JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

# Mechanism of the All-$\alpha$ to All-$\beta$ Conformational Transition of RfaH-CTD: Molecular Dynamics Simulation and Markov State Model

Shanshan Li,[†,‡,∥] Bing Xiong,[†,∥] Yuan Xu,[†] Tao Lu,[‡] Xiaomin Luo,[†] Cheng Luo,[†] Jingkang Shen,[†] Kaixian Chen,[†,§] Mingyue Zheng,*[,†] and Hualiang Jiang*[,†,§]

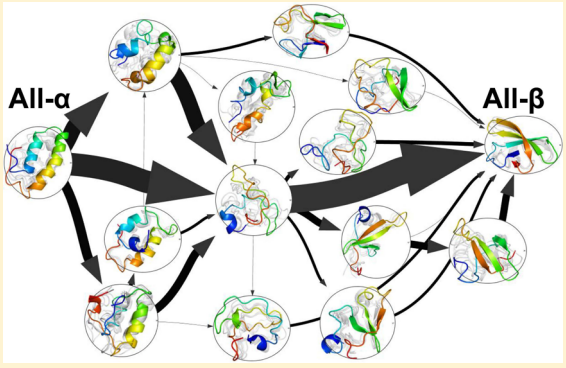
[†]State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai 201203, China

[‡]Laboratory of Molecular Design and Drug Discovery, School of Science, China Pharmaceutical University, 24 Tongjiaxiang, Nanjing 210009, China

[§]School of Life Science and Technology, Shanghai Tech University, Shanghai 200031, China

Supporting Information

**ABSTRACT:** The C-terminal domain of the bacterial transcription antiterminator RfaH undergoes a dramatic all-$\alpha$-helix to all-$\beta$-barrel transition when released from its N-terminal domain. These two distinct folding patterns correspond to different functions: the all-$\alpha$ state acts as an essential regulator of transcription to ensure RNA polymerase binding, whereas the all-$\beta$ state operates as an activator of translation by interacting with the ribosomal protein S10 and recruits ribosomal mRNA. Accordingly, this drastic conformational change enables RfaH to physically couple the transcription and translation processes in gene expression. To understand the mechanism behind this extraordinary functionally relevant structural transition, we constructed Markov state models using an adaptive seeding method. The constructed models highlight several parallel folding pathways with heterogeneous molecular mechanisms, which reveal the folding kinetics and atomic details of the conformational transition.




## 1. INTRODUCTION

The relationship between the molecular functions of proteins and their three-dimensional (3D) folds is one of the unanswered biological questions that has attracted extensive attention. Despite decades of research, protein folding continues to remain among the great challenges in biophysics.[1] Most proteins fold into a single dominant stable conformation via local fluctuations or concerted motions of entire subdomains. However, some proteins can adopt multiple stable conformations with respect to different cellular conditions. These interesting proteins are known as metamorphic proteins[2] and include the chemokine lymphotactin (Ltn),[3] the Mad2 spindle checkpoint protein,[4] and the chloride intracellular channel1 (CLIC1).[5]

Classically, a protein can adopt only one unique 3D structure dictated by its amino acid sequence to fulfill a specific function,[6] and misfolding of the protein may impede normal cellular functions. For some proteins, such as the prion protein, misfolding can even lead to a conformational infectious disease. In contrast, metamorphic proteins can fold into multiple stable conformations to adapt to different cellular conditions, providing a new perspective on protein folding. Recently, a study by Burmann et al.[7] on the transcription factor RfaH from *Escherichia coli* extended this view by demonstrating a particularly dramatic protein conformational transition. The full-length bacterial RfaH protein has been reported to exhibit an $\alpha$-helical hairpin in its C-terminal domain (CTD) that folds tightly against the N-terminal domain (NTD),[8] creating a closed conformation. These tight domain interactions are disrupted upon the binding of RfaH to the operon polarity suppressor (*ops*) site of the nontemplate DNA strand. When RfaH-CTD is released, it subsequently undergoes an unprecedented complete transition from an all-$\alpha$ to an all-$\beta$ conformation. As shown in Figure 1, the all-$\alpha$ state (domain-closed conformation) consists of two long antiparallel $\alpha$-helices that exhibit a hairpin structure, whereas the all-$\beta$ state (domain-open conformation) consists of a five-stranded antiparallel $\beta$-sheet with a strand order of $\beta_5$-$\beta_1$-$\beta_2$-$\beta_3$-$\beta_4$. Remarkably, in addition to enabling a stable $\beta$-barrel conformation, this dramatic structural rearrangement plays a key role in the control of gene expression by coupling transcription and translation. As an $\alpha$-helical hairpin, RfaH-CTD masks the RNAP-binding surface until RfaH recognizes the *ops* site, which is essential to avoid interference with gene regulation.[9] In the $\beta$-barrel conformation, RfaH-CTD interacts with the ribosomal

Received: March 18, 2014
Published: May 12, 2014

   dx.doi.org/10.1021/ct5002279 | J. Chem. Theory Comput. 2014, 10, 2255−2264

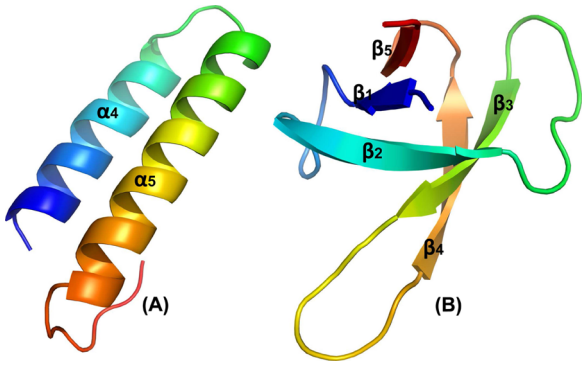



**Figure 1.** Comparison of the two conformations of RfaH-CTD: (A) the all-$\alpha$ state (domain-closed state, 2OUG) and (B) the all-$\beta$ state (domain-open state, 2LCL). These structures were rendered using PyMOL.[10]

protein S10 and recruits ribosomal mRNA, converting RfaH into a potent activator of translation. This conformational transition is biologically significant for RfaH-dependent operons: because these operons do not naturally possess canonical ribosomal binding sites, the transcription−translation coupling can substantially boost the protein output.

The complete $\alpha \rightarrow \beta$ structural transition of RfaH-CTD has significant implications for studies of the control of gene expression and protein folding. Clearly, additional proteins similar to RfaH must exist because this unique refolding ability endows proteins with a more efficient gene regulation mechanism. This paradigm is also useful in expanding the capabilities of protein engineering. Moreover, the refolding of RfaH provides a unique opportunity to study the kinetics and dynamics of the conformational conversion, which may shed light on diseases that involve misfolded proteins. It is therefore of great interest to investigate the atomic-level mechanism of the transition process. In particular, the following aspects should be considered: the folding mechanisms of the $\alpha \rightarrow \beta$ transition, the chemical details underlying the complete refolding of the CTD, the CTD refolding kinetics and time scale, and the one-sequence multistructure relationship of metamorphic proteins. One of the central goals of molecular biophysics is to understand the transition mechanisms between functionally important states of proteins.[11] Although direct molecular dynamics (MD) simulation is well suited for this aim, reaching biologically relevant time scales at an atomic level of detail continues to pose a great challenge for such a dramatic conformational rearrangement.[12]

