Network representation of conformational transitions between hidden intermediates of Rd-apocytochrome b_{562}

Mojie Duan,^{1,2,a),b)} Hanzhong Liu,^{1,a)} Minghai Li,^{1,c)} and Shuanghong Huo^{1,b)}

¹*Gustaf H. Carlson School of Chemistry and Biochemistry, Clark University, 950 Main Street, Worcester, Massachusetts 01610, USA*

²*Key Laboratory of Magnetic Resonance in Biological Systems, State Key Laboratory of Magnetic Resonance and Atomic and Molecular Physics, National Center for Magnetic Resonance in Wuhan, Wuhan Institute of Physics and Mathematics, Chinese Academy of Sciences, Wuhan 430071, China*

(Received 30 May 2015; accepted 14 September 2015; published online 1 October 2015)

The folding kinetics of Rd-apocytochrome b_{562} is two-state, but native-state hydrogen exchange experiments show that there are discrete partially unfolded (PUF) structures in equilibrium with the native state. These PUF structures are called hidden intermediates because they are not detected in kinetic experiments and they exist after the rate-limiting step. Structures of the mimics of hidden intermediates of Rd-apocytochrome b_{562} are resolved by NMR. Based upon their relative stability and structural features, the folding mechanism was proposed to follow a specific pathway (unfolded $\rightarrow$ rate-limiting transition state $\rightarrow$ PUF1 $\rightarrow$ PUF2 $\rightarrow$ native). Investigating the roles of equilibrium PUF structures in folding kinetics and their interrelationship not only deepens our understanding of the details of folding mechanism but also provides guides in protein design and prevention of misfolding. We performed molecular dynamics simulations starting from a hidden intermediate and the native state of Rd-apocytochrome b_{562} in explicit solvent, for a total of 37.18 μ s mainly with Anton. We validated our simulations by detailed comparison with experimental data and other computations. We have verified that we sampled the post rate-limiting transition state region only. Markov state model was used to analyze the simulation results. We replace the specific pathway model with a network model. Transition-path theory was employed to calculate the net effective flux from the most unfolded state towards the most folded state in the network. The proposed sequential folding pathway via PUF1 then more stable, more native-like PUF2 is one of the routes in our network, but it is not dominant. The dominant path visits PUF2 without going through PUF1. There is also a route from PUF1 directly to the most folded state in the network without visiting PUF2. Our results indicate that the PUF states are not necessarily sequential in the folding. The major routes predicted in our network are testable by future experiments such as single molecule experiment. © 2015 AIP Publishing LLC. [<http://dx.doi.org/10.1063/1.4931921>]

INTRODUCTION

Ever since Levinthal raised the pivotal question of protein folding,¹ continuous efforts have been devoted to searching for folding intermediates to characterize the folding pathways. The early experiments in this aspect can be traced back to 1970s.^{2,3} For some apparent two-state proteins, native-state hydrogen exchange experiments show that there are discrete partially unfolded (PUF) structures in equilibrium with the native state.^{4–6} It was interpreted that these PUF structures formed sequentially to define a specific folding pathway.^{7,8} Investigating PUF states is important not only for understanding the kinetic mechanism of protein folding,^{7,9,10} but also for providing guide for structure prediction¹¹ and protein design. The recent success in prediction of protein structures and folding pathways utilizing partially cooperative

unfolding/refolding units (called foldons) as a computational search strategy has proven the importance of studying PUF states.¹¹ Partially unfolded structures also have implications in protein function¹² and protein (mis)folding diseases. Increasing lines of evidence show that protein can even aggregate under native conditions starting from partially unfolded and/or misfolded conformations that are accessible from the native state through thermal fluctuations.¹³

Rd-apocytochrome b_{562} is a four-helix bundle protein with an apparent two-state folding kinetics.¹⁴ Several PUF structures are found in equilibrium with the native state.¹⁴ These PUF structures are called hidden intermediates^{15,16} because they are not detected in kinetic experiments. They are significantly more stable than the fully unfolded state, and they are even more folded than the rate-limiting transition state.^{14,17} Therefore, they exist after the rate-limiting step and in equilibrium with the native state but have never been accumulated to a dominant population. Thus, it is experimentally difficult to characterize the PUF structures because of the dynamic and transient nature of these species though promising results have been reported recently using NMR methods.^{18–20}

^{a)}M. Duan and H. Liu contributed equally to this work.

^{b)}Authors to whom correspondence should be addressed. Electronic addresses: shuo@clarku.edu and mjduan@wipm.ac.cn

^{c)}Present address: Physics Department, Northeastern University, Boston, Massachusetts 02115, USA.

Protein engineering (mutation and/or truncation) has been utilized to either selectively destabilize the native state or to stabilize hidden intermediates^{4,16,21–24} to isolate the PUF structures for several proteins.^{5,6,25–27} The structures of the mimics of three hidden intermediates of Rd-apocytochrome *b*₅₆₂ are resolved by NMR: PUF1,²² PUF2,²⁸ and N*.²⁹ Their structural features are the following: the N-terminus is unfolded and the C-terminus is partially folded in PUF1, PUF2 is unfolded at the N-terminus only, and the backbone conformation of N* is close to that of the native state, but with nonnative side-chain packing. Based upon these structural features and relative stability, it was proposed that Rd-apocytochrome *b*₅₆₂ folds in a sequential manner with the increase of folded structure in each step: unfolded → rate limiting step → PUF1 → PUF2 → N* → native state (Fig. 1(a)).^{22,28,30} The relaxation time of PUF2 mimic is about 25 μ s at 65 °C,²⁷ suggesting that there is a substantial free energy barrier between the hidden intermediate and the native state. Lattice models were used to study whether hidden intermediates form a sequential pathway,^{15,31} but controversial results were obtained. The occurrence of hidden intermediates was found to be parallel rather than sequential by solving a master equation,¹⁵ while a Monte Carlo method showed sequential ordering of late folding events.³¹

With the dramatic growth of computer power, especially long time scale computer simulations powered by Anton,³² and the recent advance in force field development, we can reinvestigate previously studied folders with all-atom models. Recent analyses of long molecular dynamics trajectories using

the Markov state model show that systems with approximate two-state behavior have remarkable hidden complexity.³³ We use Rd-apocytochrome *b*₅₆₂ as a model system to study the hidden intermediates because the structures of their mimics are resolved. We studied the folding/unfolding of PUF2 using Anton and Beowulf cluster with the AMBER99SB-ILDN force field.³⁴ We present a network model with substantial complexity relative to the proposed pathway in literature.

