THERMODYNAMICS OF PROTEIN CONFORMATIONAL CHANGES BY UMBRELLA SAMPLING AND FREE ENERGY CALCULATION

by

Seyed Hamed Meshkin

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Physics

West Lafayette, Indiana

December 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL

STATEMENT OF COMMITTEE APPROVAL

Dr. Stephen R. Wassall, Chair

Department of Physics

Dr. Fangqiang Zhu

Lawrence Livermore National Laboratory

Dr. Horia I. Petrache

Department of Physics

Dr. Ricardo Decca

Department of Physics

Dr. Sergei F. Savikhin

Department of Physics

Approved by:

Dr. John P. Finley

To my wife Donya

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Fangqiang Zhu, for his guidance through the research and writing process. I also want to thank Dr. Stephan R. Wassall for his help and support. I am appreciative to all of my committee members for their time they put into completing my dissertation.

Finally, and most importantly, a special thanks to my wife Donya for her full support and patience during my Ph.D. process.

Thank you all for your unwavering support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Spontaneous transitions between the native and non-native protein conformations are normally rare events that hardly take place in typical unbiased molecular dynamics simulations. It was recently demonstrated that such transitions could be well described by a reaction coordinate, Q, that represents the collective fraction of the native contacts between the protein atoms. Here we attempt to use this reaction coordinate to enhance the conformational sampling. We perform umbrella sampling simulations with biasing potentials on Q for two model proteins, Trp-Cage and BBA, using the CHARMM force field. Hamiltonian replica exchange is implemented in these simulations to further facilitate the sampling. The simulations appear to have reached satisfactory convergence, resulting in unbiased, free energies as a function of Q. In addition to the native structure, multiple folded conformations are identified in the reconstructed equilibrium ensemble. Some conformations without any native contacts nonetheless have rather compact geometries and are stabilized by hydrogen bonds not present in the native structure. Whereas the enhanced sampling along with Q reasonably reproduces the equilibrium conformational space, we also find that the folding of an α-helix in Trp-Cage is a slow degree of freedom orthogonal to Q and therefore cannot be accelerated by biasing the reaction coordinate. Overall, we conclude that whereas Q is an excellent parameter to analyze the simulations, it is not necessarily a perfect reaction coordinate for enhanced sampling, and better incorporation of other slow degrees of freedom may further improve this reaction coordinate.

To analyze such behavior like slow degrees of freedom, we conducted another research study. Proteins may adopt multiple conformations, and they undergo various transitions from one conformation to another. A well-defined reaction coordinate can describe these transitions. However, there is no efficient way to define the entire conformational space of a complex biological system by only one reaction coordinate. In the two-state system, a protein can adopt two different conformations, A and B. We implemented a stepwise transition model. The targeted protein starts from metastable A, and it will undergo a transition to another intermediate state, and from that intermediate state, the protein undergoes another transition to and so on. Therefore, by N transitions, we can get to the metastable state B. During each step transition, we apply a boundary potential over other degrees of freedom to keep them unchanged. With this strategy, along with a simple definition of the reaction coordinate, we have high accuracy in our

thermodynamics and protein dynamics measurements. As a case study, we implemented all-atom Umbrella Sampling simulations to characterize the conformational changes between outward-facing open (OF) and outward-facing occluded (OC) states of transmembrane protein Mhp1. For each step transition, the reaction coordinate was defined by a simple dihedral angle or a bond length. We could obtain six transition steps with five intermediate states that connect the two OF and OC stable states. We measured each step transition free energy profile from WHAM equations. We performed two independent sampling simulations with different initial structures: the transition initiates from OF state indicated *InitOF*, and the transition initiates from the OC state indicated as *InitOC* transition. By comparing the obtained free energy profiles with the stepwise model, we implied the extent of convergence in our calculations. The energy difference between OF and OC states in our study is $\Delta G = -1.02 \pm 1.1$ and $\Delta G = -1.12 \pm 1.14$ kcal/mol for *InitOF* and *InitOC* transition, respectively.

CHAPTER 1. AN INTRODUCTION TO ALL-ATOM MOLECULAR DYNAMICS SIMULATION

Most of the proteins need to undergo conformational changes to be functional [1, 2]. Thus, understanding protein's adopted conformations and the dynamics of the conformational changes between available states are essential in the studies of biomolecular systems. The variation of physicochemical properties such as temperature[3, 4], pressure, ligand binding[5], and divalent ion distribution[3] are a few reasons that cause proteins to undergo several conformational changes that are associated with their functionality. Most large biological molecules such as proteins can at least achieve two distinct metastable conformations, which is described as a two-state model. A spontaneous transition between the two conformational states is a rare random event, that happens very quickly; therefore, it is not likely to observe the conformational transitions directly in a real experiment.

All-atom Molecular Dynamics (MD) simulation provides insight into individual atomic motions to predict the detail of the structural changes caused by the forward and reverse conformational transitions[6, 7]. Recently, MD simulations are widely used to answer the various questions about the thermodynamics and kinetic properties of a biological system, often more rapidly, compared to experiments on a real system. In theory, one single long MD simulation is necessarily enough to observe several spontaneous transitions that can switch back and forth between two possible conformational states. However, in practice, such a simple approach to construct the equilibrium ensemble needs an extremely long simulation time, which is unachievable for most systems of interest. Even with the most high-performance computer cluster resources available nowadays, the reasonable simulation time might be reachable only for small biomolecular system sizes with comparably fast transition rates. Still, the equilibrium ensemble is necessary to derive the thermodynamic properties of the system; therefore, it is required to be generated despite the time scale problem.

Alternatives to a long MD simulation, various computational methods known as enhanced sampling techniques are suggested to alleviate the insufficient simulation time problem. These methods include Umbrella Sampling [2, 8], Metadynamics[9], Weighted Ensemble[10], Transition Path Sampling [11], Accelerated MD [12], String method [13, 14], Adaptively Biased MD [15], Milestoning [16], and Dynamic Importance Sampling [17]. To accelerate the dynamics of the

simulations at a reasonable time, all enhanced sampling methods implementing some form of bias potentials to measure the equilibrium ensemble. Many enhanced sampling techniques employ a Reaction Coordinate (RC) that distinguishes the two protein conformations, such that driving along the RC could enforce continuous conversions between the conformational states. The free energy profile quantifies the thermodynamics of the conformational transition and, in particular, gives the free energy difference between the two conformational states. The weighted histogram analysis method (WHAM) [18, 19] can be employed to calculate the equilibrium free energy from the trajectory of the enhanced sampling techniques such as umbrella sampling.

1.1 Free Energy

For any biological system, the equilibrium probability between states is the fundamental concept for the conformational changes. In a two-state system, a protein can adopt two alternative conformations A and B. At equilibrium, the probabilities for the two conformations are p_A and p_B, with the fact that the sum of the probabilities between the two states is equal to one.

Free energy, along with a reaction coordinate between the two alternative conformations A (reactant) and B (product), is a reversible work that is widely used in the studies of protein conformational changes. The free energy profile can be obtained by the contribution of the Entropy and enthalpy as $G = TS - H$, where T, S, and H are temperature, entropy, and enthalpy, respectively. Alternatively, to measure the free energy changes for complex systems, one can use numerical simulations by means of statistical mechanics or the Newtonian equation of motions. To that extend, the accurate measurement of the free energy changes between the two states (A and B) is to explore the conformational space of the relevant system to obtain the low-energy states between the reference and target system [18]. Therefore, the high precision of the free energy calculation relies on adequately sampling the configurational space. Molecular Dynamics [19] and Monte Carlo simulations [20] are promising tools in this respect. Fundamentally, in an unbiased simulation, the free energy is related to the probability density function $p(X)$ along with a reaction coordinate X as

$$p(X) = \frac{\int d\vec{r}\ \delta(X'(\vec{r}) - X)\, e^{-\frac{U(\vec{r})}{k_B T}}}{\int d\vec{r}\, e^{-\frac{U(\vec{r})}{k_B T}}} \qquad \text{Eq. 1-1}$$

$$G(X) = -k_B T\, ln\, p(X)$$

where k_B is the Boltzmann constant. $\vec{r}$ is a set of 3N configurational vectors. $U(\vec{r})$ represents the total energy as a function of $\vec{r}$. $X'(\vec{r})$ is the fixed conformation, and all microstate with that conformation are equally likely. The free energy equation means that events with high probability get a low free energy value along with the reaction coordinate. In contrast, rare events (transition regions) are sampled in the simulation with a low population and cause $G(X)$ to accept a relatively high value. At the transition regions, a small number of sample points can be achieved during the simulation, which results in a significant statistical error.

To reduce the calculated error, one can increase the simulation time, sometimes beyond what is practically possible. One can use enhanced sampling technics such as Umbrella Sampling (US) by employing a biased potential confining the system at the transition region. In US simulation, the entire conformational space between the two states A and B is divided into a limited sections, and each piece represents one umbrella window. An individual simulation can then be set up for each umbrella window to reduce the statistical error for the regions not being sampled adequately in an unbiased simulation.

For the thermodynamics measurement of a system of interest, free energy is depicted as a function of a reaction coordinate. Therefore, in an US simulation, the primary concern is choosing an order parameter that clearly shows the progress of the reaction in the configurational space. The fraction of native contacts, Q, a widely used reaction coordinate, is constructed by the ratio of non-native to native contacts. Besides, a native contact is formed by a pair of atoms in the native structure that belong to particular separated residues, with a cutoff distance criteria. Another reaction coordinate, close to Q, is the overlap function that shows the similarity to the native structure and is defined by the Heaviside step function [21]. Different groups propose other Continuous forms of the fraction of native contacts. The proposed Q could be: A Gaussian function with the mean value equal to the native contact distance [22-24], a continuous form of a step function analogous to the Fermi-Dirac distribution function, has recently been used with different arbitrary parameters [2, 25-29]. Alternatively, a reaction coordinate can be defined by a function of other native state geometries like the fraction of native state dihedral angels [30-32], the fraction of native hydrogen bonds [2], the number of core water molecules [33, 34], as well as holistic parameters such as radius of gyration [26] and root mean-squire deviation (RMSD).

1.2 Molecular Dynamics Simulation

Time evolution of atoms within classical mechanics concepts using numerical simulation is the Molecular Dynamics (MD) simulation methodology. Any atoms in the simulation are assigned as one particle in the simulation. MD simulation, utilizing atoms positions x and moments p at time t, predicts a new position and momentum for particles at time $t + \Delta t$. The time interval Δt must be in the order of magnitude less than the displacement frequency with the highest period of oscillation in the system. Generally, this time interval is one to two femtoseconds (fs). When simulation generates the trajectory, a more extended time step causes losing precision in the integration process and chaotic behavior in the system [35].

The calculation of the trajectories in MD simulation is through numerical integration of the Newtonian equations of motion. For a system consist of N particles in 3D conformational space with the cartesian coordinates (x, y, z), the system's potential gradient $F = -\nabla U(\vec{r})$ is the force acting on each particle i with the vector $F_i\ (r_1, r_2, \ldots, r_N, t)$. Moreover, the acceleration at each step can be measured by $F_i = m_i\, a_i$. m_i and a_i are the mass and acceleration of particle i, respectively. The initial coordinate of a protein system in the cartesian coordinates can be obtained using X-ray crystallography, NMR spectroscopy, and electron microscopy [35]. The initial velocity is randomly assigned to each atom by satisfying the Maxwell-Boltzmann distribution with an assigned temperature value. Now knowing the positions and velocity of each atom, the only remaining is to know the force acting on each atom to calculate the position at the next instant of time $t_0 + \Delta t$.

The potential energy U in a molecular system can be obtained from Quantum mechanics, which is highly accurate but too slow incredibly for a large-scale system. The Coulomb's law and Schrodinger equation are two essential calculations in the Quantum mechanics approach. Molecular mechanics with the Heuristic energy function are another alternative for the potential energy function, which are frequently used to simulate biomolecular systems such as proteins, membranes, and DNA. The classical interaction potentials, known as force fields, represent a simple way to calculate interatomic forces. However, in their functional form, they involve several parameters that must be adjusted to obtain accurate results.

1.3 CHARMM Force Field

Molecular Dynamics (MD) simulation is a powerful tool that determines the coordinates and velocity of atoms in a regular time interval. At each time interval, the recorded coordinate for all the atoms generates the trajectory of the system of interest. Then the generated trajectory is used for analyzing the thermodynamics properties of the system. This method is doable first by attaining a vast knowledge of physical interatomic forces and the physical model of our system. Because by having the positions and velocity of each atom at the time t_0, we need the force acting on each atom to predict the position at the next instant of time $t_0 + \Delta t$ by employing the Newtonian equation of motion. This force is measurable by the gradient of the potential energy function, which can be obtained from the physical property of the targeted system. Our study for all of the simulations utilized the CHARMM force field as the potential energy function. The CHARMM force field consists of two terms as bonded and nonbonded. Bonded has six terms (bonds, torsion/dihedral angle, improper dihedral angle, Urey-Bradley, and CMAP). The intermolecular nonbonded term involves electrostatic and van der Waals (vdW) interactions.

$$\sum_{bonds} k_b(b - b_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{dihedral} k_\phi(1 + \cos(n\,\phi - \delta))^2 + \sum_{impropers} k_\omega(\omega - \omega_0)^2 + \sum_{Urey-Bradley} k_u(u - u_0)^2 + \sum_{residues} u_{CMAP}(\phi, \psi) + \sum_{nonbonded}^{elec} \frac{q_i q_j}{4\,\pi\, r_{ij}} + \sum_{nonbonded}^{LJ} \epsilon_{ij}\left[\left(\frac{R_{min,ij}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{min,ij}}{r_{ij}}\right)^{6}\right] \quad \text{Eq. 1-2}$$

where b_0, θ_0, ω_0 and u_0 are the equilibrium of bond, angle, improper and Urey-Bradley, respectively. All k's ($k_b, k_\theta, k_\phi, etc$) are the various force constant. δ and n are the dihedral phase and dihedral multiplicity, respectively. Besides, the CHARMM energy function includes an energy correction map, so-called CMAP, to improve the conformational properties of the φ, ψ terms in the dihedral angle of the amino acids peptide chain. For the electrostatic term q_i and q_j are the partial atomic charge belong to atoms i and j. For the Lennard-Jones term, $\varepsilon_{ij} = \sqrt{\epsilon_i \epsilon_j}$, the

geometric mean is the depth of the potential well. $\sigma_{ij} = \frac{\sigma_i + \sigma_j}{2}$, arithmetic mean is the distance where the potential is zero. $R_{min,ij}$is the radius where the potential reaches the minimum value. Extra information about the CHARMM force field can be found in references [36-38].

1.4 Umbrella Sampling

Umbrella sampling, a widely used enhanced sampling method, is employed to overcome the energetic barrier in MD simulations. Torrie et al. were the first who developed this approach in 1977 [47]. In a general US algorithm, a series of independent simulations with a reaction coordinate as the selected collective variable (CV) was set and confined by a biased restrain to obtain a conformational transition. Each simulation refers to as one US window and is confined by a harmonic biased potential. Therefore, each window is specified to sample a narrow segment of the conformational space. Here, the entire region of the configurational space (reaction coordinate) is supposed to get covered by all the US windows, so there must be an overlap between every two neighboring windows. The biased harmonic potential is added to the Hamiltonian of the system:

$$U_i(x) = \frac{k}{2}\ \left(x - x_i^{ref}\right)^2 \qquad \text{Eq. 1-3}$$

In which i is assigned to one window. k is the spring constant and defines the strength of the harmonic potential. x_i^{ref} is the center of each harmonic biasing potential $U_i(x)$. Before starting the US simulations, a set of initial structures possessing the reaction coordinate approximately closed to each umbrella window reference value is essential to run the MD simulation.

A molecular dynamics pulling simulation is applied to capture the initial conformation for each US simulation. For this purpose, one can run a biased MD simulation with the protein crystal structure as the initial conformation. In this MD simulation, equation 3 needs to get employed with $i = 1$ as the first window reference value, which should be approximately equal to the crystal structure's reaction coordinate value. After a specific timestep, $i\ is$ incremented by one to drive the system out of the crystal structure state. This procedure is continuing until simulation lasts

enough to cover all the configurational space. Finally, the system's conformation at the vicinity of each umbrella window is taken for the initial structures of the US simulations.