Numerous theoretical models have been developed to solve these rare-event simulation problems.[13] The Markov state model (MSM) approach has emerged as an efficient solution and can be used to construct a statistical description of protein folding.[14,15] MSMs utilize large-scale statistical sampling to construct a comprehensive model of folding, which has been demonstrated to yield quantitative agreement with experimental results and has revealed the hub-like character of protein native states.[16] In addition to solely relying on geometric criteria as in several other clustering techniques, MSMs constitute a kinetic clustering of simulation data,[11,14] which means that rapidly interconverting conformations are grouped within a single state, and slowly interconverting conformations are grouped into separate states. The Markov transition matrix can then be constructed from several short simulations that only reach local equilibration,[17] allowing the extraction of thermodynamic and kinetic parameters. Although no single trajectory visits every state, MSMs demonstrate the unique strength of capturing long-time scale dynamics and reconstructing the folding pathway by taking advantage of overlap among these short simulations. Therefore, the problem of predicting long time scale behavior is transformed into sufficiently sampling a series of short trajectories and calculating the transition probabilities between states. Furthermore, MSMs can provide feedback on under-sampled areas of protein phase space. Simulations can be more appropriately allocated to increase sampling through a process called adaptive sampling, resulting in less time and lower resource cost. Furthermore, MSMs do not require the predefinition of reaction coordinates and are, thus, less prone to a biased or oversimplified view of the kinetics.[18] To date, MSMs have been successfully employed to investigate the folding mechanisms of small RNA hairpins (8-nucleotide RNA)[19] and several protein systems.[20−26]

In this study, we investigated the folding mechanism of the dramatic conformational transition of RfaH-CTD from the all-$\alpha$ state to the all-$\beta$ state. High-temperature MD simulations and nonequilibrium simulations using targeted molecular dynamics (TMD) were first performed to identify the metastable states. These states were then used as seeds to initiate equilibrium simulations using conventional molecular dynamics (CMD). Based on the adaptive seeding strategy,[17] several rounds of adaptive sampling were performed to improve the MSMs. In total, a set of 1334 CMD simulations of RfaH-CTD with an aggregate simulation time of approximately 200 $\mu$s were obtained, which were used to build the final Markov state model and determine the full ensemble of transition paths of RfaH-CTD and their fluxes. The results not only demonstrate stable intermediate conformations in a statistically significant manner but also detail the sequential formation of secondary structures with a variety of different transition paths. More importantly, this study provides insights into the conformational transition between functionally important states of biomolecular systems.

## 2. METHODS

**2.1. Structure Preparation.** Two crystal structures of the C-terminal domain of RfaH were individually obtained from the Protein Data Bank (PDB),[27] of which the PDB accession numbers for the all-$\alpha$ and all-$\beta$ structures are 2OUG[8] (residues 115−156) and 2LCL[7] (residues 115−162), respectively. The terminal residues 157−162 of the all-$\alpha$ structure were added using the Prime module of Schrödinger 2010.[28]

**2.2. Simulation Overview.** The simulation workflow consists of five steps, as shown in Figure 2. Additional details on each step are provided in the following subsections.

Step 1. Conformational Sampling. To adequately sample the conformational space surrounding the transition, two seeding simulation strategies were employed, including TMD seeding and high-temperature MD seeding.

Step 2. Molecular Dynamics Simulation. To reduce the bias caused by TMD and MD simulations under high temperature, short unbiased MD simulations at constant temperature (300 K) were performed using the previously obtained configurations as starting points. Simultaneously, long-time MD simulations beginning from the $\alpha$ state (2OUG) and the $\beta$ state (2LCL) were also performed, providing an additional valuable source of simulation trajectories.

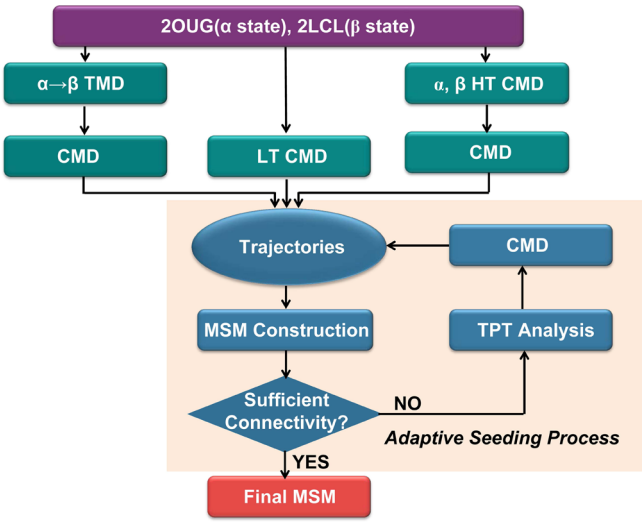



**Figure 2.** Overview of the simulation workflow. TMD, targeted molecular dynamics simulation; CMD, conventional molecular dynamics simulation; LT, long time; HT, high temperature; MSM, Markov state model; TPT, transition path theory.

Step 3. MSM Construction. The MSM approach was employed to extract equilibrium properties from the obtained simulation trajectories and to build candidate MSMs.

Step 4. Transition Path Theory Analysis. The transition path theory (TPT) approach was employed to investigate the constructed candidate MSMs and to identify the transition pathways with high folding flux.

Step 5. Adaptive Seeding Methods. To improve the quality of the resultant MSMs, an adaptive seeding strategy was used to allocate additional simulations to the metastable states along the transition pathway identified by TPT analysis. This strategy may reduce the computational resources required to build MSMs with sufficient connectivity,

**2.3. Conformational Sampling.** *Targeted Molecular Dynamics Simulations.* TMD is a method of inducing a conformational change to a known target structure by applying a time-dependent geometrical constraint. This method was employed to characterize the conformational transition from the all-$\alpha$ state to the all-$\beta$ state, with the all-$\alpha$ state as the initial structure and the all-$\beta$ as the target reference structure. In total, 200 parallel targeted MD simulations were performed with the Sander module of AMBER (version 10.0) using a force constant of 0.4 kcal mol$^{-1}$ Å$^{-2}$, a simulation time of 10 ns, and a constant temperature of 300 K. The restraint force was applied to all atoms in the initial all-$\alpha$ structure to bias the trajectories toward the target all-$\beta$ reference structure. Disregarding three outliers, the remaining 197 trajectories were clustered using MSMBuilder based on the RMSD among the CA, CB, C, N, and O atoms, for which the RMSD threshold for clustering was set to 3.9 Å. In total, 401 clusters were obtained, and the cluster centers were then extracted, yielding a total of 401 seed conformations.

*High-Temperature Molecular Dynamics Simulations.* Increasing the temperature of molecular dynamics simulations is an efficient approach to explore the accessible conformational space for pathways, which allows the possibility of sampling all possible minimum conformations along the transition without becoming trapped in a local energy minimum.[22] Beginning from each functional state, i.e., all-$\alpha$ or all-$\beta$, three 100 ns canonical MD simulations were performed at temperatures of 370 K. Snapshots were then extracted every 4 ns from the six 100 ns trajectories, providing a total of 150 seed conformations.

**2.4. Molecular Dynamics Simulation.** All of the MD simulations were performed using the Amber software package (version 10.0).[29] Specifically, the Amber ff99SB[30,31] force field was applied to the proteins, and the generalized Born/surface area (GBSA) implicit solvent model[32] was used to represent the solvent effects. During the simulation, all of the bonds involving hydrogen atoms were constrained using the SHAKE algorithm[33] with the integration step set to 2 fs. The nonbonded cutoff was set to 999.0 Å. The nonbonded pairs were updated every 25 steps, and the protein conformations were written every 20 ps. The temperature for all equilibrium production runs was set to 300 K.

