JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Article

pubs.acs.org/JACS

Multidomain Protein Solves the Folding Problem by Multifunnel Combined Landscape: Theoretical Investigation of a Y-Family DNA Polymerase

Yong Wang,[†] Xiakun Chu,[‡] Zucai Suo,[§] Erkang Wang,[†] and Jin Wang*[,†,‡,∥]

[†]State Key Laboratory of Electroanalytical Chemistry, Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun, Jilin, 130022, P.R. China

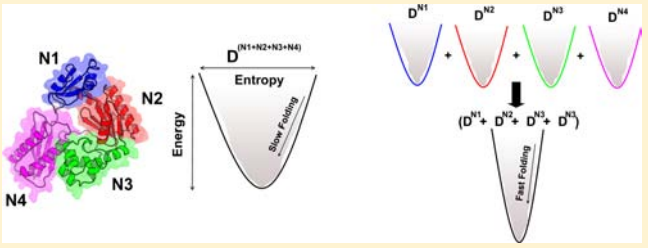
[‡]College of Physics, Jilin University, Changchun, Jilin, P.R. China

[§]Department of Biochemistry, Ohio State University, Columbus, Ohio 43210, United States

[∥]Department of Chemistry, Physics and Applied Mathematics, State University of New York at Stony Brook, Stony Brook, New York 11794-3400, United States

Supporting Information

ABSTRACT: Approximately three-fourths of eukaryotic proteins are composed of multiple independently folded domains. However, much of our understanding is based on single domain proteins or isolated domains whose studies directly lead to well-known energy landscape theory in which proteins fold by navigating through a funneled energy landscape toward native structure ensembles. The degrees of freedom for proteins with multiple domains are many orders of magnitude larger than that for single domain proteins. Now, the question arises: How do the multidomain proteins solve the "protein folding problem"? Here, we specifically address this issue by exploring the structure folding relationship of *Sulfolobus solfataricus* DNA polymerase IV (DPO4), a prototype Y-family DNA polymerase which contains a polymerase core consisting of a palm (P domain), a finger (F domain), and a thumb domain (T domain) in addition to a little finger domain (LF domain). The theoretical results are in good agreement with the experimental data and lead to several theoretical predictions. Finally, we propose that for rapid folding into well-defined conformations which carry out the biological functions, four-domain DPO4 employs a divide-and-conquer strategy, that is, combining multiple individual folding funnels into a single funnel (domains fold independently and then coalesce). In this way, the degrees of freedom for multidomain proteins are polynomial rather than exponential, and the conformational search process can be reduced effectively from a large to a smaller time scale.



INTRODUCTION

Computational and bioinformatics analysis has demonstrated that approximately two-fifths to two-thirds of prokaryotic proteins are composed of more than one domain, and the proportion can be up to 80% in eukaryotic proteins.[1,2] Further investigation revealed that about 95% of multidomain proteins contain no more than 5 domains. But, for a few proteins, the domain counts can reach up to 300.[1] Despite the fact that proteins from all three kingdoms of life predominantly fold into multidomain conformations, most protein folding studies focus on individual domains. In other words, our current knowledge of protein folding is predominantly based on studies of small or single domain proteins.[3] This raises an important question of whether the folding principle derived from single domain proteins can be extended for larger multidomain proteins and, if so, to what degree.[4] Further, we may ask whether it is true that the presence of neighboring domains has slight effects on protein folding properties[5] and whether the energy landscape of multidomain proteins is more complex than that of individual domains,[6,7] or are there more metastable intermediate states along the transition pathway in multidomain proteins?[7] Given the fact that the degrees of freedom in multidomain proteins are many orders of magnitude higher than those in single domain proteins, we may ask how do multidomain proteins solve the "protein folding problem"?[8]

Although there are abundant data available on protein folding, both experimentally and theoretically, detailed studies have only been performed with a few proteins consisting of two or three domains.[7,9−17] In contrast to typical small globular single domain proteins which were often described by simple two-state model, these multidomain proteins have been suggested to have more complex folding landscape by using state-of-the-art single molecule techniques.[6,7] Due to the fact that each domain is in close association with its neighbor in multidomain protein by interdomain interactions, it is possible that a domain folds as an isolated protein is different from its folding as part of a multidomain protein. Previous experimental

Received: May 10, 2012
Published: July 24, 2012

   dx.doi.org/10.1021/ja3045663 | J. Am. Chem. Soc. 2012, 134, 13755−13764

investigations of the folding of a protein with tandem spectrin domains have suggested that domain interfaces can significantly affect stability, folding, and unfolding rates but have trifling impact on folding pathways.[18,19] Shank et al. investigated the unfolding of T4 lysozyme with two domains by optical tweezer and found that the protein topology critically determines the folding cooperativity and communication between domains.[15] A recent theoretical study on three two-domain protein systems has shown that the native topology has a determinant role in the folding and unfolding process, and the other factors, including the domain connectivity as well as the interfacial interactions, are also important.[20] These experimental and theoretical studies suggest that multidomain proteins are not the simple addition of each single domains and that their folding is quite complicated.

In the present work, we explore the folding of DNA polymerase IV (DPO4) from *Sulfolobus solfataricus* which contains a polymerase core consisting of a palm (P), finger (F) and thumb (T) domain in addition to a fourth domain known as a little finger (LF) domain, as shown in Figure 1. We studied

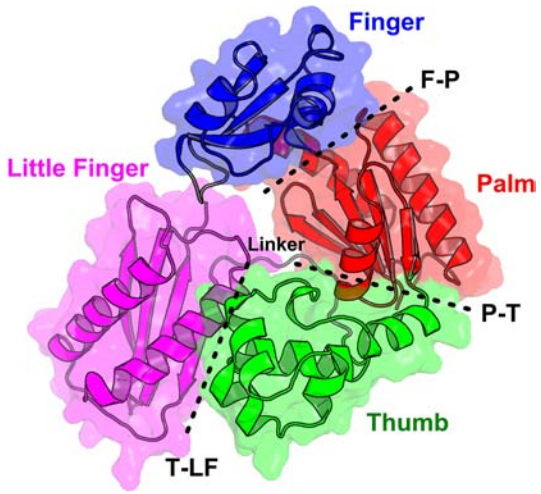


Figure 1. Structural illustration of DPO4. DPO4 has polymerase core consisting of a palm (P), finger (F), and thumb (T) domain in addition to a fourth domain known as a little finger (LF) domain. The LF domain is physically located next to the F domain and does not interact with the T domain, although it is tethered to the T domain by a 14-residue linker. The interfaces between these domains are labeled F−P, P−T and T−LF, respectively.