When the US simulations converge the system to an equilibrium state, the Weighted Histogram Analysis Method (WHAM) can be employed to combine the independent windows' statistics. The WHAM equations are the most popular method and are expressed as follows:

$$p_u(x) = \frac{\sum_{l=1}^{M} h_l(x)}{\sum_{j=1}^{M} n_j \exp\left[\frac{F_j - U(x)}{k_B T}\right]}$$

Eq. 1-4

$$F_j = -k_B T ln \sum_{q} p_u(x) \ \exp[-\frac{U(x)}{k_B T}]$$

The WHAM equation converts the biased probability is generated at each US simulation to an unbiased probability distribution. Therefore, $p_u(x)$ is the unbiased probability distribution along the reaction coordinate. M is the total number of umbrella windows in the US simulation. $h(x)$ is the count at the bin x with l shows the index of the umbrella windows. n_j is the number of data points at the window j. $p_u(x)$ and F_k are the recurrence equations and should be solved iteratively until achieving a self-consistent solution. Consequently, The free energy profile can be measured by $G(x) = k_B T ln\, p_u(x)$. More details of the WHAM equation can be found in references [48]

1.5 Aims of Research

1.5.1 Thermodynamics of Protein Folding Studied by Umbrella Sampling

Best et al. [26] analyzed the trajectories of millisecond equilibrium MD simulations [25] of some small proteins and, utilizing a reaction coordinate of the collective fraction of native contacts, characterized the folding and unfolding of such small proteins. A good reaction coordinate can facilitate the enhanced sampling technique to provide the thermodynamics of the system of interest potentially more efficiently than an unbiased simulation. This study plans to

check this approach by applying US along the Q reaction coordinate. We are employing two small proteins, Trp-Cage [50] and BBA [51], as the cases study in our research.

Our simulations serve as a case study for using the reaction coordinate based on the native contacts for sampling protein conformations. Through detailed analysis, we need to demonstrate the effectiveness and the problems with this approach. Although we specifically adopted US in this study, we note that many other enhanced sampling methods also require a pre-determined reaction coordinate and would have similar problems with the folding reaction coordinate examined here.

1.5.2 Toward Convergence in Free Energy Calculation for Protein Conformational Changes

This study aims to identify transition processes that change a system from an initial state A to another final state B along with the backward processes. Several possible dynamical pathways are suggested for proteins that could link the pair states A and B together.[52, 53] A pathway consists of a sequence of N discrete intermediate states that are in between the two metastable states of A and B. Hence, the transition of the forward and reverse pathways can be broken down into N+1 stepwise transition processes. Alternatively, an enhanced sampling method with harmonic potential on a reaction coordinate that is associated with each single-step pathway can also be applied to measure the free energy profile of each step process individually. The reaction coordinate can be defined here as any form of torsion angle, bond angle, and bond length. In this study, we implemented Umbrella Sampling (US) to characterize the conformational changes of each step process between the ligand-free outward-facing open (OF) and outward-facing occluded (OC) state of Mhp1.

1.6 Figures

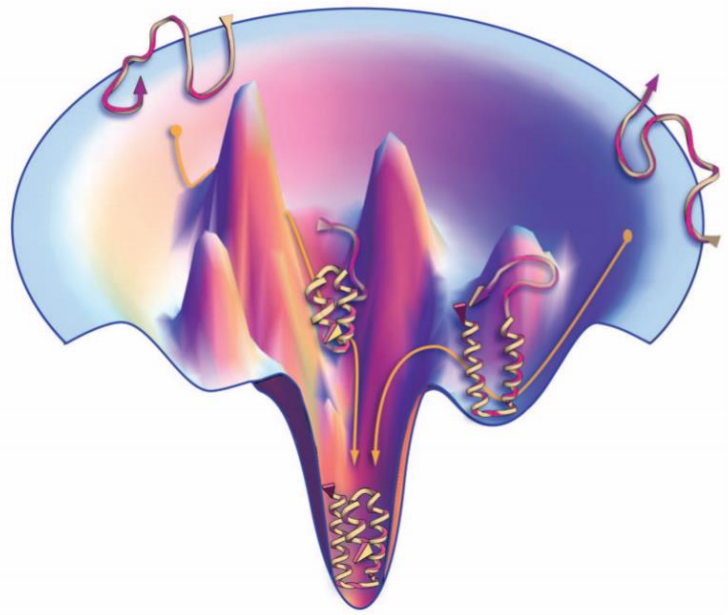


Figure 1-1 - The funnel-shaped energy landscape of proteins. Non-native structures with high energy compared to the folded state with low-energy at the bottom of the funnel. Alternative pathways drive a protein from the non-folded state to the folded state. Figure captured from Reference [41]

1.7 References

1. Karplus, M. and J. Kuriyan, *Molecular dynamics and protein function.* Proceedings of the National Academy of Sciences, 2005. **102**(19): p. 6679-6685.

2. Shea, J.-E. and C.L. Brooks III, *From folding theories to folding proteins: a review and assessment of simulation studies of protein folding and unfolding.* Annual review of physical chemistry, 2001. **52**(1): p. 499-535.

3. Gratacos-Cubarsi, M. and R. Lametsch, *Determination of changes in protein conformation caused by pH and temperature.* Meat science, 2008. **80**(2): p. 545-549.

4. Becktel, W.J. and J.A. Schellman, *Protein stability curves.* Biopolymers: Original Research on Biomolecules, 1987. **26**(11): p. 1859-1877.

5. Swain, J.F. and L.M. Gierasch, *The changing landscape of protein allostery.* Current opinion in structural biology, 2006. **16**(1): p. 102-108.

6. Shaw, D.E., et al., *Atomic-level characterization of the structural dynamics of proteins.* Science, 2010. **330**(6002): p. 341-346.

7. Khalili-Araghi, F., et al., *Molecular dynamics simulations of membrane channels and transporters.* Current opinion in structural biology, 2009. **19**(2): p. 128-137.

8. Kästner, J., *Umbrella sampling.* Wiley Interdisciplinary Reviews: Computational Molecular Science, 2011. **1**(6): p. 932-942.

9. Barducci, A., M. Bonomi, and M. Parrinello, *Metadynamics.* Wiley Interdisciplinary Reviews: Computational Molecular Science, 2011. **1**(5): p. 826-843.

10. Dickson, A. and C.L. Brooks III, *WExplore: hierarchical exploration of high-dimensional spaces using the weighted ensemble algorithm.* The Journal of Physical Chemistry B, 2014. **118**(13): p. 3532-3542.

11. Dellago, C., et al., *Transition path sampling and the calculation of rate constants.* The Journal of chemical physics, 1998. **108**(5): p. 1964-1977.

12. Hamelberg, D., J. Mongan, and J.A. McCammon, *Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules.* The Journal of chemical physics, 2004. **120**(24): p. 11919-11929.

13. Song, H.D. and F. Zhu, *Conformational changes in two inter-helical loops of Mhp1 membrane transporter.* PloS one, 2015. **10**(7).

14. Weinan, E., W. Ren, and E. Vanden-Eijnden, *String method for the study of rare events.* Physical Review B, 2002. **66**(5): p. 052301.

15. Babin, V., C. Roland, and C. Sagui, *Adaptively biased molecular dynamics for free energy calculations.* The Journal of chemical physics, 2008. **128**(13): p. 134101.

16. Faradjian, A.K. and R. Elber, *Computing time scales from reaction coordinates by milestoning.* The Journal of chemical physics, 2004. **120**(23): p. 10880-10889.

17. Dupuis, P., A.D. Sezer, and H. Wang, *Dynamic importance sampling for queueing networks.* The Annals of Applied Probability, 2007. **17**(4): p. 1306-1346.

18. Chipot, C. and A. Pohorille, *Free energy calculations.* Springer series in chemical physics, 2007. **86**: p. 159-184.

19. Allen, M.P., *Introduction to molecular dynamics simulation.* Computational soft matter: from synthetic polymers to proteins, 2004. **23**(1): p. 1-28.

20. Corti, D.S., *Monte Carlo simulations in the isothermal—isobaric ensemble: the requirement of a 'shell'molecule and simulations of small systems.* Molecular Physics, 2002. **100**(12): p. 1887-1904.

21. Shea, J.-E., J.N. Onuchic, and C.L. Brooks III, *Energetic frustration and the nature of the transition state in protein folding.* The Journal of Chemical Physics, 2000. **113**(17): p. 7663-7671.

22. Hardin, C., Z. Luthey-Schulten, and P.G. Wolynes, *Backbone dynamics, fast folding, and secondary structure formation in helical proteins and peptides.* Proteins: Structure, Function, and Bioinformatics, 1999. **34**(3): p. 281-294.

23. Eastwood, M.P. and P.G. Wolynes, *Role of explicitly cooperative interactions in protein folding funnels: a simulation study.* The Journal of Chemical Physics, 2001. **114**(10): p. 4702-4716.

24. Pogorelov, T.V. and Z. Luthey-Schulten, *Variations in the fast folding rates of the λ-repressor: A hybrid molecular dynamics study.* Biophysical journal, 2004. **87**(1): p. 207-214.

25. Lindorff-Larsen, K., et al., *How fast-folding proteins fold.* Science, 2011. **334**(6055): p. 517-520.

26. Best, R.B., G. Hummer, and W.A. Eaton, *Native contacts determine protein folding mechanisms in atomistic simulations.* Proceedings of the National Academy of Sciences, 2013. **110**(44): p. 17874-17879.

27. Sheinerman, F.B. and C.L. Brooks, *Molecular picture of folding of a small α/β protein.* Proceedings of the National Academy of Sciences, 1998. **95**(4): p. 1562-1567.

28. Chen, J. and C.L. Brooks III, *Can molecular dynamics simulations provide high-resolution refinement of protein structure?* Proteins: Structure, Function, and Bioinformatics, 2007. **67**(4): p. 922-930.

29. Best, R.B., et al., *Pulling direction as a reaction coordinate for the mechanical unfolding of single molecules.* The Journal of Physical Chemistry B, 2008. **112**(19): p. 5968-5976.

30. Levy, Y. and O.M. Becker, *Energy landscapes of conformationally constrained peptides.* The Journal of Chemical Physics, 2001. **114**(2): p. 993-1009.

31. Vengadesan, K. and N. Gautham, *Energy landscape of Met-enkephalin and Leu-enkephalin drawn using mutually orthogonal Latin squares sampling.* The Journal of Physical Chemistry B, 2004. **108**(30): p. 11196-11205.

32. Itoh, K. and M. Sasai, *Flexibly varying folding mechanism of a nearly symmetrical protein: B domain of protein A.* Proceedings of the National Academy of Sciences, 2006. **103**(19): p. 7298-7303.

33. Guo, W., S. Lampoudi, and J.-E. Shea, *Posttransition state desolvation of the hydrophobic core of the src-SH3 protein domain.* Biophysical journal, 2003. **85**(1): p. 61-69.

34. Juraszek, J. and P.G. Bolhuis, *Rate constant and reaction coordinate of Trp-cage folding in explicit water.* Biophysical journal, 2008. **95**(9): p. 4246-4257.

35. Rapaport, D.C., *The art of molecular dynamics simulation*. 2004: Cambridge university press.

36. MacKerell Jr, A.D., et al., *All-atom empirical potential for molecular modeling and dynamics studies of proteins.* The journal of physical chemistry B, 1998. **102**(18): p. 3586-3616.

37. MacKerell Jr, A.D., M. Feig, and C.L. Brooks, *Improved treatment of the protein backbone in empirical force fields.* Journal of the American Chemical Society, 2004. **126**(3): p. 698-699.

38. Best, R.B., et al., *Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles.* Journal of chemical theory and computation, 2012. **8**(9): p. 3257-3273.

39. Muirhead, H. and M. Perutz, *Structure of hæmoglobin: A three-dimensional fourier synthesis of reduced human haemoglobin at 5.5 Å resolution.* Nature, 1963. **199**(4894): p. 633-638.

40. Kendrew, J.C., et al., *A three-dimensional model of the myoglobin molecule obtained by x-ray analysis.* Nature, 1958. **181**(4610): p. 662-666.

41. Dill, K.A. and J.L. MacCallum, *The protein-folding problem, 50 years on.* science, 2012. **338**(6110): p. 1042-1046.

42. Levinthal, C., *How to fold graciously*. Mossbauer spectroscopy in biological systems, 1969. **67**: p. 22-24.

43. Dill, K.A., et al., *The protein folding problem.* Annu. Rev. Biophys., 2008. **37**: p. 289-316.

44. Pande, V.S., A.Y. Grosberg, and T. Tanaka, *Statistical mechanics of simple models of protein folding and design.* Biophysical journal, 1997. **73**(6): p. 3192-3210.

45. Fiser, A., *Template-based protein structure modeling.* Computational biology, 2010: p. 73-94.

46. Sugita, Y. and Y. Okamoto, *Replica-exchange molecular dynamics method for protein folding.* Chemical physics letters, 1999. **314**(1-2): p. 141-151.

47. Torrie, G.M. and J.P. Valleau, *Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling.* Journal of Computational Physics, 1977. **23**(2): p. 187-199.

48. Zhu, F. and G. Hummer, *Convergence and error estimation in free energy calculations using the weighted histogram analysis method.* Journal of computational chemistry, 2012. **33**(4): p. 453-465.

49. Affentranger, R., I. Tavernelli, and E.E. Di Iorio, *A novel Hamiltonian replica exchange MD protocol to enhance protein conformational space sampling.* Journal of Chemical Theory and Computation, 2006. **2**(2): p. 217-228.

50. Meuzelaar, H., et al., *Folding dynamics of the Trp-cage miniprotein: evidence for a native-like intermediate from combined time-resolved vibrational spectroscopy and molecular dynamics simulations.* The Journal of Physical Chemistry B, 2013. **117**(39): p. 11490-11501.

51. Sarisky, C.A. and S.L. Mayo, *The ββα fold: explorations in sequence space.* Journal of molecular biology, 2001. **307**(5): p. 1411-1418.

52. Shea, J.-E., J.N. Onuchic, and C.L. Brooks, *Probing the folding free energy landscape of the src-SH3 protein domain.* Proceedings of the National Academy of Sciences, 2002. **99**(25): p. 16064-16068.

53. Noé, F., et al., *Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations.* Proceedings of the National Academy of Sciences, 2009. **106**(45): p. 19011-19016.

CHAPTER 2. THERMODYNAMICS OF PROTEIN FOLDING STUDIED BY UMBRELLA SAMPLING ALONG A REACTION COORDINATE OF NATIVE CONTACTS

2.1 Introduction

The function of a protein is determined by its three-dimensional structures [1, 2]. Many proteins adopt a specific folded conformation, referred to as the native structure, under physiological conditions. Thermodynamically, the native structure typically corresponds to a minimum in the free energy surface. Early theoretical analysis suggested that the native structure would obey the minimal frustration principle [3, 4], and recent simulation studies further revealed that the native structure also serves as a kinetic hub that connects multiple highly distinct non-native conformations [5]. Indeed, the native structure is not necessarily the only conformation adopted by a protein, and there may exist an equilibrium between the native structure and the non-native (such as disordered and extended) conformations. The thermodynamics and kinetics for the transitions between the native and the non-native protein structures, such as the folding rate [6], the transition state, [7], and the intermediates states [8], have been extensively studied for decades.

Computational methods such as molecular dynamics (MD) simulations [9] are powerful tools to complement protein folding experiments.

Among all the MD methods, the most straightforward approach is to directly simulate a protein in its natural environment and observe the spontaneous transitions between the native and the non-native conformations. If the simulation is long enough such that a statistically sufficient number of transitions occur, all thermodynamic and kinetic quantities of protein folding can be directly obtained from the simulation trajectory. Thanks to the breakthrough in specialized computer hardware and algorithm, all-atom simulations of millisecond time scale have been achieved [10, 11], which allowed direct observation of folding/unfolding transitions for a number of small proteins with relatively fast kinetics. Alternatively, a variety of enhanced sampling methods have been applied to simulate protein folding [9, 12, 13]. Some of these methods, such as umbrella sampling (US) [14, 15] and metadynamics [16], employ non-Boltzmann sampling with biasing potentials to accelerate the transitions over the energy barriers. Similar acceleration can also be achieved, e.g., in weighted ensemble simulations [17, 18], by generating multiple replicas to enhance the sampling in regions with low equilibrium probabilities.

In all of the methods above, the unbiased equilibrium thermodynamics can be reconstructed from the simulation trajectories based on rigorous theories in statistical mechanics. In addition, serial or parallel tempering [19] can be employed in methods such as replica exchange MD (REMD) simulations [20, 21], in which multiple replicas are run in parallel and periodically attempt to exchange their temperatures or biasing potentials [22-25]. An exact protein conformation must be described in a multidimensional space. Indeed, the conformational space for proteins has been successfully described by Markov state models [26]. Alternatively, in many cases, it is also desirable to project the high-dimensional protein conformations onto a single reaction coordinate (or order parameter) to simplify the analysis. Once such a reaction coordinate is defined, its equilibrium probability distribution can be determined from the equilibrium ensemble of the protein conformations and will correspond to the free energy as a function of the reaction coordinate. With a "good" reaction coordinate for protein folding, the associated free energy would not only clearly distinguish the native and the non-native states, but also reflect the kinetic barrier for the transitions.