METHODS

Simulation protocol

We collected eight trajectories in total of 37.18 μ s: a 5.43- μ s trajectory initiated from the native state of Rd-apocytochrome *b*₅₆₂ (PDB ID: 1YYJ²⁸ model 2 because the coordinates of residue 81 are missing in model 1) and seven independent trajectories starting from PUF2 (Rd-apocytochrome *b*₅₆₂ with the N-terminal helix unfolded) including a 13- μ s trajectory and a 6.75- μ s using Anton, two 3- μ s and three 2- μ s trajectories using our local cluster. The partially unfolded state, PUF2, was designed by substituting the hydrophobic residues in the N-terminal (W7D/L10G/L14G/V16G/I17G) with Gly and Asp to selectively destabilize the native state.³⁵ We took the structure of PUF2 (PDB ID: 1YZA³⁵ model 1) while we mutated its sequence back to the wild type and used it as the initial structure for the simulations of PUF2.

Maestro³⁶ was used to set up the system. The all-atom AMBER99SB-ILDN force field and the TIP3P³⁷ explicit water model were utilized. Three Na⁺ ions were added as counter ions. The protein atoms were at least 16 Å away from the boundary of the water box. The whole system including protein, counter ions, and water consists of 29 851 atoms and 38 716 atoms for the simulations of native state and PUF2, respectively. The minimization is a hybrid of the steepest decent method and the limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) method. After ten steps of steepest decent minimization, LBFGS minimization was used until the maximum iteration steps reach 2000 or the convergence criterion (50 kcal/(mol Å)) is satisfied, whatever comes first. The system was first minimized with restraints on the solute followed by another round of minimization without the restraints.

Because Anton is a special machine, it requires a well equilibrated trajectory to be prepared on our local machine before the job is submitted. Desmond standard NPT relaxation protocol was used: first, the system was equilibrated for 1.2 ns using Desmond³⁸ in the NVT ensemble at 10 K using the Berendsen thermostat³⁹ with harmonic restraints (force constant = 50.0 kcal/(mol Å²)) on the protein heavy atoms; second, the system was equilibrated for 1.2 ns in the NPT ensemble at 10 K using the Berendsen thermostat with the same restraints; third, the temperature of system was increased to 300 K linearly over the time period of 1.2 ns with restraints; at last, the unrestrained system was equilibrated for 10 ns using the Berendsen thermostat in the NPT ensemble. The observation of convergence of equilibration was shown in Fig. S1.⁶⁴ The total potential energy of the system as a function of time shows that the fluctuation of potential energy in the

FIG. 1. Structures of the native state and hidden intermediates of Rd-apocytochrome *b*₅₆₂ and the secondary structure analysis along the simulations. (a) Native state (PDB ID: 1YYJ), PUF2 (PDB ID: 1YZA with native sequence), and PUF1 (PDB ID: 1YZC with native sequence). The four helical regions of the native structure are labeled as H1 (residue 2–18), H2 (residue 23–40), H3 (residue 59–78), and H4 (residue 84–104). The protein conformation is depicted with UCSF Chimera.⁵⁵ The color scheme is rainbow with the N-terminus in blue and the C-terminus in red. (b) Helicity of each residue computed over all the trajectories. The blue bars on the top of the figure show the helical regions in the native structure. The red and black lines denote the helicity averaged over the native-state and PUF2 simulations, respectively.

final step of equilibration is around 0.6%, indicating a good convergence of equilibration.

The production runs were performed under 300 K using the Berendsen thermostat in the NPT ensemble with a multiple time step approach⁴⁰ (bonded, near, and far time steps = 2 fs, 2 fs, and 6 fs). The snapshots were saved every 240 ps. The short-range forces and the long-range forces were evaluated at each time step and every three time steps, respectively. The cutoff for the van der Waals and short-range electrostatic interactions was 9.0 Å. The long-range electrostatic interactions were computed using the k-space Gaussian split Ewald method⁴¹ with the $64 \times 64 \times 64$ grid size. All the runs including those on Anton and on our local cluster were performed using the same protocol, the same force field, and the same software (Desmond or its equivalent on Anton) except the different random numbers to generate the initial velocity.

Structural analysis

The snapshots after 10 ns in each trajectory were used for the following analysis, totally 22 541 and 132 547 conformations for the native and PUF2 simulations, respectively. STRIDE⁴² implemented in Visual Molecular Dynamics (VMD)⁴³ was used for the secondary structure assignment. Helicity is defined as the helical fraction of individual residues during the simulations. For the residues that are more than three residues apart in sequence, their heavy atoms (each from a different residue) are considered to be in contact with each other if the distance between the atoms is less than 4.5 Å. The total number of the heavy-atom pairs that are in contact in the native state is N in Eq. (1). The fraction of native contacts for a given snapshot S is defined as⁴⁴

$$Q(S) = \frac{1}{N} \sum_{(i,j)} \frac{1}{1 + \exp[\beta(r_{i,j}(S) - \lambda r_{i,j}^0)]}, \quad (1)$$

where $r_{i,j}^0$ is the distance between heavy atom i and j in the native structure and $r_{i,j}(S)$ is the distance between the same atom pair in snapshot S . The smoothing parameter β is set to be 5 Å^{-1} and the contact fluctuation is taken into account by the parameter $\lambda = 1.8$. The summation runs over the N pairs. The definition of fraction of native contacts for residue k in snapshot S is similar to Q for the whole protein; however, only the contacts associated with residue k is calculated

$$Q_k(S) = \frac{1}{N_k} \sum_{(k,i,j)} \frac{1}{1 + \exp[(\beta(r_{k,i,j}(S) - \lambda r_{k,i,j}^0)]}, \quad (2)$$

where N_k is the number of heavy-atom pairs in native contact between residue k and other residues (>3 residues apart in sequence). If there is no heavy atom within 4.5 Å to the heavy atoms of residue k , then $Q_k(S) = 0$.

