Corrections

**BIOPHYSICS AND COMPUTATIONAL BIOLOGY**

The authors note that two protein names appeared incorrectly. All instances of “2out-knot” should instead appear as “2ouf-knot,” and all instances of “2out-ds” should instead appear as “2ouf-ds.” Both the online article and the print article have been corrected.

**ENVIRONMENTAL SCIENCES**

The authors note that additional GenBank accession numbers for all the mitochondrial sequences reported in their paper are JX475454–JX475467.

**IMMUNOLOGY**

The authors note that, due to a printer’s error, the statement in the Acknowledgments, “we thank the Division of Intramural Research, Eunice Kennedy Shriver National Institute of Child Health and Development, for continued support and funding” should instead appear as “we thank the Division of Intramural Research, National Institute of Dental and Craniofacial Research, for continued support and funding.”

**MEDICAL SCIENCES**

The authors note that the Acknowledgments appeared incorrectly. The corrected Acknowledgments appear below.

**PHYSIOLOGY**

The authors note that the author name Faisal Cheema should instead appear as Faisal H. Cheema. The corrected author line appears below. The online version has been corrected.

XiaoBo Gao, Dominique Eladari, Francoise Leivel, Ben Yi Tew, Cristina Miro-Julia, Faisal H. Cheema, Lance Miller, Raoul Nelson, Teodor G. Pauwescu, Mary McKee, Dennis Brown, and Qais Al-Awqati

**www.pnas.org/cgi/doi/10.1073/pnas.1010364107**
While fast folding of small proteins has been relatively well characterized by experiments and theories, much less is known for slow folding of larger proteins, for which recent experiments suggested quite complex and rich folding behaviors. Here, we address how the energy landscape theory can be applied to these slow folding reactions. Combining the perfect-funnel approximation with a multiscale method, we first extended our previous atomic-interaction based coarse-grained (AICG) model to take into account local flexibility of protein molecules. Using this model, we then investigated the energy landscapes and folding routes of two proteins with complex topologies: a multidomain protein adenylate kinase (AKE) and a knotted protein 2ouf-knot. In the AKE folding, consistent with experimental results, the kinetic free energy surface showed several substates between the fully unfolded and native states. We characterized the structural features of these substates and transitions among them, finding temperature-dependent multiroute folding. For protein 2ouf-knot, we found that the improved atomic-interaction based coarse-grained model can spontaneously tie a knot and fold the protein with a probability up to 96%. The computed folding rate of the knotted protein was much slower than that of its unknotted counterpart, in agreement with experimental findings. Similar to the AKE case, the 2ouf-knot folding exhibited several substates and transitions among them. Interestingly, we found a dead-end substate that lacks the knot, thus suggesting backtracking mechanisms.

In the last two decades, chemico-physical knowledge of protein folding has been greatly deepened primarily focusing on small and fast-folding proteins (1, 2). Many such proteins were identified as two-state folders and their transition state ensembles were experimentally characterized by site-directed mutagenesis via the ΔG-value analysis (3, 4). Ultrafast experimental technology enabled direct observation of fast-folding kinetics unambiguously (5–7). Single molecule experiments directly visualized cooperative two-state folding transitions (8–10). In parallel, statistical physics theory clarified that proteins have evolved their sequences to have minimally frustrated and globally funnel-like energy landscapes (11, 12). Many simulations inspired by the theory were compared with experimental observations and gave feedback to the theory (13–17). In particular, structure-based models that can concisely realize perfect funnel landscapes of given native topologies showed great promise to capture many essential features for small and fast-folding proteins (13, 14, 18–22). Recently, a specialized computer enabled folding simulations of small proteins with an atomistic model, which were consistent with the two-state folding on funnel-like landscapes (23).

In contrast, much less is understood for slow folding of proteins with medium or large size (e.g., proteins larger than 150 residues). Generally, the folding of such proteins is much more complex with more than two states. Do such proteins still have globally funnel-like energy landscapes? More specifically, can we use similar perfect funnel models to characterize such complex folding mechanisms of individual proteins? Here, we address these issues focusing on two specific cases: folding of a multidomain protein, adenylate kinase, and that of a designed knotted protein, 2ouf-knot.

Adenylate kinase (AKE) is a medium-sized and slow-folding protein with three domains. Very recently, its folding has been characterized by the high-throughput single molecule fluorescence resonance energy transfer (smFRET) (24). Long time smFRET trajectories were fed into the hidden Markov method to identify approximately six metastable states and state-to-state transition rates. Their analysis revealed that multiple intersecting folding routes coexist in the AKE folding, and denaturant concentration can modulate the probabilities of these folding routes. A somewhat similar combination of single molecule experiments and the learning algorithm were used to obtain a complex multi-route folding networks of calmodulin, a two-domain protein (25). Such complex folding schemes provide us an ideal opportunity to examine the view of funnel-like landscape for multidomain proteins.

The knotted proteins are another class of proteins in which the structures show complexity by making knots. About 10 knotted protein folds have been identified up to now (26, 27). Their folding involves the precise threading of a chain segment through a knotting loop and therefore poses new challenge to the current protein folding theories. Shakhnovich et al. simulated the folding of a pair of related knotted proteins, Yibk and YbeA, using a structure-based Go model (28) and suggested that addition of specific nonnative interactions can greatly facilitate knot formation and thus successful folding, which calls for some alternation in the funnel landscape. Later, Onuchic and coworkers found that the similar Go model, without nonnative interactions, is able to guide the folding of the knotted protein to the native state via slpKnot intermediate (29), although the probability of successful folding was very low (≈1–2%) even after selectively strengthening/weakening the native interactions between certain residue pairs. Apparently, further studies are necessary. Recently, by connecting two monomers with a linker or by a disulfide bond, a knotted protein 2ouf-knot and its unknotted counterpart 2ouf-ds were designed (26). The 2ouf-knot and 2ouf-ds have almost identical sequence and tertiary structures and therefore provide ideal
model systems to study the role of a knot in the protein folding and to identify the critical energetic factors contributing to the successful folding of the knotted proteins. Here, folding of this pair of proteins, 2ouf-knot and 2ouf-ds, were investigated by computer simulations.

To compare simulations vis-à-vis experiments for the complex folding of larger proteins, we need accurate yet concise simulation methodology. Here, while sticking on the globally funnel-like landscape view, we include sequence-specific and atomic interactions as much as possible. For the purpose, we previously developed an atomic-interaction-based coarse-grained (AICG) model for proteins (30), and here, we extend it to improve description of local flexibility (the extended model designed as AICG2).

In the present work, we first briefly describe the AICG2 model that takes into account sequence-specific local and nonlocal interactions and further optimizes the relative weights of them via a multiscale protocol. Using the AICG2 model, we studied the folding of AKE and a knotted protein 2ouf-knot. We were able to capture all the experimentally found major features in the folding of AKE and further provide structural characterization of and transitions among substates. For the knotted protein, AICG2 realized efficient folding of 2ouf-knot with a probability up to 96%. Thus, the present study suggests that the perfect funnel model can capture essential features of the folding of these proteins with complex topologies once native interactions were carefully tuned and local flexibility was included.

Results and Discussions

Accurate yet Concise Modelling of Folding. Previously, we developed an AICG using a multiscale protocol (30). AICG uses one-bead-per-residue representation, but its interresidue nonlocal contact energies were derived from atomic interactions. Here, we extended AICG by introducing a flexible local potential (AICG2). The local interactions are described by combination of generic statistical potentials of the residue types and local contact potentials that are based on the native structure. The relative weight and average strength of the local contact potentials were derived by a multiscale protocol, the same as for the nonlocal contact interactions. Incorporating sequence-specificity and the locally flexible nature of chains, AICG2 can reproduce folding energy landscape better than the simple Go model. As an illustration, the asymmetric folding pathways of the topologically symmetric protein G and protein L, were well reproduced (SI Appendix, Figs. S1 and S2).

High Free Energy Barrier and Substates in the AKE Folding. AKE is a 214-residue kinase that plays important roles in maintaining the balance of ATP and ADP in the cell and, more importantly here, it has been used as a model multidomain protein (31). AKE consists of three domains, i.e., LID domain, NMP domain, and CORE domain (31) (Fig. L4). AKE folds relatively slowly and a recent single molecule experiment revealed complex multiroute folding (24).

We first tried to realize reversible folding and unfolding transitions using AICG2 model and unbiased molecular dynamics (MD) simulations, which turned out to be difficult due to its intrinsically slow folding. Then, we used the modified multicanonical method (32) for better conformational sampling. The converged one-dimensional free energy profiles near the folding temperature $T_F$ (Fig. 1B) show relatively high and broad free energy barrier, consistent with slow folding in vitro (33) and in silico. Albeit no apparent intermediate states in Fig. 1B, the two shoulders labelled by the arrows suggest that the folding/unfolding is not a simple two-state transition (see SI Appendix, Fig. S3). Some substates, virtually hidden in Fig. 1B, may exist between denatured and native states.

Because the high temperature simulations tend to hide the possible substates, we next conducted folding simulations starting from fully unfolded structures at the three temperatures, 0.90$T_F$, 0.92$T_F$, and 0.94$T_F$. A representative time course of formation of native contacts, i.e., so-called Q-score, at 0.90$T_F$ (Fig. 1C) shows successive transitions among different substates. Particularly, the CORE domain encounters at least five substates beside the fully unfolded and native states. Such transitions among multiple substates closely resemble smFRET trajectories observed (24), although directly mapping the substates by the experiment with those by the simulation is difficult due to the lack of structural information in the experiment.

To identify these substates robustly and statistically, we plotted the kinetic free energy landscape on the reaction coordinates $Q_{\text{tot}}$ and $Q_{\text{core}}$ at 0.90$T_F$ and its one-dimensional projections. The seven substates and completely unfolded (U) and native states (N) are labeled. The substate 1 may correspond to the denatured state (D).

Structural Features of the Substates in the AKE Folding. An advantage of the MD simulation is the ability to provide detailed structural
information for each of the substates, Fig. 2A shows, for each substate, the average contact scores at every residue by color; continuously from blue (native-like) to red (unfolded). The corresponding contact probability maps are shown in SI Appendix, Fig. S6. Each substate has a well-defined structural feature, which further rationalizes the identified substates. For example, in substates 1, 2, 3, and 4, the LID domain is folded. These substates differ in the order of the CORE domain or the NMP domain. Because of its depth, the state 1 corresponds to the denatured state. Whereas, in substates 5, 6, and 7, the LID domain is unfolded. Interestingly, among the seven substates, the CORE domain has five distinct structural features, consistent with the observations in Fig. 1D. We note that the number of the substates found is less than the product of the number of states of individual domains, suggesting coupling between domains. We also calculated the distributions of the distance between residues 73 and 203, which correspond to the label sites in the single molecule experiment, for each of the substates. Both the peak positions and the distribution widths largely vary among the substates, and resemble the results of the single molecule experiment (SI Appendix, Fig. S7).

