Simplified Protein Models: Predicting Folding Pathways and Structure Using Amino Acid Sequences

Aashish N. Adhikari,[1,2,3] Karl F. Freed,[1,2,4,*,†] and Tobin R. Sosnick[3,4,5,*,‡]

[1]*Department of Chemistry, University of Chicago, Chicago, Illinois 60637, USA*
[2]*James Franck Institute, University of Chicago, Chicago, Illinois 60637, USA*
[3]*Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, Illinois 60637, USA*
[4]*Computation Institute, University of Chicago, Chicago, Illinois 60637, USA*
[5]*Institute for Biophysical Dynamics, University of Chicago, Chicago, Illinois 60637, USA*

(Received 27 April 2013; published 11 July 2013)

We demonstrate the ability of simultaneously determining a protein's folding pathway and structure using a properly formulated model without prior knowledge of the native structure. Our model employs a natural coordinate system for describing proteins and a search strategy inspired by the observation that real proteins fold in a sequential fashion by incrementally stabilizing nativelike substructures or "foldons." Comparable folding pathways and structures are obtained for the twelve proteins recently studied using atomistic molecular dynamics simulations [K. Lindorff-Larsen, S. Piana, R. O. Dror, D. E. Shaw, Science **334**, 517 (2011)], with our calculations running several orders of magnitude faster. We find that nativelike propensities in the unfolded state do not necessarily determine the order of structure formation, a departure from a major conclusion of the molecular dynamics study. Instead, our results support a more expansive view wherein intrinsic local structural propensities may be enhanced or overridden in the folding process by environmental context. The success of our search strategy validates it as an expedient mechanism for folding both *in silico* and *in vivo*.

DOI: 10.1103/PhysRevLett.111.028103 PACS numbers: 87.15.Cc, 87.15.hm

The discovery that a protein's structure is determined by its amino acid sequence has motivated efforts to replicate the folding process *in silico*. A successful algorithm for describing folding should enable predicting both the pathway and structure, two intertwined issues that generally have been treated separately. All-atom molecular dynamics (MD) simulations can address both issues simultaneously as demonstrated by a recent success in folding a dozen small proteins [1]. Although remarkable, the simulations require very specialized hardware and extensive amounts of computing time. Our goal is to develop an alternate approach that identifies basic folding principles and then integrates them into a rapid, accurate, and physically revealing algorithm.

Our algorithm, termed TERITFIX, is motivated by the manner in which real proteins fold. Growing evidence suggests that proteins fold along a limited number of low-energy pathways [2–8], with the order of events guided by a process termed sequential stabilization. Here, nascent nativelike substructures serve as templates for the formation of additional structure through the stepwise addition of cooperative folding subunits or "foldons" [9–15]. We explicitly implement sequential stabilization by using the information gleaned from earlier rounds of folding simulations to guide folding in subsequent rounds. The biasing is intended to assist the polypeptide up and over the major free-energy barrier between the unfolded and native states in a manner that replicates the authentic folding process [16,17].

Our initial folding round involves ~500 separate Monte Carlo simulated annealing (MCSA) trajectories that begin from a realistic denatured state ensemble (DSE) [18], rather than from a state containing, for example, biases from homology-based secondary structure predictions. The best 25% (lowest energy) structures are used to identify the preferred local and nonlocal interactions for each residue in the form of a consensus secondary structure and average inter-residue contacts and hydrogen bonds. This information from a given round is used in the next round of ~500 trajectories to restrict the sampling of backbone (φ, ψ) dihedral angles and energetically bias the formation of the tertiary contacts and hydrogen bonds. The iterative process incrementally generates additional secondary and tertiary structure and hydrogen bonds as the rounds proceed, producing a series of events that may correspond to the genuine folding pathway.

We use a representation containing all backbone atoms plus the Cβ carbons, a move set involving smart (φ, ψ) dihedral angle distributions, and a combination of single (φ, ψ) pivots and local crankshaft moves [17]. Angles are selected from a PDB-based coil library, contingent on the chemical identity of the flanking residues. As secondary structure information is deduced from prior rounds, angle selection is correspondingly biased. Three energy functions capture the chemical properties of the different amino acids [16,17,19,20]. The first function includes a pairwise additive, distance, orientation, and secondary structure dependent statistical potential designed to promote the

formation of chain topologies with hydrophobic cores. The other two statistical potentials are multibody terms designed to capture the properties of side chain burial and hydrogen bonding.