Clustering

Since experimentally PUF1 and PUF2 are classified as two distinct intermediate states after the rate-limiting step, our clustering method and criteria should be able to separate the PUF1-like and PUF2-like conformations into two clusters, while the clustering criterion solely based on RMSD (root-mean-square deviation) failed to do so (Figs. S2 and S3 in the

supplementary material⁶⁴). Information about the secondary and tertiary structure is the key as well as making connection with experimental results. This idea was suggested in the early work of Markov-based analysis.⁴⁵ Baker and coworkers showed the ineffectiveness of RMSD as a distance metric in the study of discretization of state space for Markov state model.⁴⁶ The advantage of system specific distance metrics over RMSD was seen in complex systems.^{46–48} By analyzing the secondary structural changes in the simulation and observing the tertiary structure changes along the trajectories, we find major secondary structure changes in the N-terminal part of helix 3 and near the C-terminus of helix 4, and helix 3 bends at V69 (see the helicity analysis in the section on Results and Discussion). Based on these analysis and observations, we adopted a system specific hierarchical approach to cluster the PUF2 trajectories. Because Q value is commonly used to describe the protein nativeness along the progress of folding, we first used Q value as a coarse criterion to divide the conformation space into four regions according to the distribution of Q : two main regions aside the distribution peak ($Q = 0.56$) and the two tails of the distribution (see the distribution of Q in the section on Results and Discussion). Then for each region of Q , we used three criteria to cluster: (1) whether the number of non-helical residues in the helix 3 region <2 or not; (2) whether the number of helical residues in the C-terminal part (residues 99–106) of helix 4 ≥ 3 or not; (3) whether the main-chain heavy atom RMSD of helix 3 (residue 59–78) with respect to the initial structure of PUF2 $<4.0 \text{ Å}$ or not. Finally, 11 states (or clusters) were obtained. “State” and “cluster” are used interchangeably. The structural features of each cluster are shown in the section on Results and Discussion.

Markov state model analysis

MSMBuilder2^{49,50} was used to build the transition matrices for the 11 states with different lag time (τ), ranged from 240 ps (the time interval of saving each snapshots) to 800 ns. At each lag time, the corresponding transition matrix was diagonalized. The eigenvalues (λ s) of the transition matrix are related to the relaxation time of the master equation by $-\tau/\ln \lambda_i$. The relaxation time scale as a function of lag time was used to check when the model starts to satisfy the Markov assumption. The lag time used in the Chapman–Kolmogorov test is 60 ns when the slowest relaxation time scale starts to level off and longer time at 120 ns and 240 ns. Transition matrix (T) contains the transition probability between all the pairs of states, but not all the transitions contribute to the transition to the final state; in fact, only a fraction of the transitions makes net contributions. Therefore, the transition matrix needs to be corrected according to the discrete transition-path theory,⁵¹

$$\begin{aligned} f_{ij}^+ &= \max \{0, f_{ij} - f_{ji}\}, \\ f_{ij} &= \pi_i q_i^- T_{ij} q_j^+, \end{aligned} \quad (3)$$

where f_{ij}^+ is the net effective flux from state i to j that contributes to the overall transition from source to sink, π_i is the normalized population of state i , and q_j^+ and q_i^- are the forward-committor probability and the backward-committor probability, respectively. The calculateTPT module of the

MSMBuilder2 package is employed to compute f_{ij}^+ between every pair of states that contributes to the total flow from the source (state 8) to the sink (state 1). The transition matrix used for this calculation is based on the lag time (τ) of 60 ns. The total time from state 8 to the sink is⁵²

$$t = \frac{\tau\pi_8}{F}, \quad (4)$$

where F is the total net flow leaving state 8, $F = \sum f_{8j}^+ = 4.81 \times 10^{-3}$. The population of state 8 (π_8) is 10.16% when the system reaches convergence.

RESULTS AND DISCUSSION

Helicity

The simulations started from the native-state structure as well as the PUF2 structure. The major difference between these two starting structures is that the helix 1 region is disordered in PUF2. The helix 1 region remains unstructured in the PUF2 simulations; thus, PUF2 does not fold into the native state in our simulations. This is not surprising because the estimated folding time of PUF2 analog at 65 °C is 25 μ s.²⁷ We neither expect that PUF2 folds to N* which has native fold with non-native hydrophobic interactions because the conformation of N* is closer to that of the native state than to that of PUF2. The analysis of helicity of each residue during the simulations shows that the secondary structure is well maintained during the simulation of the native state, while significantly low helicity is present in the N-terminus of helix 3 (~40% helicity) and in the C-terminus of helix 4 (<60% helicity) for the simulations of PUF2 (Fig. 1(b)). The low helicity in the N-terminus of helix 3 and in the C-terminus of helix 4 was observed in many trajectories rather than in a single trajectory (Fig. S4 in the supplementary material⁶⁴).

The order of helical stabilities in our simulations of PUF2 is in excellent agreement with those obtained by experiments and other computational methods. Native-state hydrogen exchange shows the increasing order of backbone hydrogen bonding stability: helix 1 $\cong$ F65 and D66 (that are located at the N-terminal part of helix 3), the C-terminal part of helix 4, and helix 2 $\cong$ the C-terminal part of helix 3 $\cong$ the N-terminal part of helix 4.¹⁴ Previous high-temperature simulations⁵³ also gave similar order of helical stability. The helical content predicted by the AGADIR program⁵⁴ shows the increasing order of helical content which is consistent with the experiment and simulation: the N-terminal part of helix 3 (helical content <5%), the C-terminal part of helix 4 (helical content <5%), helix 1 (helical content $\cong$ 5%), helix 2 (helical content $\cong$ 15%), the C-terminal part of helix 3 (helical content $\cong$ 18%), and the N-terminal part of helix 4 (helical content $\cong$ 22%). We realize that AGADIR helix prediction algorithm is based on helix-coil theory which does not take tertiary structures into account, but it is worth to point out the coincidence between the prediction and our simulation result. Our native-state simulation serves as a control to validate the force field; the loss in helical content in the N-terminal part of helix 3 and the C-terminal part of helix 4 regions is not due to the defect of the force field; otherwise, the same observation

should be made in the 5.43- μ s native-state simulation as well. In fact, all the secondary structures are well maintained in the native-state simulation (Fig. 1(b)). Taking together the results in the literature native-state hydrogen exchange, high-temperature simulation, and prediction of helical content, as well as our native-state simulation, we believe that the simulation of PUF2 is well validated.