**Multiroute Folding of AKE and Its Modulation by Temperature.** To characterize folding routes, we analyzed the transitions among the above defined substates, identifying five most probable folding routes (arrows of each color in Fig. 2 represent one route). In the figure, the breadth of the arrows represents the probabilities of the routes. (see also SI Appendix, Table S2 and Fig. S8 for more details). At 0.907\(T_F\) (Fig. 2A), the dominant folding route is \(U \rightarrow 1(D) \rightarrow 3 \rightarrow 4 \rightarrow N\) (green arrows). In this route, the LID domain folds first (or it is folded in the denatured state). Then the central \(\beta\) sheet of the CORE domain folds (substate 3), which is followed by folding of the NMP domain (substate 4). At last, \(N\)- and C-terminal \(\alpha\) helices of the CORE domain dock to make the native state. There was a slightly less probable alternative route, \(U \rightarrow 1(D) \rightarrow 3 \rightarrow 6 \rightarrow 7 \rightarrow N\) (black arrows), where the LID domain, which was folded in 1(D) and substate 3, transiently unfolded in substate 6, and 7.

This second route became dominant at higher temperatures, 0.92\(T_F\) and 0.947\(T_F\) (Fig. 2B and C). Thus, the AKE folding took heterogeneous routes, and the populations of the folding routes are sensitive to temperatures. These results are in agreement with the single molecule observations in which a few folding routes were identified and were found sensitive to the denaturant concentration. We note that although we can approximately model the effect of denaturant by high temperatures, the actual temperature dependence of folding can be very complicated as shown in a theoretical work (37). Both the equilibrium simulations around the folding temperature and the folding simulations at low temperatures showed that the LID domain frequently transits between the folded and unfolded states during the whole folding trajectory (Fig. 1C, SI Appendix, Fig. S9). As higher temperature tends to destabilize the LID domain (SI Appendix, Fig. S10), increasing the temperature promotes the folding route \(U \rightarrow 1(D) \rightarrow 3 \rightarrow 6 \rightarrow 7 \rightarrow N\) as observed in Fig. 2. Interestingly, such multiroute folding was also observed in some simple proteins (38), and has been generally linked to topological complexity of the folded state and the structure symmetry (39).

**Slow Folding of Knotted Protein 2ouf-knot.** We move to a protein with another type of topological complexity, namely, knot formation. Here we focus on a designed knotted protein, 2ouf-knot, and its unknotted counterpart, 2ouf-ds, which have almost identical sequences and tertiary structures, thus being ideal model systems to study the role of a knot in folding (Fig. 3A and B). These proteins were designed based on a natural dimeric protein, in which each subunit has four helices. Connecting two subunits with a nine-residue loop (L1 in Fig. 3A) led to 2ouf-knot, with left-handed trefoil knot of 28-residue depth in \(N\) terminus and 57-residue depth in \(C\) terminus. In contrast, the two subunits were connected with a disulfide bond to make 2ouf-ds that lacks a knot. Experimentally, both proteins reversibly folded, but 2ouf-knot folded an order of magnitude slower than 2ouf-ds (26).

Here, using AICG2, we conducted extensive unbiased MD simulations at a wide temperature range for both 2ouf-knot and 2ouf-ds. In simulations, both proteins exhibited reversible folding and unfolding transitions (Fig. 3C and D) and we confirmed that the knot is indeed formed for 2ouf-knot. Consistent with the experiment, the folding of the knotted one was much slower than the unknotted counterpart (Its folding rate is nearly seven times slower than that of the 2ouf-ds). The higher free energy barrier of 2ouf-knot in the one-dimensional free energy profile (Fig. 3E) also indicates slower folding of the knotted protein than that of 2ouf-ds. Apparently, such slowing of 2ouf-knot folding purely results from the knot because it is the only major difference between 2ouf-knot and 2ouf-ds.

**Interaction Heterogeneity and Chain Flexibility Promote the Knot Formation.** The reversible folding/unfolding transitions in Fig. 3C surprised us because previous folding simulations of another knotted protein with a Go model showed quite a low ratio of successful knot formation (29). To quantify the ratio of successful folding, we conducted extensive independent folding simulations in a range of temperatures below \(T_F\). The successful folding ratio
models when started from fully unfolded structures. For the details of each model). Consistent with the previous study (29), the simple Go model led to quite low successful events, the protein stayed at highly trapped structures or due to difference in topology? To this end, we conducted additional folding simulations with three different models, namely, a simple Go model and two intermediate models (Fig. 3A, B). Crystal structures of 2ouf-knot (PDB: 3mlg) (A) and 2ouf-ds (PDB: 3mli) (B). The eight helices are labeled by H1, H2, ..., H8, and the loop is labeled by L1. Two cysteines forming the disulfide bond in 2ouf-ds are explicitly shown. (C, D) Time courses of the Q score of the entire protein for the knotted protein (C) and the unknotted protein (D) at the folding temperature $T = T_f$. (E) One-dimensional free energy profiles for the knotted (solid line) and unknotted (dashed line) proteins at $T = 0.98T_f$, 1.00T_f, and 1.02T_f. (F) The ratios of the successful folding with four different models when started from fully unfolded structures.

as a function of temperature showed a bell-like shape with the peak of 96% near 0.95$T_f$ (Fig. 3F). In comparison with the previous study, is this high ratio due to interaction features in AICG2 or due to difference in topology? To this end, we conducted additional folding simulations with three different models, namely, a simple Go model and two intermediate models (Fig. 3F; see SI Appendix for the details of each model). Consistent with the previous study (29), the simple Go model led to quite low successful folding ratios (approximately 30%). For majority of the unsuccessful events, the protein stayed at highly trapped structures (SI Appendix, Fig. S11), in which the segment H6-H7 is misplaced around the H3, without forming the knot. Apparently, knotting from this trapped structure requires the detachment of H8 from the H2 and H5, and the threading of the longer C-terminal segment H7-H8 through the H2-H3-H4-L1 loop, which is difficult.

On the contrary, for all the successful folding events by simple Go model and AICG2, the protein formed a knot by threading the segment H1 through the H4-L1-H5-H6 loop, which is much easier due to the short length of H1. A chimera model where local interactions are simple Go and nonlocal interactions are AICG (designated AICG) gave an intermediate ratio of successful folding. Similarly, another chimera model where local interactions are of AICG2 type but with homogeneous strength and nonlocal interactions are from simple Go model (designated FLP) also showed intermediate ratio. Thus, we can conclude that both the sequence specificity of contact interactions and the local flexibility in AICG2 contributed to the ability of avoiding the highly trapped structure during folding, although it is not easy to trace such ability to certain specific contacts (SI Appendix, Fig. S12).

Actually, a previous study by Onuchic et al. showed that by selectively enhancing certain contact interactions, the successful ratio can be improved to a large extent (29), indicating the importance of the sequence specific interactions. Yet, it is of some surprise because 2ouf-knot is a designed protein and has not evolved to achieve folding ability.

**Folding of the 2ouf-knot Is More Topologically Frustrated Than Its Unknotted Counterpart.** To compare folding mechanisms of knotted and unknotted proteins, we next drew a number of two-dimensional free energy surfaces (Fig. 4, SI Appendix, Figs. S13 and S14). In Fig. 4, although the overall shapes of the free energy surfaces of two proteins show some similarity, the knotted one possesses some unique features. In early stages of folding ($Q_{tot} \sim 0.2$ and $Q_{tot} \sim 0.5$), the H1-H4 contacts can be formed partly (Fig. 4A). However, because the formation of such H1-H4 contacts hinders the knot formation, they need to be transiently broken up ($Q_{tot} \sim 0.35$ and $Q_{tot} \sim 0.6$). Thus, the knot formation introduces topological frustrations, which was also observed in the folding of other proteins with complex topology (32). In contrast, the free energy landscape of 2ouf-ds is much smoother (Fig. 4B). The folding order of the H4-H6 contacts also has large differences. For the 2ouf-knot, the H4-H6 contacts are predominantly formed early on (Fig. 4C), whereas for the 2ouf-ds, the H4-H6 contacts can be formed in a later stage with certain probability (Fig. 4D). In addition, for the 2ouf-ds, the folding of the two halves is symmetric due to the symmetry in sequences and the tertiary structures (SI Appendix, Fig. S14). In comparison, for the 2ouf-knot, the pathway with the H1-H4 contacts form before the H5-H8 contact formation is much more probable. Such asymmetry in folding routes is crucial for the efficient folding of the knot, which will be discussed in more detail below.

**Structural Features of the Substates in 2ouf-knot Folding.** Similar to the case of AKE, we conducted extensive folding simulations of 2ouf-knot starting from fully unfolded structures at the three temperatures, 0.90$T_f$, 0.925$T_f$, and 0.95$T_f$. Fig. 5 A and B shows two representative trajectories. The stepwise folding of different parts of the 2ouf-knot implies the existence of several substates. We plotted the kinetic free energy landscape projected onto the reaction coordinate $Q_{tot}$ and rmsd in Fig. 5C. Six substates besides the unfolded state and the native state were SI Appendix, Fig. S15). In Fig. 5D, the free energy landscape of the substates in 2ouf-knot folding at 0.95$T_f$ is plotted on the surfaces of the AICG2 (A) and AICG (B) models, respectively. In Fig. 5D, the free energy landscape of the substates in 2ouf-knot folding at 0.95$T_f$ is plotted on the surfaces of the AICG2 (A) and AICG (B) models, respectively.
Multiroute Folding of 2ouf-knot. In the same way as in AKE, we analyzed the folding routes of 2ouf-knot at \(0.90T_F\) (Fig. 6A), \(0.925T_F\) (Fig. 6B), and \(0.95T_F\) (Fig. 6C). More detailed analysis is given in SI Appendix, Table S3 and Fig. S16. Not surprisingly, multiple routes can populate with high probabilities, suggesting the heterogeneity of the 2ouf-knot folding routes. At \(0.90T_F\), the dominant folding route \(U \rightarrow 1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow N\) is the most sequential one. The major feature of this route is that the knot forms before the packing of the H8 to the H2 and H5 (Fig. 5A, Movie S1). In this case, the L1 and H5, parts of the knotting loop, can have high flexibility during the knot formation. Such high flexibility of the L1 and H5 tends to expand and tilt the knotting loop, which could make the threading of H1 much easier. Thus, the knot formation is not just the threading of H1 itself, but it is accompanied with high flexibility of L1 and H5. We thus call it “concerted threading.” Detailed analysis shows that in all the trajectories with this folding route at the temperature of \(0.90T_F\), the knot forms by the concerted threading mechanism. In comparison, if the pathway involves the substate 6, the knot mostly forms via a slipknot intermediate (Fig. 5B, Movie S2). For example, among the trajectories with the folding route \(U \rightarrow 1 \rightarrow 2 \rightarrow 6 \rightarrow N\), which is the most probable route via substate 6, approximately 80% of the trajectories follow the “slipknot intermediate” knotting mechanism. In this case, the H8 is well packed to H5 before the knot formation (substate 6), which makes the L1 and H5 well fixed. The rigid knotting loop makes the direct threading of H1 difficult. Consequently, the formation of the knot by the slipknot intermediate is preferred (Fig. 5B, Movie S2). Such results suggest that the knotting mechanism depends on the folding order of other parts of the protein. Notably, no route leading to the folded state via the substate 5, which requires the threading of the much longer C terminus (H8 and H9) through a different knotting loop (H2-H3-H4-L1), is observed. We found that most routes involving substate 5 terminate at the substate 5, which is a dead-end trap.