Figure 1 displays the most nativelike structures obtained from the TERITFIX simulations for the twelve proteins studied by Lindorff-Larsen *et al.* [1] The calculations for each protein take around 600 CPU hours on an Intel 2.6 GHz "Sandy Bridge" Xeon E5-2670 processor. Using the same processor running NAMD, a single 10 μs MD trajectory would take around 3 000 000 CPU hours/protein. TERITFIX produces an average root-mean-square deviation (RMSD) from the native structure of 2.96 ± 1.33 Å for the centroids of the largest clusters, compared to 2.07 ± 1.31 Å for the all-atom MD simulations. TERITFIX generates centroids with lower RMSDs for half of the proteins. By a significant 4.5 Å margin, TERITFIX's worst result is for NTL9, whereas this protein produces MD's best result (5.0 versus 0.5 Å for the cluster centroids). The crystal structure of NTL9 appears with an extra 12 residue helix which produces a helix-swapped dimer. Without this extra helix, the structure has an unusual, exposed hydrophobic face that probably contributes to TERITFIX's difficulty in obtaining a good prediction.

Unlike most free modeling algorithms for predicting structure, TERITFIX does not use fragments or invoke any prior assumptions (or predictions from machine learning) about the protein's secondary structure. An additional feature distinguishing TERITFIX from prior approaches is the generation of a *de facto* folding pathway that is determined from the progressive appearance of structure in the multiple rounds of MCSA simulations. As described elsewhere, TERITFIX is sensitive enough to identify the non-native interactions that lead to intermediates [21] and non-native strand registry [22], results consistent with experimental studies.

Calculations for the majority of the twelve fast-folding proteins converge within the first 2–3 rounds of TERITFIX. The folding pathway is depicted for each residue by plotting the fraction of structures from the end of each round for which the residue lies within 2 Å of the native structure when the structures are aligned to the native structure by the TM score, a global metric used to assess the quality of structures [23].

The TERITFIX (Fig. 2) and all-atom MD pathways (Fig. 3 in Ref. [1]) exhibit a similar order of structure formation. The same regions of the sequence tend to develop structure early in both classes of simulations, further validating the ability of TERITFIX to identify gross aspects of the folding pathways. Our definition of structure formation is based on a global alignment to the native structure, whereas the definition used by Lindorff-Larsen *et al.* employs a local (five residue) metric of similarity to the native structure. Consequently, their depiction of the pathway is primarily sensitive to secondary rather than tertiary structure formation. The two methods also yield similar pathways when analyzed with the same local metric (see Fig. S1 in the Supplemental Material [24]).

The order in which residues become nativelike in the all-atom MD simulations correlates with their propensity to form local nativelike secondary structure in the DSE [1]. The DSE of the MD simulations contains a high, and potentially excessive amount of secondary structure in an overly collapsed DSE [7,25,26], particularly for the λ repressor where denaturation is accompanied by the unfolding of the helices according to multiple methods