DPO4 in this work not only because it is an excellent model system as a multiple domain protein but also due to its important biological role as the most thoroughly studied member of the Y-family DNA polymerases.[21] The Y-family DNA polymerases are well-known as a class of low-fidelity DNA synthesis enzymes. Cellular DNA is frequently attacked by numerous DNA-damaging agents, leading to the formation of a myriad of DNA lesions. Unrepaired DNA lesions stall replicative A- or B-family DNA polymerases in cellular DNA replication machinery. The inability to replicate genomic material will stop cell cycles and cause cell death.[22] To rescue genomic replication, cells usually employ the Y-family DNA polymerases to bypass the unrepaired DNA lesions. For example, DPO4, the lone Y-family DNA polymerase in *S. solfataricus* likely catalyzes translesion DNA synthesis and helps this organism to survive in tough environmental conditions (80 °C and pH 2−3).[23] Although the Y-family enzymes share little sequence identity with the DNA polymerases from other families, all DNA polymerases share a structurally conserved right-handed polymerase core consisting of three domains: F, T, and P. The LF domain is unique to the Y-family DNA polymerases and has been suggested to contribute to DNA binding affinity.[24]

Here, we performed thermodynamic as well as kinetic simulations of folding/unfolding for DPO4, using the combination of a native structure-based model ("traditional SBM") and its sequence-flavored variant ("sequence-flavored SBM") based on the globally funnel-like landscape view. We will compare the results of the traditional SBM and that of the sequence-flavored model. This allows us to factor out the key finding from the simulation data and test the robustness of the different simulation models as well as investigate the sequence dependent in the folding of multidomain proteins. To the best of our knowledge, our present work provides the first report of multidomain protein folding in the context of more than three globular domains. We show that the simplified unfrustrated coarse-grained models capture well the complex folding mechanism of the four-domain DPO4. Remarkably, the results from thermodynamic and kinetic simulations are in good agreement with the unfolding experimental measurements of DPO4 (see ref 25), as evidenced from the following aspects: (1) the existence of unfolding intermediate state(s); (2) the superior stability of the LF domain than the polymerase core of DPO4; (3) the high flexibility of the T−LF interface; (4) the irreversibility of folding/unfolding process; and (5) the underlying relationship between the flexibility of the T−LF interface and the remaining polymerase activity of DPO4 at high temperature (lower than melting temperature). Interestingly, we find that DPO4 folds in a similar manner as cotranslational folding in the cell,[13,26] that is, folding one domain at a time rather than folding multiple domains as a whole. This divide-and-conquer strategy will lead to more efficient folding.

RESULTS

Identification of Intermediates. We first examined the results from simulations using a traditional SBM. These simulations were carried out at folding temperature T_f defined by the peak of the heat capacity curve (Figure S1).[27] At folding temperature, we can observe the whole process of the unfolding transition through multiple intermediate and transition states by long time simulations. Two typical unfolding trajectories are shown in Figure 2A,B. These trajectories clearly reveal the existence of multiple intermediate states which are also observed in the free energy profiles as a function of native contacts at a variety of temperatures from 0.90 to 1.05 T_f (Figure S2). Here, the free energy profile at folding temperature is shown in Figure 2C. Native state is located at the basin with ≈900 native contacts. Totally unfolded state is at the basin with ∼200 native contacts. These intermediates are denoted as I_1, I_2, I_3, and I_4, and native and unfolded states are N and U, respectively. We can see that the free energy barriers between intermediates are no more than $2k_BT$, which allows them to be overcome easily by thermal fluctuation. The low-energy barriers also imply that the intermediate states are relatively unstable.

The multiple intermediate states are also observed from simulations using a sequence-flavored model (Figure S4). The heat capacity curve of the sequence-flavored model shows three peaks (Figure S1), indicating multiple phase transitions. It also

dx.doi.org/10.1021/ja3045663 | J. Am. Chem. Soc. 2012, 134, 13755−13764

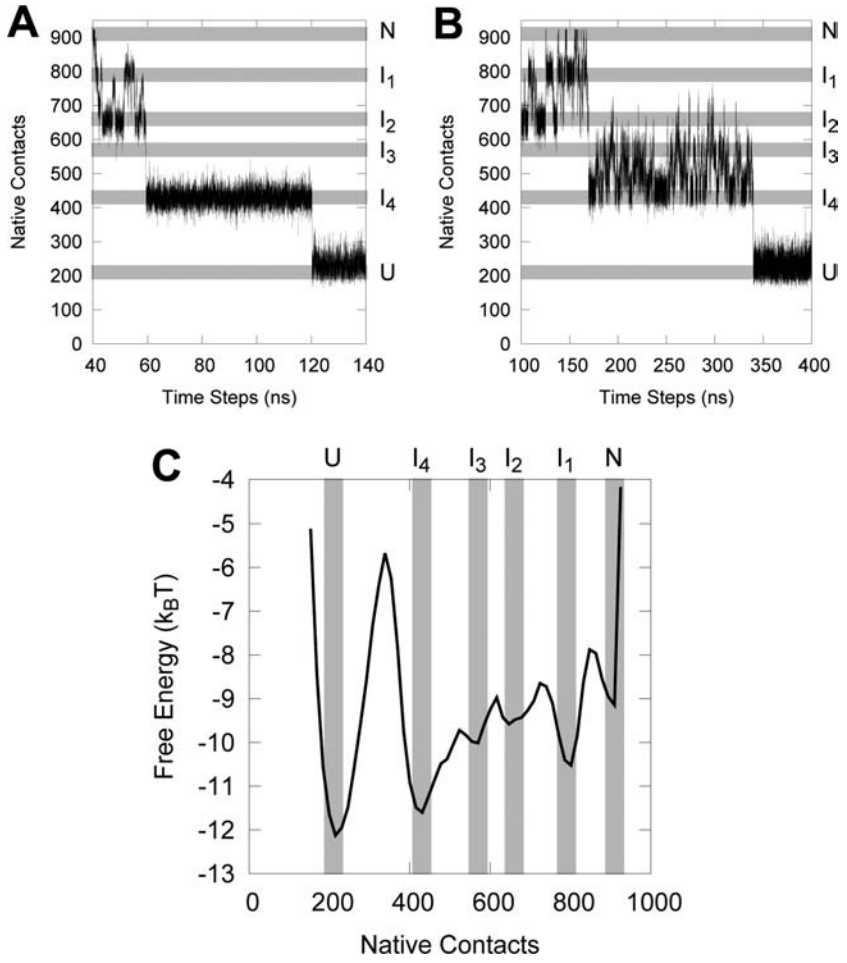


Figure 2. Multiple intermediate states. Two typical kinetic trajectories of DPO4 unfolding are shown in (A,B). The free energy profiles as a function of native contacts were calculated under a variety of temperatures from 0.90 to 1.05 T_f. (C) The free energy profile at T_f is shown. Native state (labeled by N) is located at the basin with ≈900 native contacts. Totally unfolded state (labeled by U) is at the basin with about 200 native contacts. The four intermediates are labeled I_1, I_2, I_3, and I_4 along the unfolding pathway. The free energy barriers for transitions between intermediate or transition states are low ($1-2k_BT$) during a wide temperature range.

indicates that the folding landscape of DPO4 becomes rougher by the introduction of energetic heterogeneity. Although the free energy surfaces from the two models are not in full agreement, both share a similar landscape with multiple intermediate states which have low barriers between each state. Experimentally, circular dichroism (CD) spectroscopy exhibited a three-state cooperative unfolding profile, with an unfolding intermediate clearly in existence between the native and denatured states.[25] Next, we characterized the conformational ensembles of the two intermediate states I_1 and I_4 near folding or unfolding states.

The contact maps and the corresponding typical structures of I_1 and I_4 are shown in Figure 3 (I_2 and I_3 in Figure S5). For I_1, the T−LF interface is fully broken, and the F domain is partially unfolded, while other regions of DPO4 remain folded. This implies that T−LF interface is more dynamic than the other two interfaces (F−P and P−T), and F domain exhibits more fluctuation than the other three domains. For I_4, the whole protein is unfolded except for the P domain, indicating the high stability of this domain which contains conserved catalytic carboxylates. However, we should keep in mind that these intermediates determined by 1D free energy profile are the average over multiple parallel pathways which are revealed by the following 2D surfaces.