In Fig. 6, we also observe some shortcuts of the folding routes, e.g., \(U \rightarrow 1 \rightarrow 2 \rightarrow N\) in which the placements of the N-terminal segment (H1) and the C-terminal segments (H7 and H8) are cooperative. These shortcut routes become more probable at higher temperatures, perhaps because higher temperature usually tends to make the folding transition more cooperative.

Folding of Two Other Knotted Proteins. Because folding mechanisms of natural proteins and designed proteins could be very different as clarified in ref. 40, we also conducted the folding simulations...
for two natural proteins with trefoil knot, i.e., MJ0366 and YibK (28, 29, 41), which we briefly describe here. The trefoil knot of YibK is much deeper than those of the 2ouf-knot and MJ0366 (SI Appendix, Figs. S17–S19). For the MJ0366, we found that the maximal ratio of successful folding with AICG2 was 99%, a much higher ratio than that with the simple Go model (50%). However, the maximal ratios of successful folding for YibK by both simple Go model and AICG2 were very low (0.1–0.2%), which may suggest that in addition to the interaction ingredients included in the current AICG2, more realistic interaction features, e.g., nonnative interactions (28, 40), need to be considered in order to achieve efficient folding of YibK.

**Material and Methods**

**Reaction Coordinates and Data Analysis.** In describing the folding of AKE, the $Q_{\text{core}}$, $Q_{\text{cld}}$, and $Q_{\text{nmb}}$ defined as the fraction of native contacts for the full-length protein, the CORE domain, the LID domain, and the NMP domain, respectively, were used as reaction coordinates. Similarly, in describing the folding of 2ouf-knot and its unknotted counterpart, besides $Q_{\text{ref}}$, we used $Q_{\text{H1-H2}}$, $Q_{\text{H1-H3}}$, $Q_{\text{H2-H3}}$, and $Q_{\text{H1-H4}}$, defined as the fraction of native contacts between the corresponding segments. We also used the rmsd of Cα atoms. The Koniaris–Muthukumar–Taylor algorithm was used in detecting the knot (42). In constructing the folding routes, the snapshots of the folding trajectories were assigned to the substates according to the values of the reaction coordinates. See SI Appendix for details.

**Simulation Details.** We used CafeMol (43) developed in our laboratory for general CG simulations (44). The all-atom MD simulations were conducted by AMBER 11 (45). The protein structures were visualized by VMD software (46). See SI Appendix for details.

**ACKNOWLEDGMENTS.** This work was based on a joint project between the Japan Society for the Promotion of Science and the National Natural Science Foundation of China, partly supported by Grant-in-Aid for Scientific Research on Innovative Areas “Molecular Science of Fluctuations Toward Biological Functions,” by Research and Development of the Next-Generation Integrated Simulation of Living Matter of the Ministry of Education, Culture, Sports, Science, and Technology, by the National Natural Science Foundation of China (Grants Nos. 1019023320, 31130006, 10834002), and by a project funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.
Supporting Information for

Energy landscape and multi-route folding of topologically complex proteins adenylate kinase and 2out-knot

Wenfei Li\textsuperscript{a,b,c}, Tsuyoshi Terakawa\textsuperscript{b}, Wei Wang\textsuperscript{a,1} and Shoji Takada\textsuperscript{b,c,1}

\textsuperscript{a}National Laboratory of Solid State Microstructure and Department of Physics, Nanjing University, Nanjing 210093, China; \textsuperscript{b}Department of Biophysics, Graduate School of Science, Kyoto University, Kyoto 606-8502, Japan, \textsuperscript{c}CREST, Japan Science and Technology Agency, Kawaguchi, Saitama 332-0012, Japan.

(\textsuperscript{1}Email: takada@biophys.kyoto-u.ac.jp).

Text

Atomic Interaction-based Coarse-grained Model. The structure-based models, in which the energy function is fully determined by the three-dimensional structure of the native state, are able to reproduce many essential features of the protein folding, dimerization and large-amplitude conformational changes (1-7). In our recent work, within the framework of the structure-based model, we developed an atomic interaction-based coarse-grained (AICG) model (8), in which the sequence specificity is incorporated into the native nonlocal interactions by using the multiscale protocol. AICG takes one-bead-per-residue resolution and place beads at \(C_{\alpha}\) positions. The energy function of the AICG is given by

\[
V(R | R^0) = \sum_i k_b (r^i - r^i_0)^2 + \sum_i k_a (\theta^i - \theta^i_0)^2 \\
+ \sum_i \left\{ \varepsilon_{\text{dih}}^i [1 - \cos(\phi^i - \phi^i_0)] + \varepsilon_{\text{dih}}^i [1 - \cos 3(\phi^i - \phi^i_0)]/2 \right\} \\
+ \sum_{\text{native}} \varepsilon^{\text{ij}} \left[ 5\left( r^{\text{ij}} / r^{\text{ij}0} \right)^2 - 6\left( r^{\text{ij}} / r^{\text{ij}0} \right)^3 \right] + \sum_{\text{non-native}} \varepsilon^{\text{ex}} \left( C / r^{\text{ij}} \right)^2.
\]  

(S1)

In Eq. S1, the \(R\) represents the coordinates of the simulated protein structure, and \(R^0\) represents the corresponding coordinates of the native structure. \(r^i\), \(\theta^i\), \(\phi^i\) and \(r^{\text{ij}}\) are the virtual bond length...
between \( I \) and \( I+1 \) th residues, the virtual bond-bond angle composed of \( I, I+1, \) and \( I+2 \) th residues, the dihedral angle composed of \( I, I+1, I+2, \) and \( I+3 \) th residues, and the distance between \( I \)-th and \( J \)-th residues, respectively. \( r'_0, \theta'_0, \phi'_0 \) and \( r''_0 \) are the corresponding variables at the native structure.

The last term describes the excluded volume effect of the non-native residue pairs. The parameters of the bond angle term \( k'_a \), dihedral angle term \( e'_{ab} \), and nonlocal interaction term \( \epsilon'_{IJ} \) are residue dependent. These parameters are derived by decomposing the atomic interactions into the contact energies between residue pairs, and by matching the fluctuations of atomic simulations. When the residue independent parameters are used in the above energy function, the AICG is reduced to the conventional Go model (1, 2, 4). As demonstrated in Ref. (8), the AICG model gives much improved description on the large-amplitude functional motions for many allosteric proteins including adenylate kinase (AKE).

**AICG with a flexible local potential.** The energy function given by Eq. S1 formulates a rather stiff local potential, especially for the bond angle term which uses a harmonic potential. Such rigid local interactions may lead to over-stabilization of local structures. For example, the helices of a protein often stay well folded even at relatively high temperatures with the above energy function. We anticipated that in describing the folding of proteins with complex topologies, the local potential with more realistic flexibility can be important. In this work, we thus improve the AICG model by implementing a flexible local potential. For convenience, the modified AICG model with flexible local potential is termed as AICG2 throughout the current work. The following energy function is used in the AICG2:

\[
V = \sum_I k'_b (r'_I - r''_0)^2 \\
+ \sum_I V'_a (\theta'_I) + \sum_I V'_d (\phi'_I) + \sum_{J+2 \leq I \leq J+3} e'_{ab} \exp(-\frac{(r'_{IJ} - r''_{IJ})^2}{2w^2}) \\
+ \sum_{J+3} \sum_{I> J+3} e'_{ij} [5(r'_{IJ} / r''_{IJ})^2 - 6(r'_{0IJ} / r''_{IJ})^6] + \sum_{I> J+3} e(C / r''_{IJ})^{12}
\]

(S2)

The second and the third terms are the flexible local potentials that represent generic local propensity of given amino acids, while the fourth term represents the structure-based local contact potential that represents specific local interactions of the given protein structure. The other terms are the same as those in Eq. S1.

The flexible local potentials \( V'_a (\theta'_I) \) and \( V'_d (\phi'_I) \) have recently been developed (9), and here we briefly summarize it. We constructed these local potentials by the following Boltzmann inversion,

\[
V_a (\theta) = -k_a T \ln \frac{P(\theta)}{\sin \theta}
\]

(S3)
\[ V_{\text{dih}}(\varphi) = -k_B T \ln(P(\varphi)) \]  

where \( P(\theta) \) and \( P(\varphi) \) are the distributions of the angles and dihedral angles calculated from a “loop-segment-library” given in Ref. (9). \( P(\theta) \) is calculated for the angles with different central residue types. Similarly, \( P(\varphi) \) is calculated for the dihedral angles of every combinations of the two central residues, resulting in 400 distributions. These 400 probability distributions were inspected and reduced into 23 types of \( P(\varphi) \). The \( k_B \) and \( T (=300K) \) are the Boltzmann constant and temperature, respectively. The resulted local potentials are represented by tabulated values. In the simulations, the potential for the \( I \)-th angle \( V_{\text{a}I}(\theta) \) was obtained by interpolating the tabulated potentials with spline functions. Meanwhile, the potential for the \( I \)-th dihedral angle \( V_{\text{dih}}(\varphi) \) was obtained by fitting the tabulated potentials with the following truncated Fourier series

\[ V_{\text{dih}}(\varphi) = \sum_{n=1}^{3} k_n \sin(n\varphi) + \sum_{n=1}^{3} k_n \cos(n\varphi) + C \]  

where \( k_m, k_n, \) and \( C \) are the Fourier coefficients. The above generic local potential considers the propensities of secondary structures, therefore can reasonably describe the conformational distribution of the unfolded states or intrinsically disordered proteins/regions (9). More detailed introduction to the above generic local potential can be found in Ref. (9).

The fourth term in Eq. S2 is the specific local potential in the form of contacts between the \( I, I+2 \) and \( I, I+3 \) residue pairs, which represent the local contributions to the funnel-shaped energy landscape (10). We used a Gaussian function to describe the interactions of the local contacts with the \( \varepsilon_{\text{loc}}^{IJ} \) being the strengths of the local contacting interactions between residue \( I \) and \( J \), and \( w \) being the width of Gaussian potential. In this work, a value of 0.15 Å was used for the Gaussian width \( w \).