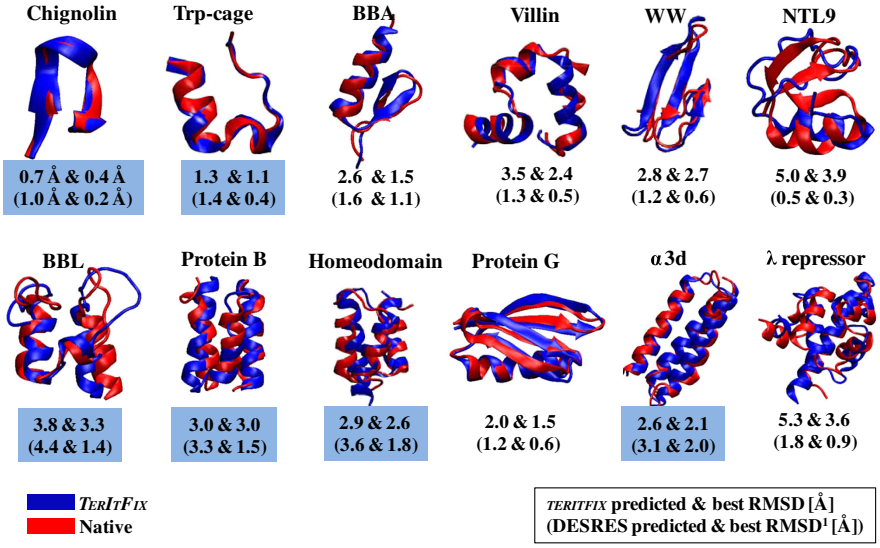

FIG. 1 (color online). Native structures produced by TERITFIX. The RMSDs of the centroids of the largest clusters and the best structures are reported (values from the MD simulations in parentheses). Blue highlighting denotes the proteins where the TERITFIX cluster centroids produce a lower RMSD.

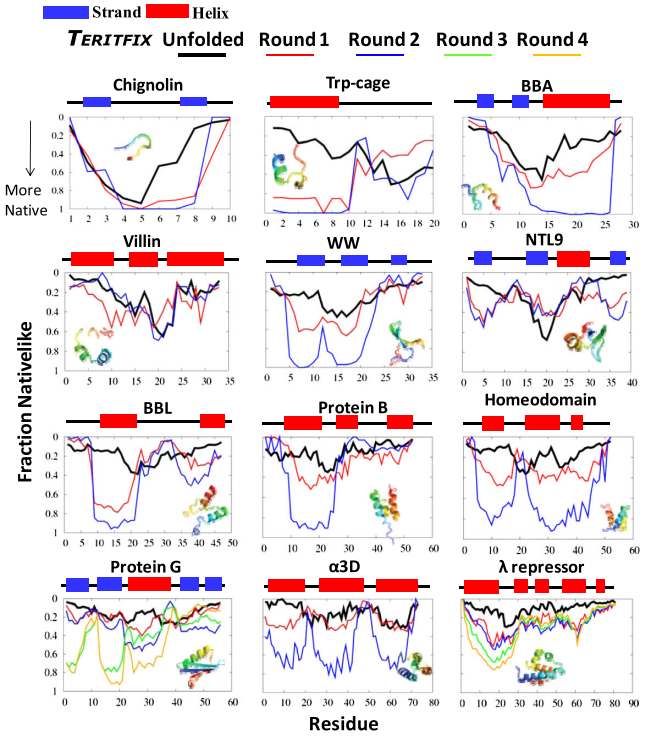

FIG. 2 (color online). Order of structure formation in TERITFIX. The fraction of structures where a residue is within 2 Å of the native structure (fraction nativelike) for the various rounds. The black line describes predictions when the starting structures are generated from the DSE ensemble of structures from the statistical coil library [17,18]. The native state secondary structure is depicted above the graphs (helix = red, strand = blue).

[17,27–30]. In contrast, TERITFIX simulations begin from an unstructured DSE that reproduces the experimentally observed NMR residual dipolar coupling patterns and dimensions of expanded chains in the DSEs [18]. Figure S2 in the Supplemental Material [24] provides a sample of five random initial structures from the TERITFIX's DSE for each of the twelve proteins. This difference likely accounts for the early portions of the pathways produced by TERITFIX having less helical structure, especially for λ repressor, Protein B and villin.

In spite of this disparity, the similarities between the methods are notable. The collapsed environment in the all-atom DSE present in the MD simulations likely promotes secondary structure in the manner similar to what TERITFIX produces as a consequence of templating onto existing structure as the temperature is lowered in the MCSA. The folding behavior is guided by similar environmental clues in both methods, a feature that may account for the general agreement between the two methods.

