How Fast-Folding Proteins Fold

Kresten Lindorff-Larsen,[1]*† Stefano Piana,[1]*† Ron O. Dror,[1] David E. Shaw[1,2]†

An outstanding challenge in the field of molecular biology has been to understand the process by which proteins fold into their characteristic three-dimensional structures. Here, we report the results of atomic-level molecular dynamics simulations, over periods ranging between 100 μs and 1 ms, that reveal a set of common principles underlying the folding of 12 structurally diverse proteins. In simulations conducted with a single physics-based energy function, the proteins, representing all three major structural classes, spontaneously and repeatedly fold to their experimentally determined native structures. Early in the folding process, the protein backbone adopts a nativelike topology while certain secondary structure elements and a small number of nonlocal contacts form. In most cases, folding follows a single dominant route in which elements of the native structure appear in an order highly correlated with their propensity to form in the unfolded state.

Protein folding is a process of molecular self-assembly during which a disordered polypeptide chain collapses to form a compact and well-defined three-dimensional structure. Hundreds of studies have been devoted to understanding the mechanisms underlying this process, but experimentally characterizing the full folding pathway for even a single protein—let alone for many proteins differing in size, topology, and stability—has proven extremely difficult. Similarly, simulating the folding of a small protein at an atomic level of detail is a daunting task. Both experimental and computational studies have thus generally focused on one protein at a time, with such studies each performed under different conditions or with different techniques. Possibly because of the resulting heterogeneity of the available data, numerous theories have been proposed to describe protein folding and no consensus has been reached on which of these theories, if any, is correct (*1*).

Our research group has developed a specialized supercomputer, called Anton, which greatly accelerates the execution of atomistic molecular dynamics (MD) simulations (*2*, *3*). In addition, we recently modified the CHARMM force field in an effort to make it more easily transferable among different protein classes (*4*). Here, we have combined these advances to study the folding process of fast-folding proteins through equilibrium MD simulations (*2*). We studied 12 protein domains (*5*) that range in size from 10 to 80 amino acid residues, contain no disulfide bonds or prosthetic groups, and include members of all three major structural classes (α-helical, β sheet and mixed α/β). Of these 12 protein domains, 9 represent the nine folds considered in a review of fast-folding proteins (*6*). As most of these nine proteins contain only α helices, we also included two additional α/β proteins and a stable β hairpin to increase the structural diversity of the set of proteins examined.

In our simulations, all of which used a single force field (*4*) and included explicitly represented solvent molecules, 11 of the 12 proteins folded spontaneously to structures matching their experimentally determined native structures to atomic resolution (Fig. 1). The native state of the 12th protein, the Engrailed homeodomain, proved unstable in simulation. We were, however, able to fold a different homeodomain (*7*) with the same overall structure; the results reported below pertain to this variant, rather than the Engrailed homeodomain.

For all 12 proteins that folded in simulation, we were also able to perform simulations near the melting temperature, at which both folding and unfolding could be observed repeatedly in a single, long equilibrium MD simulation. For each of the 12 proteins, we performed between one and four simulations, each between 100 μs and 1 ms long, and observed a total of at least 10 folding and 10 unfolding events. In total, we collected ~8 ms of simulation, containing more than 400 folding or unfolding events. For 8 of the 12 proteins, the most representative structure of the folded state fell within 2 Å root mean square deviation (RMSD) of the experimental structure (Fig. 1). This is particularly notable given that the RMSD calculations included the flexible tail residues and that, in some cases, there was no experimental structure available