Fraction of native contacts

The fraction of native contacts (Q) has been commonly used as a reaction coordinate to describe the process of folding. We analyzed the distribution of Q for the conformations sampled in our simulations (Fig. 2(a)). For the simulations starting from PUF2, the Q values distribute from 0.4 to 0.7 with the highest peak located at 0.56. For comparison, the distribution of the Q values of the native-state simulation is very narrow with a single peak at 0.79. Assume that Q is the reaction coordinate, evidently, the simulations starting from PUF2 sampled the regions between the folded and unfolded state.

The fraction of residue native contacts (Q_k) was calculated to monitor the nativeness of each residue (Figs. 2(b)–2(d)). Fig. 2(b) shows Q_k along the protein sequence averaged over the trajectories of the native-state simulation and the simulations starting from PUF2, respectively. In both the native-state and PUF2 simulations, the helix 2-helix 3 loop shows very low Q_k . In accord with the result of helicity (Fig. 1(b)), residues in the helix 1 region, residues 59–68 (in the N-terminal part of helix 3), and residues 92–104 (in the C-terminal part of helix 4) show less native contacts than other regions in the PUF2 simulations. Since the helix 1 region is unfolded in the PUF2 mimic resolved by NMR, some native contacts between the helix 1 region and the helix 2 region are lost; thus, the Q_k values for Q25, A29, K32, M33, A36, and A40 in helix 2 are relatively low. In addition, the native contact between Q71 (of helix 3) and Q93 (of helix 4) is absent in the NMR structure of the PUF2 mimic and the simulations of PUF2, while the missing native contact between R34 (of helix 2) and E73 (of helix 3) in the PUF2 mimic re-forms in the simulations.

Markov state model and network of net effective flux

To describe the complexity of conformational transitions between partially unfolded states, we built the Markov state model and used transition-path theory to calculate the net effective flux that contributes to the overall transition from source to sink. The network of the net effective flux among the eight largest states is shown in Fig. 3. Both state 1 and 2 can be considered as PUF2 from the secondary structure point of view, but more native contacts are present in state 1. Therefore, state 1 was chosen as the sink in the calculation of net effective flux. State 3 is PUF1 because the C-terminal of helix 4 is unfolded. Besides the proposed hidden intermediates, PUF1 and PUF2, other partially unfolded states are uncovered, of which the main conformational difference among the states is in the helix 3 region. State 4 is different from PUF1 and PUF2 because of the presence of the nonhelical conformations in the N-terminal part (from residue 59 to 69) of helix 3. The

FIG. 2. Native contacts. (a) Distribution of Q . (b) Q_k (residue native contacts) along the protein sequence averaged over all the trajectories. The helical regions are labeled at the top. For both (a) and (b), the red and black lines denote the native-state simulation and PUF2 simulations, respectively. (c) Mapping Q_k averaged over the PUF2 simulations onto the PUF2 initial structure. (d) Q_k of PUF2 initial structure. The major differences in Q_k between (c) and (d) are in the helix 2-helix 3 loop, the N-terminal part of helix 3, and the C-terminal part of helix 4. The legend of Q_k is shown in the figure.

helical structure in this region is further lost in state 5 and totally lost in state 8. While the helical content in the same region is high in states 6 and 7, helix 3 is bent with a kink at V69. We think states 6 and 7 are misfolded states. State 8 (population = 10.2%) is the most unfolded state among the eight largest states with a Q value less than 0.5. Even though state 9, 10, and 11 are less native than state 8, their populations are low ($<0.5\%$). Accordingly, state 8 is chosen as source in the flux calculation. In this state, besides the unfolded helix 1 region and the C-terminal of helix 4, residues 59–69 in the N-terminal part of helix 3 are also unstructured, leading to a 29-residue long loop between helix 2 and helix 3. The floppy loop flips back and forth from one side of the helix bundle to the other side and sometimes the loop points upwards.

We are interested in the region after the rate-limiting step. Both kinetics experiments and hydrogen exchange studies showed that this protein is a two-state folder.¹⁴ Namely, there is a single one rate-limiting step. The structural feature of the transition state obtained by the unfolding simulation at high temperature⁵³ is consistent with that characterized experimentally,^{14,17} indicating that folding and unfolding overcome the same rate-limiting transition state. As shown in Fig. 2(a), the simulations starting from PUF2 have not reached the unfolded state ($Q \approx 0$) because the Q -value distribution of the conformations collected in the simulations is in the range

from 0.4 to 0.7. In comparison with that of the conformations in the native-state simulations, clearly, the system samples the region between the native state and the unfolded state that has few native contacts. We notice that Fig. 2(a) provides a necessary but not sufficient condition to show that the partially unfolded states are post-transition state. But the time scale of the MD simulations starting PUF2 limits the sampling to the post-transition state region. The rate-limiting barrier for PUF1/PUF2 to reach the unfolded state must be higher than that between PUF1/PUF2 and the native state; otherwise, it is not rate-limiting. The relaxation time of PUF2 analog is about 25 μ s at 65 °C.²⁷ Therefore, it is almost impossible to observe folding to the native state from PUF1/PUF2 with regular MD simulations at 300 K within tens of microsecond, let alone to cross the rate-limiting barrier. Accordingly, regular MD simulations at 300 K within a few microsecond to 10 μ s time scale sample post-transition state region only.