Many common reaction coordinates for protein folding are based on the fraction of native contacts [27]. One contact is usually defined as a pair of residues that are spatially close (shorter than some cut-off distance) but not in sequence proximity, and all such contacts in the native structure constitute the set of native contacts. One can then examine how many of the native contacts are present or absent in any given conformation based on the inter-residue distances. As a simple criterion, a Heaviside step function [28] can be used to map a distance to a contact number, which can be either 0 or 1 as determined by the cut-off distance. Other criteria assign a non-integer contact number between 0 and 1 using a continuous function of the distance, such as Gaussian [29-31] or Fermi-Dirac distribution functions [11, 14, 32-35]. The sum of the contact numbers in the given conformation, as a fraction of the maximum possible total number (as in the native structure), can then be defined as the reaction coordinate, with a value close to 1 and 0 representing the native and the non-native states, respectively. Alternative to the native contacts, reaction coordinates can also be defined based on dihedral angles [36-38], native hydrogen bonds [14], the number of core water molecules [39, 40], as well as holistic parameters such as radius of gyration [28] and root-mean-square deviation (RMSD) [41].

Recently, Best et al. [42] analyzed the trajectories of millisecond-long unbiased MD simulations [11] of some small proteins and concluded that a reaction coordinate based on the

collective fraction of native contacts characterizes the folding/unfolding transitions remarkably well [42]. In principle, once a good reaction coordinate is identified, enhanced sampling along that coordinate could provide the conformational thermodynamics in a potentially more efficient way compared to the straightforward, unbiased simulations. Here we test this strategy by performing US along the reaction coordinate mentioned above, as similarly done in some earlier studies [14, 43-45]. Our all-atom simulations are performed with explicit solvent, and we employ the Hamiltonian REMD technique [22] to facilitate the US [14, 46] in this study. We use two small proteins, Trp-Cage [47] and zinc finger motif (BBA) [48], as the test cases here. Trp-Cage is a 20-residue protein that can fold rapidly to a stable structure. BBA is a 28-residue protein with a native structure that consists of two β-sheets and one α-helix. Both proteins have been extensively studied in previous simulations [11, 21, 47-50]. We determine the free energy profile and reconstruct the equilibrium ensemble for each protein from the simulations here.

2.2 Methods

In this study, we focus on the folding of two proteins, Trp-Cage [47] and BBA [48], which have also been extensively studied in previous simulations [5, 11, 21, 41, 47-50]. In particular, Lindorff-Larsen et al. [11] performed long unbiased simulations on the two proteins, and Best et al. [42] analyzed the simulation trajectories using a reaction coordinate representing the collective fraction of native contacts. Here we take the reaction coordinate above and perform US [14, 46] simulations with Hamiltonian Replica Exchange Molecular Dynamics (HREMD) [22] to reproduce the equilibrium ensemble for the proteins. The computational details are provided below.

2.2.1 System Setup

Both of our simulation systems are similar to the ones used in Lindorff-Larsen et al. [11]. The first protein is a Trp-Cage mutant, denoted as TC10b (PDB: 2JOF [51]), with the sequence DAYAQWLADGGPSSGRPPPS. In comparison to the wild type, residue 8 in the sequence is mutated from LYS to ALA [11]. The simulation system consists of the protein in a solution of 1639 water molecules and 65 mM NaCl. The total number of atoms in the Trp-Cage simulation system is 5230. The second protein, BBA (PDB: 1FME [48]), with the sequence EQYTAKYKGRTFRNEKELRDFIEKFKGR, was solvated with 2978 water molecules and four

Chloride ions. The simulation system for BBA consists of a total of 9442 atoms. We adopted the standard protonation state at pH 7 for all residues of the two proteins. For both proteins, the first frame in the PDB file was taken as the native structure in this study.

We adopted the CHARMM (Ver. c36, released in December 2013) protein force field [35, 52, 53] and the TIP3P water model [54] in this study. The MD simulations were carried out using the NAMD2 program [55] with a time step of 2 fs and in the NPT ensemble with the periodic boundary conditions. A constant pressure of 1 atm was obtained by applying the Nose-Hoover Langevin piston method [56], and a Langevin thermostat with a damping coefficient of 1 ps^{-1} was used to maintain the constant temperature of the system. The SHAKE [57] and SETTLE [58] algorithms were used to maintain rigid bonds involving all hydrogen atoms. We used a 12 Å cut-off for non-bonded interactions, with a smooth switching function starting at 10 Å. Full electrostatics was calculated every 4 fs using the particle mesh Ewald (PME) method [59].

The two systems were first minimized and equilibrated for a total of 10 ns. In the equilibration phase, the temperatures of the Trp-Cage and the BBA systems were 290 K and 325 K, respectively, although Trp-Cage was simulated at two additional temperatures as well, as will be described later.

2.2.2 Reaction Coordinate

We adopt the same reaction coordinate in Best et al. [42] based on the fraction of native contacts. The set of native contacts is defined from the native structure. Specifically, a pair of heavy atoms (i, j) in residues R_i and R_j is counted as a native contact if $|R_i - R_j| > 3$ and the interatomic distance r_{ij}^0 in the native structure is smaller than 4.5Å. In our case, the number of native contacts identified from the crystal structure is $N = 156$ and $N = 279$ for Trp-Cage and BBA, respectively. Assuming that the atom pair (i, j) is one of the native contacts, we use $r_{ij}(X)$ to denote the distance between the two atoms in a given protein conformation X. The reaction coordinate Q for any conformation X is then determined by the distances for the N pairs of atoms in this conformation [42]:

$$Q(X) = \frac{1}{N} \sum_{ij}^{N} \frac{1}{1 + exp[\beta\ (r_{ij}(X) - \lambda\ r_{ij}^0)]} \qquad \text{Eq. 2-1}$$

with $\lambda = 1.8$ and a smoothing parameter $\beta = 5.0\mathring{A}^{-1}$. The summand in the equation above is effectively a pairwise contact strength that approaches 1 when the distance r_{ij} is small and approaches 0 when r_{ij} is large, thus quantifying the degree of contact between the two atoms. The reaction coordinate (Q) is the average overall pairwise contact strengths, thus representing the collective fraction of the native contacts present in a given conformation. A value of Q close to 1 indicates that the protein is in the native state because all of the native contacts are intact. In contrast, $Q = 0$ corresponds to completely non-native structures with all the native contacts broken.

2.2.3 Umbrella Sampling Simulations

We employed a total of 32 umbrella windows. The biasing potential in window i is in the harmonic form:

$$U_i(X) = \frac{K}{2}\,(Q(X) - q_i)^2 \qquad \text{Eq. 2-2}$$

in which $i = 1, \ldots, 32$. The spring constant K was taken to be 1400 kcal/mol for all the simulations in this study, and q_i is the center of the harmonic biasing potential. The values of q_i $(i = 1, \ldots, n)$ cover the range from 0 to 1 with a uniform spacing of 1/31.

To start the US simulations, we need a set of initial conformations with the reaction coordinate close to the q_i in each window. One common method to generate a diverse set of conformations is to run an equilibrium simulation at high temperatures [14]. Here we instead adopted pulling simulations, similar to the steered molecular dynamics [60], for this purpose. Specifically, we performed a simulation to drive the system from the native state ($Q \sim 1$) to the non-native state ($Q \sim 0$), by sequentially applying the 32 umbrella potentials for 0.4 ns each. The simulation thus lasted for a total of 12.8 ns. From this simulation trajectory, frames with the reaction coordinate close to each q_n were then selected as the initial coordinates for the respective umbrella window.

In the US, the umbrella windows were sampled by the same number of individual simulations (each referred to as a replica), and HREMD [22] was implemented to allow two neighboring windows to swap their replicas. The exchange was attempted every 200 time steps

(i.e., 0.4 ps). Suppose that umbrella windows i and j are a pair of neighbors, and that at the time of an exchange attempt, the current reaction coordinates are Q_i and Q_j, respectively. A swap would thus change the combined Hamiltonian by $\Delta \mathrm{E} = K\left[Q_i - Q_j\right][q_i - q_j]$, in which q_i, q_j, and K are from the harmonic biasing potential (Eq. 2-2). We accept the exchange with a probability of $min[\exp(-\frac{\Delta \mathrm{E}}{k_B T}),1]$ according to the Metropolis Criterion [22]. If the exchange is accepted, the two umbrella windows will swap their replicas, thus effectively exchanging the system microstates (coordinates, velocities, etc.).

We performed a total of four sets of US simulations, including the Trp-Cage system at 270 K, 280 K, and 290 K, and the BBA system at 325 K. Each simulation of Trp-Cage was run for 3.00 μs per window or a total of 96.00 μs for the 32 windows. The simulation of BBA was run for 1.01 μs per window or 32.32 μs in total. The initial coordinates for the Trp-Cage simulation at 290 K and the BBA simulation were taken from the pulling simulations described earlier. The last frames of the Trp-Cage simulation (290 K) were then used to initiate the US simulations at 280 K and 270K.

2.2.4 Analysis

The second half of the trajectories was used for the analysis of each simulation. Due to replica exchange, each umbrella window may be sampled by different replicas at different times of the simulation. We thus first reassembled the trajectories for each umbrella window. From these trajectories, we constructed the histograms of Q for each window, using a uniform bin width of $\Delta Q = 1.1 \times 10^{-4}$ for the Trp-Cage simulations and $\Delta Q = 2.0 \times 10^{-4}$ for the BBA simulation. Then the weighted histogram analysis method (WHAM) [61, 62] was used to calculate the equilibrium free energy as a function of Q.

With the equilibrium probability distribution of Q and the trajectories from the US simulations, we can reconstruct the equilibrium ensemble and obtain the probability distribution for any given parameter R, such as RMSD or radius of gyration. Specifically, we first group all frames in the simulation trajectories according to their values of Q. For each set of frames with the same Q, we construct the histogram for R as an estimate for the conditional probability $P(R|Q)$. In addition, $P_Q(Q)$, the marginal distribution for Q, is directly obtained from WHAM or the free

energy $G(Q)$. The joint equilibrium probability for R and Q is therefore given by $P(R,Q) = P(R|Q)P_Q(Q)$.

2.3 Results

As described in Methods, we performed US simulations with HREMD [22] on the Trp-Cage [47] and BBA [48] systems, using a reaction coordinate [42] Q based on the native contacts. The Trp-Cage system was simulated at three different temperatures.

2.3.1 Equilibrium Distributions Along With the Reaction Coordinate

Figure 2-1a shows the free energy profiles as a function of the reaction coordinate Q, obtained from the US simulation trajectories by WHAM [61, 62]. The statistical errors were estimated from the uncertainties of the mean forces at each window [62]. Overall, the free energy profiles here do not appear to describe a typical two-state system that has two major metastable states separated by a prominent energetic barrier. Instead, the profiles feature multiple minima and peaks with magnitudes not significantly larger than k_BT, thus indicating a continuous spectrum of intermediate conformations at equilibrium. In general, the locations of the major free energy barriers in our profiles are qualitatively similar to those reported by Best et al. [42] for the long unbiased simulations [11], although the magnitudes are not in good agreement. We caution that the two studies are not expected to yield similar quantitative results due to the different force fields adopted. Figure 2-2a shows the cumulative distribution function (CDF) that integrates the equilibrium probability along Q. For Trp-Cage at the three temperatures, the free energies, and the CDFs show that the equilibrium populations of the native (with large Q) and the non-native (with small Q) states are roughly comparable. For BBA at 325 K, in contrast, the vast majority of the equilibrium population is in the non-native state.

Any MD sampling has to start with some initial coordinates of the system, and convergence is only achieved when the "memory" has been completely lost, and the results become independent of the initial state. In our case here, although we discarded the first half of the trajectories in our analysis, slow equilibration in degrees of freedom orthogonal to the reaction coordinate could still potentially give rise to convergence issues. For umbrella sampling, one way to detect such issues is to examine the consistency between the histograms from neighboring windows. As described in

Ref. [62], the two neighboring histograms should ideally predict a consensus probability distribution for the overlapping region. An insufficient sampling of the orthogonal degrees of freedom,or hysteresis often manifests itself as an inconsistency between the histograms [62]. Therefore, for every pair of adjacent umbrella windows, we compared their consensus probability distributions (under a common potential) reconstructed from the two histograms. For such comparison, we adopted the inconsistency coefficient $\theta_{i,i+1}$ defined in Ref. [62] based on the Kolmogorov-Smirnov test. A θ value much larger than one would indicate an abnormal inconsistency between the two histograms. Figure 2-2b shows that all θ values from our simulations are below 1.05, and therefore no major inconsistency is detected. This analysis thus suggests that the calculated statistical errors here are reasonable estimates for the actual sampling errors.

HREMD [22] was implemented in our simulations, with the exchange rates between neighboring windows in the range of $20\% - 40\%$. In this scheme, the biasing potential on each replica undergoes a discrete random walk during the simulation [63]. The behavior of such random walk, quantified by parameters such as the transmission factors [63], could also potentially reveal regions with slow relaxation in the degrees of freedom orthogonal to the reaction coordinate [63]. The calculated transmission factors for our simulations did not exhibit significant variations [63] across different regions of Q, and thus did not indicate any particularly problematic region for the sampling. Figure 2-3 shows the umbrella windows sampled by each replica during the simulations. The sampled ranges for the individual replicas are clearly very different. The majority of the replicas visited a substantial range of the umbrella windows, with few covering almost the entire Q-range while some only covering a narrow section. It is well known that due to the effect of replica sorting [63], the replicas in HREMD simulations tend to be trapped in local regions.

The ultimate validation of an enhanced sampling method (such as US) would be a direct comparison to ideally long unbiased simulations. Although millisecond simulations [11] were not affordable here, we performed unbiased simulations from the native state of Trp-Cage at 280 K as an additional test. Specifically, we took a total of 32 frames in the US trajectories, with the reaction coordinate Q ranging from 0.94 to 0.98. From each frame, we initiated an unbiased simulation (without any restraint) for 344 ns. The histograms of each simulation from the second half (172 ns) of the trajectory are shown in Figure 2-4 (*dotted* lines). Remarkably, the histograms from these individual simulations are still significantly different from each other after 344 ns, thus indicating

that the equilibration is not very fast even when the protein is near the local free energy minimum for the native conformation, presumably due to the effects of other degrees of freedom. Whereas the protein in most unbiased simulations stayed in the native conformation during the 344 ns, we also observed a single spontaneous partial unfolding transition in one simulation, with the protein converted to some intermediate conformations with $Q\sim0.4$. Overall, despite the large variations among the individual histograms, their average is in reasonable agreement with the prediction from the US simulations (Figure 2-4).

In principle, with the knowledge of the free energy and the diffusion coefficients along the reaction coordinate, one may further obtain the kinetics of the transition [64, 65]. Although we performed some additional US simulations to calculate the diffusion coefficients [66], the statistical uncertainties appeared to be very large. Furthermore, the thermodynamics here does not indicate a two-state transition, as mentioned earlier. Therefore, we did not further estimate the folding/unfolding rates for the transition as in other studies [64, 65].

2.3.2 Energetics of the Conformational Space

The Gibbs free energy (G) can be decomposed as the enthalpy (H) and the entropy (S): $G = H - TS$. Our US simulations could provide these thermodynamic quantities for different conformational states (described by Q). As discussed earlier, the free energy as a function of Q was calculated by WHAM [61, 62]. Furthermore, we calculated the enthalpy for each frame in the simulation trajectories as $H = U + PV$, in which U is the potential energy for the underlying atomic interactions, V is the volume of the simulation system, and P is the pressure. Under the constant pressure of 1 atm, the variations in the PV term are much smaller than in the potential energy U. We took the average for all snapshots with the same Q as the enthalpy value at that Q. The entropy was then determined from the difference between the free energy and the enthalpy.

The enthalpy and entropy of each system are shown in Figure 2-1 b and c, along with the free energy. In general, the variations in the enthalpy here are larger than in the free energy. For BBA, as expected, the minimum enthalpy is at large Q representing the native state. For Trp-Cage, surprisingly, the enthalpy for the native state is actually not the global minimum. Instead, the enthalpy minimum for Trp-Cage is at $Q\sim0.5$, thus suggesting that some intermediate conformations, as will be described in more details later, actually have even more favorable potential energies than the native structure.

We also attempted to calculate the heat capacity for the conformations at different Q, obtained from the equilibrium energy fluctuation. However, the statistical uncertainties in this calculation are too large to reveal any clear difference of the heat capacity across the range of Q.

2.3.3 Stability of the Native Contacts

The Trp-Cage crystal structure consists of a short α-helix (residues 2-9) and a Polyproline-II segment (residues 16-19) connected by a loop (residues 10-15) that contains a 3_{10}-helix. The indole ring of the tryptophan residue (W6) is located at the center of the protein and makes contact with all of the three segments. Our simulation trajectories reveal different degrees of stability for the three segments, as shown in Figure 2-5 for the average fraction of the native contacts between each pair of protein residues for conformations at different Q. The contact maps for all three temperatures are quite similar, with the ones for 270 K and 290 K shown in the figure. Whereas the reaction coordinate Q is essentially an aggregate of the pairwise contacts, the maps indicate that the individual contact strengths do not simply increase linearly with Q from the non-native to the native states. Instead, the pairwise contacts are formed in different stages, thus implying different stabilities for the three segments.