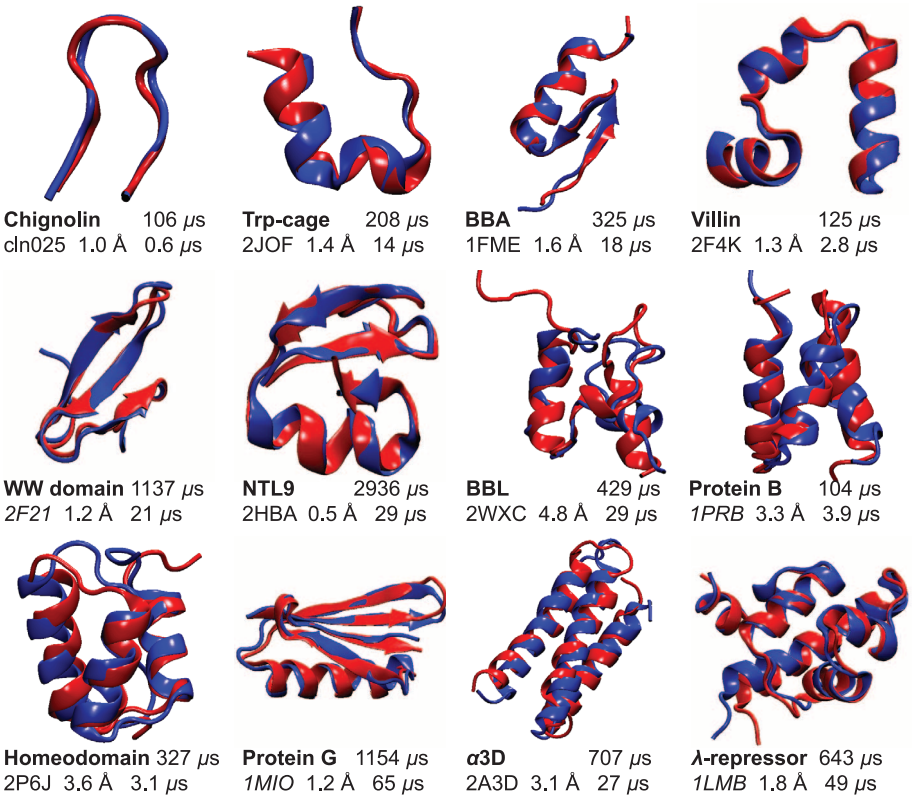

Fig. 1. Representative structures of the folded state observed in reversible folding simulations of 12 proteins. For each protein, we show the folded structure obtained from simulation (blue) superimposed on the experimentally determined structure (red), along with the total simulation time, the PDB entry of the experimental structure, the Cα-RMSD (over all residues) between the two structures, and the folding time (obtained as the average lifetime in the unfolded state observed in the simulations). Each protein is labeled with a commonly used name, although in several cases, we studied mutants of the parent sequence [amino acid sequences of the 12 proteins and simulation details are presented in (*5*)]. PDB entries in italics indicate that the structure has not been determined for the simulated sequence and that, instead, we compare it with the structure of the closest homolog in the PDB. The calculated structure was obtained by clustering the simulations (*26*) to avoid bias toward the experimentally determined structure.

[1]D. E. Shaw Research, New York, NY 10036, USA. [2]Center for Computational Biology and Bioinformatics, Columbia University, New York, NY 10032, USA.

*These authors contributed equally to the manuscript.
†To whom correspondence should be addressed. E-mail: david.shaw@DEShawResearch.com (D.E.S.); kresten.lindorff-larsen@DEShawResearch.com (K.L.-L.); stefano.piana-agostinetti@DEShawResearch.com (S.P.)

for the exact sequence that we simulated (we instead calculated the RMSD to the structure of the protein with the most similar sequence). The proteins exhibiting the largest deviations from their experimental structures (BBL, protein B, and the homeodomain) are all three-helix bundles; this finding hints at a minor residual force-field deficiency. It has been argued, however, that the native state for at least one of these three proteins may depend on experimental conditions (*8*); it is thus possible that these deviations might reflect genuine differences between the protein's structure at the simulated temperature and at the lower temperatures used for experimental structure determination. Overall, comparison with available experimental data indicates that the force field provides a reasonable description of the structure, thermodynamics, and kinetics of the 12 proteins [see (*5*) for a more detailed comparison], which affords some confidence in the accuracy of the folding mechanisms observed in simulation.

Among the many analyses that can be performed on this data set, we focus here on elucidating the general principles that underlie protein folding and do not discuss in detail the properties of each individual system. In particular, we have used this data set to examine several important and unresolved general questions (*1*): (i) What is the general nature and order of events that lead to folding? (ii) What role, if any, is played by the residual structure in the unfolded state? (iii) How many distinct folding pathways are present, and how different are they from one another? and (iv) Is there a free-energy barrier for folding, and what is its magnitude?