Committer probability can be used to verify that our observed partially unfolded states occur after the rate-limiting step. States 1, 2, and 3 are close to experimentally measured structures of hidden intermediates that are present after the rate-limiting step. If state 8 occurs before the rate-limiting step, then the conformational transitions from state 8 to states 1, 2, and 3 must overcome the rate-limiting barrier. If it is after the rate-limiting transition state, then the barrier that

FIG. 3. Network of net effective flux that contributes to the overall transition from source (state 8) to sink (state 1). For clarity, the low population states (state 9-11) are neglected. The width of the edge is proportional to $\ln(Nf_{ij}^+)$, where N is the total number of conformations, $N = 132\,547$, and f_{ij}^+ is the net effective flux between the corresponding nodes. The edges with $Nf_{ij}^+ < 1$ are ignored for the sake of clarity. The color of each state is associated with the average Q value of the conformations in the corresponding state. Both states 1 and 2 can be considered as PUF2 from the secondary structure point of view, but more native contacts are present in state 1. For each state, the conformation with its Q closest to the average value of the state is chosen as a representative conformation of the state and depicted in the plot. Each residue of the representative conformations is color coded by Q_k which has the same color scheme as Q . The heavy-atom RMSD between the representative conformations of two states is presented above the corresponding edge. The legends of the size of each node, Q/Q_k , and the width of edge are shown in the figure. The value of f_{ij}^+ is given in Table I. The PDB file of the center of each state in the network is available in the supplementary material.⁶⁴

separates state 8 and states 1, 2, and 3 must be more native-like than the rate-limiting transition state. In the network (Fig. 3), the state with committor probability equal to 0.5 is the transition state between state 8 and PUF2, which means that the conformations in this state have equal chance to go forward to PUF2 or backward to state 8. If the state with committor probability equal to 0.5 has the same structural feature as that of the rate-limiting transition state which was characterized experimentally using Φ values,^{14,17} then this state is the rate-limiting transition state. If this state is more native-like than the rate-limiting transition state, then this state is post-transition and state 8 must be after the rate-limiting step, too. The committor probability of state 8 and PUF2 was set to zero and one, respectively.

The helical structures formed in the rate-limiting transition state were investigated experimentally by Ala/Gly pair mutations at 13 solvent-accessible positions.¹⁷ The rate-limiting transition state was estimated to have 51% native helical content. The calculated committor value of state 4 ($=0.43$) is the closest value to 0.5 among the states. The mean helical content of state 4 is 69%. We calculated the nativeness of the residues (Q_k) in state 4 and state 8 that have experimental Φ values^{14,17} in helix 2, helix 3, and the N-terminal part of helix 4. A Φ value near 0 suggests that the local

structure surrounding the residue is relatively unstructured in the transition state, while a value near 1 implies that the residue and its surrounding region already adopts native conformation in transition state. As shown in Fig. 4, most of the residues show very low experimental Φ value (<0.2), indicating that most of the residues are unstructured in the transition state. On the contrary, only four residues (M33, A40, F61, and I67) and seven residues show low nativeness (<0.6) in state 4 and state

FIG. 4. Residue nativeness (Q_k) for the residues with experimental Φ value in helix 2, helix 3, and the N-terminal part of helix 4 in state 4 and state 8. The black line denotes the experimental Φ values. The red line and blue line denote Q_k for state 4 and state 8, respectively.

FIG. 5. Relaxation time scale and the test of Markovianity. (a) Relaxation time scale as a function of lag time (τ). The five largest eigenvalues (λ) are used in the calculation of the relaxation time scale ($= -\tau / \ln \lambda$). (b) Chapman-Kolmogorov test of the Markovianity for the eleven states at different lag time. The errors of $T(k\tau)$ are shown as error bars. (c) Time evolution of the populations of state 1-8.

8 (Fig. 4), respectively, which suggests that the structures of most residues in these helical regions of state 4 and state 8 are near native. In conclusion, the structures of state 8 and state 4 are more native-like than the experimentally characterized transition state structure. Therefore, state 8 and other partially unfolded states are located after the rate-limiting step.

To validate the assumption of Markovian, we tested the convergence of the slowest relaxation time scales (Fig. 5(a)). Initially, the relaxation time scale sharply increases with the lag time. When the lag time is near 60 ns, the slowest relaxation time scale starts to level off. However, the constant slowest relaxation time scale after a certain lag time is a necessary but not sufficient condition of Markovian dynamics.⁵⁶ We, therefore, further validate the Markovianity using the Chapman-Kolmogorov test (Fig. 5(b)). If the k -step transition matrix with a lag time of τ , $[T(\tau)]^k$, where k is an integer, is equivalent to the transition matrix of a Markov state model with the lag time of $k\tau$, then the dynamics is Markovian with this lag time and with the partition of conformation space. We tested the lag time of 60 ns when the slowest relaxation time scale starts to level off and longer time at 120 ns and 240 ns for the discretization of state space resulted from the hierarchical approach. The equation of $[T(\tau)]^k = T(k\tau)$ holds within the statistical error for all the tested lag times (Fig. 5(b)). Therefore, our hierarchical approach of partition of conformation space is validated by the fact of constant relaxation time after a certain lag time and the satisfaction of Chapman-Kolmogorov test. Fig. 5(c) presents the time evolution of the populations of state 1-8, given that the initial population of state 8 is 1 and zero population for all other states. After about 1 μ s, the population of states reaches

statistical convergence. Using Eq. (4), the total time from state 8 to state 1 is 1.27 μ s. It is not clear why PUF1 and PUF2 are not the most populated states. Possibly, it is due to the fact that other partially unfolded states are harder to be resolved experimentally than PUF1 and PUF2 because their conformations are more flexible than PUF1 and PUF2.

We applied transition-path theory to calculate the net effective flux between any two states that makes contributions to the transition from source ($\equiv$ state 8) to sink ($\equiv$ state 1). In the network (Fig. 3), the nodes are the eight largest states and each edge weight is the net effective flux between the associated nodes. Note that the net effective flux is different from direct transitions. Two nodes could have a large amount of direct transitions observed in the simulations, but these transitions do not necessarily contribute to the transition from source to sink, namely, they could be misfolded states. In the network, the major routes (with large flux) are state 8 $\rightarrow$ state 5 $\rightarrow$ state 4 $\rightarrow$ state 1 or from state 4 via state 2 to state 1. The routes that via state 3 to state 1 are not major routes as shown in Table I because the flux from state 3 to state 1 is about one order of magnitude less than those from state 4/state 2 to state 1. Note that from the secondary structure point of view only, state 2 and state 1 are both considered as PUF2.

In the network shown in Fig. 3, from state 8 to state 5 the native contacts increase but no substantial growth in the secondary structure. From state 5 to state 4, the Q value (nativeness) increases to the range of 0.56–0.66 from less than 0.56, but helix 3 is still partially unfolded. Then state 4 can fold directly into PUF2 with well folded helices 2, 3, and 4, or fold to PUF1 with the C-terminus of helix 4 unfolded. Following a minor route, the conformations in state 5 can directly transit

TABLE I. Net effective flux (f_{ij}^+) between states. Because flux $i \rightarrow j$ is equal to flux $j \rightarrow i$, only half of the matrix is shown.