Stability of Domains and Domain Interfaces. The relative stability of domain interfaces is further characterized by 1D free energy profiles as a function of Q scores (see definition in Materials and Methods) at lower temperature $T = 0.95\ T_f$. At this temperature, DPO4 not only avoids totally unfolding but also exhibits certain flexibility, allowing us to measure the local stability of the protein. Q score has been suggested to be a good reaction coordinate in the description of protein folding.[28] Note that for $Q = 1$, the conformation is native folded, and for $Q = 0$, it is in totally unfolded state. Figure 4A indicates that the T−LF interface is more flexible than that of the F−P interface, as evident from the wider free energy basin for Q score of T−LF interface. It also indicates that the P−T interface is more stable than other interfaces. The high flexibility of T−LF interface is consistent with the CD spectroscopy analysis and fluorescence-based thermal scanning (FTS) assay.[25] Overall, the results in simulation support the stability of domain interfaces in this order: first P−T, then F−P, and finally T−LF.

We next measure the stability of four individual domains as shown in Figure 4B. It shows that the P and LF domains stay in the folded state ($Q = 0.8$) at this temperature, but the other two domains can fluctuate to unfolded states ($Q = 0.2$). For the F domain, a significant transition between the folded and unfolded states was observed. Interestingly, X-ray crystallo-

dx.doi.org/10.1021/ja3045663 | J. Am. Chem. Soc. 2012, 134, 13755−13764

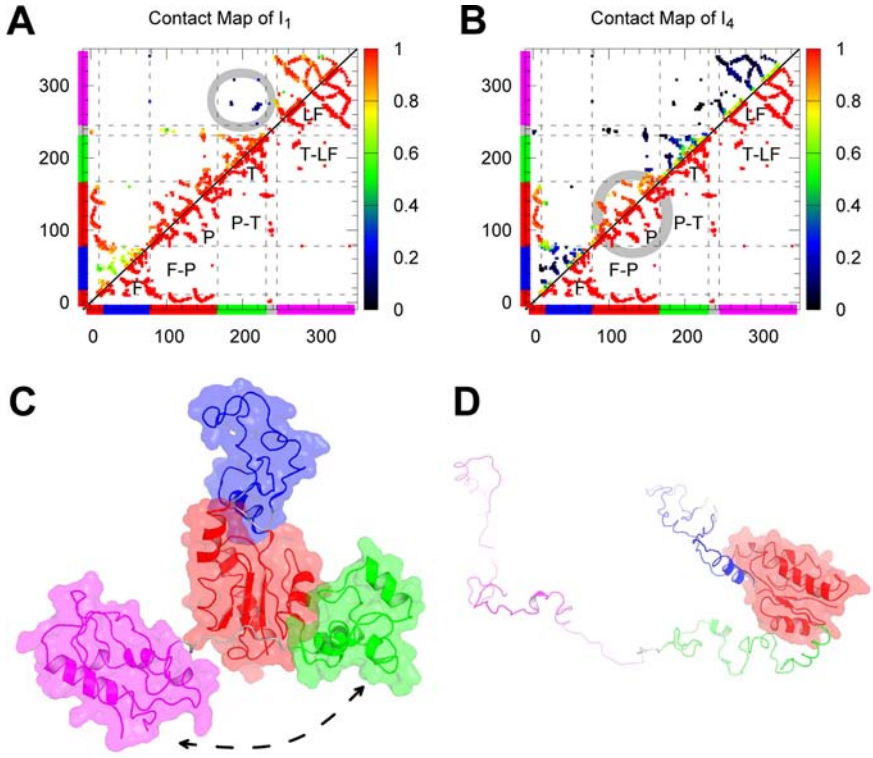


Figure 3. Contact maps of intermediate states. Above the diagonal of the contact map corresponds to the probability map of native contacts formed in an intermediate state. For better structural characterization of the intermediate state, the contact map of native structure of DPO4 is shown below the diagonal. The corresponding probability for a particular residue pair forming two-body native contact is illustrated according to side color bar in which red means high probability and blue means low probability. The regions corresponding to the F (11−77 in blue), P (1−10 and 78−166 in red), T (167−233 in green) and LF (244−341 in magenta) domains are labeled in the *X*- and *Y*-axes with different colors. Additionally, the linker region (234−243) between T and lF domains is colored in gray. Contact probability maps of I_1 and I_4 are shown in (A) and (B) with the corresponding typical structures shown in (C) and (D), respectively. For I_1, the T−LF interface is fully broken, which is highlighted by gray spherical circle, and F domain is partially unfolded. For I_4, the whole protein is unfolded except for the P domain.

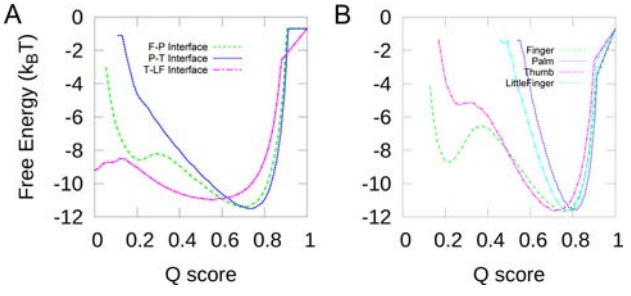


Figure 4. The stability of domains and domain interfaces derived from free energy profiles as a function of Q scores at temperature 0.95 T_f. The free energy profiles as a function of Q scores of three domain−domain interfaces are shown in (A) and of four domains in (B).

graphic studies suggested that a loop region (residues 33−40) in the F domain is disordered in the absence of DNA binding.[21] This intrinsically disordered region may contribute to the fluctuation of the F domain. Although the stability of the P domain is comparable to that of the LF domain, the LF domain is much more stable than the other two domains in the polymerase core. This leads to that the overall stability of the LF domain is higher than the polymerase core, consisting with the experimental observation that the LF domain is the most stable domain during the thermal denaturation process reflected by the apparent T_m (the melting temperature) values and the FTS analysis.[25]

To sample the entire accessible conformational space, including native and totally unfolded states, and the intermediate and transition states, we further performed long time equilibrium thermodynamic simulations at folding temperature T_f. The 2D free energy profiles are shown in Figure 5. Clearly, it can be seen that the unfolding of F and T domains is prior to that of LF and P domains, as labeled by white arrows in free energy surfaces.

However, the relative folding/unfolding order between F and T domains is not obvious from the thermodynamic landscape analysis as well as between the P and LF domains, as shown in Figure 5B,E. To lend support to the mechanism determined by the equilibrium simulations at folding temperature, we further performed 1600 independent kinetic folding simulations, either starting from a random coil unfolded structure under lower temperature (0.90 and 0.95 T_f) or initiating from the native folded structure with different initial velocities at higher temperature (1.02 and 1.05 T_f). Note that in simulation, temperature was used as a control parameter to fold or denature the native state.