As a first step, we partitioned all trajectories into folded, unfolded, and transition-path segments (*5*). Unfolding transitions were analyzed in reverse so that all transitions could be treated as folding events. For each folding and unfolding event, we quantified the formation of the native topology (*9*, *10*), secondary structure, and nonlocal native contacts along the transition path (Fig. 2). We found that the formation of a native topology and secondary structure begins earlier than the formation of most nonlocal contacts. Whereas most contacts are formed late, a few specific key contacts are formed relatively early in the transition paths (*5*). In most cases, formation of secondary structure appears to decrease the solvent-exposed area of the protein (fig. S2), in line with experimental observations (*1*).

Analysis of the unfolded state observed in the simulations reveals the presence of both native and non-native secondary structure elements. On average, the 12 proteins contain 16% helical and 5% sheet structure in the unfolded state (*5*). These secondary structure elements form transiently (partially or completely), but are typically only marginally stable in the absence of the stabilizing tertiary interactions, and they persist for tens to hundreds of nanoseconds in the unfolded state (Fig. 3 and fig. S6). The propensity to form local nativelike structure in the unfolded state correlates strongly with the order of formation of local nativelike structure along the transition path (Fig. 3). In particular, initiation sites for folding are preferentially located in regions that have a high propensity to form native structure in the unfolded state (*11*). In helical proteins, these regions often correspond to individual helices, and we find that the helices with the highest stability in the unfolded state generally form first during folding (Fig. 3). These observations support a mechanism for protein folding in which the formation of a subset of key long-range native contacts early