Our finding that TERITFIX produces similar folding pathways despite starting with a much less structured DSE cautions against overemphasizing the importance of the unfolded state propensities in determining folding pathways. TERITFIX's success provides support to a more expansive view where intrinsic local structural propensities (which often favor non-native polyproline II conformations [18,31]) may be overridden by environmental context as stable motifs sequentially interact and stabilize the formation of additional structure in an incremental fashion. For example, an otherwise unstable amphipathic helix or hairpin is stabilized in the presence of hydrophobic surfaces, whether they arise semirandomly or through specific interactions. We believe this view more accurately reflects folding behavior [32] than one that emphasizes the formation of local nativelike structure in the DSE.

The TERITFIX algorithm contains all six of the necessary physical interactions neeeded for a successful algorithm, as recently proposed by Dill, [33] namely, hydrogen bonds, van der Waals interactions, backbone angle preferences, electrostatic interactions, hydrophobic interactions, and chain entropy. In addition, we include a backbone desolvation term to reflect the observation that buried hydrogen bond donors and acceptors essentially always form hydrogen bonds in native structures [34]. Accordingly, the term we introduce to recognize this 7th feature penalizes buried amide nitrogens and carbonyl oxygens with unsatisfied hydrogen bonds. This burial term also serves to inhibit an unphysical, early, nonspecific hydrophobic collapse [25,26,35].

However, these seven energetic terms alone are woefully insufficient to locate the "needle in a haystack" native structure within the vastness of conformational space. Early folding steps in the cooperative folding of proteins are uphill in free energy, and even productive conformations unfold faster than they form on the reactant side of the kinetic barrier. Hence, an explicit search strategy is essential to guide the uphill exploration through conformational space. Accordingly, we reinforce the process of sequential stabilization by biasing (rather than enforcing) the backbone sampling, hydrogen bonding, and tertiary contacts in an iterative manner intended to mimic how real proteins traverse the energy surface on the route(s) to the native state. The success of this approach provides strong support for the contention that sequential stabilization provides an expedient mechanism for folding proteins both *in vitro* and *in silico*. The natural process of stepwise assembly in protein folding might also explain successes of previous protein modeling methods based on a combination of building blocks [36] and evolutionary algorithms to increase native content [37].

We produce encouraging results for both native structures and folding pathways for a variety of proteins without utilizing homology-based information. The overall simplicity of the Cβ-level model (lacking side chain rotameric states) decreases computational requirements by orders of magnitude and provides a direct route to apply and validate our understanding of the fundamental principles relevant to protein folding, principles also expected to be crucial for predicting protein recognition and conformational changes.

We thank the Freed/Sosnick groups, S. Piana and D. Shaw for helpful discussions and sharing of results, M. Wilde (Argonne National Laboratory, ANL) for computing assistance, and the ANL/UC CI for computing resources. This work was supported by grants from the NIH (Grant No. GM55694) (computational work and comparisons with experiment) and the U.S. Department of Energy, Office of Basic Energy Sciences, Division of Materials Sciences and Engineering under Award No. DE-SC0008631 (theory and algorithm development). Computations were also performed on the Midway and Beagle clusters at the University of Chicago.

*Authors to whom all correspondence should be addressed.
†trsosnic@uchicago.edu
‡freed@uchicago.edu