Unfolding Order by Kinetic Analysis. Again, the Q score was employed to monitor the folding/unfolding process of kinetic simulations. Considering that there are four domains and three corresponding domain−domain interfaces, and a linker region between T and LF domains, we artificially monitored the folding/unfolding order of eight specific regions by eight kinetic steps and calculated their corresponding probability of folding/unfolding at each step. Note that the

dx.doi.org/10.1021/ja3045663 | J. Am. Chem. Soc. 2012, 134, 13755−13764

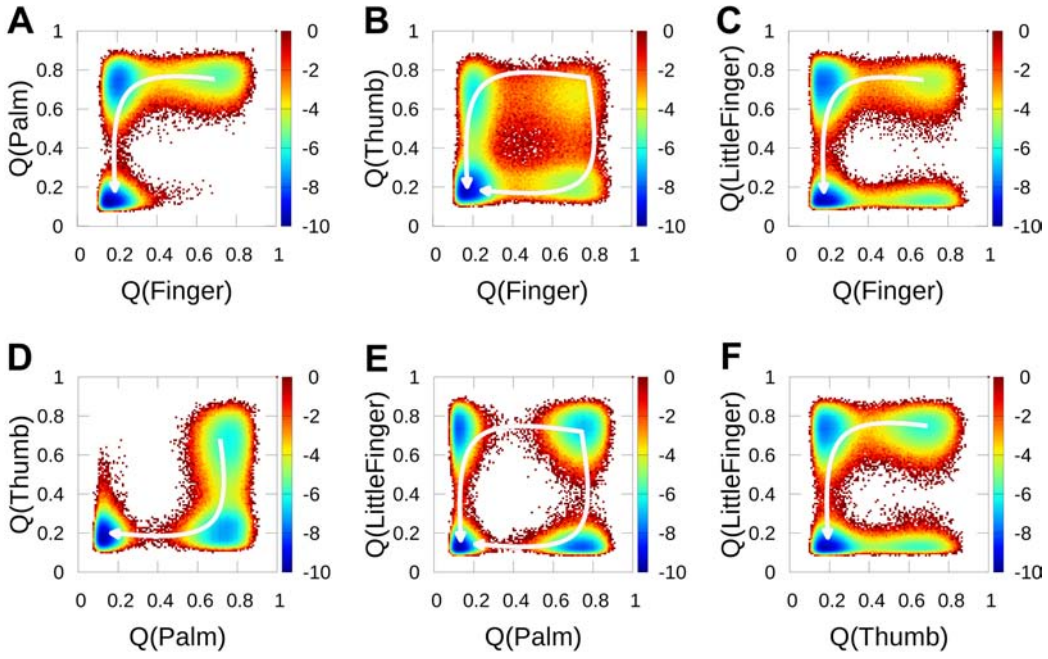


Figure 5. Sequential folding mechanism. Two-dimensional free energy profiles are plotted as a function of Q scores at T_f. Unfolding of F and T domains precedes that of LF and P domains. It is remarkable that there is no pathway along the diagonal lines in these surfaces, indicating that the four domains in DPO4 fold in a sequential order, without coupling.

number of kinetic steps used for measurement is not important to determine the relative order of these regions.

First, the 400 unfolding trajectories at T = 1.05 T_f were averaged, and the unfolding probability of different regions in DPO4 along the unfolding pathway are summarized in Table S1 and shown in Figure 6A. Another 400 unfolding trajectories at T = 1.02 T_f were computed to investigate the temperature dependence (summarized in Table S2 and Figure S6).

Clearly, our results show that the probability of T−LF interface unfolding is above 0.8 in the first step and that of F domain unfolding is above 0.5 in the second step. This indicates that the F domain and T−LF interface are unstable and unfold prior to other parts of DPO4, consistent with the contact map analysis on intermediate I_1 (Figure 3A) and the aforementioned free energy profiles (Figure 4). In addition, it also shows that LF and P unfold with high probability during the final steps, but P domain unfolds with a probability 0.74 which is higher than the probability of 0.24 of LF during the last step of unfolding. This supports the conclusions made from the contact probability map of I_4 (Figure 3B) and also free energy profiles, which suggests that both LF and P domains are thermal stable relative to F and T domains, but P has a little advantage over LF. Previous studies have suggested that the P is the most structurally conserved domain for several DNA polymerase families including A-, B-, X-, and Y-families.[24]

Moreover, we may conclude that unfolding of DPO4 begins from the domain rigid sliding from the kinetic analysis, that is, the movement of the interface between T and LF domain. As the early event of DPO4 unfolding, this stage does not involve the breaking of secondary structure and can be characterized by an opening movement of LF relative to the core of DPO4. This conclusion is also supported by the experimental analysis of thermal denaturation of the truncation fragments and point mutants of DPO4 in the linker which identifies that the flexibility stems from the linker between T and LF domains,

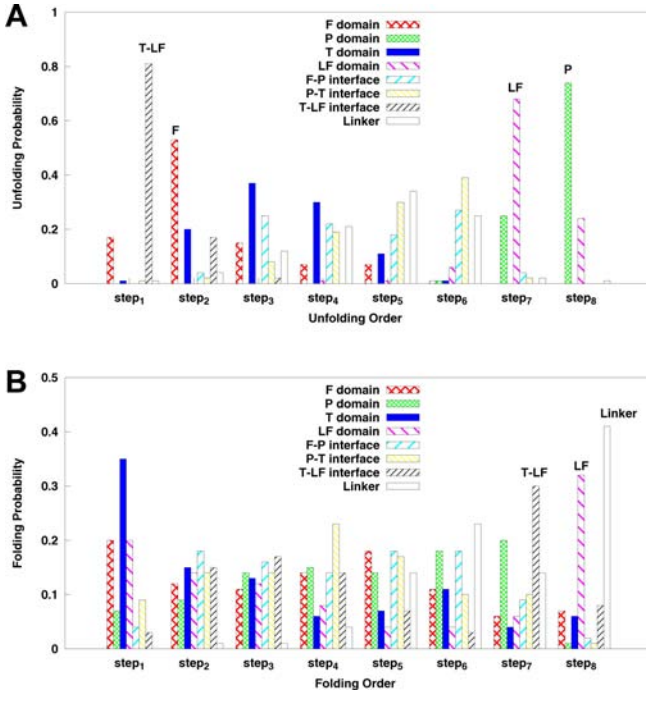


Figure 6. Unfolding order and folding order by kinetic analysis. (A) 400 unfolding trajectories were collected at T = 1.05 T_f. (B) 400 folding trajectories were collected at T = 0.95 T_f.

rather than from the independent unfolding of the four domains of DPO4.[25]

Overall, the most probable unfolding order revealed from the ensemble kinetic simulations is in a good agreement with that from thermodynamic landscape calculations.

Folding Order. The folding process was also measured by kinetic simulations performed at T = 0.95 T_f. Similarly, another 400 kinetic trajectories at 0.9 T_f (Figure S7) were collected for

dx.doi.org/10.1021/ja3045663 | J. Am. Chem. Soc. 2012, 134, 13755−13764

a robustness test. Interestingly, the folding probability distribution is wide (with the highest probability less than 0.5) by comparison with the unfolding probability distribution. This indicates that the folding pathway is dispersive in the energy landscape and that the dominant pathway is not obvious. The folding probability distribution, in Figure 6B and Table S3, shows that LF and its interface as well as the linker fold in the last step with a relative high probability (>0.30). This means that for the folding process of DPO4, the most probable pathway is that the core domain assembles as a whole, subsequently LF folds with coupled binding to core domain. Note again, this folding pathway is not absolutely dominant due to the relatively low probability.

Diverse Parallel Folding Pathways. Whether protein folding follows multiple pathways is still under debate.[7,29,30] It is generally accepted that proteins with symmetric topology are prone to folding via multiple pathways.[31] The pathway diversity is also expected to exist in folding of multidomain proteins, because there are a couple of possible combinations of domain folding and binding. We next illustrate the pathway diversity of DPO4 on thermodynamic 2D free energy surfaces, as shown in Figure 7. Our results indicate that there are six routes bridging the unfolded (U) and native folded (N) states. These routes correspond different parallel folding pathways. The parallel pathways are also observed by projecting the free energy surfaces into other folding reaction coordinates (Figure S8).