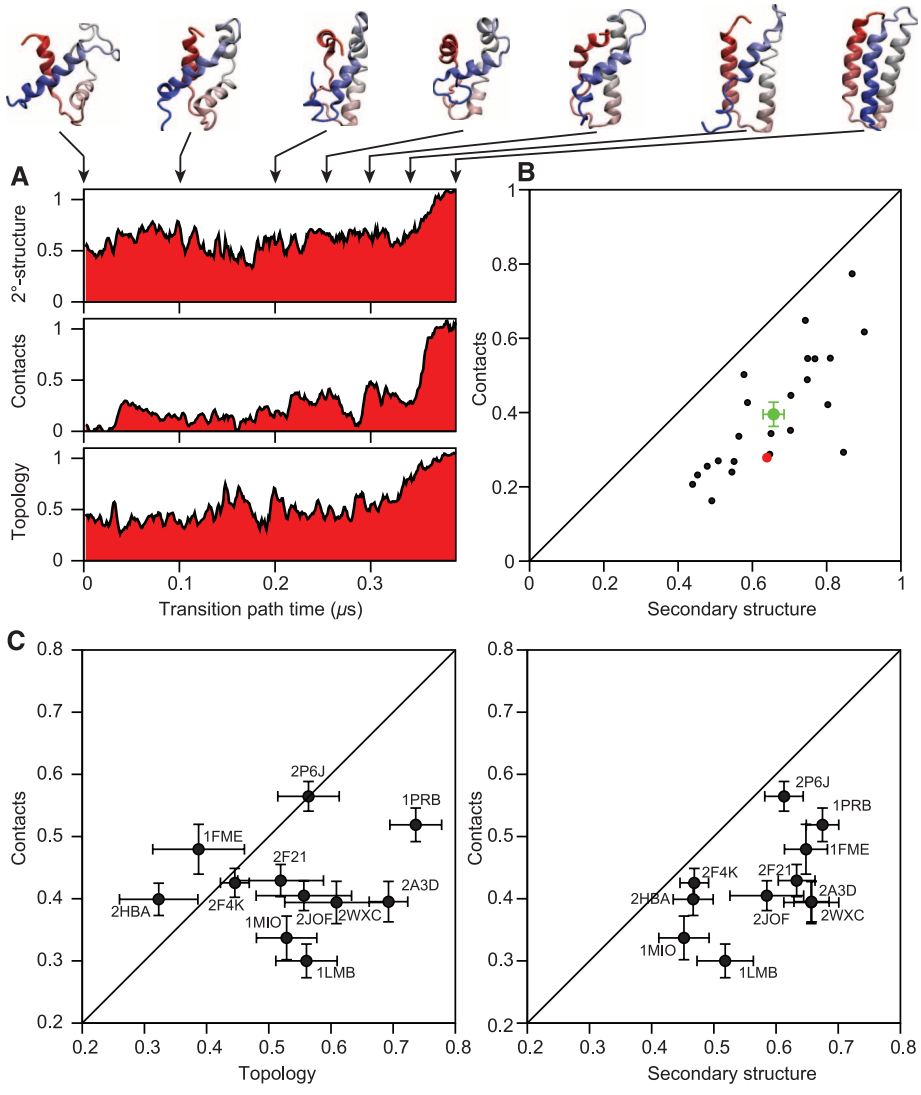

Fig. 2. Formation of topology, native contacts, and secondary structure during protein folding. (**A**) The three panels show the accumulation of native secondary structure, nonlocal native contacts, and native topology during a single folding event for α3D. Each of the three quantities was normalized such that the average value in the unfolded state was zero, and the average value in the folded state was one. Above the three panels we show seven representative structures from this transition path, with the corresponding time points shown with arrows. This analysis was repeated for each of the 24 folding and unfolding events observed for this protein, and for each of these transitions, the relative orders of formation of secondary structure, contacts, and topology were quantified by integration of these time series (with the resulting integrals, corresponding to the area under the curves, here represented by the area of the red shading). High values of this integral thus correspond to early formation of the corresponding quantity during a folding event. (**B**) The 24 transitions of α3D in a scatter plot are represented, with each of the black points corresponding to the time series integral for a single folding event (unfolding events were analyzed in reverse). The red point corresponds to the folding event shown in (A), and the green point represents the average of the time series integrals over all 24 transitions (error bars represent SEM). (**C**) We repeated this analysis for 11 of the 12 proteins (chignolin was omitted because of its small size). Each point shows the average value over all folding and unfolding events observed for one protein [as described above for the green point in (B)]. Each point is labeled with the PDB code of the relevant protein (see also Fig. 1). Most proteins fall below the diagonal in these plots, showing that topology and secondary structure develop earlier than the full set of native contacts.

in the folding process is sufficient to establish a nativelike topology and to stabilize the native secondary structure elements that are only transiently formed in the unfolded state (*9*).

To quantify the heterogeneity of the folding process, we examined the order of formation of structural elements in the transition paths of the 12 proteins. Each individual transition path is, of course, different from all the others at a sufficiently detailed level of resolution, but transition paths where structural elements are formed in a similar order are typically defined as belonging to a common "pathway." Theory and previous simulations suggest that protein folding may be a highly heterogeneous process, with multiple such pathways each accounting for a small fraction of the total flux (*1*, *12*). There is experimental evidence for heterogeneous folding mechanisms in only a few two-state folding proteins (*1*, *13*), but this could be attributed to the difficulty of distinguishing similar pathways in experiments. Indeed, at a coarser level, different pathways may still share a large fraction of structural features, and we define such pathways as belonging to the same folding "route." For each of the 12 proteins, we address the questions of how many folding pathways are present and how many different routes these pathways represent.

As a first step, we determined for each protein how many folding pathways are traversed that are distinct in the sense that native interactions are formed in different orders and that the pathways do not interconvert on the transition path time scale (allowing individual transition paths to be robustly assigned to a specific pathway) (*14*, *15*). In particular, we defined a metric to calculate the distances between two individual transition paths and used these distances to cluster the folding and unfolding transitions for each protein into structurally distinct pathways (*5*). We find that for most proteins, the transition paths can be robustly assigned to two or three well-defined clusters (*5*), which reveals the existence of a small number of parallel pathways.

We then examined whether these parallel pathways arise because of "structural noise" superimposed on a single, well-defined folding route (*16*) or because the protein does in fact fold along multiple, distinct routes. To distinguish between these two scenarios, we quantified the similarity of the different pathways (as defined in the previous paragraph) by calculating the fraction of the native contacts they share at various intermediate points. We find that in most cases the order of formation of the native structure is similar in the different pathways (Fig. 3); for 9 out of 11 proteins considered, the pathways belonging