	S1	S2	S3	S4	S5	S6	S7	S8
S1	0	2.88×10^{-3}	2.93×10^{-4}	1.46×10^{-3}	2.42×10^{-5}	0	0	0
S2		0	4.38×10^{-4}	2.03×10^{-3}	3.22×10^{-4}	0	0	7.66×10^{-6}
S3			0	5.80×10^{-4}	1.33×10^{-4}	5.91×10^{-7}	0	5.41×10^{-6}
S4				0	3.76×10^{-3}	1.48×10^{-4}	1.22×10^{-4}	2.98×10^{-4}
S5					0	1.78×10^{-4}	6.60×10^{-5}	4.52×10^{-3}
S6						0	5.23×10^{-5}	2.87×10^{-5}
S7							0	4.06×10^{-6}
S8								0

to PUF1 or PUF2. It is clear from the network that the sequential transitions from state 8 to PUF1 then to PUF2 do contribute to the source-to-sink transition, but it is not the greatest contribution because one order of magnitude greater flux is observed from the direct transition to PUF2 without going through PUF1 (Table I). The protein can also be trapped in the misfolded states (states 6 and 7) after leaving state 5. The paths going through state 7 and/or state 6 make minimal contribution to the overall transition from state 8 to state 1.

Our network model is consistent with the funnel concept of parallel events^{57–59} as well as the recent observations using Markov state model.^{52,60} As schematically illustrated by Bryngelson *et al.* using a network model,⁶¹ in the denatured state, there could be many accessible paths leading to the native state. But these many different paths eventually merge to a few routes after the rate-limiting step which can be considered as a bottleneck of processes in the perspective of energy landscape. Since our focus is the post-transition intermediates, many parallel pathways initiated from the denatured state have already begun to merge. The emergence of folding pathways is vividly shown in the folding of PinWW⁵² and NTL9(1–39).⁶⁰ The folding flux of PinWW appears as many parallel pathways with crossovers, while the folding flux of NTL9(1–39) resembles a network with a major route to the native state guided by populated intermediate states. Both our work of Rd-

apocytochrome *b*₅₆₂ and NTL9(1–39) show the existence of folding networks. The organization of folding routes can also be represented by a tree structure.^{62,63} Networks allow loops between nodes, while trees do not. When greedy algorithms are used, the folded structures increase in every step, leading to a tree structure which excludes loops starting from a node and then back to the same node along the path.

Residue-residue contact maps

To analyze the roles of residue-residue interactions in the conformational transitions, we calculated the residue contact maps for the conformations in different intermediate states. The contact maps of state 1 and state 8 are shown in Figures 6(a) and 6(b), respectively. The contact maps of other states are shown in Fig. S5 in the supplementary material.⁶⁴ Comparing the contact map of state 1 with that of the NMR native structure, clearly, their patterns show high similarity while the low occupancy is seen in some regions, particularly, helix 1-helix 2 and helix 1-helix 4 contact regions, because of the partially formed helix 1. Nonnative contacts are nonnegligible, especially the numerous contacts between helix 2/part of helix 2-helix 3 loop and helix 4 which are observed in state 1 and also present in state 8. There are 11, 4, and 9 native inter-helical contacts for helix 2-helix 3, helix 2-helix 4,

FIG. 6. Residue-residue contacts. (a) and (b) Occupancy of residue-residue contacts for the conformations in state 1 and state 8, respectively. The contact map of the NMR native structure is shown in the lower triangle in both (a) and (b). The residue-residue contacts include the heavy atoms of both main chains and side chains. Red circles highlight the native contacts between F65 and helix 2 and helix 4. The helical regions are labeled at the top. The legend of native-contact occupancy (%) is shown in the figure. (c) The hydrophobic core centered at F65 in the native structure, including the residues within 4.5 Å of F65. For clarity, the sequence neighbors up to three residues apart of F65 are not shown. (d) F65 residue nativeness (Q_{65}) and the solvent accessible surface area (SASA) of the residues within 4.5 Å of F65 including F65 itself of state 1 to state 8.

and helix 3–helix 4, respectively. In state 8, 9 out of 11 helix 2–helix 3 native contacts and 4 out of 9 helix 3–helix 4 contacts are maintained (with >50% occupancy). And all native contacts between helix 2 and helix 4 are preserved state 8. However, the key contacts involving F65 in the hydrophobic core are absent in state 8 (see the red circles in state 1 and state 8 in Fig. 6).

In the native state, F65 and its surrounding residues form a hydrophobic cluster.²⁸ As shown in Fig. 6(c), F65 is buried in the hydrophobic core, surrounded by A37, A40, Q41, F61, V69, Y101, and K104. F65 acts as a key contact point to glue helix 2, helix 3, and helix 4 together, maintaining the compactness of the helix bundle. As shown in Fig. 6(d), it appears that F65 also plays a critical role in the folding. Like the native structure, F65 is in contact with A37, A40, Q41, F61, L68, V69, and Y101 in state 1 (PUF2), state 2 (PUF2), and state 3 (PUF1) (Fig. 6(d)). However, the nonlocal contacts ($>i + 4$) are partially lost in states 4 and 5 and nearly completely lost in state 8 (Fig. 6(d)). The decrease in Q_k indicates that the hydrophobic contacts around F65 are gradually lost starting from state 4, and meanwhile, the increase in the surface area of the residues within 4.5 Å of F65 shows that the hydrophobic core opens up (Fig. 6(d)). The disruption of the hydrophobic core accompanies the unwinding of the N-terminus of helix 3.