**2.5. Markov State Model Construction.** The MSMs for RfaH-CTD were constructed using the MSMBuilder package (version 2.5)[14,34] and the following steps: (1) conversion of the trajectory data to MSMBuilder format and alignment of all configurations saved from the trajectories to the native structure 2OUG (via RMSD fit on the backbone atoms) to remove translation and rotation of the entire system; (2) clustering of the conformations to divide the geometrically similar conformations into small sets called microstates; (3) calculation of the transition probability matrix of these microstates and plotting the implied time scale plots to determine the lag time; and (4) use of the PCCA+ algorithm[35] implemented in MSMBuilder to combine kinetically related microstates into metastable states (macrostates). More methodology details of MSM are provided in the Supporting Information (SI).

**2.6. Transition Path Theory Analysis.** TPT is a theoretical framework to analyze folding pathways with high reactive flux between conformations, which can be used to gain insight into the microscopic complexity of protein folding.[20,36,37] Given a MSM, the ensemble of protein folding pathways and their probabilities are computed based on TPT. First, two subsets of the state space U (unfolded) and F (folded) are assigned to specify the transition process of interest, whereas the remaining states are considered "intermediate" states. Subsequently, questions concerning the folding pathway can be answered by determining the typical sequence of "intermediate" states along the transition from U to F. In this study, U and F correspond to the states that most closely resemble the known all-$\alpha$ and all-$\beta$ crystal structures, respectively. The script "DOTPT.py" in MSMBuilder was used to calculate the transition path. The $P_{fold}$ value was computed from the macrostate transition matrix,[20] and the application MSMExplorer[38] was used to provide a network visualization of the resultant MSMs.

**2.7. Adaptive Seeding Strategy.** To improve the initial MSM constructed by the aforementioned seeding trajectories, the adaptive seeding method[17] was employed to incorporate microstate characteristics into an estimation of the conformational space of the transition. This method involves the following steps: (1) construction of an initial MSM using the TMD and high-temperature CMD seeding trajectories; (2) calculation of the folding pathways based on the constructed MSM using TPT; (3) extraction of the cluster center and two random conformations from each microstate of the top three transition pathways (i.e., pathways with the highest folding flux) and seeding new simulations from these points; and (4) combination of these trajectories with the previous trajectories to construct a new MSM. Steps 2−4 are repeated until the

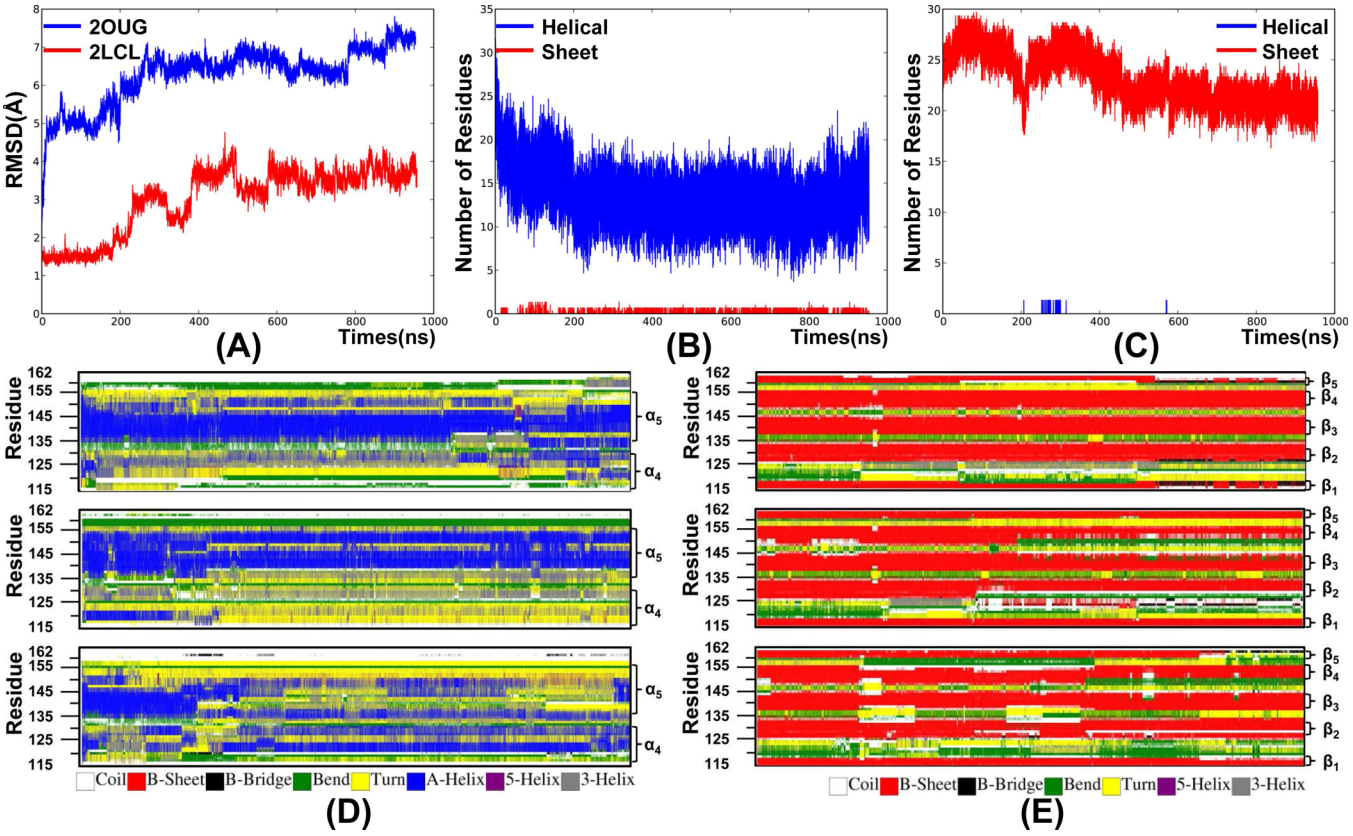



**Figure 3.** (A) Time dependencies of the RMSD averages of the three LT MD trajectories started from domain-closed state (blue) and domain-open (red) state. The backbone RMSD values were monitored throughout the MD process using the first frame as the reference. (B) The average number of $\alpha$-helical and $\beta$-sheet residues during three LT MD simulations of the domain-closed state. (C) The average number of $\alpha$-helical and $\beta$-sheet residues during three LT MD simulations of the domain-open state. (D) Secondary structures as a function of time for three LT MD simulations of the domain-closed state. (E) Secondary structures as a function of time for three LT MD simulations of the domain-open state. The corresponding secondary structures were also labeled in D and E.

connectivity of the constructed MSM is sufficient. Here, the termination criterion was set to the percentage of conformations within the nonconnected microstates reaching less than 5%.

**2.8. Structural Analysis of Macrostate Ensembles.** Because the macrostate conformational ensemble can be heterogeneous and diffuse, we used the metric $Q$-value defined by Voelz et al. to quantify the extent of its similarity to a native state.[23] Given a protein structure with $m$ residues, a, $m \times m$ matrix $\mathbf{C}$ was created to represent the contact profile of each macrostate, where the element $C_{i,j}$ represents the fraction of the conformation ensemble whose $\alpha$-carbons of the $i$th and $j$th residues are closer than 7 Å. The $Q$-value of a given macrostate is then calculated as its projection onto the "native" crystal structure (i.e., the all-$\alpha$ or all-$\beta$ native state), $C_{\text{nat}}$.

$$Q = \frac{C \cdot C_{\text{nat}}}{C_{\text{nat}} \cdot C_{\text{nat}}} \quad (1)$$

In this study, we calculated the $Q_\alpha$ value by projecting onto the all-$\alpha$ native state. To examine the secondary structural changes along the folding reaction, we defined six $Q$-values for specific structural elements in the all-$\alpha$ and all-$\beta$ native states, as described in section 3.3.C. We also plotted the contact profile for each macrostate based on the previously defined matrix $\mathbf{C}$ to analyze the secondary structure characteristics of different macrostates.