In particular, the α-helix appears to have the most stable secondary structure. At a relatively low Q (0.3 or 0.4), the signature contacts within the α-helix already become prominent. In contrast, contacts involving the Polyproline-II and the loop segments appear to be less stable. For example, the native contacts between W6 and those two segments only start to form at $Q = 0.7$. Finally, some native contacts are quite weak even in the highly native conformations. For instance, the average contact strength for the D9-R16 salt bridge is smaller than 0.3 among the conformations at $Q = 0.9$.

Some insight on the relative stability can also be gained from the spontaneous transition away from the native structure observed in the unbiased simulation described earlier. In this transition, the α-helix remained essentially unchanged, whereas the loop and the Polyproline-II segment underwent large deviations from the initial native conformation. At the end of the partial transition, the protein is in a partly native conformation with an intact α-helix. This observation is consistent with our conclusion of a more stable α-helix and suggests that the unfolding of the α-helix would be the last step in reaching the completely non-native conformation.

2.3.4 Radius of Gyration

The free energies discussed above are directly related to the marginal probability distribution of Q at equilibrium, with all other degrees of freedom integrated out. It is thus possible that highly distinct conformations are mapped to the same value of Q. In the meantime, other parameters can be introduced to represent the equilibrium ensemble from different angles. As described in Methods, we can project the equilibrium ensemble onto any parameters and obtain the joint probability distribution. The free energy as a function of those relevant parameters may then reveal conformational states that otherwise cannot be distinguished by Q alone.

One relevant order parameter is the radius of gyration, Rg, which measures the geometric extendedness of the protein conformation [4, 28]. Figure 2-6 shows two-dimensional free energies as a function of Q and Rg, obtained from their joint probability distribution in the equilibrium ensemble. Qualitatively, the free energy maps for all simulations exhibit some common features. At large Q, the protein is in the native state, and Rg is therefore narrowly distributed around the value for the crystal structure. As Q decreases, the sampled range of Rg becomes increasingly larger, indicating the presence of more extended conformations. However, all major free energy minima, regardless of Q, are located at small values of Rg, and therefore the vast majority of the equilibrium population has Rg values similar to the crystal structure. Even for the non-native state near $Q = 0$ with all the native contacts completely lost, highly extended conformations (with large Rg) only represent a very small fraction of the population. These observations indicate that the non-native states here, albeit completely different from the crystal structure, are still folded in fairly compact geometries.

The two-dimensional free energy maps reveal a number of metastable conformations that are not clearly distinguishable in the one-dimensional profile. Some of the conformations are shown in Figure 2-6 for Trp-Cage at 270 K. At $Q\sim1$, conformation A is the native state as defined by the crystal structure. Around $Q\sim0.5$, conformations B-D are partly native conformations with the α-helix similar to the crystal structure, but the loop region highly different, especially for conformations C and D. In conformation B, the R16 guanidinium group simultaneously forms salt bridges with the carboxylate groups of both D1 and D9. In conformation C, the Polyproline-II segment contacts the α-helix, and the W6 indole ring forms an H-bond with the backbone carbonyl group of P12 or S13. Conformation D is similar to conformation C, except that the W6 indole ring H-bonds with the backbone carbonyl group of S14, G15, or R16, or with the sidechain of S13. At

low values of Q, conformations E-I correspond to completely non-native structures. Among them, conformation I is a fully extended structure with the maximum Rg (17 Å). The equilibrium population of this extended conformation, however, is small in comparison to other non-native conformations. Those conformations (E-H) have lost almost all of the native contacts but nonetheless are almost as compact (with Rg 7-9 Å) as the native structure (with Rg 6.9 Å). They are mainly stabilized by a different set of H-bonds that are not present in the native structure, as will be further discussed later.

Overall, the $Q - Rg$ free energy maps (Figure 2-6) of Trp-Cage at the three temperatures are qualitatively similar. The average Rg in the entire equilibrium ensemble is 8.1 Å, 7.8 Å, and 8.0 Å at 270 K, 280 K, and 290 K, respectively. However, the free energy minima corresponding to the distinct conformations discussed above are most prominent at 270 K, although those conformations can indeed be found (with somewhat lower probabilities) in the equilibrium ensembles at 280 K and 290 K as well. In addition, the relative free energy at small Q for 270 K is lower than that for the other two temperatures, thus indicating that the equilibrium population of the non-native conformations (such as the fully extended conformation) is higher at 270 K. For protein BBA, the two-dimensional free energy map indicates that the non-native state (with low Q) is more predominant than the other states (Figure 2-6), also consistent with its one-dimensional $G(Q)$ profile (Figure 2-1a). Similar to the case of Trp-Cage, the majority of the non-native BBA conformations are relatively compact, with Rg comparable to its native structure.

2.3.5 Hydrogen Bonds

H-bonds are believed to play important roles in the stability of protein conformations [39, 67]. We identified all H-bonds in the simulation trajectories, using a criterion that the donor-acceptor (which can be N or O atoms) distance be smaller than 4.0 Å and the donor-H-acceptor angle be larger than 140^0. The identified H-bonds are classified as native hydrogen bond (NHB) or non-native hydrogen bond (N-NHB), depending on whether they are present in the native crystal structure or not. By using the criteria above, there are a total of 12 NHBs in the crystal structure. One NHB is actually a salt bridge between the guanidinium group of R16 and the carboxylate group of D9. Another NHB is between the sidechain indole ring of W6 and the backbone carbonyl

group of R16. The other 10 NHBs are between the backbone amide N-H and the carbonyl C=O groups in residues 1-15.

Figure 2-7 shows two-dimensional free energy maps determined from the joint probability distribution of Q and the number of NHBs or N-NHBs in the equilibrium ensemble of Trp-Cage. As expected, the number of NHBs strongly correlates with Q, the fraction of the native contacts. For the free energy basin corresponding to the native state, most conformations have at least 7 NHBs. There are typically 4-6 NHBs in the intermediate conformations with Q between 0.3 and 0.7, whereas the non-native conformations have no more than 3 NHBs. In contrast, the number of N-NHBs does not appear to depend on Q. Even for the completely non-native conformations with $Q \sim 0$, the number of N-NHBs is similar to that in the native conformations. As discussed earlier, most conformations at Q still have folded geometries that are almost as compact as the native structure. Results here thus suggest that these compact non-native conformations are stabilized by different sets of H-bonds that are not present in the native structure.

2.3.6 Folding of the α-helix in Trp-Cage

As described earlier, the α-helix at the N-terminal of Trp-Cage is largely intact in the partly native conformations, suggesting that the formation of this α-helix would be an important step in the folding transition. We calculated the RMSD values (denoted as $RMSD_{hx}$) of the Cα atoms in the α-helix for all conformations in the simulation trajectories, using the native α-helix structure as the reference. Figure 2-8 displays the two-dimensional free energy maps as a function of Q and $RMSD_{hx}$ for the equilibrium ensemble. Interestingly, the free energies exhibit a more prominent two-state signature along with the $RMSD_{hx}$ parameter than along Q. There are two major minima along $RMSD_{hx}$: the minimum at $RMSD_{hx} \sim 0$ corresponds to the folded α-helix (such as conformations A-D in Figure 2-6), and the minimum around 3-5Å corresponds to the completely unfolded helix (such as conformations E-I in Figure 2-6). Some intermediate conformations (with $RMSD_{hx}$ around 2 Å) of a partially folded α-helix also exist, but only with minority populations. Overall, there is an energetic barrier along $RMSD_{hx}$, as identified in an earlier REMD study [68]. The free energy maps also show that the transitions along $RMSD_{hx}$ would occur when Q is around 0.3.

2.3.7 Folding/Unfolding Transition of the α-helix in Trp-Cage

Due to the biasing potentials, US simulations cannot directly reveal spontaneous transitions. However, with HREMD [22], each replica may sample multiple umbrella windows and thus a wide range of the reaction coordinate. An examination of the individual replicas in the Trp-Cage simulations (Figure 2-3) shows that although they all sampled a number of umbrella windows during the 1.5 μs simulation time, few replicas covered the entire range of Q. Moreover, the replicas can be roughly divided into two groups based on the α-helix conformation. The group of replicas with an unfolded α-helix segment mainly sampled the low-Q range, whereas the replicas with an intact and folded α-helix mainly sampled the high-Q range. Although the replicas in the same group had frequent exchanges with each other during the simulations, exchanges between replicas from different groups only occurred near the boundary ($Q \sim 0.3$) between the two ranges, almost without any replica moving far into the opposite range. Furthermore, complete transitions between the folded and unfolded α-helix were very rare, as we only observed two unfolding and two folding events among all simulation trajectories. Figure 2-9 shows the trajectories of two replicas in which a complete folding or unfolding of the α-helix occurred. Given that the helical conformation is maintained by the typical backbone H-bonds between residues 2–5 and residues 6–9, we display in the figure the time evolutions of each canonical H-bond in the trajectories.

In addition, the folding/unfolding of the α-helix was accompanied by large rotations of the backbone torsions, especially the ψ angles, and we therefore also show the time evolution of these dihedral angles in Figure 2-9. The four folding/unfolding events for the α-helix, including the two shown in Figure 2-9, share some common features. Overall, all transitions followed similar pathways in the Q-$RMSD_{hx}$ plane. Furthermore, from the unfolded to the folded conformation, the H-bonds (A2-W6, Y3-L7) in the N-terminal half of the α-helix were always formed earlier than those (A4-A8, Q5-D9) in the C-terminal half. In addition, other parts of the protein also underwent some conformational changes along with the folding of the α-helix. In particular, the Polyproline-II segment tended to move away from the α-helix during the folding/unfolding transitions, thus resulting in intermediate protein conformations with more extended geometry and higher radius of gyration in comparison to the structures at the two ends. Similar intermediate states were also reported by Juraszek and Bolhuis for a Trp-Cage of a slightly different sequence [49]. Other than the commonalities above, the four folding/unfolding events differed in the order and timing of the individual changes in the H-bonds and the torsions.

2.4 Discussion

In this study, using a reaction coordinate representing the collective fraction of the native contacts, we carried out US [46] simulations in combination with HREMD [22] to sample the protein conformational space. Overall, the free energy calculation (Figure 2-1a) appears to have converged, and the consistency test (Figure 2-2b) suggests that the statistical errors in the free energy have been reasonably estimated. The equilibrium ensemble of protein conformations thus appears to be satisfactorily reconstructed from these simulations.

The reconstructed equilibrium ensembles reveal multiple folded conformations for the two proteins here, Trp-Cage and BBA. The reaction coordinate Q only quantifies the resemblance to the native structure but does not describe the compactness of the conformation. In fact, the non-native state does not merely consist of disordered or extended conformations. Even at $Q\sim0$, with all native contacts completely broken, the majority of the populations are still comprised of well-defined conformations almost as compact as the native structure (Figure 2-6), and these folded conformations are stabilized by some H-bonds (Figure 2-7) not present in the native structure. For Trp-Cage, some alternatively folded conformations have even lower enthalpy than the native structure (Figure 2-1b). In the presence of such conformations [68], therefore, the conformational space would not be described by a simple two-state model with a folded conformation and an unfolded state of disordered conformations.

For Trp-Cage, the α-helix at the N-terminal plays an important role in the folding of this protein. UV resonance Raman spectroscopy [69] detected in the unfolded ensemble the presence of compact intermediate conformations with the intact α-helix, and concluded that the Trp-Cage is not a two-state folder [69]. Infrared spectroscopy also indicated that the α-helix is fully formed in the folding transition state [70]. These conclusions were further supported by recent simulations [71]. Our simulations here showed that the α-helix is more stable than other parts of the protein (Figure 2-5) and is largely intact in the intermediate conformations at $Q\sim0.5$. Furthermore, the spontaneous partial unfolding transition in one of our unbiased simulations showed that the α-helix remained intact when other parts of the protein deviated from the native conformation. Therefore, our simulations are fully consistent with the previous findings that the α-helix is formed at the early folding stage, although we caution that the Trp-Cage sequences in those experimental studies [69, 70] are slightly different from ours. Importantly, our reconstructed equilibrium ensemble revealed that the transition between the folded and unfolded α-helix is almost orthogonal to the

reaction coordinate Q (Figure 2-8). Consequently, the restraint on Q in the US simulations cannot enhance the sampling of the α-helix conformations, which would thus compromise the sampling efficiency and contribute to the statistical errors in the free energy. Furthermore, the one-dimensional free energy as a function of Q does not reflect the energetic barrier between the folded and unfolded conformations of the α-helix (Figure 2-8). In fact, the folding of the α-helix resembles a two-state process more than the folding of the entire Trp-Cage does, as also noted in previous experiments [69].

Trp-Cage at various temperatures has been studied in NMR experiments [68, 72]. Here we carried out simulations at three different temperatures (270 K, 280 K, and 290 K) for this protein. Whereas the reconstructed equilibrium ensembles at these temperatures are qualitatively similar to each other, it is notable that the non-native state turns out to have a higher equilibrium probability at the lowest temperature (270 K) than at the other temperatures (Figure 2-2a). This somewhat unexpected result may be attributed to several factors. First, given the relatively small magnitude of the free energies here, the statistical errors in our calculation are relatively large. Consequently, the differences in the calculated equilibrium probabilities at the three temperatures are not much larger than the estimated statistical uncertainty. More importantly, as discussed earlier, the equilibrium ensembles consist of multiple folded conformations. Some alternatively folded conformations are enthalpically even more favorable than the native structure (Figure 2-1b). Consequently, lowering the temperature is not guaranteed to shift the equilibrium toward the native structure and away from other folded conformations. In fact, at the lowest temperature (270 K) here, the completely non-native conformations (at $Q\sim0$) have even lower relative enthalpies, which could be responsible for their higher equilibrium populations than at the other temperatures. Finally, some Trp-Cage mutant was found to exhibit cold denaturation at low temperatures [73, 74], and this mechanism remains a possibility in our case as well.

Despite some qualitative agreement, our results considerably deviate from previous simulations [11, 42]. Most notably, here some compact non-native Trp-Cage conformations have even lower enthalpies than the native structure does, which is clearly unexpected. For BBA, moreover, our free energy profile (Figure 2-1a) indicates that the non-native state is significantly more stable (by ~4 kcal/mol) than the native state, which is also different from previous simulations [11]. Such discrepancies are most likely due to the force field issues. First, our version of the CHARMM36 force field was retrieved before the most recent updates for improving the

sampling of disordered protein states. More importantly, oversampling of compact conformations have been identified as a common deficiency of some force fields [75-78], and the high populations of compact conformations in our equilibrium ensemble may well be due to such artifacts. In addition, the CHARMM force field is known to over-stabilize the interaction between the guanidinium and the carboxylate groups [79, 80], thus very likely responsible for the unexpectedly low enthalpy for the Trp-Cage conformations at $Q\sim0.5$, some of which (Figure 2-6}, conformation B) are indeed stabilized by salt bridges between the ARG and ASP residues. Although the optimized CHARMM22* force field [79] appears to produce excellent results in folding simulations [11], the predicted enthalpy for Trp-Cage still has a large discrepancy [11] with experiments. In light of such problems, it should be worthwhile to use the many available NMR data [68, 72] on small model proteins such as Trp-Cage to validate and calibrate the force fields [81].

As mentioned before, with a good reaction coordinate, many enhanced sampling methods, including the US simulations adopted in this study, can be applied to sample the protein conformations. Here we demonstrated that using Q as the reaction coordinate, US in combination with HREMD [22] could reasonably sample the protein conformational space and reconstruct the equilibrium ensemble. The efficiency of such methods relative to the unbiased simulations, however, clearly depends on the underlying kinetics. For the Trp-Cage with relatively fast transition rates here, given the aggregated simulation times one could alternatively obtain multiple spontaneous transitions in unbiased simulations. The advantage of the US approach is therefore not prominent here (other than a technical gain of much better parallel efficiency). However, the required sampling time for unbiased simulations may increase by many orders of magnitude for proteins with slow kinetics. Even for BBA, a fast-folding protein, because the system is not at the melting temperature [10, 11] here, in unbiased simulations the protein would predominantly stay in the non-native state and the spontaneous transitions will be significantly less frequent, thus requiring much longer simulation times. In contrast, with a good reaction coordinate, the computational cost for the US [14, 46] and other enhanced sampling methods would not be nearly as sensitive to the height and skewness of the underlying free energy, and they have been routinely used to calculate free energies with high barriers in many applications. Furthermore, unlike the temperature replica exchange simulations which typically require more replicas for systems of

higher atom count (such as in the explicit-solvent simulations), the enhanced sampling methods based on a reaction coordinate can be readily applied to systems of any size.