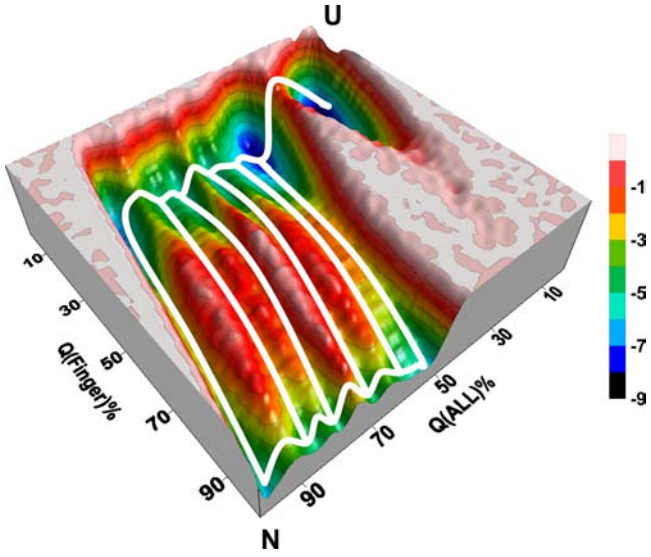


Figure 7. Diverse parallel pathways. The free energy profile is plotted as a function of total native contacts (Q(ALL)) and native contacts in F domain (Q(Finger)); 0% means no native contacts, and 100% means forming all native contacts. There are six routes bridging the unfolded (U) and native folded (N) states. These routes correspond different parallel folding pathways.

We further investigated all possible folding pathways by checking the kinetic trajectories. We found that there are ample possible routes linking the U and N states. Of course, these routes hold with different probabilities. The routes with highest probabilities correspond well to the pathways on the thermodynamic free energy surfaces. Among unfolding pathways, almost all follow the trend that T−LF interface unfolds first and LF and P domains unfold last. For folding kinetic trajectories, folding in the first step has no dominant routes, but in most trajectories, the T−LF interface, LF, and the linker fold in the last steps. The pathway analysis has been summarized by probability distribution in the above sections. All in all, the diversity of these pathways indicates that the folding kinetics for DPO4 are controlled by multiple transition-state ensembles, and the underlying landscape indicates the process is quite complex.

Folding Mechanism. The folding reaction coordinates for all four domains in DPO4 were projected into 2D free energy surfaces, as shown in Figure 5. It is shown that all free energy minima are located at the corner of the 2D free energy surfaces, corresponding to folding or unfolding states. It is remarkable that there is no pathway along the diagonal lines in these surfaces, indicating that the four domains in DPO4 fold in a sequential order, without coupling. Note that the analysis of the relative order is given in the aforementioned section of Stability of Domains and Domain Interfaces.

Although the order of domains folding is not exclusively determined (such as, between P and LF), it is robust to conclude that DPO4 folds into its apo conformation domain by domain and a folded domain can serve as the template of another. However, it is important to clarify that there is a coupling between the formation of interfaces and the folding of domains, as shown in Figure S9. Our results indicate that the formation of the F−P interface is coupled with the folding of the F domain (Figure S9A) and the formation of the P−T interface is coupled with the folding of the T domain (Figure S9B). Figure S9C,D further supports that the P domain folds first, followed by the folding of the F and T domains coupled with their binding to the P domain. However, the coupling is not significant for the formation of the T−LF interface and the folding of T domain (Figure S9E), especially for the formation of the T−LF interface and the folding of LF domain (Figure S9F). It indicates that the folding of the LF domain is relatively independent of the DPO4 core. This may be responsible by the 14-residue linker between the T and LF domain which has been identified to be the source of the protein flexibility at the temperature lower than T_m by CD spectroscopy and FTS.[25]

All in all, our results from the simulation strongly support the conclusion that the four-domain DPO4 folds into its native conformation through a mechanism that individual domains fold independently with each other. However, the processes of domain assembly exhibit dependence and diversity. For the F, P, and T domains in the DPO4 core, folding and binding are coupled, whereas for LF domain, there is no coupling between its folding and the formation of the T−LF interface.

DISCUSSIONS

Three-State Model (with Multiple Metastable States). Experimentally, the thermal denaturation of DPO4 monitored by CD spectroscopy supports the existence of an unfolding intermediate state and a three-state cooperative unfolding process.[25] The CD spectroscopy of several DPO4 mutants generated through site-directed mutagenesis suggests that the LF and Core domains both remained well folded in the intermediate state. And further analysis supports the notion that the formation of the unfolding intermediate is due to the disruption of the salt-bridges formed between the residues in the linker and the polymerase core.

Although the existence of intermediate state(s) is also well supported by simulation, the predicted intermediates are numerous. In fact, more free energy minima were revealed by projecting reaction coordinates into multidimensional free energy surfaces (Figures S3 and S8), indicating more possible

intermediate and transition states in the folding landscape of DPO4. The multiple intermediate states were also observed from simulations using the sequence-flavored model. Although the free energy surfaces from the two models are not in full agreement, both share a similar landscape with multiple intermediate states which have low barriers between each state. With the advances in experimental techniques, such as multiple probes in fluorescence and high-dimensional spectroscopy, the metastable intermediates predicted in simulation may become detectable. The intermediates in simulation have a Q score ranging from 0.4 to 0.9, suggesting large structural differences. In contrast, the intermediate in experiment is observed to show no change in the overall secondary structure content of DPO4.[25] In other words, most of native contacts except T−LF interfacial contacts are not broken. Thus, it should correspond to conformational region with Q score near 1. Further analysis suggests that the experimental intermediate may correspond to intermediate I_1 in which the T−LF interface is fully broken, while the DPO4 core remains well folded.

Taken together, our experimental[25] and theoretical results prompt us to propose a macroscopic three-state model to describe the folding mechanism of DPO4, as shown in Figure 8. The features of the energy landscape described by this diagram serve as a basis for the interpretation of the experimental and simulation data. In this model, there are three global states for DPO4: U and N represent totally unfolded and native states, respectively, while I represents the intermediate ensemble state which is comprised of multiple unstable intermediates separated by low-energy barriers. These metastable states visited during the folding process are difficult to distinguish in bulk experiments which only identify the ensemble of states. This is especially true for CD spectroscopy, one of the most general and basic tools to study protein folding, which has been widely used to monitor the degree of foldedness in an ensemble of proteins. However, its application is limited to the detection of conformational changes in secondary structure. To this end, MD simulations can reveal details that cannot be determined experimentally.

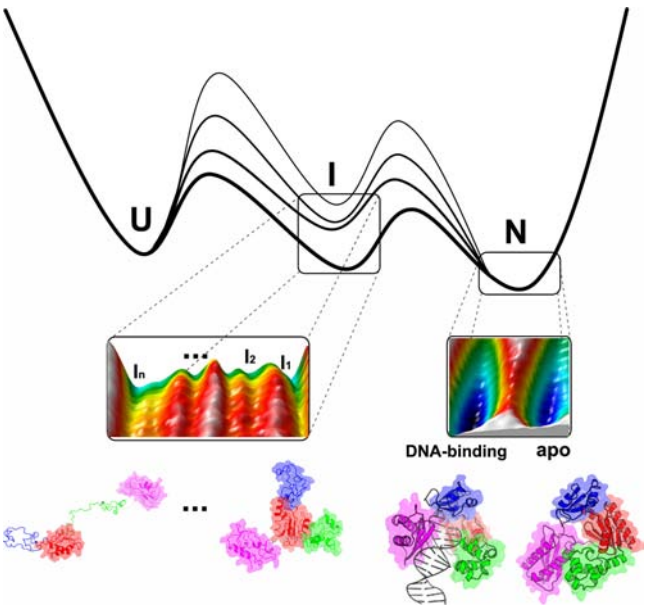


Figure 8. Three-state model of DPO4 folding represented by a schematic one-dimensional free energy landscape. In this model, there are three global states for DPO4: U and N represent totally unfolded and native states, respectively, while I represents the intermediate ensemble state which is comprised of multiple unstable intermediates with low barriers. The relative height of barriers is arbitrary, and there are multiple parallel pathways between U and N. Under conditions promoting the native state, such as low temperature, apo-DPO4 stays at the N basin with closed conformation. Slightly elevating temperature or adding DNA, LF domain moves, and the T−LF interface unfolds, leading to the opening of DPO4 which may facilitate DNA binding. Under conditions unfavoring the native state, DPO4 fluctuates at the I basin during multiple metastable states. And under conditions that intensely unfavor folding, DPO4 unfolds totally and stays at the U basin. In addition, it is important to emphasize that there are multiple parallel pathways bridging U and N.