## 3. RESULTS AND DISCUSSION

**3.1. The Domain-Open State Is More Stable than the Domain-Closed State.** Detailed studies of proteins that can reversibly fold between structures have enabled the general description of several fundamental features of refolding mechanisms, including measurements of secondary structure formation and folding rates.[1] Unfortunately, conventional MD simulations are unable to capture the conformational transition of RfaH-CTD, as reflected by the three approximately 1-$\mu$s-long time MD simulations performed on the crystal structures of both the all-$\alpha$ state (2OUG) and the all-$\beta$ state (2LCL). However, from these simulations, we were able to determine that the domain-open state of RfaH-CTD is more stable than the domain-closed state. As shown in Figure 3A, the domain-closed all-$\alpha$ state exhibits a sudden increase in the backbone RMSD values at the beginning of the simulations, and the largest RMSD value is 7.5 Å. In comparison, the domain-open all-$\beta$ state is more stable, exhibiting RMSD values below 2 Å for the first 200 ns and a largest RMSD value of only 4 Å. Analyzing all three long-simulation trajectories of the all-$\alpha$ state, we observed that helix 4 is always the first helix to shorten, which is consistent with results from a solution nuclear magnetic resonance (NMR) analysis reported by Burmann et al.[7] In the crystal structure, helix 4 of RfaH-CTD is stabilized by packing interactions with the NTD but is intrinsically unstable, as the chemical shifts for K102-G121 are close to the random coil value. Thus, the tail of helix 4 exhibits high conformational


dx.doi.org/10.1021/ct5002279 | J. Chem. Theory Comput. 2014, 10, 2255−2264

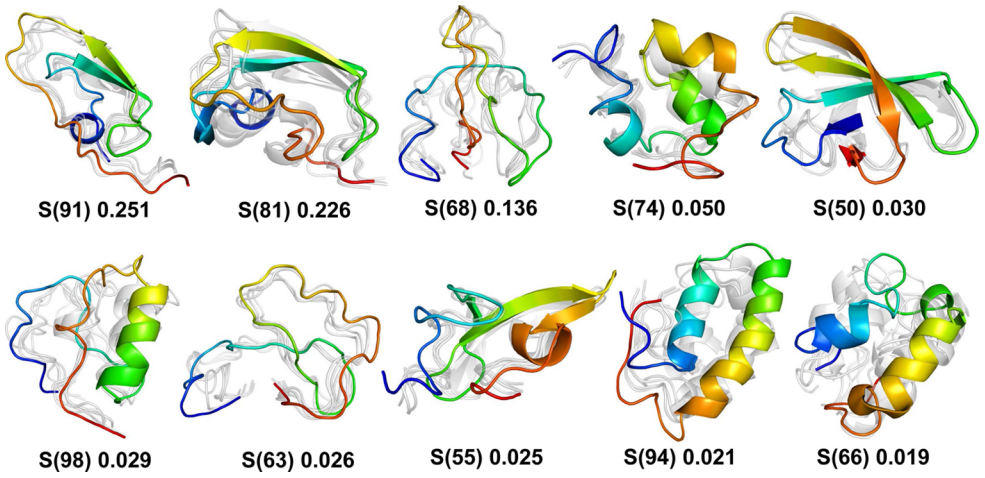



**Figure 4.** The 10 most populated macrostates from our coarse-grained MSM with their equilibrium populations.

flexibility and was destabilized upon separation from the NTD, leading to a strong propensity to adopt an unstructured conformation.

Residues located in a hydrogen bond network are commonly validated for their roles in protein stability. Table S1 summarizes the hydrogen bonds and their occupancies during the long-time MD simulations. Clearly, the three all-$\alpha$ simulations exhibit fewer hydrogen bonds and lower occupancies. There are six shared pairs of hydrogen bonds in the three all-$\beta$ simulations (shown in bold), whereas for the all-$\alpha$ simulations, no common hydrogen bonds exist. Figure 3B and C show the average number of $\alpha$-helical and $\beta$-sheet residues calculated using DSSP[39] during the simulations. For the all-$\beta$ simulations, the number of $\beta$-sheet residues varies between 17 and 30 with an average value of 23. In contrast, the number of $\alpha$-helix residues varies more widely for the all-$\alpha$ simulations. We also calculated the transformation of secondary structure along the trajectories of the all-$\alpha$ and all-$\beta$ simulations. As shown in Figure 3D and E, the profile of the secondary structure transformation indicated that in the all-$\alpha$ simulations, helix 4 and the end of helix 5 were less stable and demonstrated a propensity to unwind during the simulations. In contrast, in the all-$\beta$ simulations, with the exception of strands 1 and 5 located at the termini of the protein, the $\beta$-strands are stable and remain intact during the simulations. Overall, both the hydrogen bond analysis and the secondary structure analysis confirmed that the all-$\alpha$ state is less stable, which indicated that when released from RfaH-NTD, RfaH-CTD demonstrates the propensity to spontaneously fold into the all-$\beta$ state. Additionally, we analyzed the distribution of hydrophobic residues, and this analysis is described in the SI.

**3.2. Construction of the Final Markov State Model.** Although the LT MD simulations revealed the instability of the all-$\alpha$ state, they could not describe the details of the transition from the all-$\alpha$ to the all-$\beta$ state. To overcome the time scale limitations of the CMD simulations, we constructed MSMs based on the physical simulations. In addition to the six approximately 1-$\mu$s MD simulations initiated from both crystal structures, we run TMD simulations to generate the initial conformational transition pathways and HT MD simulations to enrich the conformational space. In this study, 401 seed conformations extracted from the TMD simulations and 150 conformations from the HT simulations were collected, and for each, a 150 ns unbiased MD simulation at 300 K was performed. In total, 557 unbiased MD trajectories with an aggregate simulation time of approximately 88 $\mu$s were obtained for the construction of the first MSM. However, the connectivity of this MSM is less than favorable: 223 of the 774 microstates were excluded due to disconnection, which accounts for 15% of all of the conformations. To further improve the quality of the MSM, 10 rounds of adaptive seeding followed by CMD simulations were performed, which yielded a total of 777 trajectories, each performed for 150 ns at 300 K. The final simulation data set contains 1334 implicit MD trajectories with an aggregate simulation time of approximately 200 $\mu$s. Because the beginning of a trajectory may be biased by an off-equilibrium starting condition, the first 2 ns were removed, leaving approximately 10 million total snapshots. It should be noted that none of the TMD trajectories or high temperature MD trajectories were included in constructing MSMs.

The resultant conformations were clustered into 815 microstates using the hybrid k-centers k-medoids clustering methods,[34] and 131 disconnected microstates were removed to avoid noise and insufficient sampling. The connectivity of the resultant MSM was significantly improved, with only 5% of the total conformation snapshots excluded. The resultant kinetically connected, well-populated 684 microstates possess an average radius of approximately 2.2 Å, which ensures that the equilibrium properties were preserved and maximizes the resolution of the model. A transition count matrix and the corresponding transition probability matrix were then constructed based on the 684 microstates. Subsequently, the implied time scale plots were calculated for lag times from 1 to 32 ns at 4 ns intervals, which were found to plateau at approximately 12 ns, suggesting a 12 ns Markov time.[40] Finally, the kinetically related microstates were combined into 100 macrostates using the PCCA+ algorithm, and the resultant 100-macrostate MSM was used to analyze the distribution of folding pathway fluxes from the all-$\alpha$ state to the all-$\beta$ state. The details of the MSM validation can be found in the SI.