The importance of such local contacts was motivated from a previous work by Eswar and Ramakrishnan in which the authors conducted a detailed statistical survey to the hydrogen bonds in globular protein structures, and revealed that there are significant hydrogen bonds even between the \( I, I+2 \) or \( I, I+3 \) residue pairs (11). Such local hydrogen bonds may play crucial roles for the folding and stability of proteins. Compared to the spring-type local potential, the Gaussian potential, together with the generic statistical potential, given in Eq. S2 represent a more realistic description to the flexibility of the local interactions, especially for the unfolded states or intrinsically disordered proteins/regions.

Following the Ref. (8), the coefficients \( \varepsilon_{\text{loc}}^{IJ} \) and \( \varepsilon_{\text{nloc}}^{IJ} \), which are the strengths of the local and nonlocal interactions, are derived by the multiscale protocol. Detailed procedure of the multiscale protocol can be found in Ref. (8). Briefly, the \( \varepsilon_{\text{loc}}^{IJ} \) (\( \varepsilon_{\text{nloc}}^{IJ} \)) in Eq. S2 is written as \( \varepsilon_{\text{loc}}^{IJ} = \varepsilon_{\text{loc}}^{IJ} W_{\text{loc}}^{IJ} \) (\( \varepsilon_{\text{nloc}}^{IJ} = \varepsilon_{\text{nloc}}^{IJ} W_{\text{nloc}}^{IJ} \)), where the \( \varepsilon_{\text{loc}} \) (\( \varepsilon_{\text{nloc}} \)) represents the average strength of the local (nonlocal) interactions. The \( W_{\text{loc}}^{IJ} \) (\( W_{\text{nloc}}^{IJ} \)) is the relative weight of the local and nonlocal interactions. The
$W_{\text{loc}}^{IJ}$ and $W_{\text{nloc}}^{IJ}$ are given by $W_{\text{loc}}^{IJ} = E_{\text{loc}}^{IJ} / \langle E_{\text{loc}}^{IJ} \rangle$ and $W_{\text{nloc}}^{IJ} = E_{\text{nloc}}^{IJ} / \langle E_{\text{nloc}}^{IJ} \rangle$, respectively. Here, $E_{\text{loc}}^{IJ}$ is the contact energy between the natively interacting residue pair $I$ and $J$. Similarly, $E_{\text{nloc}}^{IJ}$ is the contact energy between the local residue pair $I$ and $J$ with $J + 2 \leq I \leq J + 3$. $\langle \cdots \rangle$ represents the average of the contact energies. The contact energies are derived by decomposing the all atom (AA) interactions into pair-wise residue-residue interactions by (12-15)

$$E^{IJ} = \sum_{i \in I} \sum_{j \in J} u_{AA}^{ij}$$

(S6)

where $u_{AA}^{ij}$ is the decomposed AA interactions between atom $i$ and $j$. AMBER11 with force field ff99SB was used for the AA energy calculations (12, 16, 17). In calculating the contributions of the water molecules to the pairwise interactions of protein atoms, the GB/SA implicit solvation model was used with the solvent accessible surface area (SASA) dependent term being treated by the LCPO method (12, 16, 17). Before the energy decomposition, the PDB structures of the studied proteins were firstly minimized for 700 steps to remove possible bad contacts. With the obtained contact energies, we can calculate the relative weight of the local and nonlocal interactions, $W_{\text{loc}}^{IJ}$ and $W_{\text{nloc}}^{IJ}$. As in Ref. (8), in calculating the $W_{\text{loc}}^{IJ}$ and $W_{\text{nloc}}^{IJ}$, the lower and upper cutoffs of the contact energies are used, which were set as -5.0 kcal/mol and -0.5 kcal/mol, respectively.

The average strengths of the local and nonlocal interactions, $\varepsilon_{\text{nloc}}$ and $\varepsilon_{\text{loc}}$, are optimized by fitting the calculated fluctuations to those calculated by AA molecular dynamics (MD) simulations. The AA MD simulations were conducted at 300K with the AMBER force field ff99SB and explicit water, and the lengths of the simulation time are 50.0ns. The CG simulations were conducted by Langevin dynamics at 300K for $2 \times 10^6$ MD steps with friction coefficient $\gamma = 0.25$. The fluctuations corresponding to the bond lengths, the distances of the local contacts, and the distances of the native nonlocal contacts are given by $F_b = \sum_1 (< r_{ij}^2 > - < r_i^2 >)$, $F_{\text{loc}} = \sum_{J+2 \leq I \leq J+3} (< r_{IJ}^2 > - < r_{ij}^2 >)$ and $F_{\text{nloc}} = \sum_{I,J} (< r_{IJ}^2 > - < r_{ij}^2 >)$, respectively. The three $I, I+2 (I, I+3)$ pairs in both termini of the amino acid chain were not included in the summation. Starting from a set of initial values which are listed in the Table S1, the CG parameters $k_b$, $\varepsilon_{\text{loc}}^{IJ}$, and $\varepsilon_{\text{nloc}}^{IJ}$ are optimized iteratively by

$k_b (i + 1) = k_b (i) \cdot (1 + (1 - F_{\text{AA}}^{AA} / F_{\text{CG}}^{CG} (i)) \cdot f)$,

$\varepsilon_{\text{loc}} (i + 1) = \varepsilon_{\text{loc}} (i) \cdot (1 + (1 - F_{\text{loc}}^{AA} / F_{\text{loc}}^{CG} (i)) \cdot f)$, and

$\varepsilon_{\text{nloc}} (i + 1) = \varepsilon_{\text{nloc}} (i) \cdot (1 + (1 - F_{\text{nloc}}^{AA} / F_{\text{nloc}}^{CG} (i)) \cdot f$, respectively. Here, the index $i$ represents the iteration step. The parameter $f$ ($0.0 < f < 1.0$) is a scaling factor used to control the convergence speed. The superscripts AA and CG represent that the fluctuations are calculated by AA MD simulations and CG MD simulations, respectively. More details of the fluctuation matching are given in Ref. (8). As
demonstrated in Ref. (8), with the above optimization algorithm, we can derive the CG parameters for each target protein. We repeated the same optimization procedure for all the proteins in the dataset list below, and the averaged parameters are used as generic parameters. The averages and the standard errors of the derived parameters are given in Table S1.

**Dataset.** A dataset containing 44 single domain proteins with sizes ranging from 45 to 193 residues was used in training the AICG2 parameters by the fluctuation matching. These proteins were taken from Ref. (8, 18). The PDB codes of these proteins are: 1amx, 1bm8, 1bv1, 2cpl, 1ctf, 2end, 1fna, 1gpr, 1ifc, 2igd, 1cll, 1msi, 1npk, 1pdo, 1pne, 2pth, 2rn2, 3tss, 1vie, 1whi, 1who, 1xnb, 1aps, 1coa, 1csp, 1fmk, 1hz6, 1imq, 1n88, 1o6x, 1pgb, 1rfa, 1shf, 1shg, 1ss1, 1sso, 1st7, 1ubq, 1w4e, 1w4j, 1yyj, 2abd, 1div (N domain), 1div (C domain).

**Details of the CG models used in this work.** In this work, to decipher the contributions of different interaction factors to the protein folding mechanisms, we used four models with different interaction features, including the conventional Go model (Go), the AICG2 (AICG2), the flexible local potential model (FLP), and the AICG developed in our previous work (AICG) (8). By comparing the simulation results of these models, we can decipher the effects of the sequence specificity of the interactions and the flexibility of the local potentials. The interaction features for each of the above models are described in details below.

In the Go model, the energy function given in Eq. S1 was used. The parameters in Eq. S1 are constants and independent of the residues. All the parameters are taken from the default values in CafeMol (19) \(k_b=100\text{kcal/mol/Å}^2, k_a=20.0\text{kcal/mol/rad}^2, \varepsilon_{\text{dih}}=1.0\text{kcal/mol}, \) and \(\varepsilon_{\text{nloc}}=0.3\text{kcal/mol}\). With the default parameters, the folding temperatures for some typical small proteins can be reasonably reproduced (19). As discussed above, the Go model formulates a stiff local interaction. Meanwhile, the contact energies are homogeneous, therefore are sequence independent.

In the AICG2, the energy function given in Eq. S2 was used. The averages of the interaction strength were taken from Table S1. The relative weights of the local contacts and the nonlocal contacts were calculated according to the atomic interactions by using the energy decomposition described above. Therefore, compared the Go model, the AICG2 includes both realistic flexibility in the local interactions and the sequence specificity in the local and nonlocal contact interactions.

In the FLP model, the energy function Eq. S2 was used. However, compared to the AICG2, all the local and nonlocal contact interactions are independent of the residues, and the interaction strengths were taken from Table S1. Therefore, by comparing the results from AICG2 simulations and FLP simulations, we can analyze the effects of the sequence specificity of the contact interactions.
In the AICG, the energy function Eq. S1 was used. The averages of the interaction strength were extracted by using the fluctuation matching in Ref. (8). Similar to AICG2, the relative weights of the nonlocal contacts were calculated according to the atomic interactions by using the energy decomposition. Therefore, by comparing the results from AICG2 simulations and AICG simulations, we can decipher the effects of the realistic flexibility of the local interactions. The intermediate models AICG, FLP were referred to as chimera models in the main text. We emphasize that all these chimera models were purely used as a technique of simulations, and they cannot be realized in experiment.

**CG simulations for the folding of AKE.** In constructing the free energy landscapes for the folding of the AKE, we used the modified multicanonical method to sample the conformational space since the unbiased MD trajectories could not sample the folded and unfold states reversibly within reasonable simulation time. The modified multicanonical method, which was developed by Onuchic et al (20), can efficiently speed up the reversible folding/unfolding transitions even for quite large proteins. The CG simulations were conducted by Langevin dynamics for $2 \times 10^9$ MD steps with friction coefficient $\gamma = 0.25$ and temperatures ranging from 300K to 420K. Both the Go model and AICG2 were used in the simulations. The folding temperature $T_F$ was determined from these simulations.

We also conducted separate folding simulations at the temperatures of $0.90T_F$, $0.92T_F$, and $0.94T_F$ starting from fully unfolded structures using unbiased MD. At each of the temperatures, 200 independent folding simulations with the length of $2 \times 10^7$ MD steps were conducted. The initial structures were prepared by MD simulations at 1000K, and each of the 200 folding simulations used a different initial structure. Both the Go model and AICG2 were used in the folding simulations.