From the kinetic perspective, the relative instability of intermediate states along the folding/unfolding pathway can avoid trapping the protein in energy minima and accelerate the protein folding/unfolding process. In addition, these on-pathway intermediates allow folding to occur in a stepwise manner and effectively reduce the conformational search process from a large to a smaller time scale.[32] Especially for multidomain proteins, it is expected that the multiple step folding mechanism is common and necessary for rapid assembly and disassembly.

Remarkably, a three-domain protein, adenylate kinase (ADK), has six metastable states on the folding landscape, as recently revealed by high-throughput single-molecule fluorescence studies.[7] In addition, the single-molecule fluorescence also suggested the existence of multiple intersecting folding pathways whose weights were modulated by denaturant concentration. Another recent study carried out by Stigler et al. using ultrastable high-resolution optical tweezers has identified four intermediates for the folding of two-domain calmodulin.[6] Our present work suggests the existence of eight intermediate states for the folding of the four-domain DPO4 (Figure 7). Thus, we propose that a landscape with multiple intermediate states is a common characterization for multi-domain proteins. We believe that more cases will be found soon.

How to Solve the Levinthal Paradox for Multidomain Proteins? If we define that the degrees of freedom per residue are D (about 900 from ref 33), then for a typical single domain protein with the size of N (≈100) amino acids the degrees of freedom are $N_s = D^N = 900^{100}$. This is an astronomical number. Despite the astronomical number of possible conformational states, the fact is that naturally occurring single domain proteins generally fold into well-defined native conformation in times on the order of μs to s.[32] This leads to the well-known "Levinthal paradox".[8] It is now widely accepted that small single domain proteins solve the Levinthal paradox through a funnel-like energy landscape.[8,34] Based on the estimate of the degrees of freedom for single domain with N residues, they for a multidomain protein with M domains (assuming each domain has the same size) are $N_m = D^{N\times M} = N_s^M$ which is an exponential function of that for a typical small protein. This indicates that the degrees of freedom for proteins with multiple domains are many orders of magnitude larger than that for single domain proteins. Now, the question arises: How do the multidomain proteins solve the "protein folding problem"?

The present work gives us a chance to address this issue. Our results from simulation support that DPO4 folds by a stepwise assembly process along multiple parallel pathways with a number of unstable intermediate states. Based on the similarity between these results from recent and current works on the

dx.doi.org/10.1021/ja3045663 | J. Am. Chem. Soc. 2012, 134, 13755−13764

folding of multidomain proteins,[6,7] we expect that the other multidomain proteins share a similar folding mechanism. We further propose that, for rapid folding into well-defined conformations which carry out the biological functions, multidomain proteins should achieve this within a reasonable time through a divide-and-conquer strategy. That is, they fold one domain at a time, then assemble these folded domains into completely large proteins. In this way, the degrees of freedom for multidomain proteins are $N_m = D^N \times M = N_s \times M$. In other words, the Levinthal paradox can be solved through the polynomials rather than exponential: the degrees of freedom N_s^M now becomes $N_s \times M$. Indeed, to solve the Levinthal paradox for the multidomain proteins whose lengths are usually far beyond the upper limit of theoretically foldable proteins,[35] the solution has to be achieved by dividing these large-size proteins into multiple independent folding units whose folding problem has been solved with funnel-shape energy landscape. Overall, we propose that single domain proteins solve the Levinthal paradox by funnelling their energy landscapes, while multidomain proteins overcome the folding problem through combining these individual folding funnels.[36,37]

A similar folding pattern was found in a large repeat protein[30] that shows the existence of sequential folding pathways initiating from the different folding nucleation of the protein. It is remarkable that a divide-and-conquer strategy is also employed for such a large protein despite the fact it does not belong to the multidomain category. Interestingly, there is compelling evidence to suggest that multidomain proteins in vivo also adopt such a strategy to fold, that is, domain by domain.[13,38] Different from in vitro, folding of proteins in vivo is coupled directly to their synthesis in the ribosome; such a process also is termed "cotranslational folding".[26] In cotranslational folding, proteins fold in a sequential way, especially for multidomain proteins, which is vital to the efficient folding so as to avoid misfolding and aggregation.[39] We propose that the divide-and-conquer strategy is used for folding of a typical multidomain protein in vitro. It is interesting that such strategies can also be realized in other large proteins and in a more complicated in vivo environment. This suggests the divide-and-conquer strategy may be quite common in nature to achieve efficient folding.

What Happens Under the Temperature Ranging From 56 to 80 °C? A published biochemical study demonstrates that DPO4 maintained significant polymerase activity after being heated for 5 min at temperatures up to 95 °C.[40] Moreover, we have found that the nucleotide incorporation fidelity of DPO4 is almost unchanged from 2 to 56 °C,[41] suggesting there is no substantial structural change in DPO4 under this temperature range. From 80 to 90 °C, however, our experimental data demonstrate that a conformational change may have occurred after the disruption of two salt bridges and other interactions between residues in the linker and the P domain of DPO4.[25] What happens during an increase in temperature from 56 to 80 °C?

In simulation, we found that there were no significant conformational changes at lower temperatures than 0.90 T_f. And at temperatures higher than 0.95 T_f, the partially unfolding of domains was observed. At the temperature region between 0.90 and 0.95 T_f, our results indicate that the main flexibility of DPO4 is reflected by the motions of LF domain which results in the formation and breaking of the T−LF interface.

It is worth noting that because of the coarse-grained feature and the use of reduced units in our simulations, the simulation temperature is hard to exactly correlate with actual temperature in Kelvins. It is, however, possible to make an estimate based on the melting temperature between simulations and experiments. Accordingly, it may be reasonable that T_f in simulation corresponds to the thermal denaturation temperature T_m = 96 °C (369 K) measured by our folding experiments.[25] Therefore 0.90 T_f may correspond to 59 °C (332 K), 0.95 T_f to 78 °C (351 K). This comparison should be taken as a very crude estimate.

Thus, we may infer that under a middle temperature range from 56 to 80 °C, the major conformational change of DPO4 occurs at the T−LF interface. Our X-ray crystallographic and tryptophan fluorescence studies have revealed that from the apo form of DPO4 to the binary complex between DPO4 and DNA, the LF domain undergoes dramatic rotation relative to the core domain upon DNA binding.[21,42] In the apo structure of DPO4, the LF and T domains bind to and interact with each other physically.[21] In contrast, the DPO4-DNA binary structure reveals that the LF domain is physically located next to the F domain and does not interact with the T domain, although it is tethered to the T domain by the linker.[21] Together, the above analysis leads to a view that DPO4 populates dynamically between a closed conformation (corresponding to the native conformation in low temperature) and an open conformation (may facilitate DNA binding) under a temperature range from 56 to 80 °C. The dynamic equilibrium occurs in native states, as illustrated at the N basin in Figure 8. Under conditions promoting the native state, apo DPO4 resides at the closed basin while in the presence of DNA, DPO4 will stay at the open basin. Under conditions not favoring the native state, DPO4 fluctuates at the I basin during multiple metastable states. When the conditions strongly disfavor folding, DPO4 unfolds completely and stays at the U basin.