**3.3. Insights into the Folding Mechanisms.** *A. Analysis of the 10 Most Populated States.* To gain a mesoscopic view of the folding pathway, we combined our 684 microstates into a MSM with 100 macrostates with sufficiently high resolution to yield more detailed insights into the folding mechanism. Figure 4 shows the 10 most populated states from our coarse-grained MSM, listed in descending order of their equilibrium


dx.doi.org/10.1021/ct5002279 | J. Chem. Theory Comput. 2014, 10, 2255−2264

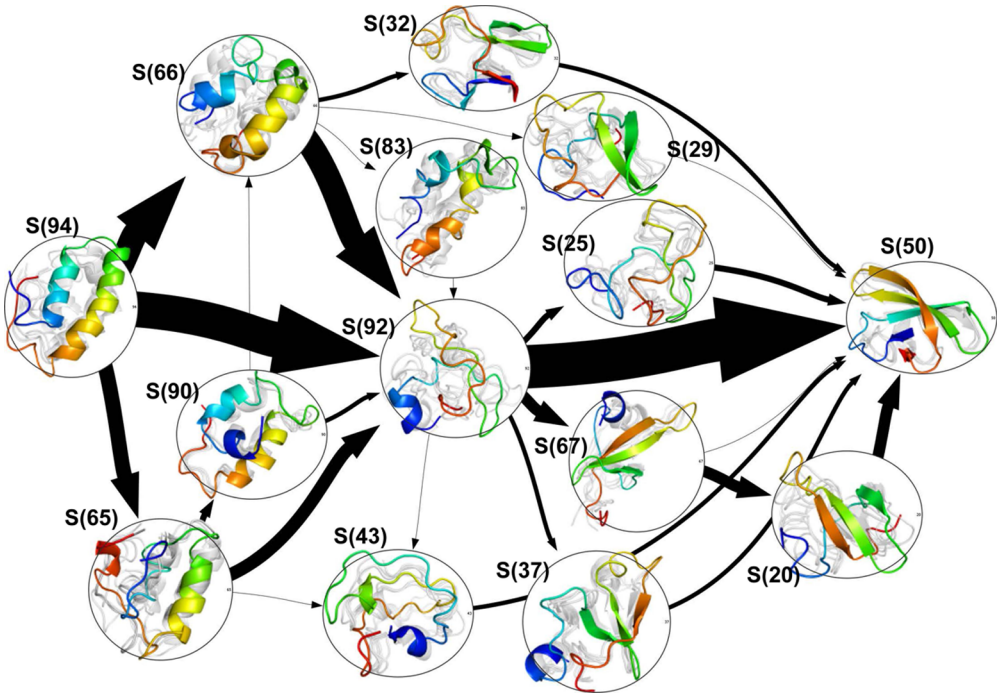



**Figure 5.** Superposition of the top 10 folding fluxes calculated using TPT. The representative structures were calculated from five random conformations from each state, and the arrow size is proportional to the interstate flux. The flux percentages for the top 10 pathways are 53.58%, 26.07%, 9.42%, 3.16%, 1.52%, 0.98%, 0.73%, 0.55%, 0.41%, and 0.01%, respectively.

probabilities. First, none of these states resemble the all-$\alpha$ native state. The macrostate **S(94)** exhibits a high ratio of helical content, from which we can observe that both ends of the two helices collapse to some extent. This helix-rich macrostate has a relatively small population of 0.021, again, confirming the instability of the all-$\alpha$ state inferred from our LT MD simulations. The macrostate **S(50)** most resembles the all-$\beta$ native state, which exhibits the highest ratio of $\beta$-sheet content. In comparison to the domain-open state crystal structure, the conformation of strands 1 and 5 fluctuated in our model, whereas the remaining three strands were similar to their native state conformations. For the states **S(66)**, **S(74)**, and **S(98)**, most of the helical structures are unwound, and both ends of the structures are curled. Because the conformational transition is from the all-$\alpha$ to the all-$\beta$ state, this result suggests that "helix unwinding" should represent the first step of the process, and these three macrostates may represent important intermediate states during the onset of the conformational transition. The macrostates **S(55)**, **S(63)**, and **S(68)** exhibit folding patterns that are similar to **S(50)**, suggesting that these three states may represent intermediate states late in the conformational transition. The macrostates **S(91)** and **S(81)** have the highest populations. One common feature of these macrostates is the formation of strands 2 and 3, and considering the large populations, we can speculate that strands 2 and 3 are relatively stable. Furthermore, in **S(68)**, the residues at the positions of strands 3 and 4 are very close, and in **S(55)**, the corresponding residues have already adopted a $\beta$-sheet structure, suggesting that strands 3 and 4 are also stable and exhibit the propensity to fold earlier. Notably, Figure 4 indicates that most of the populated states in our model exhibit a significant proportion of random coil content, suggesting its essential role in the dramatic all-$\alpha$ to all-$\beta$ conformational transition.

*B. How Does RfaH-CTD Fold in Our Simulations?* One of the most important strengths of MSMs is their ability to extract interpretable details of the folding mechanism from the coarse-grained macrostates. Figure S6 shows the 100-macrostate MSM built from the 684 microstates using a lag time of 12 ns. From this view, we can determine that the macrostates are diffuse collections of conformational states, including both the near all-$\alpha$ domain-closed macrostate and the near all-$\beta$ domain-open macrostate. Clearly, the diffuse collections of the conformational states of the 100 macrostates indicated that the all-$\alpha$ state transforms to the all-$\beta$ state through multiple parallel pathways. Next, we can perform an in-depth analysis of these pathways by considering protein kinetics as a set of states and the rates of exchange between these states.

The characterization of protein folding mechanisms is typically performed in terms of describing the order of events. Figure 5 shows the 10 pathways with the highest flux calculated from the macrostate transition matrix using TPT, indicating that the transition process proceeds through multiple parallel pathways. These top-ranked pathways involve 14 of the 100 macrostates and account for approximately 96.44% of the total flux. **S(94)** and **S(50)**, which are the macrostates that most resemble the all-$\alpha$ and all-$\beta$ crystal structures, were chosen as the initial and final points of the transition pathways, respectively; these macrostates are also among the top 10 most populated macrostates shown in Figure 4, suggesting that proteins tend to transition into near native states. Among these macrostates, **S(66)**, **S(90)**, and **S(65)** lie at the beginning of the transition process, and **S(66)** is also among the top 10 macrostates, confirming the speculation that "helix unwinding" is the first step during the conformational transition. The central macrostate **S(92)**, which is composed of "random coil" residues at the position of strands 2−4 and "helix" residues at the end of helix 4, may represent an important metastable state. The different transition processes are completed via different