On the other hand, the success of the US as well as many other methods critically depend on the quality of the adopted reaction coordinate. An ideal reaction coordinate should ensure that all orthogonal motions can be well equilibrated within the simulation time. A poor reaction coordinate could severely compromise the sampling efficiency as well as cause other problems. The fraction of the native contacts, Q, appears to be a reasonable reaction coordinate, as we could generate the non-native states and reproduce the equilibrium distribution by applying restraints on Q alone in the simulations. On the other hand, Q is probably not always a perfect reaction coordinate for enhanced sampling, as we also identified slow equilibration of an orthogonal degree of freedom, i.e., the folding/unfolding of the α-helix, for the protein Trp-Cage here. In such cases, Hamiltonian replica exchange could somewhat alleviate the problem of slow orthogonal relaxations and facilitate the sampling along an imperfect reaction coordinate [63]. We also note that the identified problems with Q may be partly due to the force field issues discussed earlier, as Q was shown to be a very good reaction coordinate [42] to analyze folding simulations [11] using the CHARMM22* force field. Nonetheless, our finding in this study suggests that the reaction coordinate Q could be improved, e.g., by better incorporating the slow degrees of freedom representing the α-helix conformation for Trp-Cage, and that an improved reaction coordinate should further enhance the sampling efficiency.

2.5 Figures

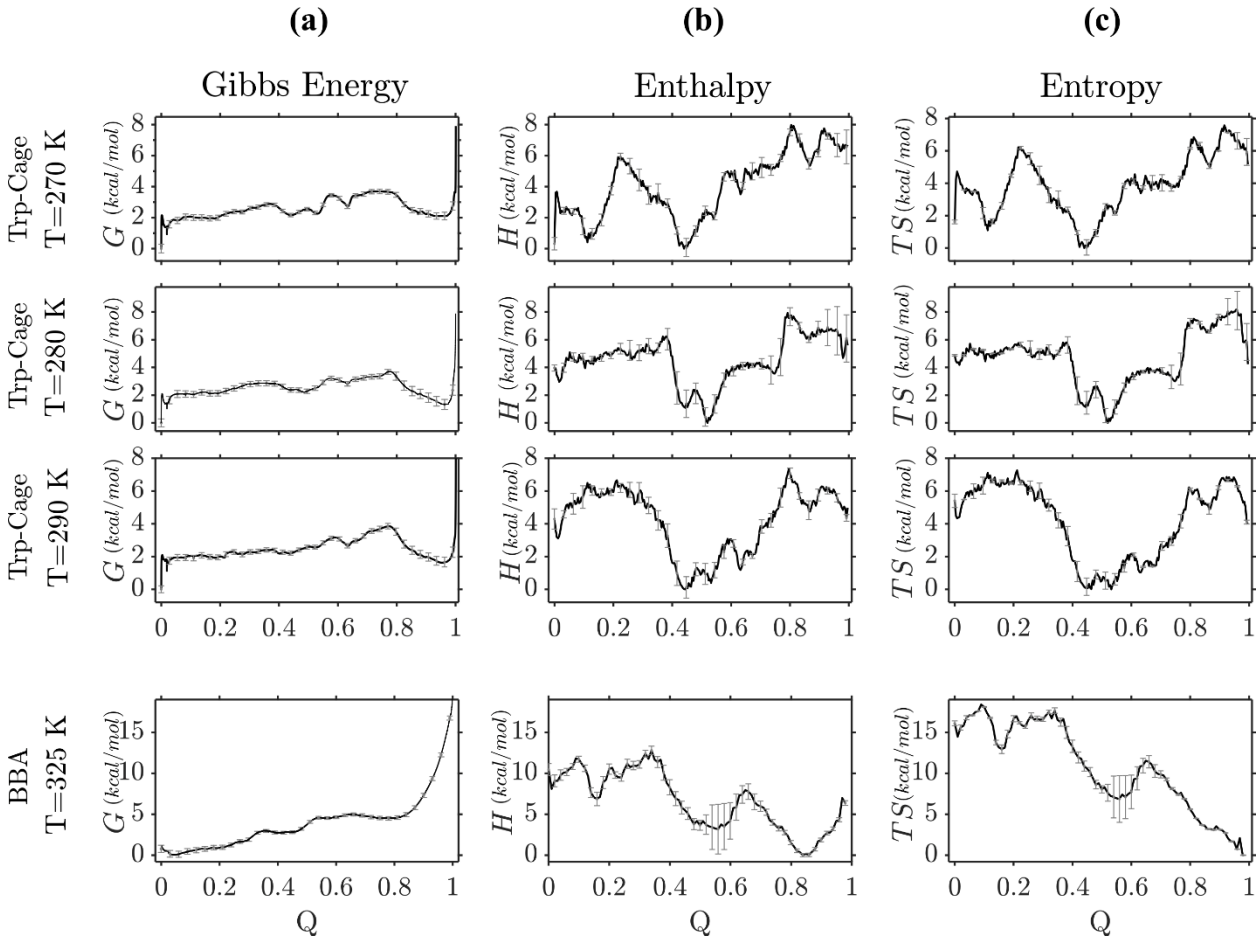


Figure 2-1. Energetics along with the reaction coordinate Q from the US simulations. a) The free energy profiles calculated from the WHAM [61, 62] equations. The statistical errors are with respect to the difference between the free energy value at the given position and the average value of the entire profile and were estimated from the uncertainties in the mean force at each umbrella window [62]. b) The profile of average enthalpy along with Q. c) The entropy multiplied by the temperature.

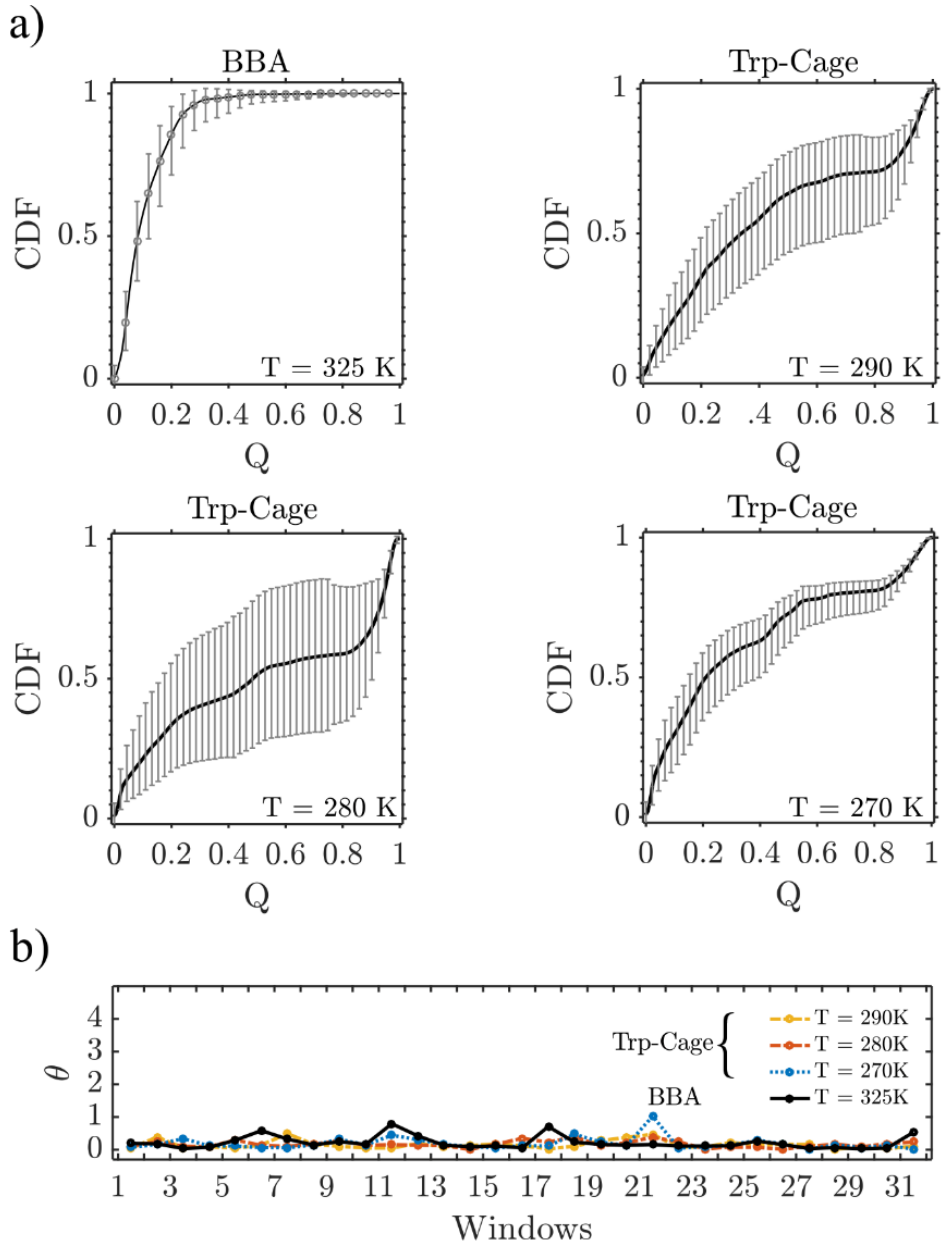


Figure 2-2. a) Cumulative distribution function obtained by integrating the equilibrium probability distribution along with Q. The error bars at each data point were estimated separately. For any given point Q_i, the upper and lower bounds (taken as ±1 standard deviation) for the profile of the free energy differences relative to Q_i were obtained (similarly from the statistical errors in the mean force for each window) and used to calculate the upper and lower limits for the cumulative probability at Q_i. b) Inconsistency coefficient θ for pairs of histograms in the adjacent umbrella windows [62].

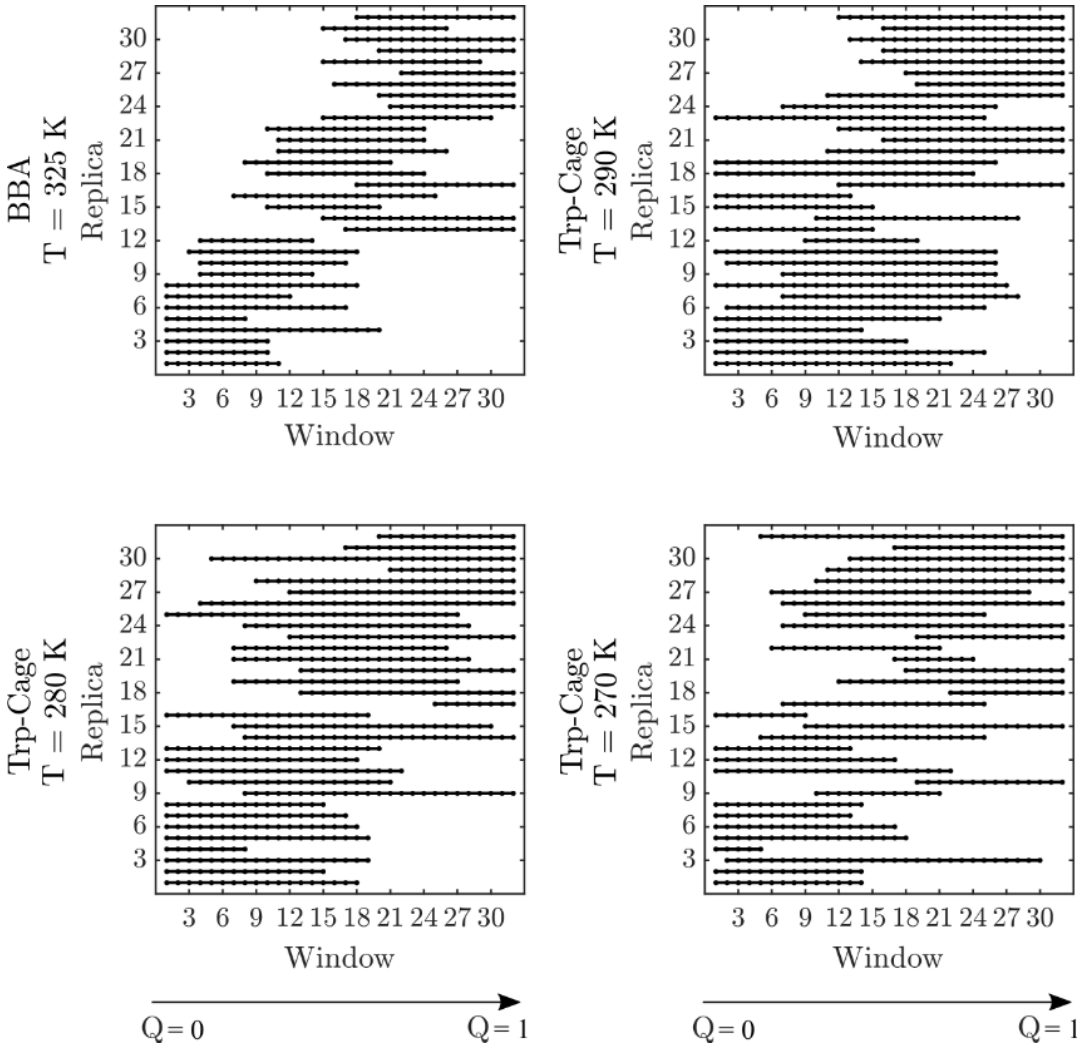


Figure 2-3. The umbrella windows that each replica sampled during the second half of the US simulations.

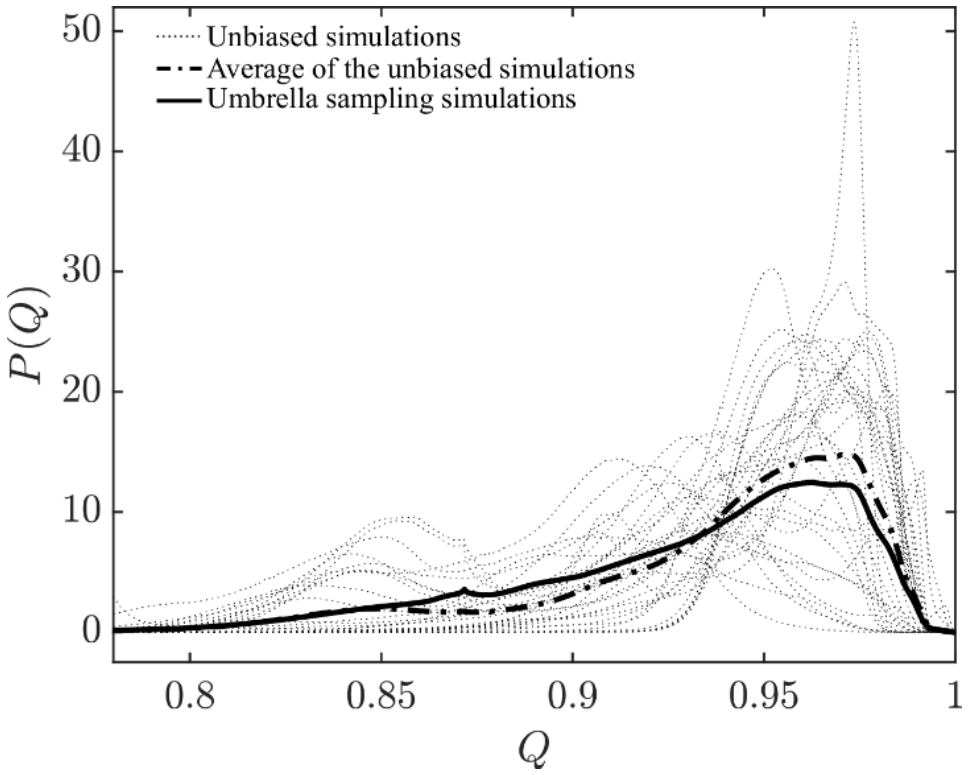


Figure 2-4. Data from the 32 unbiased simulations (344 ns each) at the native state of Trp-Cage at 280 K. For each unbiased simulation, the histogram from the second half (172 ns) of the trajectory is shown as a dotted line. The average of the 32 histograms is shown as the dashed line. The solid line shows the normalized equilibrium probabilities for the range of Q representing the native conformation, which was calculated from the US simulations (cf. . Energetics along with the reaction coordinate Q from the US simulations. a) The free energy profiles calculated from the WHAM [61, 62] equations. The statistical errors are with respect to the difference between the free energy value at the given position and the average value of the entire profile and were estimated from the uncertainties in the mean force at each umbrella window [62]. b) The profile of average enthalpy along with Q. c) The entropy multiplied by the temperature.a).