CONCLUSIONS

We performed more than 37 μ s MD simulations of the native state of Rd-apocytochrome b_{562} and the partially unfolded state (PUF2) after the rate-limiting step. The stability of secondary structures observed in our simulations is validated by the native state hydrogen exchange data. Although the PUF conformations are not observed to fold to the native state, in fact, they are not supposed to do so within our simulation time scale, the flux that represents the net effective contribution to the transition to the sink which is the most folded state in the network still shows the interrelationship between the different PUF states and has indications to the roles of the PUF states in folding. Clearly, there are major routes guided by some PUF states. The difference in flux between the major routes and minor routes is about an order of magnitude. The proposed pathway with PUF1 and PUF2 as sequential folding intermediates based on protein engineering and D-H exchange is one of the routes, but it is not the dominant one. The dominant path visits PUF2 without going through PUF1. There is also a route from PUF1 directly to the most folded state in the network without visiting PUF2. Our results indicate that the PUF states are not necessarily sequential in the folding. This observation is in line with the result of master equation.¹⁵ The major routes predicted in our network are testable by future experiments such as single molecule experiment. Our network model provides more detailed and complete description of the complexity of conformational changes between hidden intermediates than the simplistic pathway model. We also observed some misfolded states. Near native misfolded states may be common for all proteins. Therefore, stabilizing native state or preventing locally unfolding of the native state could be a general strategy against misfolding.

ACKNOWLEDGMENTS

We acknowledge the Anton Award Nos. MCB110061P and PSCA12013P through National Resource for Biomedical Supercomputing and Pittsburgh Supercomputing Center. Anton computer time was provided by the National Center for Multiscale Modeling of Biological Systems (MMBioS) through Grant No. P41GM103712-S1 from the National Institutes of Health and the Pittsburgh Supercomputing Center (PSC). The Anton machine at PSC was generously made available by D.E. Shaw Research. This work is funded by National Institutes of Health (Grant No. R01-GM088326) to S.H. M.D. is partially supported by National Science Foundation of China (Grant No. 21403291). We thank Dr. Yawen Bai for discussion.

- ¹C. Levinthal, *J. Chim. Phys.* **65**, 44–45 (1968).
- ²A. Ikai and C. Tanford, *Nature* **230**(5289), 100–102 (1971).
- ³T. Y. Tsong, R. L. Baldwin, and E. L. Elson, *Proc. Natl. Acad. Sci. U. S. A.* **68**(11), 2712–2715 (1971).
- ⁴Y. Bai, T. R. Sosnick, L. Mayne, and S. W. Englander, *Science* **269**(5221), 192–197 (1995).
- ⁵J. Cellitti, M. Llinas, N. Echols, E. A. Shank, B. Gillespie, E. Kwon, S. M. Crowder, F. W. Dahlquist, T. Alber, and S. Marqusee, *Protein Sci.* **16**(5), 842–851 (2007).
- ⁶J. Cellitti, R. Bernstein, and S. Marqusee, *Protein Sci.* **16**(5), 852–862 (2007).
- ⁷S. W. Englander, L. Mayne, and M. M. Krishna, *Q. Rev. Biophys.* **40**(4), 287–326 (2007).
- ⁸J. Rumbley, L. Hoang, L. Mayne, and S. W. Englander, *Proc. Natl. Acad. Sci. U. S. A.* **98**(1), 105–112 (2001).
- ⁹G. C. Rollins and K. A. Dill, *J. Am. Chem. Soc.* **136**(32), 11420–11427 (2014).
- ¹⁰N. P. Schafer, R. M. Hoffman, A. Burger, P. O. Craig, E. A. Komives, and P. G. Wolynes, *PLoS One* **7**(12), e50635 (2012).
- ¹¹A. N. Adhikari, K. F. Freed, and T. R. Sosnick, *Proc. Natl. Acad. Sci. U. S. A.* **109**(43), 17442–17447 (2012).
- ¹²B. F. Volkman, D. Lipson, D. E. Wemmer, and D. Kern, *Science* **291**(5512), 2429–2433 (2001).
- ¹³F. Chiti and C. M. Dobson, *Nat. Chem. Biol.* **5**(1), 15–22 (2009).
- ¹⁴R. Chu, W. Pei, J. Takei, and Y. Bai, *Biochemistry* **41**, 7998–8003 (2002).
- ¹⁵S. B. Ozkan, K. A. Dill, and I. Bahar, *Protein Sci.* **11**(8), 1958–1970 (2002).
- ¹⁶Y. Bai, *Biochem. Biophys. Res. Commun.* **305**, 785–788 (2003).
- ¹⁷Z. Zhou, Y. Huang, and Y. Bai, *J. Mol. Biol.* **352**, 757–764 (2005).
- ¹⁸D. M. Korzhnev, T. L. Religa, and L. E. Kay, *Proc. Natl. Acad. Sci. U. S. A.* **109**(44), 17777–17782 (2012).
- ¹⁹D. M. Korzhnev and L. E. Kay, *Acc. Chem. Res.* **41**(3), 442–451 (2008).
- ²⁰D. M. Korzhnev, T. L. Religa, W. Banachewicz, A. R. Fersht, and L. E. Kay, *Science* **329**(5997), 1312–1316 (2010).
- ²¹A. Borgia, D. Bonivento, C. Travaglini-Allocatelli, A. Di Matteo, and M. Brunori, *J. Biol. Chem.* **281**(14), 9331–9336 (2006).
- ²²H. Feng, Z. Zhou, and Y. Bai, *Proc. Natl. Acad. Sci. U. S. A.* **102**(14), 5026–5031 (2005).
- ²³T. L. Religa, J. S. Markson, U. Mayor, S. M. Freund, and A. R. Fersht, *Nature* **437**(7061), 1053–1056 (2005).
- ²⁴S. B. Whittaker, G. R. Spence, J. G. Grossmann, S. E. Radford, and G. R. Moore, *J. Mol. Biol.* **366**(3), 1001–1015 (2007).
- ²⁵H. Kato, H. Feng, and Y. Bai, *J. Mol. Biol.* **365**(3), 870–880 (2007).
- ²⁶H. Kato, N. D. Vu, H. Feng, Z. Zhou, and Y. Bai, *J. Mol. Biol.* **365**(3), 881–891 (2007).
- ²⁷T. Wang, Z. Zhou, M. R. Bunagan, D. Du, Y. Bai, and F. Gai, *Protein Sci.* **16**(6), 1176–1183 (2007).
- ²⁸H. Feng, J. Takei, R. Lipsitz, N. Tjandra, and Y. Bai, *Biochemistry* **42**(43), 12461–12465 (2003).
- ²⁹H. Feng, N. D. Vu, and Y. Bai, *J. Mol. Biol.* **343**(5), 1477–1485 (2004).
- ³⁰H. Feng, N. D. Vu, and Y. Bai, *J. Mol. Biol.* **346**(1), 345–353 (2005).
- ³¹H. Kaya and H. S. Chan, *Proteins* **58**(1), 31–44 (2005).
- ³²D. E. Shaw, R. O. Dror, J. K. Salmon, J. P. Grossman, K. M. Mackenzie, J. A. Bank, C. Young, M. M. Deneroff, B. Batson, K. J. Bowers, E. Chow, M. P. Eastwood, D. J. Ierardi, J. L. Klepeis, J. S. Kuskin, R. H. Larson, K. Lindorff-Larsen, P. Maragakis, M. A. Moraes, S. Piana, Y.