dx.doi.org/10.1021/ct5002279 | J. Chem. Theory Comput. 2014, 10, 2255−2264

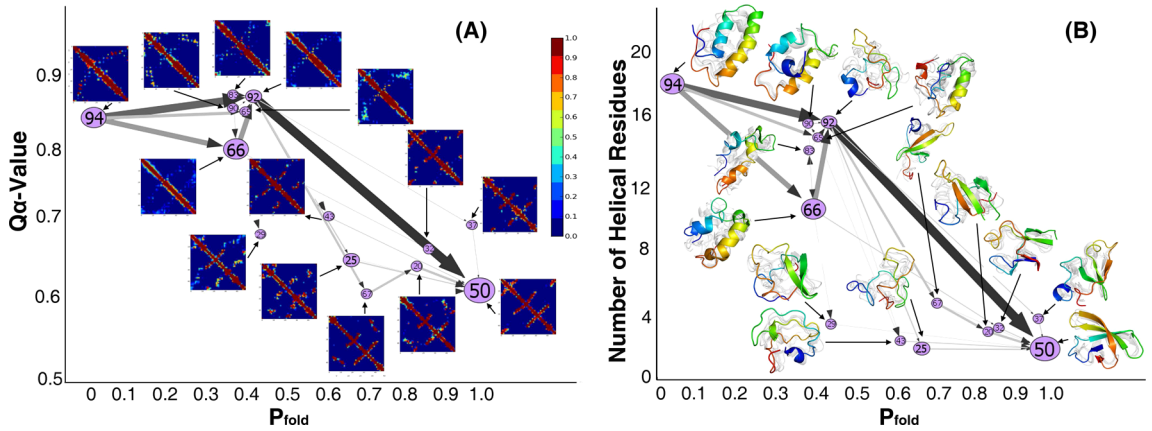



**Figure 6.** The 14 macrostates involved in the top 10 folding pathways plotted against the $P_{fold}$ values (horizontal axis). (A) The vertical axis represents the $Q_\alpha$ value, which is calculated using the all-$\alpha$ native state as a reference, and the 14 macrostates are represented as the contact profiles calculated from 100 randomly selected conformations from each macrostate. (B) The vertical axis represents the average number of $\alpha$-helical residues, and the 14 macrostates are represented as the representative structures aligned from five randomly selected conformations from each macrostate.

states, including **S(20)**, **S(29)**, **S(32)**, and **S(37)**. However, the topologies (folding patterns) of these states all resemble the topology of **S(55)** and **S(63)**, highlighting the importance of a "similar folding pattern" during the late stage of the transition. Furthermore, strands 2−4 all have already folded in these states, again, confirming our speculation that these three strands are relatively stable and form earlier during the transition process.

Figure 6A shows change of the secondary structure formed in a given state versus its position along the reaction (conformation transition) coordinate, where $Q$ is a structural metric defined to quantify the likeness of a macrostate to the initial all-$\alpha$ state, and $P_{fold}$ is a kinetic metric computed from the macrostate transition matrix[20,36,41] to quantify the probability of folding toward all-$\beta$ before folding toward all-$\alpha$ macrostates. Overall, this plot revealed that the possible pathways for folding are heterogeneous and may involve different ordering of native-like $\alpha$-helix to $\beta$-sheet transition. There are some macrostates showing split $Q$ values around a specific position of the transition coordinate, e.g., **S(83)**, **S(90)**, **S(65)**, and **S(92)** located around the $P_{fold}$ value of 0.35. This observation suggested that these states may contain similar amount of native-like helical elements when their reaction progress of corresponding pathways is around 30−40%. However, these states have diverse secondary structures, as reflected by the associated contact maps. Generally, across the 14 macrostates studied, the $Q_\alpha$ value decreases as $P_{fold}$ increases, suggesting that during the conformational transition, the structures gradually lose their helical character and become more and more $\beta$-sheet-like. Along the primary pathway, the contact profiles remain highly similar to the all-$\alpha$ native state during the early phase of the folding transition, indicating that the protein begins to unfold while retaining some of its secondary structure during this stage. During the later stage of the folding transition, the contact profiles with an all-$\alpha$ conformation rapidly decline, suggesting that the protein may undergo a relatively rapid "helix-unfolding" process. During the final stage of the folding transition, the contact profiles gradually resemble the all-$\beta$ state, and strands 2−4 gradually gain their secondary structural features. These three stages approximately describe the conformational transition process. In other parallel pathways, although some macrostates may share similar $P_{fold}$ and $Q_\alpha$ values, their interconversion is significantly different. For some pathways, the secondary structural change from all-$\alpha$ to all-$\beta$ is even multidirectional, e.g., the pathways of **S(94)** → **S(66)** → **S(92)** and **S(92)** → **S(67)** → **S(20)**. The existence of these alternative pathways indicates that the folding of RfaH-CTD is a complex process that may involve a relatively rough energy landscape.

To present more detailed information on the secondary structural changes along the folding reactions, we also calculated the average number of $\alpha$-helical residues (calculated from 100 random conformations from each of these 14 macrostates). Figure 6B shows the correlation between the number of helical residues and the $P_{fold}$ value, which indicates that when the $P_{fold}$ value exceeds 0.4, the number of $\alpha$-helical residues sharply decreases, and when the $P_{fold}$ value approaches 0.85, there are almost no $\alpha$-helical residues. The contact profiles for these 14 macrostates were also plotted, as shown in Figure S7, and each contact profile was calculated from 100 random conformations from each state. The contact profiles for the two native states are shown in Figure S8 as a reference, from which we can observe that the contact profile of **S(50)** resembles the contact profile of the all-$\alpha$ state, and the contact profile of **S(94)** resembles the contact profile of the all-$\beta$ state.

*C. Folding Rate and Rate-Limiting Step.* After clarifying the folding mechanism, we now turn to the question of folding-time prediction. In comparison to traditional methods,[42] in which the prediction of the folding rate heavily depends on predefined states, the MSM provides an unambiguous solution to the folding problem. By analyzing the transition matrix of the MSM, we can obtain information on the dynamics of our system. In this study, the predicted time scale of our system is approximately 0.1 s. As shown in Figure S5, this predicted time scale was preserved in the bootstrap analysis, indicating a robust feature of the RfaH-CTD system. This relatively long time scale may be explained by the conformational exchange between the crystallographic all-$\alpha$ state and the compact all-$\beta$ state through multiple parallel pathways.

To identify the structural features that limit the overall folding rate, we further defined several $Q$ values for important structural elements by restricting the contact profile to a particular subspace of contacts. For example, $Q_{\beta 12}$ is the $Q$ value when $C$ is restricted to a subspace where $x \in \beta_{12}$, $x$ is a set of


dx.doi.org/10.1021/ct5002279 | J. Chem. Theory Comput. 2014, 10, 2255−2264

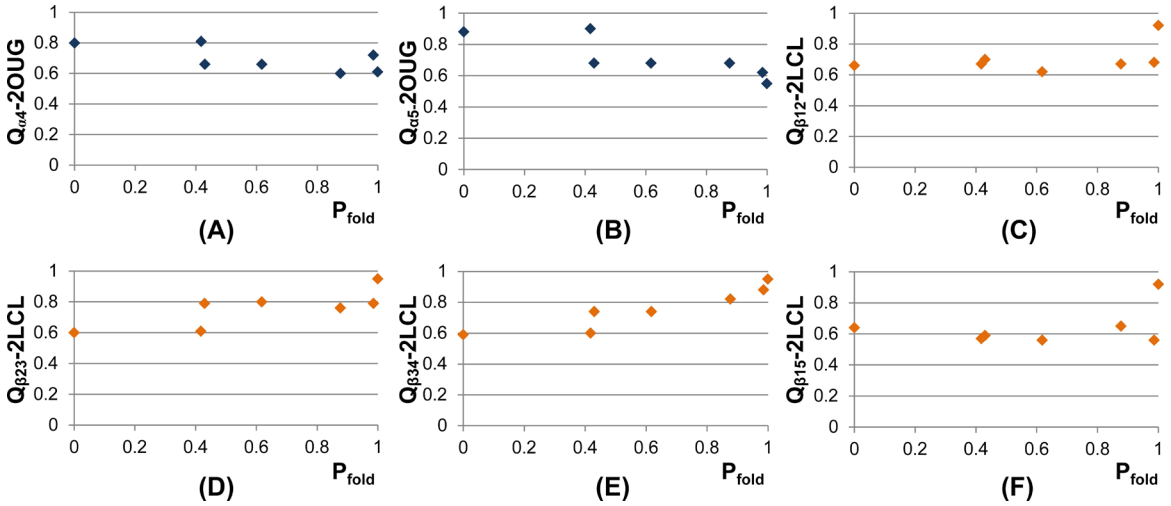



**Figure 7.** Six defined $Q$ values that quantify the extent of the similarity of the structures to a native state plotted against $P_{fold}$ values.