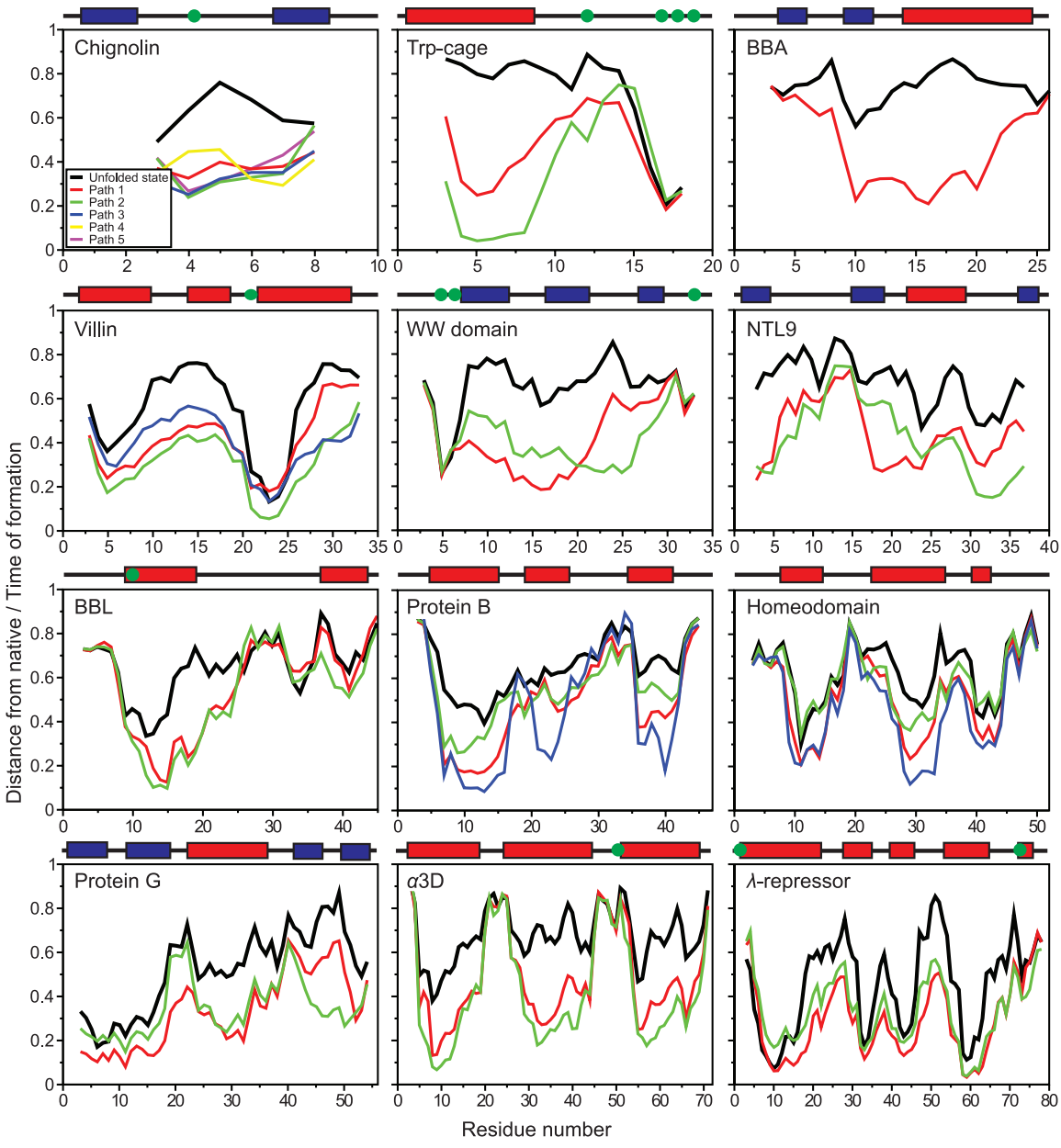

Fig. 3. Order of native structure formation along the transition pathway and the average distance from the native conformation in the unfolded state. The colored lines represent a quantity that measures when an amino acid residue adopts a nativelike structure (with a small value indicating early formation); the different colors represent the results for the different folding pathways that we obtained, as described in the main text. The average fraction of native structure in the unfolded state is shown by the black lines. The positions of helices (red) and sheets (blue) in the native state are shown above each graph together with the location of proline residues (green circles). Note that proline residues are often located at initiation sites; we speculate that this observation can be explained by the fact that proline has a restricted conformational space available and thus facilitates the local ordering of the polypeptide backbone.

to different clusters share more than 60% of the native contacts formed at any given time during the folding process (fig. S5) (chignolin was excluded from this analysis because of its small number of contacts). Thus, for each of these nine proteins, the distinct pathways appear to share a largely common folding nucleus, and we suggest that they are best considered to be variations of a single folding route. In these cases, we expect that it would be difficult to distinguish the different pathways using currently available ensemble experiments, as the pathways would be similarly affected by most mutations or changes in temperature or solvent composition. Although the exact number of pathways and routes determined by our analysis is dependent on the detailed criteria used to categorize the folding transitions, the overall picture that emerges is one where folding is usually a relatively homogeneous process in which individual structural elements tend to form in a well-defined order (*17*).