- Shan, and B. Towles, *presented at the Proceedings of the Conference on High Performance Computing, Networking, Storage and Analysis (SC09), Portland, OR, USA* (2009).
- ³³K. A. Beauchamp, R. McGibbon, Y. S. Lin, and V. S. Pande, *Proc. Natl. Acad. Sci. U. S. A.* **109**(44), 17807–17813 (2012).
- ³⁴K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, *Proteins: Struct., Funct., Bioinf.* **78**(8), 1950–1958 (2010).
- ³⁵J. Takei, W. Pei, D. Vu, and Y. Bai, *Biochemistry* **41**(41), 12308–12312 (2002).
- ³⁶Schrödinger LLC, Maestro: A Powerful, All-Purpose Molecular Modeling Environment, 2001, New York, NY.
- ³⁷W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, *J. Chem. Phys.* **79**, 926–935 (1983).
- ³⁸K. J. Bowers, E. Chow, H. Xu, R. O. Dror, M. P. Eastwood, B. A. Gregersen, J. L. Klepeis, I. Kolossváry, M. A. Moraes, F. D. Sacerdoti, J. K. Salmon, Y. Shan, and D. E. Shaw, *presented at the Proceedings of the ACM/IEEE Conference on Supercomputing (SC06), Tampa, Florida* (2006).
- ³⁹H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, *J. Chem. Phys.* **81**, 3684–3690 (1984).
- ⁴⁰R. A. Lippert, C. Predescu, D. J. Ierardi, K. M. Mackenzie, M. P. Eastwood, R. O. Dror, and D. E. Shaw, *J. Chem. Phys.* **139**(16), 164106 (2013).
- ⁴¹Y. Shan, J. L. Klepeis, M. P. Eastwood, R. O. Dror, and D. E. Shaw, *J. Chem. Phys.* **122**(5), 54101 (2005).
- ⁴²D. Frishman and P. Argos, *Proteins* **23**(4), 566–579 (1995).
- ⁴³W. Humphrey, A. Dalke, and K. Schulten, *J. Mol. Graphics* **14**, 33–38 (1996).
- ⁴⁴R. B. Best, G. Hummer, and W. A. Eaton, *Proc. Natl. Acad. Sci. U. S. A.* **110**(44), 17874–17879 (2013).
- ⁴⁵N. V. Buchete and G. Hummer, *J. Phys. Chem. B* **112**(19), 6057–6069 (2008).
- ⁴⁶E. H. Kellogg, O. F. Lange, and D. Baker, *J. Phys. Chem. B* **116**(37), 11405–11413 (2012).
- ⁴⁷K. J. Kohlhoff, D. Shukla, M. Lawrenz, G. R. Bowman, D. E. Konerding, D. Belov, R. B. Altman, and V. S. Pande, *Nat. Chem.* **6**(1), 15–21 (2014).
- ⁴⁸R. D. Malmstrom, C. T. Lee, A. T. V. Wart, and R. E. Amaro, *J. Chem. Theory Comput.* **10**(7), 2648–2657 (2014).
- ⁴⁹K. A. Beauchamp, G. R. Bowman, T. J. Lane, L. Maibaum, I. S. Haque, and V. S. Pande, *J. Chem. Theory Comput.* **7**(10), 3412–3419 (2011).
- ⁵⁰G. R. Bowman, X. Huang, and V. S. Pande, *Methods* **49**(2), 197–201 (2009).
- ⁵¹W. E and E. Vanden-Eijnden, *Annu. Rev. Phys. Chem.* **61**, 391–420 (2010).
- ⁵²F. Noe, C. Schutte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, *Proc. Natl. Acad. Sci. U. S. A.* **106**(45), 19011–19016 (2009).
- ⁵³H. Tjong and H. X. Zhou, *Biophys. J.* **98**(10), 2273–2280 (2010).
- ⁵⁴V. Munoz and L. Serrano, *Nat. Struct. Biol.* **1**(6), 399–409 (1994).
- ⁵⁵E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, *J. Comput. Chem.* **25**(13), 1605–1612 (2004).
- ⁵⁶J. H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schutte, and F. Noe, *J. Chem. Phys.* **134**(17), 174105 (2011).
- ⁵⁷P. E. Leopold, M. Montal, and J. N. Onuchic, *Proc. Natl. Acad. Sci. U. S. A.* **89**(18), 8721–8725 (1992).
- ⁵⁸P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, *Science* **267**, 1619–1620 (1995).
- ⁵⁹K. A. Dill and H. S. Chan, *Nat. Struct. Biol.* **4**(1), 10–19 (1997).
- ⁶⁰V. A. Voelz, G. R. Bowman, K. Beauchamp, and V. S. Pande, *J. Am. Chem. Soc.* **132**(5), 1526–1528 (2010).
- ⁶¹J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, *Proteins* **21**(3), 167–195 (1995).
- ⁶²J. Hockenmaier, A. K. Joshi, and K. A. Dill, *Proteins* **66**(1), 1–15 (2007).
- ⁶³J. Fan, M. Duan, D. W. Li, H. Wu, H. Yang, L. Han, and S. Huo, *Biophys. J.* **100**(10), 2457–2465 (2011).
- ⁶⁴See supplementary material at <http://dx.doi.org/10.1063/1.4931921> for Figures S1–S5 presenting the observation of convergence in the last step of equilibration, the clustering results solely based on RMSD, helicity of each residue in different trajectories, and the contact maps for the conformations in state 2 to state 7. PDB files of the center of cluster 1–8.