contact residues corresponding to the parings between $\beta$ strands $\beta_1$ and $\beta_2$. In total, we examined six $Q$ values for seven native structural elements: $Q_{\alpha4}$, $Q_{\alpha5}$, $Q_{\beta12}$, $Q_{\beta23}$, $Q_{\beta34}$, and $Q_{\beta15}$. The reference structure for $Q_{\alpha4}$ and $Q_{\alpha5}$ is the all-$\alpha$ native structure 2OUG, and the reference structure for $Q_{\beta12}$, $Q_{\beta23}$, $Q_{\beta34}$, and $Q_{\beta15}$ is the all-$\beta$ native structure 2LCL. Figure 7 plots the six individual $Q$ values, which quantify the extent of native-like structures versus $P_{fold}$ values. For clarity, seven scattered points corresponding to the seven macrostates on the top five pathways with higher flux are shown in each $Q$ plot. Generally, the $Q_{\alpha4}$ and $Q_{\alpha5}$ values decrease as $P_{fold}$ values increase, and values of $Q_{\beta12}$, $Q_{\beta23}$, $Q_{\beta34}$, and $Q_{\beta15}$ increase as $P_{fold}$ values increase.

As shown in Figure 7A and B, the $Q$ value of helix 4 is smaller than the $Q$ value of helix 5 when $P_{fold}$ is less than 0.4, which is consistent with our speculation that helix 4 expands first during the conformational transition, after which $Q_{\alpha4}$ and $Q_{\alpha5}$ fluctuate at approximately 0.6. Comparing helix 4 and helix 5 with the top 10 macrostates and TPT pathways, we may further infer the following: (1) Although helix 4 exhibits the propensity to undergo "helix unfolding" in the early phase of the folding transition, the end of helix 4 also undergoes "helix folding" during the late stage of the transition, e.g., **S(37)**, **S(43)**, **S(67)**, **S(81)**, and **S(91)**. (2) During the "helix unfolding" process of helix 5, this helix also exhibits the propensity to bend in the middle via, e.g., **S(63)**, **S(68)**, **S(83)**, **S(92)**, and **S(98)**, which is critical for the formation of strands 3 and 4. By analyzing Figure 7D and E, we observe that $Q_{\beta23}$ and $Q_{\beta34}$ exhibit a similar tendency. When $P_{fold}$ exceeds 0.4, $Q_{\beta23}$ and $Q_{\beta34}$ rapidly increase, reflecting the formation of strands 2−4 (**S(37)** and **S(67)**). Notably, the increase in $Q_{\beta23}$ is slightly higher than $Q_{\beta34}$ when $P_{fold}$ exceeds 0.4, suggesting that following the collapse of helices 4 and 5, strands 2 and 3 may form first. This observation confirmed the previous speculation based on the representative conformations of the two largest population states, **S(81)** and **S(91)**. As shown in Figure 7C and F, both $Q_{\beta12}$ and $Q_{\beta15}$ remain approximately 0.6 when $P_{fold}$ is less than 0.85, which is consistent with the results of TPT analysis that strands 1 and 5 are relatively difficult to form.

In summary, TPT analysis reveals several parallel, heterogeneous folding pathways through high-probability macrostates. Generally, macrostates with low $P_{fold}$ values exhibit greater helical content, whereas macrostates with high $P_{fold}$ values are predominantly composed of $\beta$-sheet structures. Despite the heterogeneity, some similarities exist among the parallel folding pathways. Early in the folding process, the end of helix 4 of the compact all-$\alpha$ structure expands first, as is also reflected in the LT CMD simulations, to form random coils. Subsequently, the end of helix 5 begins to extend together with helix 4, resulting in a globally collapsed conformation without significant secondary structure. One evident feature of these states is the short distance between the residues at the positions of strands 2−4, suggesting that packing precedes the formation of secondary structures in the late stage of the transition. Following this stage, an all-$\beta$ state-like topology forms, and the sheet consisting of strands 2−4 also begins to fold. Finally, with the formation of strands 1 and 5, the protein completes the all $\alpha \rightarrow \beta$ conformational transition. Minimizing the number of hydrophobic residues exposed to water is generally considered as an important driving force behind the protein folding process. As shown in Figure S1, the distribution of hydrophobic residues when RfaH-CTD is in the all-$\alpha$ state is completely different from that in the all-$\beta$ state: there are some hydrophobic residues exposed to solvents in the all-$\alpha$ state but hidden inside in the all-$\beta$ state, forming a hydrophobic core composed of F130, L141, L143, I150, V154, V116, I118, and F159. Among them, F130 is located on strand 2, L141 and L143 on strand 3, I150 and V154 on strand 4, V116 and V118 on strand 1, and F159 on strand 5. From our analyses of the top ranked macrostates and pathways, we may find that during the conformational transition, these five strands tend to first pack into a tertiary folding similar to that in the native all-$\beta$ state and then gradually establish their secondary structures. Specifically, the formation of strands 2−4 is prior to strands 1 and 5, which is consistent with our long time MD simulation results of the all-$\beta$ state structure that the strands 1 and 5 are less stable and tend to unfold earlier. Based on the above observations, we may infer that the reorganization of six hydrophobic residues F130, L141, L143, V154, I150, and V154 in strands 2−4 to form a hydrophobic core initiates the conformation transition; the hydrophobic core is further strengthened by the contribution of the residues V116, I118, and F159 in strands 1 and 5. Overall, the aggregation of the hydrophobic core may drive the all-$\alpha$ to all-$\beta$ conformational transition process.


dx.doi.org/10.1021/ct5002279 | J. Chem. Theory Comput. 2014, 10, 2255−2264

## 4. CONCLUDING REMARKS AND FUTURE PERSPECTIVES

RfaH CTD is able to transform from an $\alpha$-helical hairpin into a $\beta$-barrel. These two alternative folds of RfaH-CTD possess completely distinct secondary structures, and each plays a specific regulatory role in gene expression. Although the $\alpha$-helical state restricts RfaH recognition of *ops*-containing operons and avoids interference with the ubiquitous NusG, the $\beta$-barrel state interacts with the ribosomal S10 protein and enables translation in the absence of canonical ribosome recruitment elements. As an operon-specific transcription factor, RfaH greatly enhances the expression of horizontally transferred operons in *E. coli* and several other human pathogens. Several of these genes are located in pathogenicity islands or on plasmids and encode different virulence factors, such as hemolysin and lipopolysaccharides in proteobacteria. Nagy and collaborators[43] reported that the inactivation of RfaH dramatically attenuates the virulence of uropathogenic *E. coli*, making the RfaH mutant a potential live attenuated vaccine. In this respect, the design of small molecules that prevent or interfere with the conformational transition may exhibit potential for drug development. Moreover, RfaH demonstrates important implications for other proteins that mediate transcription and translation; its behavior indicates that other regulatory proteins in the NusG family or in other protein families may also undergo this type of structural and functional transformation.[7] In this study, we constructed MSMs based on both biased and unbiased MD simulations of the all-$\alpha$ and all-$\beta$ states of RfaH-CTD. This approach overcomes the time scale limitation of conventional approaches for the investigation of complex conformational transitions, and the resultant MSM provides a detailed mesoscopic view of the transformation pathway. We expect the current study to deepen our understanding of the refolding mechanism of RfaH-CTD, which can not only provide new insights into the physical process of protein folding and unfolding but also shed light on the functional interactions of other "transformer proteins"[44] with diverse cellular targets.