**Data analysis for the folding of AKE.** In describing the folding of the AKE, the reaction coordinates $Q_{\text{tot}}$, $Q_{\text{CORE}}$, $Q_{\text{LID}}$, and $Q_{\text{NMP}}$, which are defined as the fraction of native contacts for the full protein, the CORE domain, the LID domain, and the NMP domain, were used. We also used the root mean square deviation (RMSD) of $\alpha$ atoms as a reaction coordinate. Throughout this work, the residue pairs are considered to be in native contacts if they are separated by at least four residues in sequence and the distance between the closest heavy atoms of the two residues are within 6.5 Å in the native structure. During the CG simulations, a contact between the natively contacting residues $I$ and $J$ is considered to be formed if their distance is smaller than 1.2 times of the $\alpha$ distance of the corresponding residue pair in the native structure. In constructing the free energy landscapes around the folding temperatures, the sampled structures by modified multicanonical method at different temperatures were reweighed by the weighted histogram analysis method (WHAM). In identifying the sub-states of the AKE folding, we constructed the kinetic free energy landscapes, which were
used in characterizing the folding/unfolding pathways of proteins and RNA (21-23). The kinetic free energy landscapes were constructed based on the sampled structures of the 200 separate folding simulations at the temperature of $0.90T_F$, and were represented by the negative natural logarithm of the histogram on the reaction coordinates $Q_{tot}$ and $Q_{CORE}$.

In characterizing the structural features for each of the sub-states identified in Fig. 1D, the residue-resolved contact scores and the contact probability map were used. The residue-resolved contact scores, defined as the average contact scores at every residue, were calculated by the averages of the formation probabilities of the native contacts between a given residue and all of its contacting partners. Similarly, the contact probability is defined as the formation probability for each of the native contacts. The structures near the centers of the sub-states were used in calculating the residue-resolved contact scores and the contact probabilities.

To construct the folding routes, one needs unambiguous definition for each of the sub-states. Assigning the sub-states only on the basis of the locations of the free energy basins in the kinetic free energy landscapes may not be appropriate since the snapshots locating at the same basin may have different structural features due to the stochastic nature of the trajectories. Detailed analysis to the residue-resolved contact scores and the contact probability maps given in Fig. 2 and Fig. S4 showed that different sub-states have very well defined structure features. Therefore, we can assign the snapshots sampled in the folding simulations to each of the sub-states according to the distinct structure features revealed in the Fig. 2 and Fig. S4 by measuring the folding extent for each of the three domains. In this work, when constructing the folding routes of the folding trajectories, the snapshots were assigned to the sub-states according to the following criteria:

State U: $Q_{CORE}<0.2$, $Q_{LID}<0.5$, and $Q_{NMP}<0.5$;
Substate 1(D): $0.2<Q_{CORE}<0.55$, $Q_{LID}>0.7$, and $Q_{NMP}<0.5$;
Substate 2: $0.3<Q_{CORE}<0.55$, $Q_{LID}>0.7$, and $Q_{NMP}>0.7$;
Substate 3: $0.55<Q_{CORE}<0.71$, $Q_{LID}>0.7$, and $Q_{NMP}<0.5$;
Substate 4: $0.71<Q_{CORE}<0.81$, $Q_{LID}>0.7$, and $Q_{NMP}>0.7$;
Substate 5: $0.81<Q_{CORE}<0.91$, $Q_{LID}<0.5$, and $Q_{NMP}>0.7$;
Substate 6: $0.81<Q_{CORE}<0.91$, $Q_{LID}<0.5$, and $Q_{NMP}<0.5$;
Substate 7: $Q_{CORE}>0.91$, $Q_{LID}<0.5$, and $Q_{NMP}>0.7$
State N: $Q_{CORE}>0.91$, $Q_{LID}>0.7$, and $Q_{NMP}>0.7$.

Following the above criteria, we can unambiguously assign the snapshots to each of the sub-states, therefore, construct the folding routes. As observed in the single molecule experiment for AKE, the folding/unfolding trajectories may involve the back-and-forth transitions, which lead to loops in the folding routes (24). For simplicity of the figure, these loops were removed from the folding routes.
The assignment of the sub-states depends on the cutoffs listed above. However, changing these cutoffs within reasonable range does not affect the qualitative conclusions. For example, we changed the above cutoffs from $Q_{\text{LID}(\text{QNM})} < 0.5$ to $Q_{\text{LID}(\text{QNM})} < 0.6$, and $Q_{\text{LID}(\text{QNM})} > 0.7$ to $Q_{\text{LID}(\text{QNM})} > 0.6$ for the simulations at 0.90$T_F$, and the five most probable folding routes keep unchanged.

To describe the transitions from an initial sub-state to all the other possible sub-states, the branch ratios were calculated. The branch ratio is defined as the transition probability from the initial state to the different final states. For each initial state, the transition probability is normalized. All the transitions between the sub-states were included in calculating the branch ratios.

Throughout this work, the protein structures were visualized with the VMD software (25).

**CG simulations for the folding of 2out-knot and 2out-ds.** Before the simulations, the missing residues 93-100 in the crystal structure were modelled by the web server ModLoop (26, 27), which uses the modelling routine in MODELLER (28). In constructing the free energy landscapes for the folding of 2out-knot and 2out-ds, we used the unbiased MD to sample the conformational space except for the folding of 2out-knot with Go model, for which we had to use the modified multicanonical method. The CG simulations were conducted by Langevin dynamics for $2 \times 10^9$ MD steps with friction coefficient $\gamma = 0.25$ and temperatures ranging from 320K to 420K. We also conducted separate folding simulations at a number of temperatures below the folding temperature for 2out-knot using AICG2, Go model, FLP, and AICG. For each potential at each of the temperatures, 100 folding simulations with the length of $2 \times 10^7$ MD steps were conducted.

**Data analysis for the folding of 2out-knot and 2out-ds.** In describing the folding of 2out-knot and 2out-ds, the eight helices were labelled as H1, H2, ..., H8, and the loop in the 2out-knot was labelled as L1 (Fig. 3A). In addition to $Q_{\text{tot}}$, the reaction coordinates $Q_{H1-H4}$, $Q_{H2-L1}$, $Q_{H2-H6}$, $Q_{H2-H7}$, $Q_{H2-H8}$, $Q_{H3-H6}$, $Q_{H4-H6}$, which are defined as the fraction of native contacts between the H1 and H4, H2 and L1, H2 and H6, H2 and H7, H2 and H8, H3 and H6, and H4 and H6, respectively, were used. We also used the root mean square deviation (RMSD) of $C_\alpha$ atoms as a reaction coordinate. In constructing the free energy landscapes around the folding temperatures, the sampled structures at different temperatures were reweighed by the WHAM. The kinetic free energy landscapes were constructed based on the sampled structures of the 100 separate folding simulations at the temperature of 0.90$T_F$, and were represented by the negative natural logarithm of the histogram on the reaction coordinates $Q_{\text{tot}}$ and RMSD. The residue-resolved contact scores and the contact probability maps were calculated for 2out-knot by the similar methods described above. In detecting the knot for a given structure, we used an in-house software based on the KMT algorithm (29, 30). The knotting status is denoted by $N_{\text{knot}}$ ($N_{\text{knot}} = 1$ for knotted, and $N_{\text{knot}} = 0$ for unknotted).

Similarly to the folding of AKE, when constructing the folding routes for the folding simulations of
2 out-knot, the snapshots were assigned to the sub-states according to the following criteria:

State U: \( N_{\text{knot}} = 0, \)
- \( Q_{H4-H6} < 0.5, \ Q_{H3-H6} < 0.5, \ Q_{H2-H7} < 0.5, \)
- \( Q_{H2-H6} < 0.5, \)
- \( Q_{H1-H4} < 0.5, \ Q_{H2-L1} < 0.5, \)
- \( Q_{H2-H8} < 0.5; \)

Substate 1: \( N_{\text{knot}} = 0, \)
- \( Q_{H4-H6} > 0.8 \) or \( Q_{H3-H6} > 0.8 \) or \( Q_{H2-H7} > 0.8, \)
- \( Q_{H2-H6} < 0.5, \)
- \( Q_{H1-H4} < 0.5, \ Q_{H2-L1} < 0.5, \)
- \( Q_{H2-H8} < 0.5; \)

Substate 2: \( N_{\text{knot}} = 0, \)
- \( Q_{H4-H6} > 0.8 \) or \( Q_{H3-H6} > 0.8 \) or \( Q_{H2-H7} > 0.8, \)
- \( Q_{H2-H6} > 0.8, \)
- \( Q_{H1-H4} < 0.5, \ Q_{H2-L1} < 0.5, \)
- \( Q_{H2-H8} < 0.5; \)

Substate 3: \( N_{\text{knot}} = 0, \)
- \( Q_{H4-H6} > 0.8, \ Q_{H3-H6} > 0.8, \ Q_{H2-H7} > 0.8, \)
- \( Q_{H2-H6} > 0.8, \)
- \( Q_{H1-H4} > 0.8, \ Q_{H2-L1} < 0.5, \)
- \( Q_{H2-H8} < 0.5; \)

Substate 4: \( N_{\text{knot}} = 1, \)
- \( Q_{H4-H6} > 0.8, \ Q_{H3-H6} > 0.8, \ Q_{H2-H7} > 0.8, \)
- \( Q_{H2-H6} > 0.8, \)
- \( Q_{H1-H4} > 0.8, \ Q_{H2-L1} > 0.8, \)
- \( Q_{H2-H8} < 0.5; \)

Substate 5: \( N_{\text{knot}} = 0, \)
- \( Q_{H4-H6} > 0.8, \ Q_{H3-H6} < 0.5, \ Q_{H2-H7} > 0.8, \)
- \( Q_{H2-H6} > 0.8, \)
- \( Q_{H1-H4} > 0.8, \ Q_{H2-L1} > 0.8, \)
- \( Q_{H2-H8} > 0.8; \)

Substate 6: \( N_{\text{knot}} = 0, \)
- \( Q_{H4-H6} > 0.8, \ Q_{H3-H6} > 0.8, \ Q_{H2-H7} > 0.8, \)
- \( Q_{H2-H6} > 0.8, \)
- \( Q_{H1-H4} > 0.8, \ Q_{H2-L1} < 0.5, \)
- \( Q_{H2-H8} > 0.8; \)
State N: \( N_{\text{knot}}=1, \text{RMSD}<3.0, \)
\( Q_{H4-H6}>0.8, Q_{H3-H6}>0.8, Q_{H2-H7}>0.8, \)
\( Q_{H2-H6}>0.8, \)
\( Q_{H1-H4}>0.8, Q_{H2-L1}>0.8, \)
\( Q_{H2-H8}>0.8. \)

Based on the above assignment of sub-states, we can construct the folding routes and calculate their abundance. Similarly, we can calculate the branch ratios of transitions for the folding of 2out-knot.

