**, *86*, 4983–4987.
(20) Panchenko, A. R.; Luthey-Schulten, Z.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 2008–2013.
(21) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. *J. Chem. Phys.* **2007**, *126*, 155101.
(22) Noé, F.; Fischer, S. *Curr. Opin. Struct. Biol.* **2008**, *18*, 154–162.
(23) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Dill, K. A. *Multiscale Model. Simul.* **2006**, *5*, 1214–1226.
(24) Singhal, N.; Snow, C. D.; Pande, V. S. *J. Chem. Phys.* **2004**, *121*, 415–425.
(25) Huang, X.; Bowman, G. R.; Bacallado, S.; Pande, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 19765–19769.
(26) Bowman, G. R.; Huang, X.; Pande, V. S. *Methods* **2009**, *49*, 197–201.
(27) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. *J. Chem. Phys.* **2009**, *131*, 124101.
(28) Dasgupta, S.; Long, P. M. *J. Comput. Syst. Sci.* **2005**, *70*, 555–569.
(29) Metzner, P.; Schütte, C.; Vanden-Eijnden, E. *Multiscale Model. Simul.* **2009**, *7*, 1192–1219.
(30) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 19011–19016.
(31) Weikl, T. R. *Arch. Biochem. Biophys.* **2008**, *469*, 67–75.
(32) Snow, C. D.; Rhee, Y. M.; Pande, V. S. *Biophys. J.* **2006**, *91*, 14–24.

JA9090353