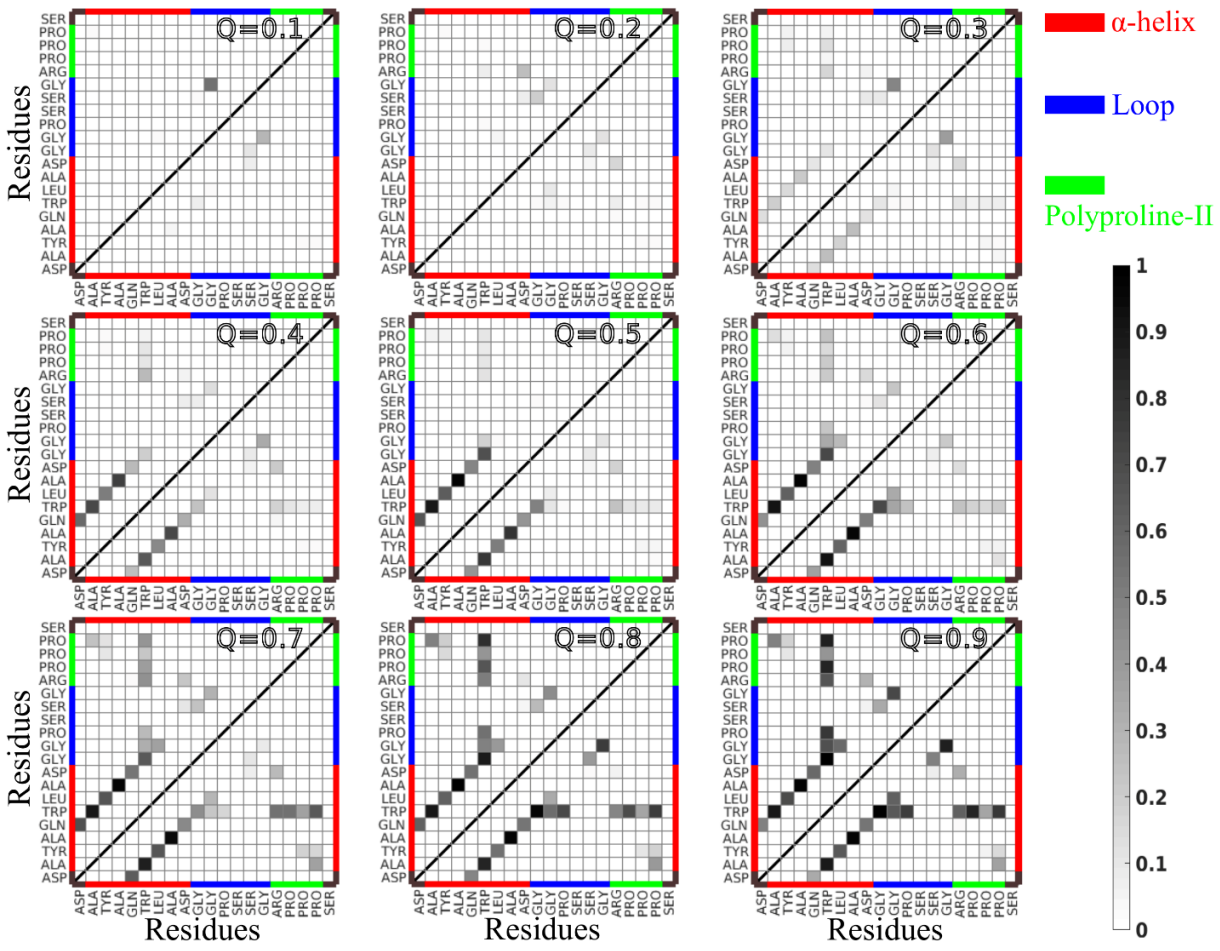


Figure 2-5. The fraction of the native contacts (or the average contact strength) between each pair of residues in the Trp-Cage conformations with different Q values at 270 K (*upper left*) and 290 K (*lower right*). For each Q value, conformations within $Q \pm 0.01$ were taken to calculate the average contact strength between every residue pair in the protein.

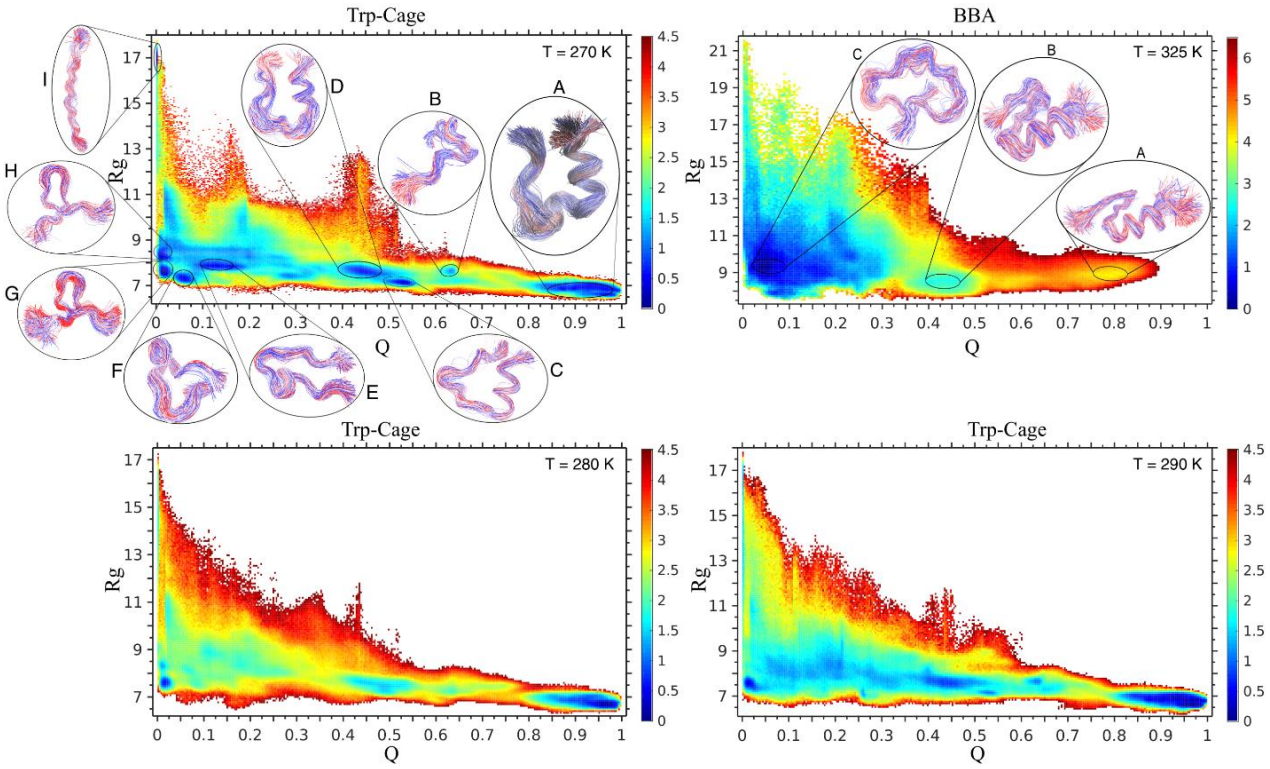


Figure 2-6. Two-dimensional free energy (in unit of kcal/mol) maps as a function of the reaction coordinate (Q) and the radius of gyration (Rg, in unit of Å) of the protein conformation, for Trp-Cage at 270 K, 280 K, and 290 K and BBA at 325 K. The free energies were determined from the joint probability distribution of Q and Rg in the equilibrium ensemble. Some representative conformations at various free energy minima are also shown in the figure.

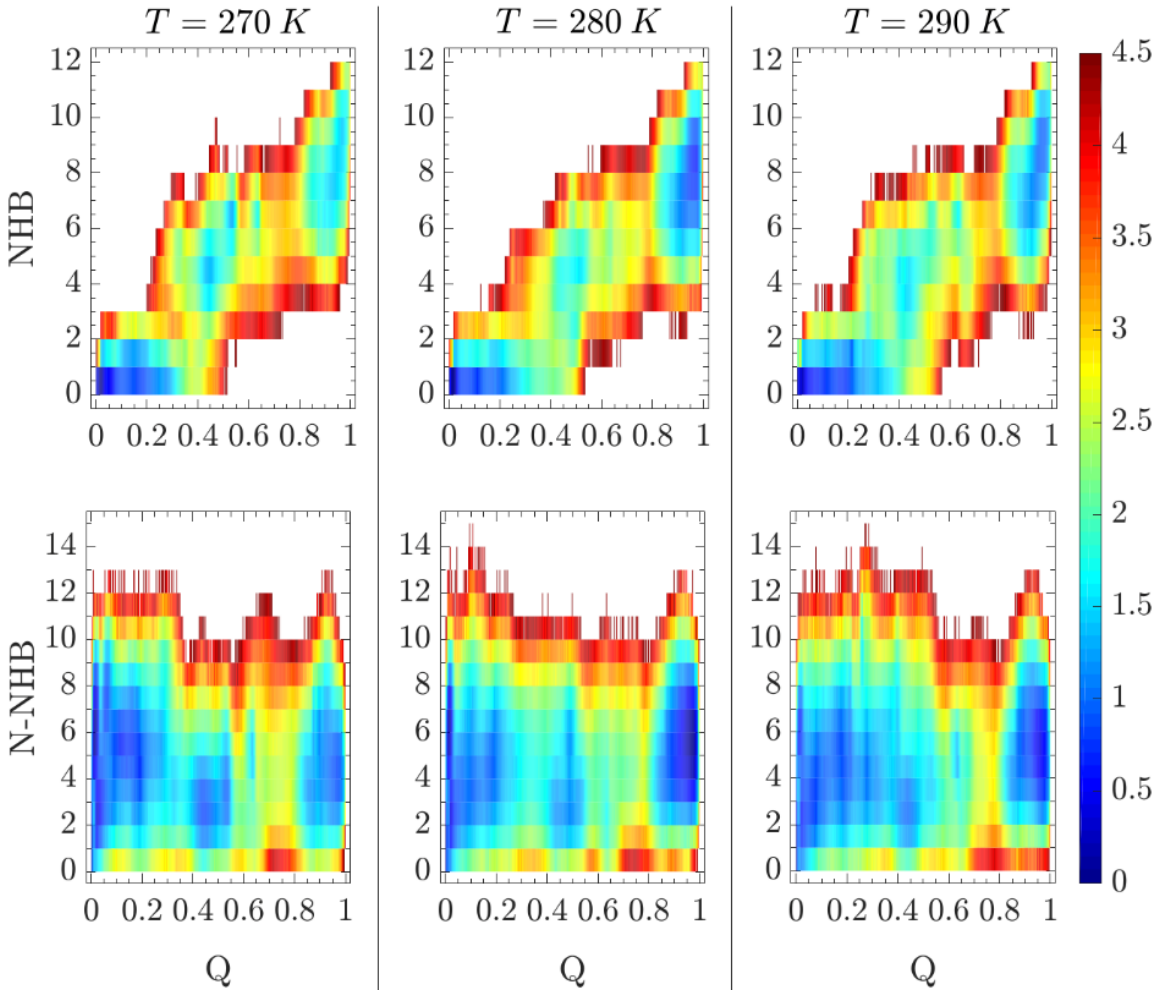


Figure 2-7. Two-dimensional free energy (in unit of kcal/mol) maps as a function of Q and the number of NHBs (*first row*) or the number of N-NHBs (*second row*) for Trp-Cage at 270 K (*left*), 280 K (*middle*) and 290 K (*right*). The free energies were determined from the joint probability distribution of Q and the H-bond count in the equilibrium ensemble.

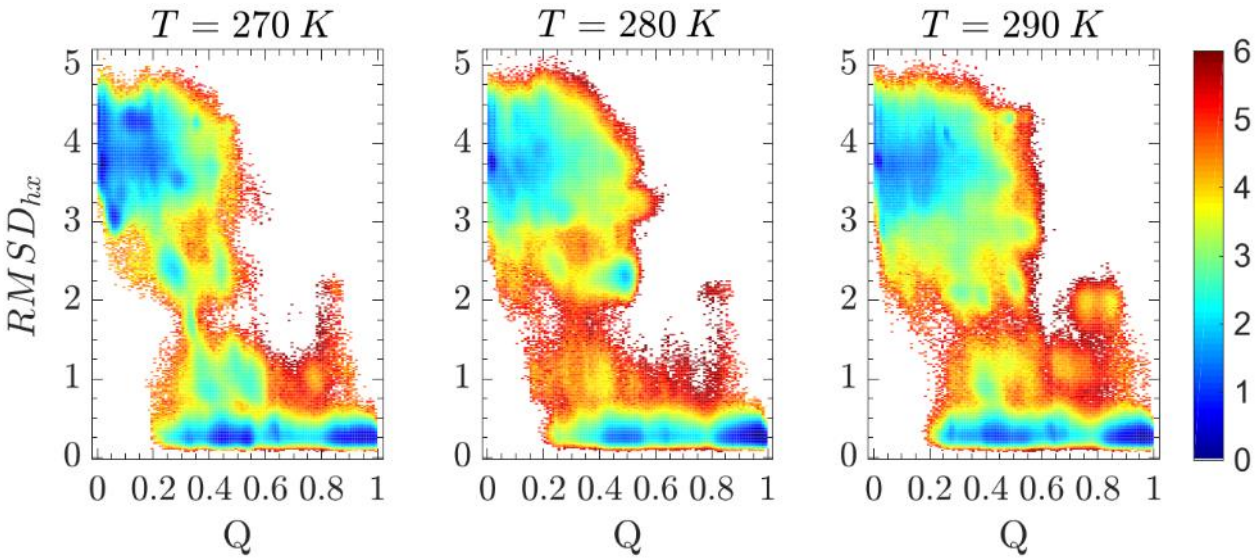


Figure 2-8. Two-dimensional free energy (in the unit of kcal/mol) maps as a function of Q and the C_α RMSD (in unit of Å) for the α-helix (residue 2-9) in Trp-Cage at 270 K (*left*), 280 K (*middle*) and 290 K (*right*). The free energies were determined from the joint probability distribution of Q and the RMSD in the equilibrium ensemble.

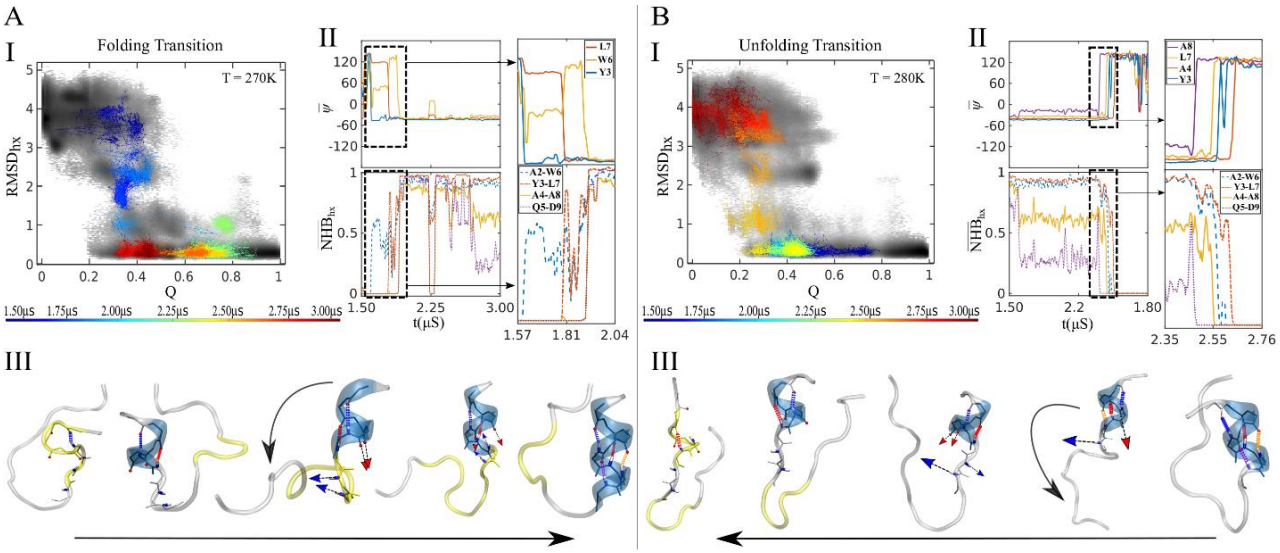
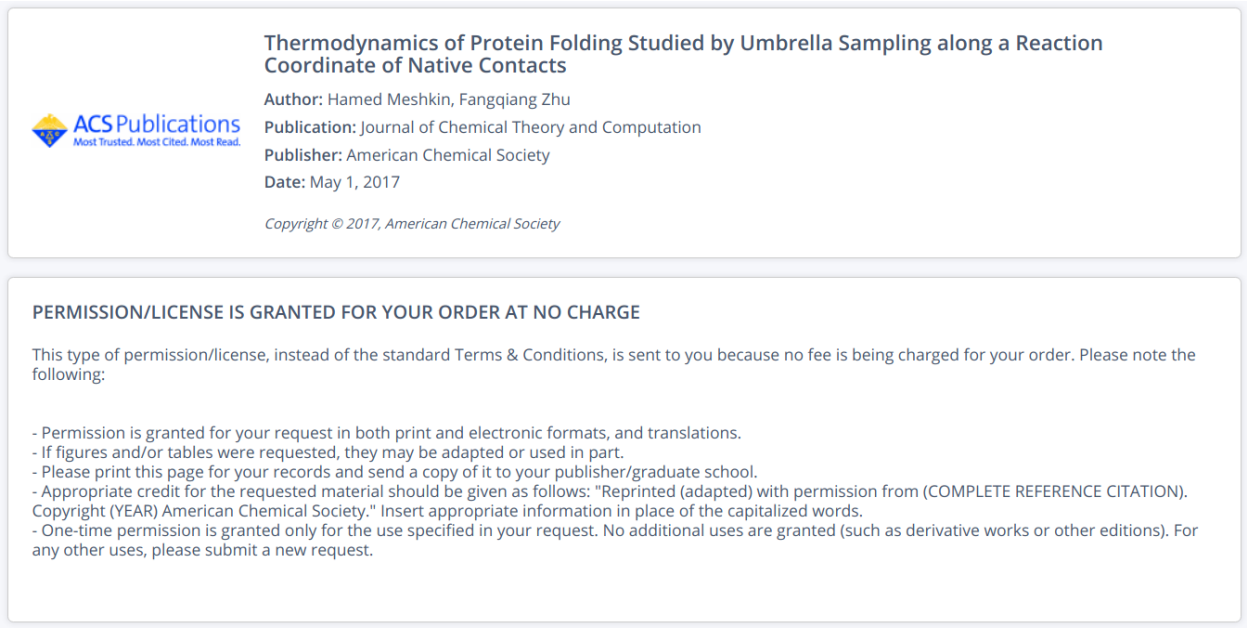


Figure 2-9. Time evolution in two of the replicas in which folding (A) or unfolding (B) of the α-helix in Trp-Cage occurred. Panel I shows the trajectory projected onto the Q-RMSD hx plane, colored by the progression in time (with a total of 1.5 μs). The equilibrium free energy (as in Figure 2-8) is displayed in the background in grayscale. Panel II shows the trajectories for some NHBs (with 1 and 0 representing formed and not formed, respectively) and backbone ψ angles in the α-helix, after being smoothed by time-averaging in intervals of 3 ns. The part of the trajectories in which the transition occurs is indicated by the dashed rectangles and also shown in zoom-in plots. Panel III shows some snapshots before, during and after the transition in the trajectory. The red and blue arrows indicate the directions of the A8/D9 amino (N−H) groups and the A4/Q5 carbonyl (C − O) groups, respectively.