The remaining two simulated proteins, NTL9 and the protein G variant, clearly exhibit two structurally distinct folding routes. Both are α/β proteins of moderate size, and the difference between the routes is a different order of formation of the β strands. (The third α/β protein in our set of proteins, BBA, which is only 20 amino acid residues long and has only two β strands, folds via a single pathway.) In the case of the redesigned variant of protein G that we studied, the principal difference between the two routes is the order in which the two hairpins form. This observation is in line with experimental results on wild-type protein G and its redesigned NuG2 variant, which share the same fold as the protein G variant considered in this study (*18*). In particular, in wild-type protein G, hairpin 2 folds first, whereas in NuG2, hairpin 1 folds first. The protein G variant that we simulated (*5*), which is intermediate in sequence between wild-type protein G and NuG2, populates both wild-type protein G–like and NuG2-like pathways. Although most of the proteins we considered fold with a single folding nucleus that is shared among the different pathways, our results for NTL9 and protein G suggest that caution should be exercised in generalizing this finding; larger proteins, particularly those with β-sheet structure, may indeed be characterized by multiple folding nuclei and truly distinct folding routes (*12*).

Finally, we examined the thermodynamics and kinetics of the folding process, and in particular the existence and size of the free-energy barrier for folding. Some of the proteins we have simulated have been suggested to fold in a downhill fashion, defined by the absence of a distinct free-energy barrier for folding. To explore such issues, we first used a previously established method (*2*, *19*) to project the folding free-energy surfaces along an optimized reaction coordinate. In all 12 cases, the application of this method yielded folding free-energy barriers smaller than 4.5 k_BT (where k_B is Boltzmann's constant and T is temperature), consistent with the fact that all these proteins are fast folders. For three proteins (BBL, protein B, and the homeodomain), we were unable to identify a free-energy barrier separating the folded and unfolded states (an observation that proved robust against changes in the parameterization of the reaction coordinate). The lack of a free-energy barrier in these cases may, in principle, be due to an inability to properly separate the folded and unfolded basins by using a single reaction coordinate. The presence of a substantial free-energy barrier, however, would also be expected to give rise to a separation of time scales between the folding process and relaxation within the folded and unfolded states. A calculation of the dynamical content (*2*, *5*) shows that, for all proteins where the calculated barrier is smaller than 1.5 k_BT, there is little or no separation of time scales between overall folding and faster local relaxation. This provides support for the notion that the folding rate of these proteins is not determined by a single, well-defined free-energy barrier, at least under experimental conditions corresponding to those used in our simulations. For these proteins, the time scales for the formation of individual structural elements overlap with those for folding, giving rise to more complex kinetic behavior that we do not expect to be satisfactorily described by a single exponential relaxation.

In addition to providing information about the height of the free-energy barrier, the analysis in the previous section identifies structures whose formation appears to be rate-limiting for folding. The structures that lie between the unfolded and folded states and have equal probability to fold or unfold are in each case compact and nativelike; they contain 60 to 97% of the native secondary structure, and their contact order is 60 to 100% of that of the native state (*5*). Earlier work based on combining simulations and experiments found that the transition state ensemble for folding has a contact order that is ~70% that of the native state (*20*, *21*); we speculate that the slightly higher value found here (the average over the 12 proteins is 85%) may be caused by a Hammond shift due to the high temperatures at which the simulations were performed.

The results presented here provide a unified picture of the folding of 12 small proteins. We find that elements of local nativelike structure are transiently formed in the unfolded state; the formation of a few additional key contacts provides further stabilization for these structural elements and initiates the folding transition. In most cases, folding then proceeds along a single, dominant route, where additional structural elements are formed in a well-defined sequence, with "optional" noise (*16*). For two proteins, however, we find clear evidence for heterogeneous folding mechanisms with differing transition state "classes" (*12*). The ensemble of structures found on the free-energy barrier has nativelike topologies with partial formation of secondary structure and tertiary contacts, in line with conclusions drawn from experiments (*1*, *22*, *23*). Also notable is the fact that a single force field was able to consistently fold a substantial number of proteins, spanning all three of the major structural classes, to their native states. The results of this rather stringent test (*24*, *25*) suggest that current molecular mechanics force fields are sufficiently accurate to make long–time scale MD simulation a powerful tool for characterizing large conformational changes in proteins.