The detailed dynamic equilibrium that occurs at N basin cannot be captured by our current models, however, they can be further investigated by multibasin models.[10,43,44] We are currently investigating the process of open−close transitions of DPO4 in the presence of DNA.

Folding is Not the Reverse of Unfolding. The kinetic analysis from simulation indicates that the folding order is not consistent with the reverse order of the unfolding. It is a little surprising that LF domain always responds in the final step, for both the folding and unfolding process. The LF domain changes its conformation following conformational change of the core domain, during both the folding and unfolding processes. A similar asymmetric motion of protein domains has been observed in other proteins.[45,46]

In general, it is expected that folding is the reverse of unfolding.[47,48] However, some recent studies have suggested that the folding and unfolding do not necessarily follow the same dominant routes or share the same mechanism.[49,50] As noted by Finkelstein et al,[47] although the principle of detailed balance indicates that proteins must fold along the reverse of unfolding pathways under the same conditions; their folding pathways under strong folding conditions are not necessarily identical to that under strong unfolding conditions. Notably, once DPO4 was thermally unfolded, it cannot be refolded when the temperature was decreased and the protein precipitated as white powder as observed in our unfolding experiments.[25]

Important Roles of the Linker. The analysis from the thermodynamic free energy profiles reveals that the folding of the LF domain is relatively independent of its interactions with the DPO4 core. Our simulation in this paper and unfolding

experiments[25] both support that the motion of the LF domain in the absence of the loss of secondary structural elements determines the enzymatic activity of DPO4 at temperatures lower than T_m. This led us to infer that the linker region plays an important role in modulating the motion of LF domain, because it bridges the LF domain and DPO4 core, and its small-scale rotation can result in the large-scale motion of LF domain. In fact, the unfolding data with various DPO4 mutants in the linker region have suggested that the linker plays an important role in the existence of the unfolding intermediate of DPO4.[25] Changing amino acid residues within the linker region indeed does not affect the secondary structure elements of DPO4 at ambient conditions but has great impacts on the overall folding stability for DPO4.[25] This is a result of strong electrostatic interactions between the linker region and the core.

CONCLUSION

The complex folding process of multidomain DPO4 was investigated by a native topology-based model and its variant by the introduction of energetic heterogeneity (sequence-flavored model). The results of the kinetic and the equilibrium analysis can provide an excellent view of the folding landscape of DPO4. The fact that DNA polymerases from different families have a highly conserved structure but relatively low sequence homology[21] highlights the important role of protein topology to enzymatic function. This makes it reasonable to study the structure−function relationship of the Y-family DNA polymerases using the native topology-based models. The simulation results are in good agreement with the experimental data[25] published during the review of this work. It also indicates that the complex folding landscape of the multidomain DNA polymerases can be captured well by such simplified models despite using at a coarse-grained level as a first approximation. In addition, we have made a number of testable predictions including the existence of multiple metastable intermediates and parallel pathways, the hysteretic conformational change of LF domain during the folding/unfolding process, the highest stability of the P domain among the domains in the polymerase core, and the domain-by-domain folding mechanism of multidomain proteins.

Overall, the underlying folding landscape of DPO4 is quite complex like other multidomain proteins.[6,7,16] The existence of multiple transition-state ensembles and parallel pathways is likely common for multidomain proteins, and the trend is more significant for proteins with more domains. This feature allows the folding of multidomain protein to occur in a stepwise manner and effectively reduce the conformational search process from an exceedingly longer time scale to a significantly shorter time scale. To achieve efficient folding, DPO4 employs a divide-and-conquer strategy, that is, combining multiple individual folding funnels into a single funnel (domains fold independently then coalesce). In this way, the degrees of freedom for multidomain proteins are polynomial rather than exponential in protein size. The theoretical study of the folding of multidomain proteins is important in bridging the gap in the knowledge of protein folding between small single domain proteins and the more common multidomain systems. Deeper insights into the folding mechanism of multidomain proteins still require further theoretical and experimental efforts.

METHOD AND MATERIALS: COMPUTATIONAL DETAILS

We use the coarse-grained native SBM (for details, see SI text and refs 51 and 52) in which each residue in the polypeptide chain is represented by one bead. The sequence-favored model treats the native interactions using the Miyazawa−Jernigan (MJ) statistical potential.[53] The underlying idea is that protein sequence encodes the strength of native interactions with different statistical weights to modulate the stability of structural elements. The combination of coarse-grained model and native topology-based potential provides a suitable tool to explore the folding of large multidomain proteins, at present. Coarse-grained representations of the polypeptide chain are employed to reduce the number of degrees of freedom, and the native topology-based potential is used to construct a smooth energy landscape. The main advantage of the coarse-grained SBM is the ability to efficiently sample the conformational space. The common limitations of the native SBMs are that they lack atomic details and transferability. Hybrid all-atom models[54,55] and multiscale protocols[51,56] can be developed to extend the application of the pure native SBMs.

A basic assumption in native SBM is that natural proteins are minimally frustrated, and the folding mechanisms are majorly determined by the geometric constraints of the native structure.[52] However, the role of other factors, such as the sequence[16,57,58] and non-native interactions,[6,51] which are typically ignored remains evident in numerous cases. We plan to incorporate non-native interactions, in particular, the electrostatic interactions, into the model in the future work. The present work employed the combination of the SBM and its sequence-dependent variant with the ability to capture the subtle differences in the folding mechanism.[59,60] The combination models allow us to test the robustness of the results and investigate the sequence effects in the folding of multidomain proteins.

Another issue that should be considered here is the functional landscape of DPO4 (located at the bottom of the energy funnel, N basin in Figure 8) which contains multiple folded states. It is difficult to be reproduced by current single structure-based model. We have developed a multibasin model which can capture the functional transitions that occur at native basins in functional landscape. This model is currently being developed to investigate the detailed conformational change of DPO4 responding to the binding of DNA.

ASSOCIATED CONTENT

Supporting Information

Computational details, supplemental Tables S1−S3, and Figures S1−S10. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

jin.wang.1@stonybrook.edu

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We acknowledge the High Performance Computing Center (HPCC) of Jilin University for supercomputer time. Y.W., X.K.C., and E.K.W. acknowledge support from the National Science Foundation of China (grant nos. 21190040 and 11174105). Z.S. acknowledges support from the National Science Foundation (grant no. MCB-0960961). J.W. thanks National Science Foundation for support.

REFERENCES

(1) Han, J. H.; Batey, S.; Nickson, A. A.; Teichmann, S. A.; Clarke, J. *Nat. Rev. Mol. Cell Biol.* **2007**, *8*, 319−30.

(2) Batey, S.; Nickson, A. A.; Clarke, J. *HFSP J* **2008**, *2*, 365−77.