Acknowledgment

This work is republished from Chemical Theory and Computation Journal, Meshkin, Hamed, and Fangqiang Zhu. "Thermodynamics of protein folding studied by umbrella sampling along a reaction coordinate of native contacts." Journal of chemical theory and computation 13.5 (2017): 2086-2097. Copyright (2017), with permission from ACS Publications.



Thermodynamics of Protein Folding Studied by Umbrella Sampling along a Reaction Coordinate of Native Contacts

Author: Hamed Meshkin, Fangqiang Zhu

Publication: Journal of Chemical Theory and Computation

Publisher: American Chemical Society

Date: May 1, 2017

2.6 References

1. Chan, et al., *Local-Nonlocal Coupling, and Nonnative Interactions: Principles of Protein Folding from Coarse- Grained Models.* Annual Review Physics Chemistry, 2011. **62**: p. 301-326.

2. Dobson, C.M., *Protein Folding and Misfolding.* Nature, 2003. **426**: p. 884-890.

3. Bryngelson, J.D. and P.G. Wolynes, *Spin Glasses and the Statistical Mechanics of Protein Folding.* Proceedings of the National Academy of Sciences, 1987. **84**: p. 7524-7528.

4. Bryngelson, J.D., et al., *Funnels, pathways, and the energy landscape of protein folding: a synthesis.* Proteins: Structure, Function, and Bioinformatics, 1995. **21**: p. 167-195.

5. Dickson, A. and C.L. Brooks III, *Native states of fast-folding proteins are kinetic traps.* Journal of the American Chemical Society, 2013. **135**: p. 4729-4734.

6. Plaxco, K.W., K.T. Simons, and D. Baker, *Contact Order, Transition State Placement and the Refolding Rates of Single Domain Proteins.* Journal of Molecular Biology, 1998. **277**: p. 985-994.

7. Alm, E. and D. Baker, *Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures.* Proceedings of the National Academy of Sciences, 1999. **96**(20): p. 11305-11310.

8. Clementi, C., H. Nymeyer, and J.N. Onuchic, *Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? an investigation for small globular proteins.* Journal of Molecular Biology, 2000. **937-953**: p. 5.

9. Klenin, K., et al., *Modelling proteins: Conformational sampling and reconstruction of folding kinetics.* Biochimica et Biophysica Acta - Proteins and Proteomics, 2011. **1814**: p. 977-100.

10. Shaw, D.E., et al., *Atomic-level characterization of the structural dynamics of proteins.* Science, 2010. **330**(6002): p. 341-346.

11. Lindorff-Larsen, K., et al., *How fast-folding proteins fold.* Science, 2011. **334**(6055): p. 517-520.

12. Adcock, S.A. and J.A. McCammon, *Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins.* Chemical reviews, 2006. **106**: p. 1589-1615.

13. Abrams, C. and G. Bussi, *Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration.* Entropy, 2013. **16**: p. 163-199.

14. Shea, J.-E. and C.L. Brooks III, *From folding theories to folding proteins: a review and assessment of simulation studies of protein folding and unfolding.* Annual review of physical chemistry, 2001. **52**(1): p. 499-535.

15. Kastner, J., *Umbrella sampling.* Wiley Interdisciplinary Reviews: Computational Molecular Science, 2011. **1**: p. 932-942.

16. Barducci, A., M. Bonomi, and M. Parrinello, *Metadynamics.* Wiley Interdisciplinary Reviews: Computational Molecular Science, 2011. **1**(5): p. 826-843.

17. Dickson, A. and C.L. Brooks III, *WExplore: hierarchical exploration of high-dimensional spaces using the weighted ensemble algorithm.* The Journal of Physical Chemistry B, 2014. **118**(13): p. 3532-3542.

18. Zwier, M.C., et al., *WESTPA: An Interoperable, Highly Scalable Software Package for Weighted Ensemble Simulation and Analysis.* Journal of Chemical Theory and Computation, 2015. **11**: p. 800-809.

19. Machta, J. and R.S. Ellis, *Monte Carlo Methods for Rough Free Energy Landscapes: Population Annealing and Parallel Tempering.* Journal of Statistical Physics, 2011. **144**: p. 341-553.

20. Zhang, J., et al., *All-atom replica exchange molecular simulation of protein BBL.* Proteins: Structure, Function, and Bioinformatics, 2008. **72**: p. 1038-1047.

21. Kannan, S. and M. Zacharias, *Folding simulations of Trp-cage mini protein in explicit solvent using biasing potential replica-exchange molecular dynamics simulations.* Proteins: Structure, Function, and Bioinformatics, 2009. **76**: p. 448-460.

22. Fukunishi, H., O. Watanabe, and S. Takada, *On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction.* The Journal of chemical physics, 2002. **116**(20): p. 9058-9067.

23. Affentranger, R., I. Tavernelli, and E.E. Di Iorio, *A novel Hamiltonian replica exchange MD protocol to enhance protein conformational space sampling.* Journal of Chemical Theory and Computation, 2006. **2**(2): p. 217-228.

24. Kannan, S. and M. Zacharias, *Enhanced sampling of peptide and protein conformations using replica exchange simulations with a peptide backbone biasing-potential.* Proteins: Structure, Function, and Bioinformatics, 2007. **66**: p. 697-706.

25. Sabri Dashti, D. and A.E. Roitberg, *Optimization of Umbrella Sampling Replica Exchange Molecular Dynamics by Replica Positioning.* Sabri Dashti, Danial and Roitberg, Adrian E, 2013. **9**: p. 4692-4699.

26. Pande, V.S., K. Beauchamp, and G.R. Bowman, *Everything you wanted to know about Markov State Models but were afraid to ask.* Methods, 2010. **52**(1): p. 99-105.

27. Shakhnovich, E., et al., *Protein folding bottlenecks: A lattice Monte Carlo simulation.* Physical review letters, 1991. **67**: p. 1665.

28. Shea, J.-E., J.N. Onuchic, and C.L. Brooks III, *Energetic frustration and the nature of the transition state in protein folding.* The Journal of Chemical Physics, 2000. **113**(17): p. 7663-7671.

29. Hardin, C., Z. Luthey-Schulten, and P.G. Wolynes, *Backbone dynamics, fast folding, and secondary structure formation in helical proteins and peptides.* Proteins: Structure, Function, and Bioinformatics, 1999. **34**: p. 281-294.

30. Eastwood, M.P. and P.G. Wolynes, *Role of explicitly cooperative interactions in protein folding funnels: a simulation study.* The Journal of Chemical Physics, 2001. **114**(10): p. 4702-4716.

31. Pogorelov, T.V. and Z. Luthey-Schulten, *Variations in the fast folding rates of the λ-repressor: A hybrid molecular dynamics study.* Biophysical journal, 2004. **87**(1): p. 207-214.

32. Sheinerman, F.B. and C.L. Brooks, *Molecular picture of folding of a small α/β protein.* Proceedings of the National Academy of Sciences, 1998. **95**(4): p. 1562-1567.

33. Chen, J. and C.L. Brooks III, *Can molecular dynamics simulations provide high-resolution refinement of protein structure?* Proteins: Structure, Function, and Bioinformatics, 2007. **67**(4): p. 922-930.

34. Best, R.B., et al., *Pulling direction as a reaction coordinate for the mechanical unfolding of single molecules.* The Journal of Physical Chemistry B, 2008. **112**(19): p. 5968-5976.

35. Best, R.B., et al., *Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles.* Journal of chemical theory and computation, 2012. **8**(9): p. 3257-3273.

36. Levy, Y. and O.M. Becker, *Energy landscapes of conformationally constrained peptides.* The Journal of Chemical Physics, 2001. **114**(2): p. 993-1009.

37. Vengadesan, K. and N. Gautham, *Energy landscape of Met-enkephalin and Leu-enkephalin drawn using mutually orthogonal Latin squares sampling.* The Journal of Physical Chemistry B, 2004. **108**(30): p. 11196-11205.

38. Itoh, K. and M. Sasai, *Flexibly varying folding mechanism of a nearly symmetrical protein: B domain of protein A.* Proceedings of the National Academy of Sciences, 2006. **103**(19): p. 7298-7303.

39. Guo, W., S. Lampoudi, and J.-E. Shea, *Posttransition state desolvation of the hydrophobic core of the src-SH3 protein domain.* Biophysical journal, 2003. **85**(1): p. 61-69.

40. Juraszek, J. and P.G. Bolhuis, *Rate constant and reaction coordinate of Trp-cage folding in explicit water.* Biophysical journal, 2008. **95**(9): p. 4246-4257.

41. Jiang, F. and Y.-D. Wu, *Folding of fourteen small proteins with a residue-specific force field and replica-exchange molecular dynamics.* Journal of the American Chemical Society, 2014. **136**: p. 9536-9539.

42. Best, R.B., G. Hummer, and W.A. Eaton, *Native contacts determine protein folding mechanisms in atomistic simulations.* Proceedings of the National Academy of Sciences, 2013. **110**(44): p. 17874-17879.

43. Guo, W., S. Lampoudi, and J.E. Shea, *Temperature dependence of the free energy landscape of the src-SH3 protein domain.* Proteins: Structure, Function, and Bioinformatics, 2004. **55**(2): p. 395-406.

44. Bursulaya, B.D. and C.L. Brooks, *Folding free energy surface of a three-stranded β-sheet protein.* Journal of the American Chemical Society, 1999. **121**(43): p. 9947-9951.

45. Sun, L., et al., *Connecting thermal and mechanical protein (un) folding landscapes.* Biophysical journal, 2014. **107**(12): p. 2950-2961.

46. Kästner, J., *Umbrella sampling.* Wiley Interdisciplinary Reviews: Computational Molecular Science, 2011. **1**(6): p. 932-942.

47. Meuzelaar, H., et al., *Folding dynamics of the Trp-cage miniprotein: evidence for a native-like intermediate from combined time-resolved vibrational spectroscopy and molecular dynamics simulations.* The Journal of Physical Chemistry B, 2013. **117**(39): p. 11490-11501.

48. Sarisky, C.A. and S.L. Mayo, *The ββα fold: explorations in sequence space.* Journal of molecular biology, 2001. **307**(5): p. 1411-1418.

49. Juraszek, J. and P. Bolhuis, *Sampling the multiple folding mechanisms of Trp-cage in explicit solvent.* Proceedings of the National Academy of Sciences, 2006. **103**: p. 15859-15864.

50. Marinelli, F., et al., *A kinetic model of trp-cage folding from multiple biased molecular dynamics simulations.* PLoS Comput Biol, 2009. **5**: p. e1000452.

51. Barua, B., et al., *The Trp-cage: optimizing the stability of a globular miniprotein.* Protein Engineering Design and Selection, 2008. **21**: p. 171-185.

52. MacKerell Jr, A.D., et al., *All-atom empirical potential for molecular modeling and dynamics studies of proteins.* The journal of physical chemistry B, 1998. **102**(18): p. 3586-3616.

53. MacKerell Jr, A.D., M. Feig, and C.L. Brooks, *Improved treatment of the protein backbone in empirical force fields.* Journal of the American Chemical Society, 2004. **126**(3): p. 698-699.

54. Jorgensen, W.L., et al., *Comparison of simple potential functions for simulating liquid water.* The Journal of chemical physics, 1983. **79**(2): p. 926-935.

55. Phillips, J.C., et al., *Scalable molecular dynamics with NAMD.* Journal of computational chemistry, 2005. **26**: p. 1781-1802.

56. Feller, S.E., et al., *Constant pressure molecular dynamics simulation: the Langevin piston method.* The Journal of chemical physics, 1995. **103**(11): p. 4613-4621.

57. Ryckaert, J.-P., G. Ciccotti, and H.J. Berendsen, *Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes.* Journal of computational physics, 1977. **23**(3): p. 327-341.

58. Miyamoto, S. and P.A. Kollman, *Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models.* Journal of computational chemistry, 1992. **13**(8): p. 952-962.

59. Darden, T., D. York, and L. Pedersen, *Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems.* The Journal of chemical physics, 1993. **98**(12): p. 10089-10092.

60. Park, S. and K. Schulten, *Calculating potentials of mean force from steered molecular dynamics simulations.* The Journal of chemical physics, 2004. **120**: p. 5946-5961.

61. Kumar, S., et al., *The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method.* Journal of computational chemistry, 1992. **13**(8): p. 1011-1021.

62. Zhu, F. and G. Hummer, *Convergence and error estimation in free energy calculations using the weighted histogram analysis method.* Journal of computational chemistry, 2012. **33**(4): p. 453-465.

63. Neale, C., et al., *Accelerating convergence in molecular dynamics simulations of solutes in lipid membranes by conducting a random walk along the bilayer normal.* Journal of chemical theory and computation, 2013. **9**(8): p. 3686-3703.

64. Zhu, F. and G. Hummer, *Theory and simulation of ion conduction in the pentameric GLIC channel.* Journal of chemical theory and computation, 2012. **8**(10): p. 3759-3768.

65. Song, H.D. and F. Zhu, *Finite temperature string method with umbrella sampling: Application on a side chain flipping in Mhp1 transporter.* The Journal of Physical Chemistry B, 2017. **121**(15): p. 3376-3386.

66. Hummer, G., *Position-dependent diffusion coefficients and free energies from Bayesian analysis of equilibrium and replica molecular dynamics simulations.* New Journal of Physics, 2005. **7**(1): p. 34.

67. Bolen, D.W. and G.D. Rose, *Structure and energetics of the hydrogen-bonded backbone in protein folding.* Annu. Rev. Biochem., 2008. **77**: p. 339-362.

68. Hałabis, A., et al., *Conformational dynamics of the Trp-cage miniprotein at its folding temperature.* The Journal of Physical Chemistry B, 2012. **116**(23): p. 6898-6907.

69. Ahmed, Z., et al., *UV− resonance Raman thermal unfolding study of Trp-cage shows that it is not a simple two-state miniprotein.* Journal of the American Chemical Society, 2005. **127**(31): p. 10943-10950.

70. Culik, R.M., et al., *Achieving Secondary Structural Resolution in Kinetic Measurements of Protein Folding: A Case Study of the Folding Mechanism of Trp-cage.* Angewandte Chemie International Edition, 2011. **50**(46): p. 10884-10887.

71. Marinelli, F., *Following easy slope paths on a free energy landscape: the case study of the Trp-cage folding mechanism.* Biophysical journal, 2013. **105**(5): p. 1236-1247.

72. Neidigh, J.W., R.M. Fesinmeyer, and N.H. Andersen, *Designing a 20-residue protein.* Nature structural biology, 2002. **9**(6): p. 425-430.

73. Paschek, D., S. Hempel, and A.E. Garcia, *Computing the stability diagram of the Trp-cage miniprotein.* Proceedings of the National Academy of Sciences, 2008. **105**: p. 17754-17759.

74. Day, R., D. Paschek, and A.E. Garcia, *Microsecond simulations of the folding/unfolding thermodynamics of the Trp-cage miniprotein.* Proteins: Structure, Function, and Bioinformatics, 2010. **78**(8): p. 1889-1899.