## ■ ASSOCIATED CONTENT

### Supporting Information

The distribution of the hydrophobic residues in the all-$\alpha$ and all-$\beta$ native states (text and Figure S1), the validation of the final MSM (text and Figures S2–S5), the 100 macrostates consisting of 684 microstates (Figure S6), the contact profiles of 14 macrostates and two native states (Figures S7 and S8), and the hydrogen bonds and their occupancies during the six approximately 1-$\mu$s simulations (Table S1). This material is available free of charge via the Internet at http://pubs.acs.org.

## ■ AUTHOR INFORMATION


### Corresponding Authors

*Tel.: 86-21-508066-1308. E-mail: myzheng@mail.shcnc.ac.cn (M.Z.)

*Tel.: 86-21-508066-1303. E-mail: hljiang@mail.shcnc.ac.cn (H.J.).

### Author Contributions

‖These authors contributed equally to this work.


### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

This work was supported by the Hi-TECH Research and Development Program of China (Grant 2012AA020302), the National Science and Technology Major Project "Key New Drug Creation and Manufacturing Program" (Grants 2013ZX09507001 and 2014ZX09507002), and the National Natural Science Foundation of China (Grants 81230076 and 21210003).

## ■ REFERENCES


(1) Dill, K. A.; MacCallum, J. L. *Science* **2012**, *338*, 1042–1046.
(2) Murzin, A. G. *Science* **2008**, *320*, 1725–1726.
(3) Tuinstra, R. L.; Peterson, F. C.; Kutlesa, S.; Elgin, E. S.; Kron, M. A.; Volkman, B. F. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 5057–5062.
(4) Luo, X.; Tang, Z.; Xia, G.; Wassmann, K.; Matsumoto, T.; Rizo, J.; Yu, H. *Nat. Struct. Mol. Biol.* **2004**, *11*, 338–345.
(5) Littler, D. R.; Harrop, S. J.; Fairlie, W. D.; Brown, L. J.; Pankhurst, G. J.; Pankhurst, S.; DeMaere, M. Z.; Campbell, T. J.; Bauskin, A. R.; Tonini, R.; Mazzanti, M.; Breit, S. N.; Curmi, P. M. G. *J. Biol. Chem.* **2004**, *279*, 9298–9305.
(6) Anfinsen, C. B. *Science* **1973**, *181*, 223–230.
(7) Burmann, B. M.; Knauer, S. H.; Sevostyanova, A.; Schweimer, K.; Mooney, R. A.; Landick, R.; Artsimovitch, I.; Rösch, P. *Cell* **2012**, *150*, 291–303.
(8) Belogurov, G. A.; Vassylyeva, M. N.; Svetlov, V.; Klyuyev, S.; Grishin, N. V.; Vassylyev, D. G.; Artsimovitch, I. *Mol. Cell* **2007**, *26*, 117–129.
(9) Belogurov, G. A.; Mooney, R. A.; Svetlov, V.; Landick, R.; Artsimovitch, I. *EMBO J.* **2009**, *28*, 112–122.
(10) *The PyMOL Molecular Graphics System*, version 1.4.1; Schrödinger, LLC: Portland, OR, 2010.
(11) Noé, F.; Fischer, S. *Curr. Opin. Struct. Biol.* **2008**, *18*, 154–162.
(12) Elber, R. *Curr. Opin. Struct. Biol.* **2005**, *15*, 151–156.
(13) Bowman, G. R.; Voelz, V. A.; Pande, V. S. *Curr. Opin. Struct. Biol.* **2011**, *21*, 4–11.
(14) Bowman, G. R.; Huang, X.; Pande, V. S. *Methods* **2009**, *49*, 197–201.
(15) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. *J. Chem. Phys.* **2011**, *134*, 174105–174127.
(16) Bowman, G. R.; Pande, V. S. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 10890–10895.
(17) Huang, X.; Bowman, G. R.; Bacallado, S.; Pande, V. S. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 19765–19769.
(18) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 14766–14770.
(19) Huang, X.; Yao, Y.; Bowman, G. R.; Sun, J.; Guibas, L. J.; Carlsson, G.; Pande, V. S. *Pac. Symp. Biocomput.* **2010**, *15*, 228–239.
(20) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 19011–19016.
(21) Lane, T. J.; Bowman, G. R.; Beauchamp, K.; Voelz, V. A.; Pande, V. S. *J. Am. Chem. Soc.* **2011**, *133*, 18413–18419.
(22) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. *J. Chem. Phys.* **2009**, *131*, 124101–124111.
(23) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. *J. Am. Chem. Soc.* **2010**, *132*, 1526–1528.
(24) Bowman, G. R.; Voelz, V. A.; Pande, V. S. *J. Am. Chem. Soc.* **2010**, *133*, 664–667.
(25) Kasson, P. M.; Kelley, N. W.; Singhal, N.; Vrljic, M.; Brunger, A. T.; Pande, V. S. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 11916–11921.
(26) Kelley, N. W.; Vishal, V.; Krafft, G. A.; Pande, V. S.; Kelley, N.; Krafft, G.; Pande, V. *J. Chem. Phys.* **2008**, *129*, 214707.
(27) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *Eur. J. Biochem.* **2008**, *80*, 319–324.
(28) *Suite 2010:Prime*, version 2.2; Schrödinger, LLC: New York, 2010.

(29) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668−1688.
(30) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1996**, *118*, 2309−2309.
(31) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 712−725.
(32) Tsui, V.; Case, D. A. *Biopolymers.* **2000**, *56*, 275−291.
(33) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327−341.
(34) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. *J. Chem. Theory Comput.* **2011**, *7*, 3412−3419.
(35) Deuflhard, P.; Weber, M. *Linear Algebra Appl.* **2005**, *398*, 161−184.
(36) Metzner, P.; Schütte, C.; Vanden-Eijnden, E. *Multiscale Model. Simul.* **2009**, *7*, 1192−1219.
(37) Weinan, E.; Vanden-Eijnden, E. *Annu. Rev. Phys. Chem.* **2010**, *61*, 391−420.
(38) Cronkite-Ratcliff, B.; Pande, V. *Bioinformatics* **2013**, *29*, 950−952.
(39) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577−2637.
(40) Swope, W. C.; Pitera, J. W.; Suits, F. *J. Phys. Chem. B* **2004**, *108*, 6571−6581.
(41) Singhal, N.; Snow, C. D.; Pande, V. S. *J. Chem. Phys.* **2004**, *121*, 415−425.
(42) Ensign, D. L.; Kasson, P. M.; Pande, V. S. *J. Mol. Biol.* **2007**, *374*, 806−816.
(43) Nagy, G.; Dobrindt, U.; Schneider, G.; Khan, A. S.; Hacker, J.; Emödy, L. *Infect. Immun.* **2002**, *70*, 4406−4413.
(44) Knauer, S. H.; Artsimovitch, I.; Rösch, P. *Cell Cycle* **2012**, *11*, 4289−4290.