References
33. Z. Zhang, H. S. Chan, Competition between native topology and nonnative interactions in simple and complex folding kinetics of natural and designed proteins. *Proc Natl Acad Sci USA* 107, 2920 (Feb 16, 2010).
Table S1. The averages (Av.) and standard errors (S.E.) of the derived AICG2 parameters for the proteins in the dataset. The corresponding initial values (Ini.) of the parameters in the fluctuation matching are also listed.

<table>
<thead>
<tr>
<th>param</th>
<th>$k_b$</th>
<th>$\varepsilon_{loc}$</th>
<th>$\varepsilon_{adloc}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ini.</td>
<td>200.00</td>
<td>5.00</td>
<td>2.00</td>
</tr>
<tr>
<td>Av.</td>
<td>110.40</td>
<td>1.46</td>
<td>0.60</td>
</tr>
<tr>
<td>S.E.</td>
<td>0.21</td>
<td>0.05</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Table S2. The ten most probable folding routes at three temperatures for protein AKE. The numbers in the bracket of the first column is the percentage of successful folding events at each temperatures. See Fig. 1D for the numbering of the states.

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Folding route</th>
<th>Probability %</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.90TF (68.5%)</td>
<td>$U \rightarrow I(D) \rightarrow 3 \rightarrow 4 \rightarrow N$</td>
<td>16.1</td>
</tr>
<tr>
<td></td>
<td>$U \rightarrow I(D) \rightarrow 3 \rightarrow 6 \rightarrow 7 \rightarrow N$</td>
<td>11.7</td>
</tr>
<tr>
<td></td>
<td>$U \rightarrow I(D) \rightarrow 2 \rightarrow 4 \rightarrow N$</td>
<td>9.5</td>
</tr>
<tr>
<td></td>
<td>$U \rightarrow I(D) \rightarrow 2 \rightarrow 4 \rightarrow 5 \rightarrow 7 \rightarrow N$</td>
<td>7.3</td>
</tr>
<tr>
<td></td>
<td>$U \rightarrow I(D) \rightarrow 3 \rightarrow 4 \rightarrow 5 \rightarrow N$</td>
<td>6.5</td>
</tr>
<tr>
<td></td>
<td>$U \rightarrow I(D) \rightarrow 3 \rightarrow 6 \rightarrow N$</td>
<td>3.6</td>
</tr>
<tr>
<td></td>
<td>$U \rightarrow I(D) \rightarrow 3 \rightarrow 4 \rightarrow 5 \rightarrow 7 \rightarrow N$</td>
<td>2.9</td>
</tr>
<tr>
<td></td>
<td>$U \rightarrow I(D) \rightarrow 3 \rightarrow N$</td>
<td>2.9</td>
</tr>
<tr>
<td></td>
<td>$U \rightarrow I(D) \rightarrow 3 \rightarrow 2 \rightarrow 4 \rightarrow N$</td>
<td>2.2</td>
</tr>
<tr>
<td></td>
<td>$U \rightarrow I(D) \rightarrow 3 \rightarrow 5 \rightarrow 7 \rightarrow N$</td>
<td>2.2</td>
</tr>
</tbody>
</table>

| 0.92TF (72.0%) | $U \rightarrow I(D) \rightarrow 3 \rightarrow 6 \rightarrow 7 \rightarrow N$ | 19.4 |
|               | $U \rightarrow I(D) \rightarrow 3 \rightarrow 4 \rightarrow N$ | 10.4 |
|               | $U \rightarrow I(D) \rightarrow 3 \rightarrow 6 \rightarrow N$ | 6.9 |
|               | $U \rightarrow I(D) \rightarrow 3 \rightarrow N$ | 6.3 |
|               | $U \rightarrow I(D) \rightarrow 2 \rightarrow 4 \rightarrow 5 \rightarrow 7 \rightarrow N$ | 5.6 |
|               | $U \rightarrow I(D) \rightarrow 3 \rightarrow 2 \rightarrow 4 \rightarrow N$ | 4.2 |
|               | $U \rightarrow I(D) \rightarrow 6 \rightarrow 7 \rightarrow N$ | 3.5 |
|               | $U \rightarrow I(D) \rightarrow 2 \rightarrow 4 \rightarrow N$ | 3.5 |
|               | $U \rightarrow I(D) \rightarrow 5 \rightarrow 7 \rightarrow N$ | 2.8 |
|               | $U \rightarrow I(D) \rightarrow 3 \rightarrow 7 \rightarrow N$ | 2.1 |

| 0.94TF (57.0%) | $U \rightarrow I(D) \rightarrow 3 \rightarrow 6 \rightarrow 7 \rightarrow N$ | 28.1 |
|               | $U \rightarrow I(D) \rightarrow 3 \rightarrow 6 \rightarrow N$ | 7.0 |
|               | $U \rightarrow I(D) \rightarrow 3 \rightarrow 4 \rightarrow N$ | 6.1 |
|               | $U \rightarrow I(D) \rightarrow 6 \rightarrow 7 \rightarrow N$ | 6.1 |
|               | $U \rightarrow I(D) \rightarrow 2 \rightarrow 4 \rightarrow 5 \rightarrow 7 \rightarrow N$ | 4.4 |
|               | $U \rightarrow I(D) \rightarrow 2 \rightarrow 4 \rightarrow N$ | 4.4 |
|               | $U \rightarrow I(D) \rightarrow 3 \rightarrow N$ | 3.5 |
|               | $U \rightarrow I(D) \rightarrow 2 \rightarrow 5 \rightarrow 7 \rightarrow N$ | 2.6 |
|               | $U \rightarrow 2 \rightarrow 4 \rightarrow N$ | 2.6 |
|               | $U \rightarrow I(D) \rightarrow 2 \rightarrow 4 \rightarrow 5 \rightarrow N$ | 1.8 |
Table S3. The ten most probable folding routes at three temperatures for protein 2out-knot. The numbers in the bracket of the first column is the percentage of successful folding events at each temperatures. See Fig. 3A for the numbering of the states.

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Folding route</th>
<th>Probability %</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.90(T_F) (91.0%)</td>
<td>U (\rightarrow) 1 (\rightarrow) 2 (\rightarrow) 3 (\rightarrow) 4 (\rightarrow) N</td>
<td>36.3</td>
</tr>
<tr>
<td></td>
<td>U (\rightarrow) 1 (\rightarrow) 2 (\rightarrow) 4 (\rightarrow) N</td>
<td>14.3</td>
</tr>
<tr>
<td></td>
<td>U (\rightarrow) 1 (\rightarrow) 2 (\rightarrow) 6 (\rightarrow) N</td>
<td>13.2</td>
</tr>
<tr>
<td></td>
<td>U (\rightarrow) 1 (\rightarrow) 2 (\rightarrow) N</td>
<td>6.6</td>
</tr>
<tr>
<td></td>
<td>U (\rightarrow) 1 (\rightarrow) 2 (\rightarrow) 3 (\rightarrow) 6 (\rightarrow) N</td>
<td>5.5</td>
</tr>
<tr>
<td></td>
<td>U (\rightarrow) 2 (\rightarrow) N</td>
<td>4.4</td>
</tr>
<tr>
<td></td>
<td>U (\rightarrow) N</td>
<td>4.4</td>
</tr>
<tr>
<td></td>
<td>U (\rightarrow) 1 (\rightarrow) N</td>
<td>3.3</td>
</tr>
<tr>
<td></td>
<td>U (\rightarrow) 1 (\rightarrow) 6 (\rightarrow) N</td>
<td>3.3</td>
</tr>
<tr>
<td></td>
<td>U (\rightarrow) 2 (\rightarrow) 3 (\rightarrow) 4 (\rightarrow) N</td>
<td>3.3</td>
</tr>
<tr>
<td>0.925(T_F) (95.0%)</td>
<td>U (\rightarrow) 1 (\rightarrow) 2 (\rightarrow) 6 (\rightarrow) N</td>
<td>24.2</td>
</tr>
<tr>
<td></td>
<td>U (\rightarrow) 1 (\rightarrow) 2 (\rightarrow) 3 (\rightarrow) 4 (\rightarrow) N</td>
<td>22.1</td>
</tr>
<tr>
<td></td>
<td>U (\rightarrow) 1 (\rightarrow) 2 (\rightarrow) N</td>
<td>10.5</td>
</tr>
<tr>
<td></td>
<td>U (\rightarrow) 1 (\rightarrow) 2 (\rightarrow) 4 (\rightarrow) N</td>
<td>6.3</td>
</tr>
<tr>
<td></td>
<td>U (\rightarrow) 2 (\rightarrow) N</td>
<td>6.3</td>
</tr>
<tr>
<td></td>
<td>U (\rightarrow) 1 (\rightarrow) 6 (\rightarrow) N</td>
<td>6.3</td>
</tr>
<tr>
<td></td>
<td>U (\rightarrow) N</td>
<td>5.3</td>
</tr>
<tr>
<td></td>
<td>U (\rightarrow) 2 (\rightarrow) 6 (\rightarrow) N</td>
<td>5.3</td>
</tr>
<tr>
<td></td>
<td>U (\rightarrow) 1 (\rightarrow) 2 (\rightarrow) 3 (\rightarrow) 6 (\rightarrow) N</td>
<td>4.2</td>
</tr>
<tr>
<td></td>
<td>U (\rightarrow) 1 (\rightarrow) 2 (\rightarrow) 3 (\rightarrow) N</td>
<td>3.2</td>
</tr>
<tr>
<td>0.95(T_F) (96.0%)</td>
<td>U (\rightarrow) 1 (\rightarrow) 2 (\rightarrow) N</td>
<td>21.9</td>
</tr>
<tr>
<td></td>
<td>U (\rightarrow) 1 (\rightarrow) 2 (\rightarrow) 3 (\rightarrow) 4 (\rightarrow) N</td>
<td>17.7</td>
</tr>
<tr>
<td></td>
<td>U (\rightarrow) 2 (\rightarrow) N</td>
<td>13.5</td>
</tr>
<tr>
<td></td>
<td>U (\rightarrow) 1 (\rightarrow) 2 (\rightarrow) 6 (\rightarrow) N</td>
<td>9.4</td>
</tr>
<tr>
<td></td>
<td>U (\rightarrow) 1 (\rightarrow) 2 (\rightarrow) 4 (\rightarrow) N</td>
<td>9.4</td>
</tr>
<tr>
<td></td>
<td>U (\rightarrow) 2 (\rightarrow) 6 (\rightarrow) N</td>
<td>7.3</td>
</tr>
<tr>
<td></td>
<td>U (\rightarrow) 2 (\rightarrow) 4 (\rightarrow) N</td>
<td>4.2</td>
</tr>
<tr>
<td></td>
<td>U (\rightarrow) 2 (\rightarrow) 3 (\rightarrow) 4 (\rightarrow) N</td>
<td>4.2</td>
</tr>
<tr>
<td></td>
<td>U (\rightarrow) 1 (\rightarrow) N</td>
<td>3.1</td>
</tr>
<tr>
<td></td>
<td>U (\rightarrow) N</td>
<td>2.1</td>
</tr>
</tbody>
</table>
Fig. S1. Cooperative folding/unfolding of protein L (a small protein with 62 residues), by AICG2. (A) Representative trajectory of the folding/unfolding transitions at folding temperature. (B) Heat capacity as a function of temperature. (C) Free energy as a function of $Q$ score. (D) Average of $Q$ score as a function of temperature. These results demonstrate that the AICG2 is able to reproduce the cooperative folding/unfolding transitions for small protein. In calculating the heat capacity, the free energy and the average of $Q$ score, the data from the simulations by unbiased MD at different temperatures were used, and reweighted by the weighted histogram analysis method (WHAM).
Fig. S2. (A, B) Three dimensional structure of protein G (A) and protein L (B); (C, D) Free energy landscapes projected onto the $Q$ scores of the N-terminal $\beta$-hairpin and C-terminal $\beta$-hairpin for protein G (C) and protein L (D) with AICG2 at folding temperature. The red dash lines with arrows are plotted to guide the eyes; (E, F) Free energy landscapes projected onto the $Q$ scores of the N-terminal $\beta$-hairpin and C-terminal $\beta$-hairpin for protein G (E) and protein L (F) with simple Go model at folding temperature. Although the two proteins have similar symmetric topologies, experimental data suggested that the C-terminal hairpin folds before the N-terminal hairpin for protein G (31). On the contrary, the N-terminal hairpin folds before the C-terminal hairpin for protein L (32). Then simple Go model gives almost symmetric folding pathways for both proteins. In comparison, such asymmetry of the folding pathways can be well reproduced by the AICG2.
Fig. S3. Free energy profiles along $Q_{tot}$ at temperatures $T=0.98T_F, 0.99T_F, 1.00T_F, 1.01T_F, \text{ and } 1.02T_F$. We separated the sampled structural data into two halves, and calculated the free energy profiles by using the first half (solid lines) and second half (dashed lines), respectively. We can see that the free energy profiles calculated based on the separate data sets are very similar. Only around the transition region, the free energy profiles show slight differences. Particularly, the shoulders labelled by the arrows can be identified in the energy profiles from both data sets. Such results suggest that the sampling is reasonably converged.
Fig. S4. Kinetic free energy landscape projected onto the reaction coordinates RMSD and $Q_{\text{CORE}}$ constructed based on 200 separate folding simulations with AICG2 at $0.90 T_f$ starting from fully unfolded structures. The unit of free energy is $k_B T$. Similar to the kinetic free energy landscape projected onto the reaction coordinates $Q_{\text{tot}}$ and $Q_{\text{CORE}}$ given in Fig. 1D, the sub-states can be clearly identified except that the sub-state 2 is overlapped with the edge region of the basin 1(D). The qualitative similarity between the free energy profiles on the reaction coordinates (RMSD, $Q_{\text{CORE}}$) and ($Q_{\text{tot}}$ and $Q_{\text{CORE}}$) suggests that the sub-states identified in this work are robust. We noticed that the sub-state 2 is well separated from the basin 1(D) when $Q_{\text{tot}}$ is used as reaction coordinate instead of RMSD. Therefore, the $Q_{\text{tot}}$ is superior to RMSD in characterizing the structure features around the denatured state for this protein.
Fig. S5. Kinetic free energy landscape projected onto the reaction coordinates $Q_{\text{tot}}$ and $Q_{\text{CORE}}$ constructed based on 200 separate folding simulations with AICG2 (A) and simple Go model (B) at 0.90$T_F$ starting from fully unfolded structures. The unit of free energy is $k_B T$. In addition to the fully unfolded and native states, seven other sub-states can be identified in the kinetic free energy landscape by AICG2. In comparison, the kinetic free energy landscape by simple Go model was much smoother, suggesting that introducing the sequence specificity and realistic flexibility to the local potential are responsible for making the sub-states in AKE.
Fig. S6. Contact probability maps for all the sub-states of AKE folding shown in Fig. 1D based on 200 separate folding simulations with AICG2. The contact probability is defined as the probability for each pair of residues to form the native contacts among the structures within the sub-states. The structures near the centers of the sub-states are used for the calculations of the contact probabilities. For comparison, the contact probability maps for the unfolded state and the native state are also shown. Color code: red, unstructured; blue: fully structured. From this figure, we can clearly assign the structure features for each of the sub-states, which are summarized below. The LID domain and NMP domain are labeled by dark blue and purple arrows at x- and y-axis of the left-top panel.
state U: The three domains are fully unfolded.