(3) Onuchic, J.; LutheySchulten, Z.; Wolynes, P. *Annu. Rev. Phys. Chem.* **1997**, *48*, 545−600.
(4) Fitter, J. *Cell. Mol. Life Sci.* **2009**, *66*, 1672−1681.
(5) Bhaskara, R. M.; Srinivasan, N. *Sci. Rep.* **2011**, *1*, 40.
(6) Stigler, J.; Ziegler, F.; Gieseke, A.; Gebhardt, J. C.; Rief, M. *Science* **2011**, *334*, 512−6.
(7) Pirchi, M.; Ziv, G.; Riven, I.; Cohen, S. S.; Zohar, N.; Barak, Y.; Haran, G. *Nat. Commun.* **2011**, *2*, 493.
(8) Karplus, M. *Folding Des.* **1997**, *2*, S69−75.
(9) Ganesh, C.; Shah, A. N.; Swaminathan, C. P.; Surolia, A.; Varadarajan, R. *Biochemistry* **1997**, *36*, 5020−8.
(10) Wang, Y.; Tang, C.; Wang, E.; Wang, J. *PLoS Comput. Biol.* **2012**, *8*, e1002471.
(11) Wilson, C. J.; Das, P.; Clementi, C.; Matthews, K. S.; Wittung-Stafshede, P. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 14563−8.
(12) Das, P.; Wilson, C. J.; Fossati, G.; Wittung-Stafshede, P.; Matthews, K. S.; Clementi, C. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 14569−14574.
(13) Elcock, A. H. *PLoS Comput. Biol.* **2006**, *2*, e98.
(14) Rundqvist, L.; Aden, J.; Sparrman, T.; Wallgren, M.; Olsson, U.; Wolf-Watz, M. *Biochemistry* **2009**, *48*, 1911−27.
(15) Shank, E. A.; Cecconi, C.; Dill, J. W.; Marqusee, S.; Bustamante, C. *Nature* **2010**, *465*, 637−640.
(16) Borgia, M. B.; Borgia, A.; Best, R. B.; Steward, A.; Nettels, D.; Wunderlich, B.; Schuler, B.; Clarke, J. *Nature* **2011**, *474*, 662−5.
(17) Porter, L. L.; Rose, G. D. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 9420−9425.
(18) Batey, S.; Scott, K. A.; Clarke, J. *Biophys. J.* **2006**, *90*, 2120−2130.
(19) Batey, S.; Clarke, J. *J. Mol. Biol.* **2008**, *378*, 297−301.
(20) Itoh, K.; Sasai, M. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 13865−70.
(21) Wong, J. H.; Fiala, K. A.; Suo, Z.; Ling, H. *J. Mol. Biol.* **2008**, *379*, 317−30.
(22) Fiala, K. A.; Suo, Z. *J. Biol. Chem.* **2007**, *282*, 8199−206.
(23) Sherrer, S. M.; Brown, J. A.; Pack, L. R.; Jasti, V. P.; Fowler, J. D.; Basu, A. K.; Suo, Z. *J. Biol. Chem.* **2009**, *284*, 6379−88.
(24) Ling, H.; Boudsocq, F.; Woodgate, R.; Yang, W. *Cell* **2001**, *107*, 91−102.
(25) Sherrer, S. M.; Maxwell, B. A.; Pack, L. R.; Fiala, K. A.; Fowler, J. D.; Zhang, J.; Suo, Z. *Chem. Res. Toxicol.* **2012**, *25*, 1531−1540.
(26) Fedorov, A. N.; Baldwin, T. O. *J. Biol. Chem.* **1997**, *272*, 32715−32718.
(27) Angelani, L.; Ruocco, G. *Europhys. Lett.* **2009**, 87.
(28) Cho, S. S.; Levy, Y.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 586−91.
(29) Sosnick, T. R.; Barrick, D. *Curr. Opin. Struct. Biol.* **2011**, *21*, 12−24.
(30) Zhou, R.; He, Y.; Xiao, Y. *Comput Biol Chem* **2011**, *35*, 169−173.
(31) Nauli, S.; Kuhlman, B.; Baker, D. *Nat. Struct. Biol.* **2001**, *8*, 602−5.
(32) Bai, Y. *Biochem. Biophys. Res. Commun.* **2003**, *305*, 785−8.
(33) Plotkin, S. S.; Onuchic, J. N. *Q. Rev. Biophys.* **2002**, *35*, 111−67.
(34) Onuchic, J. N.; Wolynes, P. G. *Curr. Opin. Struct. Biol.* **2004**, *14*, 70−5.
(35) Lin, M. M.; Zewail, A. H. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 9851−6.
(36) Tsai, C. J.; Kumar, S.; Ma, B.; Nussinov, R. *Protein Sci.* **1999**, *8*, 1181−90.
(37) Wang, J.; Xu, L.; Wang, E. *Biophys. J.* **2007**, *92*, L109−11.
(38) Lee, W.; Zeng, X.; Zhou, H.-X.; Bennett, V.; Yang, W.; Marszalek, P. E. *J. Biol. Chem.* **2010**, *285*, 38167−38172.
(39) Komar, A. A. *Trends Biochem. Sci.* **2009**, *34*, 16−24.
(40) Boudsocq, F.; Iwai, S.; Hanaoka, F.; Woodgate, R. *Nucleic Acids Res.* **2001**, *29*, 4607−16.
(41) Fiala, K. A.; Sherrer, S. M.; Brown, J. A.; Suo, Z. *Nucleic Acids Res.* **2008**, *36*, 1990−2001.
(42) Xu, C.; Maxwell, B. A.; Brown, J. A.; Zhang, L.; Suo, Z. *PLoS Biol.* **2009**, *7*, e1000225.
(43) Okazaki, K.; Koga, N.; Takada, S.; Onuchic, J. N.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 11844−11849.
(44) Lu, Q.; Wang, J. *J. Am. Chem. Soc.* **2008**, *130*, 4772−83.
(45) Whitford, P. C.; Gosavi, S.; Onuchic, J. N. *J. Biol. Chem.* **2008**, *283*, 2042−8.
(46) Bhatt, D.; Zuckerman, D. M. *J. Chem. Theory Comput.* **2011**, *7*, 2520−2527.
(47) Finkelstein, A. V. *Protein Eng.* **1997**, *10*, 843−5.
(48) Dinner, A. R.; Karplus, M. *J. Mol. Biol.* **1999**, *292*, 403−19.
(49) Finke, J. M.; Onuchic, J. N. *Biophys. J.* **2005**, *89*, 488−505.
(50) Klimov, D. K.; Thirumalai, D. *J. Mol. Biol.* **2005**, *353*, 1171−1186.
(51) Wang, J.; Wang, Y.; Chu, X.; Hagen, S. J.; Han, W.; Wang, E. *PLoS Comput. Biol.* **2011**, *7*, e1001118.
(52) Clementi, C.; Nymeyer, H.; Onuchic, J. N. *J. Mol. Biol.* **2000**, *298*, 937−953.
(53) Miyazawa, S.; Jernigan, R. L. *J. Mol. Biol.* **1996**, *256*, 623−644.
(54) Sutto, L.; Mereu, I.; Gervasio, F. L. *J. Chem. Theory Comput.* **2011**, *7*, 4208−4217.
(55) Chen, K.; Eargle, J.; Lai, J.; Kim, H.; Abeysirigunawardena, S.; Mayerle, M.; Woodson, S.; Ha, T.; Luthey-Schulten, Z. *J. Phys. Chem. B* **2012**, *116*, 6819−6831.
(56) Li, W.; Yoshii, H.; Hori, N.; Kameda, T.; Takada, S. *Methods* **2010**, *52*, 106−114.
(57) Karanicolas, J.; Brooks, C. L., III *Protein Sci.* **2002**, *11*, 2351−2361.
(58) Cho, S. S.; Levy, Y.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 434−9.
(59) Karanicolas, J.; Brooks, r., C. L. *J. Mol. Biol.* **2003**, *334*, 309−25.
(60) Hills, J., R. D.; Kathuria, S. V.; Wallace, L. A.; Day, I. J.; Brooks, r., C. L.; Matthews, C. R. *J. Mol. Biol.* **2010**, *398*, 332−50.