References and Notes

1. T. R. Sosnick, D. Barrick, *Curr. Opin. Struct. Biol.* **21**, 12 (2011).
2. D. E. Shaw *et al.*, *Science* **330**, 341 (2010).
3. D. E. Shaw *et al.*, Millisecond-scale molecular dynamics simulations on Anton. *Proceedings of the ACM/IEEE Conference on Supercomputing (SC09)*, Portland, OR, 14 to 20 November 2009 (ACM, New York, 2009); 10.1145/1654059.1654099.
4. S. Piana, K. Lindorff-Larsen, D. E. Shaw, *Biophys. J.* **100**, L47 (2011).
5. Materials and methods are available as supporting material on *Science* Online.
6. J. Kubelka, J. Hofrichter, W. A. Eaton, *Curr. Opin. Struct. Biol.* **14**, 76 (2004).
7. P. S. Shah *et al.*, *J. Mol. Biol.* **372**, 1 (2007).
8. G. Settanni, A. R. Fersht, *J. Mol. Biol.* **387**, 993 (2009).
9. K. Lindorff-Larsen, P. Røgen, E. Paci, M. Vendruscolo, C. M. Dobson, *Trends Biochem. Sci.* **30**, 13 (2005).
10. P. Røgen, *J. Phys. Condens. Matter* **17**, S1523 (2005).
11. K. Modig *et al.*, *FEBS Lett.* **581**, 4965 (2007).
12. V. S. Pande, Grosberg AYu, T. Tanaka, D. S. Rokhsar, *Curr. Opin. Struct. Biol.* **8**, 68 (1998).
13. C. F. Wright, K. Lindorff-Larsen, L. G. Randles, J. Clarke, *Nat. Struct. Biol.* **10**, 658 (2003).
14. P. Lenz, S. S. Cho, P. G. Wolynes, *Chem. Phys. Lett.* **471**, 310 (2009).
15. B. C. Gin, J. P. Garrahan, P. L. Geissler, *J. Mol. Biol.* **392**, 1303 (2009).
16. S. W. Englander, L. Mayne, M. M. Krishna, *Q. Rev. Biophys.* **40**, 287 (2007).
17. A. D. Pandit, A. Jha, K. F. Freed, T. R. Sosnick, *J. Mol. Biol.* **361**, 755 (2006).
18. B. Kuhlman, D. Baker, *Curr. Opin. Struct. Biol.* **14**, 89 (2004).
19. R. B. Best, G. Hummer, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 6732 (2005).
20. E. Paci, K. Lindorff-Larsen, C. M. Dobson, M. Karplus, M. Vendruscolo, *J. Mol. Biol.* **352**, 495 (2005).
21. T. R. Sosnick, *Protein Sci.* **17**, 1308 (2008).
22. A. R. Fersht, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 10869 (1995).
23. K. Lindorff-Larsen, M. Vendruscolo, E. Paci, C. M. Dobson, *Nat. Struct. Mol. Biol.* **11**, 443 (2004).
24. P. L. Freddolino, C. B. Harrison, Y. Liu, K. Schulten, *Nat. Phys.* **6**, 751 (2010).
25. J. C. Faver *et al.*, *PLoS ONE* **6**, e18868 (2011).
26. X. Daura *et al.*, *Angew. Chem. Int. Ed.* **38**, 236 (1999).

Acknowledgments: We thank K. Palmo, B. Gregerson, and J. Kiepeis for their input during the development of the CHARMM22* force field; P. Røgen for providing us with software to calculate generalized Gauss integrals; J. Salmon and R. Dirks for developing and testing simulation software; and R. Kastleman and M. Kirk for editorial assistance.

Supporting Online Material

www.sciencemag.org/cgi/content/full/334/6055/517/DC1
Materials and Methods
Figs. S1 to S8
Tables S1 to S3
References (*27*–*60*)

13 May 2011; accepted 1 September 2011
10.1126/science.1208351