[1] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, Science **334**, 517 (2011).
[2] J. C. Martinez and L. Serrano, Nat. Struct. Biol. **6**, 1010 (1999).
[3] M. Baxa, K. F. Freed, and T. R. Sosnick, J. Mol. Biol. **381**, 1362 (2008).
[4] A. R. Viguera and L. Serrano, Nat. Struct. Biol. **4**, 939 (1997).
[5] V. P. Grantcharova, D. S. Riddle, and D. Baker, Proc. Natl. Acad. Sci. U.S.A. **97**, 7084 (2000).
[6] A. T. Shandiz, M. C. Baxa, and T. R. Sosnick, Protein Sci. **21**, 819 (2012).
[7] T. R. Sosnick and D. Barrick, Curr. Opin. Struct. Biol. **21**, 12 (2011).
[8] T. R. Sosnick and J. R. Hinshaw, Science **334**, 464 (2011).
[9] A. K. Chamberlain, T. M. Handel, and S. Marqusee, Nat. Struct. Biol. **3**, 782 (1996).
[10] W. Hu, B. T. Walters, Z. Y. Kan, L. Mayne, L. E. Rosen, S. Marqusee, and S. W. Englander, Proc. Natl. Acad. Sci. U.S.A. **110**, 7684 (2013).
[11] S. Bedard, L. C. Mayne, R. W. Peterson, A. J. Wand, and S. W. Englander, J. Mol. Biol. **376**, 1142 (2008).
[12] Z. Zheng and T. R. Sosnick, J. Mol. Biol. **397**, 777 (2010).
[13] H. Maity, M. Maity, M. M. G. Krishna, L. Mayne, and S. W. Englander, Proc. Natl. Acad. Sci. U.S.A. **102**, 4741 (2005).
[14] Y. Bai, T. R. Sosnick, L. Mayne, and S. W. Englander, Science **269**, 192 (1995).
[15] B. A. Krantz, R. S. Dothager, and T. R. Sosnick, J. Mol. Biol. **337**, 463 (2004).
[16] J. DeBartolo, A. Colubri, A. K. Jha, J. E. Fitzgerald, K. F. Freed, and T. R. Sosnick, Proc. Natl. Acad. Sci. U.S.A. **106**, 3734 (2009).
[17] A. Adhikari, K. F. Freed, and T. R. Sosnick, Proc. Natl. Acad. Sci. U.S.A. **109**, 17 442 (2012).
[18] A. K. Jha, A. Colubri, K. F. Freed, and T. R. Sosnick, Proc. Natl. Acad. Sci. U.S.A. **102**, 13 099 (2005).
[19] J. E. Fitzgerald, A. K. Jha, A. Colubri, T. R. Sosnick, and K. F. Freed, Protein Sci. **16**, 2123 (2007).
[20] A. N. Adhikari, J. Peng, M. Wilde, J. Xu, K. F. Freed, and T. R. Sosnick, Protein Sci. **21**, 107 (2012).
[21] V. L. Morton, C. T. Friel, L. R. Allen, E. Paci, and S. E. Radford, J. Mol. Biol. **371**, 554 (2007).
[22] T. Y. Yoo, A. Adhikari, Z. Xia, T. Huynh, K. F. Freed, R. Zhou, and T. R. Sosnick, J. Mol. Biol. **420**, 220 (2012).
[23] Y. Zhang and J. Skolnick, Proteins **57**, 702 (2004).
[24] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.111.028103 for materials and methods, Figs. S1 and S2.
[25] J. Jacob, B. Krantz, R. S. Dothager, P. Thiyagarajan, and T. R. Sosnick, J. Mol. Biol. **338**, 369 (2004).
[26] T. Y. Yoo, S. P. Meisburger, J. Hinshaw, L. Pollack, G. Haran, T. R. Sosnick, and K. Plaxco, J. Mol. Biol. **418**, 226 (2012).
[27] P. Chugha, H. J. Sage, and T. G. Oas, Protein Sci. **15**, 533 (2006).
[28] B. A. Krantz, A. K. Srivastava, S. Nauli, D. Baker, R. T. Sauer, and T. R. Sosnick, Nat. Struct. Biol. **9**, 458 (2002).
[29] G. S. Huang and T. G. Oas, Biochemistry **34**, 3884 (1995).
[30] R. E. Burton, G. S. Huang, M. A. Daugherty, T. L. Calderone, and T. G. Oas, Nat. Struct. Biol. **4**, 305 (1997).
[31] A. K. Jha, A. Colubri, M. H. Zaman, S. Koide, T. R. Sosnick, and K. F. Freed, Biochemistry **44**, 9691 (2005).
[32] W. K. Meisner and T. R. Sosnick, Proc. Natl. Acad. Sci. U.S.A. **101**, 13 478 (2004).
[33] K. A. Dill and J. L. MacCallum, Science **338**, 1042 (2012).
[34] P. J. Fleming and G. D. Rose, Protein Sci. **14**, 1911 (2005).
[35] T. R. Sosnick, L. Mayne, and S. W. Englander, Proteins **24**, 413 (1996).
[36] M. L. Azoitei *et al.*, Science **334**, 373 (2011).
[37] A. Schug and W. Wenzel, Biophys. J. **90**, 4273 (2006).