sub-state 1(D): The LID domain is folded, and other parts are unfolded. This sub-state may correspond to the denatured state.

sub-state 2: Both the LID domain and the NMP domain are folded. In the CORE domain, the contacts adjacent to the NMP domain are also formed to some extent.

sub-state 3: The LID domain is folded, but the NMP domain is unfolded. In the CORE domain, the contacts adjacent to the LID domain are also partly formed.

sub-state 4: Both the LID domain and the NMP domain are folded. In the CORE domain, the contacts adjacent to the LID domain or the NMP domain are also formed.

sub-state 5: The LID domain is mostly unfolded, and the NMP domain is folded. In the CORE domain, the region adjacent to the LID domain is partly unfolded, and other regions are mostly folded.

sub-state 6: Both the LID domain and the NMP domain are unfolded. The CORE domain is mostly folded.

sub-state 7: The LID domain is mostly unfolded. Other parts of the protein are well folded.

State N: The three domains are fully folded.
**Fig. S7.** Distributions of the distances between the C$_\alpha$s of the residues 73 and 203, which correspond to the label sites in the single molecule experiment, for each of sub-states at the temperatures of $0.90T_F$ (A), $0.92T_F$ (B), and $0.94T_F$ (C) based on the 200 separate folding simulations with AICG2. The results for the unfolded state and the native state are also shown for comparison.
Fig. S8. Branch ratio of the transitions from one initial state to all the final states of AKE folding at the temperatures of $0.90T_F$ (A), $0.92T_F$ (B), and $0.94T_F$ (C) based on 200 separate folding simulations starting from fully unfolded structures with AICG2. The branch ratio is defined as the transition probability from the initial state to the different final states. For each initial state, the transition probability is normalized. Color code: light gray, low branch ratio; black, high branch ratio.
Fig. S9. (A) Two-dimensional free energy landscapes on the reaction coordinates $Q_{\text{tot}}$ and $Q_{\text{LID}}$ at folding temperature based on modified multi-canonical sampling and WHAM reweighting with AICG2. (B) Two-dimensional kinetic free energy landscape on the reaction coordinates $Q_{\text{tot}}$ and $Q_{\text{LID}}$ constructed based on 200 separate folding simulations with AICG2 at $0.90T_F$ starting from fully unfolded structures. The unit of free energy is $k_B T$. Both the equilibrium free energy landscape and the kinetic free energy landscape show that the LID domain can fold and unfold in both the folded (large $Q_{\text{tot}}$) and the unfold (small $Q_{\text{tot}}$) basins, suggesting that the coupling between the folding of the LID domain and other domains is relatively weak.
Fig. S10. One-dimensional free energy landscape on the reaction coordinate $Q_{\text{LID}}$ constructed based on the folding simulations with AICG2 at the temperatures of $0.90T_F$, $0.92T_F$, and $0.94T_F$. At each temperature, 200 separate folding simulations with the length of $2 \times 10^7$ MD steps were conducted starting from fully unfolded structures. The unit of free energy is $k_BT$. As shown in Fig. 1C, even at the temperature of $0.90T_F$, the LID domain can frequently hop between the folded and unfolded state. Therefore, the free energy shown in this figure is equilibrium free energy, and the depths of the free energy basins can be reliably used to estimate the relative stability of the LID domain. One can see that with the increasing of temperatures from $0.90T_F$ to $0.94T_F$, the LID domain becomes less stable.
Fig. S11. (A, B) Native structure of 2out-knot and its schematic representation; (C, D) Highly trapped structure of 2out-knot folding with simple Go model simulations and its schematic representation. The native structure is featured by a trefoil knot with the segments H7-H8 crossing through the H2-H3-H4-L1 loop and the segment H1 crossing though the H4-L1-H5-H6 loop. The N-terminal depth (defined as the sequence length from N-terminus to knot) and the C-terminal depth (defined as the sequence length from C-terminus to knot) of the trefoil knot are 28 and 57, respectively. From Fig. 3F of the main text, the maximal ratio of successful folding is around 33% with the simple Go model simulations. In comparison, with the AICG2, the maximal ratio of successful folding can be as high as 96%. It is interesting to investigate the features of the trapped structures in the unsuccessful trajectories with simple Go model simulations. We found that majority of the unsuccessful trajectories with simple Go model simulations are trapped in the structure shown in (C) and (D), in which the segment H6-H7 is misplaced around the H3 as labelled by the red arrow, without forming knot. For example, at the temperature with the maximal ratio of successful folding, around 95% of the unsuccessful trajectories are trapped in this structure. Apparently, knotting from this trapped structure requires the detachment of H8 from the H2 and H5, and the threading of the segment H7-H8 through the H2-H3-H4-L1 loop, which is difficult. On the contrary, all the successful folding events form knot by threading the segment H1 though the H4-L1-H5-H6 loop, which is much easier since the threaded segment H1 is shorter than the segment H7-H8. The high ratio of successful folding with AICG2 simulations shown in Fig. 3F suggests that the AICG2 can largely include the necessary interaction features to eliminate such highly trapped structure (See Fig. 3F and Fig. S12, and the related discussions).
Fig. S12. (A) Contact energies of the nonlocal native contacts calculated by the energy decomposition introduced in the SI text. In panel (A), the contact energies were subtracted by their average. Therefore, the negative contact energies (blue) represent that the interactions are stronger than the average, and positive contact energies (red) represent that the interactions are weaker than the average. (B) The ratios of the successful folding of the AICG2 simulations (black, AICG2) and those of a control simulation by replacing the interaction strength of the nonlocal native contacts by the average (red, homo-nloc). From Panel (A), we can see that distribution of the contact energies can be highly heterogeneous. Although it is not easy to directly map the high ratio of successful folding of the AICG2 to certain specific contacts, both panel (B) and the Fig. 3F suggest that the sequence information contained in both the nonlocal native contacts and the local contacts largely contribute to the successful folding of the 2out-knot.
**Fig. S13.** Two-dimensional free energy landscapes on the reaction coordinates \((Q_{\text{tot}}, Q_{H1-H4})\), \((Q_{\text{tot}}, Q_{H5-H8})\), \((Q_{\text{tot}}, Q_{H4-H6})\), \((Q_{\text{tot}}, Q_{H2-H6})\), for the 2out-knot \((A, C, E, G)\) and the 2out-ds \((B, D, F, H)\) at folding temperature with AICG2. Unbiased MD simulations were used for the equilibrium simulations, and WHAM was used in the data analysis. Here the reaction coordinates \(Q_{H1-H4}\), \(Q_{H5-H8}\), \(Q_{H4-H6}\), \(Q_{H2-H7}\), represent the fractions of the formed native contacts between the helices \((H1, H4)\), \((H5, H8)\), \((H4, H6)\), and \((H2, H6)\),
respectively. The unit of the free energy is $k_B T$. Compared to the 2out-ds, the folding of the 2out-knot encounters more frustration due to the knotting. For example, the preformed contacts between the H1 and H4, and between the H5 and H8 need to be broken during the folding since these contacts can hinder the efficient knotting. In comparison, the free energy landscapes of 2out-ds are much smoother, and the folding encounters less frustration. The free energy landscapes on the reaction coordinates ($Q_{\text{tot}}$, $Q_{\text{H4-H6}}$) also have large difference. For the 2out-knot, the H4-H6 contacts predominantly form in the early stage of the folding. Such early formation of the H4-H6 is crucial for the knotting since it corresponds to the loop formation for the knotting. Whereas for the 2out-ds, the H4-H6 contacts can be formed in a later stage with certain probability. In both proteins, the H2-H6 contacts are formed in the early stage of the folding ($G$, $H$). However, the contacts between the H2 and the H6 keep partly formed even in the fully unfolded state for the 2out-ds, due to the restraint arising from the disulfide bond.
Fig. S14. Two-dimensional free energy landscapes on the reaction coordinates ($Q_{H1-H4}$, $Q_{H5-H8}$) for the 2out-knot (A) and the 2out-ds (B) at folding temperature with AICG2. The unit of the free energy is $k_B T$. From the crystal structure (Fig. 3), the H1, H4 helices and the H5, H8 helices correspond to the terminal helices of the two monomers of the 2out-ds. As shown by the symmetric free energy landscape on the reaction coordinates ($Q_{H1-H4}$, $Q_{H5-H8}$) in (B), the folding behaviours of the two monomers in 2out-ds are quite similar, due to their symmetric sequences and tertiary structures. In comparison, the corresponding parts in the 2out-knot show different folding behaviours due to the presence of the knot. The folding pathway with the H1-H4 contacts forming before the H5-H8 contacts is much more probable. Such asymmetry of the folding pathway is crucial for the efficient knotting of the 2out-knot, as discussed in the main text.
Fig. S15. Contact probability maps for all the sub-states of 2out-knot folding shown in Fig. 5C based on 100 separate folding simulations with AICG2. The contact probability is defined as the probability for each pair of residues to form the native contacts among the structures within the sub-states. The structures near the centers of the sub-states are used for the calculations of the contact probabilities. For comparison, the contact probability maps for the unfolded state and the native state are also shown. Color code: red, unstructured; blue: fully structured. In the contact probability map for the native state (state N), some representative inter-segment contacts are circled and labelled. The structure features for each of the sub-states are summarized below.

**state U:** All the inter-segment contacts are broken.

**sub-state 1:** The knotting loop consisting of H4, L1, H5, and H6 is formed by H4-H6 or H3-H6 contacts.
**sub-state 2:** In addition to the H3 and H4, the H2 is also packed to the H6, which correctly aligns the H2 to adopt an orientation helpful for knotting.

**sub-state 3:** Compared to the sub-state 2, the H1-H4 contacts are also formed. However, the contact probability between the H2 and L1 is very low, indicating that the H1 and H2 are not correctly passed through the H4-L1-H5-H6 loop, or the L1 and H5 are not correctly placed. In this sub-state, the chain is unknotted. When the H1 and H2 incorrectly passed through the H4-L1-H5-H6 loop, the H1-H4 contacts have to be broken in the later stage of the folding, which represents topological frustration arising from the knot structure.

**sub-state 4:** Different from the sub-state 3, in this sub-state, the chain is correctly knotted as shown by the high contact probability between the H1 and H4, and between the H2 and L1. In this sub-state, the H7 and H8 are not yet packed to the final positions.

**sub-state 5:** Compared to the sub-state 4, the H7 and H8 are also packed to the final positions. However, in this sub-state, the H6 and H7 are not correctly placed around H3 as shown by the low contact probability between the H3 and H6, therefore the knot is absent. For the protein 2out-knot with a trefoil knot, there are two ways to make the knot; i) threading H1 through the H4-L1-H5-H6 loop, ii) threading H7 and H8 through the H2-H3-H4-L1 loop. The second way, which involves the sub-state 5, is much more difficult due to the much longer length of the H7 and H8 to be threaded. In fact, no folding route leading to the folded state via the sub-state 5 is observed in the current simulation work, suggesting that the sub-state 5 corresponds to a highly trapped sub-state.

**sub-state 6:** Compared to the sub-state 5, the H7 and H8 are well packed, while the H1 is not correctly threaded through L1, as shown by the low contact probability between the H2 and L1. Therefore the knot is not formed in this sub-state. Since the H1-H4 contacts can largely form in this sub-state, they need to be broken before the further folding.

**state N:** The protein is fully folded correctly.
Fig. S16. Branch ratio of the transitions from one initial state to all the final states of 2out-knot folding at the temperatures of $0.90T_f$ (A), $0.925T_f$ (B), and $0.95T_f$ (C) based on 100 separate folding simulations starting from fully unfolded structures by AICG2. The branch ratio is defined as the transition probability from the initial state to the different final states. For each initial state, the transition probability is normalized. Color code: light gray, low branch ratio; black, high branch ratio.
Fig. S17. (A, B) Native structure of MJ0366 (PDB code: 2efv) and its schematic representation; (C) Free energy profiles along $Q_{\text{tot}}$ at temperatures $T=0.98T_F$, $1.00T_F$, and $1.02T_F$. The unit of free energy is $k_B T$; (D) Ratios of the successful folding of the AICG2 simulations (black, AICG2) and those of the simple Go model simulations (red, Go). The MJ0366 is a nature protein with trefoil knot. It has 82 residues. The N-terminal depth and C-terminal depth of the trefoil knot are 15 and 12, respectively. One can see that the maximal ratio of successful folding with AICG2 (99%) is much higher than that with simple Go model simulations (50%). Recent theoretical studies showed that the folding behaviors for the natural proteins and designed proteins can be very different (33). The high ratio of successful folding for the natural protein MJ0366 suggests that the success of AICG2 is not limited to designed protein.
Fig. S18. Two-dimensional free energy landscapes on the reaction coordinates ($Q_{\text{tot}}$, $Q_{\text{B1-B2}}$) and ($Q_{\text{tot}}$, $Q_{\text{H2-H4}}$) for the MJ0366 at folding temperature with AICG2. Unbiased MD simulations were used for the equilibrium simulations, and WHAM was used in the data analysis. Here the reaction coordinates $Q_{\text{B1-B2}}$ and $Q_{\text{H2-H4}}$ represent the fractions of the formed native contacts between the $\beta$-strands B1 and B2, and between the helices H2 and H4, respectively. The unit of the free energy is $k_B T$. One can see that the $\beta$-sheet containing the strands B1 and B2 forms prior to the transition region. The formation of the $\beta$-sheet leads to the B1-L1-H1-H2-B2 loop. The threading of the H4 through the B1-L1-H1-H2-B2 loop, which results in a trefoil knot, is the last step of the folding, as shown in panel (B). The free energy landscape in panel (B) also suggests that the contacts between H2 and H4 can form to some extent at the early stage of the folding. However, these contacts have to be broken for the successful knotting in the later stage of the folding, suggesting the backtracking mechanism. Similar folding pathway was also observed in a recent study with all atom Go model simulations (34).
Fig. S19. Native structure of YibK (PDB code: 1j85) and its schematic representation. The YibK is a nature protein with trefoil knot, and has 156 residues. The N-terminal depth and C-terminal depth of the trefoil knot are 75 and 36, respectively. We conducted the folding simulations for the YibK at wide temperature range with both the simple Go model and the AICG2. However, it is very difficult to fold correctly with both the simple Go model and AICG2. For example, with the simple Go model, we only observed one successful folding event among the 1000 trials at the temperature of ~0.85\(T_F\) (Here, \(T_F\) is obtained by equilibrium simulations starting from unfolded structure. Since the state with high faction of native contacts is not the folded state due to the lack of knot, the \(T_F\) is not real folding temperature.), namely, the maximal ratio of successful folding is ~0.1%. With the AICG2, the maximal ratio of successful folding is ~0.2% at the temperature of 0.93\(T_F\). Such very low ratio of successful folding, i.e., ~0.1-0.2%, is consistent with the previous studies by using the simple Go model (35, 36). In Ref. (35), Onuchic et al studied the folding of the YibK and YbeA, and showed that the ratio of successful trajectories is very small (~0.1%). After selectively enhancing certain contact interactions, the successful ratio can be improved, but is still very low (~1-2%). Compared to the protein 2out-knot and MJ0366, the knot of the protein YibK is much deeper. Here, the similarly low ratios of successful folding by using simple Go model and AICG2 suggest that more realistic interaction features, e.g., non-native interactions, need to be included in the model in order to achieve efficient folding for the protein with deeper knot. Actually, in Ref. (36), Shakhnovich et al simulated the folding of the same proteins, Yibk and YbeA, using the simple Go model and suggested that addition of specific non-native interactions can greatly facilitate knot formation and thus successful folding.
**Movie Legends**

**Movie S1** Movie illustrating a folding trajectory of the protein 2out-knot by “concerted threading” knotting mechanism. The trajectory shown in this movie is the same as that in Fig. 5A, and it follows the folding route of U→1→2→3→4→N. The MD step is shown at the right bottom of the screen with the unit of 100.

**Movie S2** Movie illustrating a folding trajectory of the protein 2out-knot by “slipknot intermediate” knotting mechanism. The trajectory shown in this movie is the same as that in Fig. 5B, and it follows the folding route of U→1→2→6→N. The MD step is shown at the right bottom of the screen with the unit of 100.
**Supporting Information**

Li et al. 10.1073/pnas.1201807109

**Movie S1.** Movie illustrating a folding trajectory of the protein 2out-knot by “concerted threading” knotting mechanism. The trajectory shown in this movie is the same as that in Fig. 5A, and it follows the folding route of U → 1 → 2 → 3 → 4 → N. The MD step is shown at the right bottom of the screen with the unit of 100.

**Movie S1 (AVI)**

**Movie S2.** Movie illustrating a folding trajectory of the protein 2out-knot by “slipknot intermediate” knotting mechanism. The trajectory shown in this movie is the same as that in Fig. 5B, and it follows the folding route of U → 1 → 2 → 6 → N. The MD step is shown at the right bottom of the screen with the unit of 100.

**Movie S2 (AVI)**