75. Nettels, D., et al., *Single-molecule spectroscopy of the temperature-induced collapse of unfolded proteins.* Proceedings of the National Academy of Sciences, 2009. **106**(49): p. 20740-20745.

76. Best, R.B., W. Zheng, and J. Mittal, *Balanced protein–water interactions improve properties of disordered proteins and non-specific protein association.* Journal of chemical theory and computation, 2014. **10**(11): p. 5113-5124.

77. Piana, S., J.L. Klepeis, and D.E. Shaw, *Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations.* Current opinion in structural biology, 2014. **24**: p. 98-105.

78. Henriques, J., C. Cragnell, and M. Skepö, *Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment.* Journal of chemical theory and computation, 2015. **11**(7): p. 3420-3431.

79. Piana, S., K. Lindorff-Larsen, and D.E. Shaw, *How robust are protein folding simulations with respect to force field parameterization?* Biophysical journal, 2011. **100**(9): p. L47-L49.

80. Debiec, K.T., A.M. Gronenborn, and L.T. Chong, *Evaluating the strength of salt bridges: a comparison of current biomolecular force fields.* The Journal of Physical Chemistry B, 2014. **118**(24): p. 6561-6569.

81. Pietrucci, F., L. Mollica, and M. Blackledge, *Mapping the native conformational ensemble of proteins from a combination of simulations and experiments: new insight into the src-SH3 domain.* The journal of physical chemistry letters, 2013. **4**(11): p. 1943-1948.

CHAPTER 3. TOWARD CONVERGENCE IN FREE ENERGY CALCULATIONS FOR PROTEIN CONFORMATIONAL CHANGES: A CASE STUDY ON THE THIN GATE OF MHP1 TRANSPORTER

3.1 Introduction

Proteins are among the most important players in living systems. Most proteins adopt specific conformations (structures) that are closely related to their biological functions. Moreover, many proteins may adopt multiple conformations that can be reversibly converted from one to another. Some proteins, such as membrane transporters, must properly change their conformations to perform the physiological functions. Biophysically, equilibrium probabilities, and transition rates are the fundamental thermodynamic and kinetic properties, respectively, for conformational changes. In the simplest two-state case, a protein has two alternative conformations A and B. At equilibrium, the probabilities for the two conformations are P_A and P_B, with $P_A + P_B = 1$. These equilibrium probabilities are determined by the free energy difference between the two states:

$$G_B - G_A = -k_B T \ln(P_B/P_A) \qquad \text{Eq. 3-1}$$

where k_B is the Boltzmann constant, and T is the temperature. The spontaneous transition rates between the two conformations, $k_{A\to B}$ and $k_{B\to A}$, are also related to the free energy difference or the equilibrium probabilities:

$$k_{B\to A}/k_{A\to B} = P_A/P_B \qquad \text{Eq. 3-2}$$

In addition to the thermodynamic and kinetic quantities above, it is highly desirable to gain a detailed mechanistic understanding of how the spontaneous transitions between the conformations occur. Although experiments could measure the equilibrium probabilities and the transition rates for protein conformational changes, the spontaneous transitions are normally difficult to observe in detail since they are rare and transient events.

Molecular dynamics (MD) simulations are potentially a powerful technique to study conformational transitions since they could reveal molecular processes in atomic resolution1. Ideally, sufficiently long MD simulations could sample all the relevant conformations of the

protein along with a large number of spontaneous transitions between different conformations. These extremely long MD trajectories can thus be used to generate the equilibrium ensemble of the protein conformations and direct obtain all the thermodynamic and kinetic quantities. Unfortunately, although such a straightforward approach is conceptually simple and robust, it typically requires prohibitively long simulation times. Even with the most powerful computational resource nowadays, the currently affordable simulation times are only sufficient for systematically characterizing small proteins with fast kinetics[1, 2]. For most proteins, in contrast, alternative techniques need to be designed to reproduce the relevant properties of the equilibrium ensemble as ideally in the long unbiased simulations.

Many computational methods introduce various forms of bias in the simulations to circumvent the insufficient simulation times with aiming to calculate the thermodynamic and kinetic properties from the biased simulations of affordable sampling times. These methods, often collectively called enhanced sampling techniques, include umbrella sampling[3, 4], transition path sampling [5], metadynamics [6], accelerated MD [7], adaptive biasing force [8], milestoning [9], dynamic importance sampling [10], weighted ensemble [11], steered MD [12], string method [13-17], among many others. Many enhanced sampling techniques employ a reaction coordinate (RC) that distinguishes the two protein conformations, such that driving along the RC could enforce continuous conversions between the conformational states.

Importantly, a free energy profile as a function of the RC can be calculated, which determines the probability distribution along with the RC in the equilibrium ensemble. The free energy profile quantifies the thermodynamics of the conformational transition and, in particular, gives the free energy difference between the two conformational states. All the enhanced sampling techniques above are based on rigorous theories in statistical mechanics and work perfectly well for model systems. However, when applied to real protein conformational changes, the success became much more limited, partly due to the reliability of the sampling in highly complex systems. In some cases, the free energies for the same conformational transition could differ by tens of kcal/mol in different publications. Indeed, it is well recognized that achieving true convergence in such free energy calculations remains a major challenge. In this study, we propose a set of strategies to alleviate the difficulty in the sampling and to achieve consistent and convergent free energies for protein conformational changes in practical applications.

Our approach differs in the following three aspects from the common practice in calculating conformational free energy. First, we recognize that two different protein conformations normally differ in multiple degrees of freedom, which might not necessarily undergo transitions simultaneously during spontaneous conformational changes. Correspondingly, unlike many studies in which a single collective RC describes the entire conformational change, in our approach, the complete conformational change consists of a series of transition steps. Each step connects two metastable states, which can be the end states or intermediate states. Accordingly, each individual transition step involves the conversion of certain distinct degrees of freedom and is described by a distinct RC. Second, to gauge the convergence of calculated free energies, we carry out two independent sampling simulations with different initial structures as the transition initiates from state A and the transition initiates from state B, respectively. A comparison of the two obtained free energy profiles will then indicate the extent of convergence in such calculations. We believe that this protocol for examining the convergence is much more stringent than other common methods and is more likely to uncover hysteresis problems in the sampling. Third, we introduce flat-bottom restraints in dimensions perpendicular to the RC to prevent the protein from being trapped in undesired conformations. These restraints effectively confine the sampling in desired conformational space and significantly reduce the complexity and difficulty of the sampling. Together, our approaches make it practically more feasible to obtain consistent and reliable free energies for protein conformational changes.

Protein MHP1 structure

We apply the proposed approaches above to elucidate a conformational change in the bacterial hydantoin transporter Mhp1, which is a symporter that co-transports a Na^+ ion and a substrate molecule[18-20] (i.e., a hydantoin or its analog). Structurally, Mhp1 is formed by twelve transmembrane helices (TMs). The first ten TMs are arranged in two repeating units (i.e., TMs1-5 and TMs6-10) that are related by pseudosymmetry. Like all membrane transporters, Mhp1 has an outward-facing (OF) and an inward-facing (IF) conformation, with the interior of the protein exposed to the extra- and intracellular sides of the membrane, respectively. In addition, when the substrate is bound from the extracellular side, Mhp1 is found to adopt an outward occluded (OC) conformation. The conformational change between the OF and IF states is through a so-called "thick gate" that involves global rotations of TM bundles. In contrast, the transition between the

OF and OC states is through a "thin gate" that mainly involves local movement of residues around TM10. In this study, we only focus on the thin gate, or the transition between the OF and the OC conformations, in the ligand-free condition.

Overall, the OF and OC structures share a similar scaffold, with a Cα RMSD of 1.25 Å for residues 11 to 470. However, as shown in Figure 3-1, the two structures have a significant difference in TM10 and its preceding loop, which together form the thin gate that opens or closes the outward-facing binding pocket for the substrate. The thin gate has a major difference in its position and secondary structure between the OF and OC conformations (Figure 3-1). In particular, the C-terminal half of the loop adopts a distorted helical conformation in the OF structure, in contrast to a partially extended form in the OC structure. Consequently, the Cα-Cα distance between residues Leu359 and Phe355 is 9.5 Å in the OC crystal structure but only 5.5 Å in the OF structure. Also, due to different backbone conformations of the loop, the sidechain of Val358 points to almost opposite directions in the OF and OC states.

Furthermore, the rotational position of TMH 10 relative to other helices is substantially different in the OF and OC structures, as can be clearly seen from the sidechain position of Leu363 on TMH 10 (Figure 3-2). In the OF conformation, the Leu363 sidechain appears to interact closely with both Phe116 and Trp117 on TMH 3. However, in the OC state, this side chain is rotated away from Phe116 and only contacts Trp117 while being closer to TMH 6. In addition, the Leu366 side chain is on either side of the Leu113 side chain in the OF and OC structures (Figure 3-2). Therefore, Leu366 would need to cross Leu113 from one side to the other during the conformational change.

Even without the bound ion and substrate, we found that both the OF and the OC conformations remain stable in our unbiased equilibrium MD simulations, thus qualifying them as well-defined metastable states. As described above, the two conformations differ in the number of degrees of freedom, and the interconversion between them is far from trivial. Here we take this conformational change as a case study to demonstrate our computational approach, attempting to obtain the relative equilibrium probabilities (free energies) for the OF and OC states and a plausible pathway for the conformational transitions.

3.2 Method

3.2.1 Simulation Systems and Protocols

We built two systems from the crystal structures for the OF (PDB: 2JLN[18]) and the OC (PDB: 4D1B[18, 19]) states, respectively, each consisting of residues 11 to 470 of Mhp1. All ligands, including the bound ions and substrate, were removed from the crystal structures. All histidine residues were neutral with protonation on the ε nitrogen. In each system, the protein was embedded in a bilayer of 200 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoethanolamine (POPE) lipid molecules. We used VMD[21] to manually position the protein to match its hydrophobic surface to the hydrophobic interior of the lipid bilayer. The system was then solvated by 15953 water molecules. Four chloride ions were added to neutralize the simulation system. Each system consists of 79986 atoms in total. We note that the OF and OC systems have identical composition and atom counts, despite different conformations of the protein.

Our MD simulations were run by NAMD2[22] (v 2.13) and NAMD3[23] (v alpha7), using the CHARMM (vc36) force field[24] for the protein[24-26] and lipids[27] and the TIP3P model[28] for water molecules. The simulations were run with a 2.0 fs time step and under periodic boundary conditions, with the unit cell of dimensions ~ 96 Å $\times \sim 96$ Å $\times \sim 120$ Å. All bond lengths involving hydrogen atoms were constrained to their equilibrium values using the SHAKE[29] and SETTLE[30] algorithms. Nonbond interactions were calculated with a cutoff distance of 12 Å. A smooth switching function takes effect at 10 Å for the van der Waals interactions. Full electrostatics was estimated every 4 fs applying the particle mesh Ewald method[31]. All the simulations in this study were performed in an NPT ensemble. The Langevin dynamics method with a damping coefficient of 0.1 ps^{-1} was applied to maintain the temperature at 300 K. A constant pressure of 1 atm was maintained using the Nose-Hoover Langevin piston method[32].

For each system, we first fixed the protein and equilibrated the lipid, water, and ions for 2 ns after a conjugate-gradient minimization of 2000 steps. We then relaxed the protein and performed an equilibrium simulation for 100 ns without any restraint or bias.

3.2.2 Overall Scheme

As mentioned in the introduction, we model a complex protein conformational change as a sequence of multiple transition steps. In our case here, between the OF and OC conformations

of Mhp1, we introduce five intermediate metastable states, labeled M_1-M_5Which defines six transition steps labeled 1-6, as shown in Figure 3-3. Specifically, the OF conformation would undergo steps 1-6 sequentially to achieve a complete transition to the OC conformation. Reversely, the OC state would undergo steps 6-1 through intermediate states M_5-M_1 to reach OF. Each transition step involves the changes in a subset of degrees of freedom and is described by a distinct RC, as will be described in detail later.

We carried out two independent groups of simulations to calculate the free energies for the conformational change. One simulation group, termed *InitOF* here, started with the OF crystal structure and never directly used the OC crystal structure. Similarly, the other simulation group, *InitOC*, started with the OC crystal structure without any direct involvement of the OF crystal structure. Theoretically, in the limit of infinite sampling times, the long-term behaviors of the simulations should not depend on their initial states, and in such limits, the two simulation groups should result in identical free energies. In practice, a comparison between these two independent groups of simulations offers valuable indication for the convergence of the sampling.

3.2.3 Collective Variables and Reaction Coordinate

As described above, each transition step involves a two-state transition between adjacent metastable conformations, which we denote as A and B here for the sake of convenience. To distinguish the two states, we first introduce some collective variables (CVs). Each CV is a function of the atomic coordinates (denoted as X) of the protein, i.e., $CV(X)$. The CVs involved in this study are torsion-based, angle-based, or distance-based. A torsion-based CV is defined as the dihedral angle formed by four atoms. Examples of such CV in this study include the backbone φ and ψ torsions and the sidechain χ_1 (N-C_α-C_β-C_γ) and χ_2 (C_α-C_β-C_γ-$C_{\delta 2}$) torsions. An angle-based CV is an angle formed by three atoms. A distance-based CV is a distance between individual atoms or points or a linear combination of such distances. We note that all the CVs here are invariant to the rigid-body rotation or translation of the protein. For each CV, we take its two typical values cv_A and cv_B when the protein is in conformations A and B, respectively, and define a reduced CV (denoted as CV^*):

$$CV^*(X) \equiv \frac{CV(X) - cv_A}{cv_B - cv_A} \qquad \text{Eq. 3-3}$$

In the reduced form, the CV^* is dimensionless and has values around 0 and 1 for protein conformations A and B, respectively. If we adopt N such CVs for the particular transition step, the RC for this two-state transition is simply defined as the average of the reduced CVs:

$$RC(X) = \sum_{k=1}^{N} CV_k^*(X) \Big/ N \qquad \text{Eq. 3-4}$$

Similar to each reduced CV, the RC has values around 0 and 1 for the two metastable conformations as well.

3.2.4 Restraining Potential on RC in the Umbrella Sampling

To obtain a free energy profile as a function of the RC, we employ umbrella sampling[3, 4] (US). We first specify a range $[\alpha_A, \alpha_B]$ to place the umbrella windows, with α_A and α_B near 0 and 1, respectively. We evenly divide this range into $M-1$ sections, thus resulting in a total of M evenly-spaced points α_i^{ref} (i=1,2,…,M), with $\alpha_1^{ref} = \alpha_A$ and $\alpha_M^{ref} = \alpha_B$. These reference points are used to define the harmonic restraints on the RC for individual umbrella windows:

$$U_i(X) = \frac{K}{2} \left(RC(X) - \alpha_i^{ref}\right)^2 \qquad \text{Eq. 3-5}$$

for i=1,2,…,M, where K is the spring constant.

Furthermore, we place two additional umbrella windows, labeled 0 and $M+1$, at the two ends, respectively, to better sample the two metastable states A and B. In these two windows, the RC is subject to a flat-bottom potential:

$$U(X) = \begin{cases} \frac{K}{2} \left(|RC(X) - \alpha_c| - \Delta\right)^2 & |RC(X) - \alpha_c| > \Delta \\ 0 & Otherwise \end{cases} \qquad \text{Eq. 3-6}$$

The potential is 0 when the RC is in the range of $[\alpha_c - \Delta, \alpha_c + \Delta]$. Here we set $\alpha_c = \alpha_A - \Delta$ for window 0, and $\alpha_c = \alpha_B + \Delta$ for window $M + 1$. These two windows thus allow a largely unbiased sampling of the two metastable states. Therefore, the single transition step considered here is covered by a total of $M + 2$ umbrella windows, each sampled by an individual simulation (replica). The windows at the two ends with the flat-bottom potential could facilitate Hamiltonian replica exchange[33] across multiple transition steps, as will be described later.

3.2.5 Boundary Restraints in the Umbrella Sampling

In addition to the restraint on the RC, we also apply a number of restraints on some other degrees of freedom (i.e., CVs), which we call "boundary restraints" here, to facilitate proper sampling. A boundary restraint has the form of a flat-bottom harmonic potential. Therefore, it has no effect on the concerned CV when it is within the specified range (i.e., the flat portion of the potential), and only acts to pull the CV back when it exceeds the range. The boundary restraints thus mainly serve to prevent undesired transitions in the concerned CVs but will not affect the normal dynamics of the system.

The first type of boundary restraint is on the CVs that define the RC. As described earlier, in general, the RC is the average over multiple reduced CVs. By definition, all these reduced CVs and the RC have values ~0 for one metastable state and values ~1 for the other. We thus expect the CVs to also change collectively with the RC and with each other for the transition between the two metastable states. To enforce this, we apply the following boundary potential on any pair of reduced CVs (denoted as k and j here) to prevent too much separation of their values:

$$U_{b1}(X) = \begin{cases} \frac{K_{b1}}{2} \left[\left| CV_k^*(X) - CV_j^*(X) \right| - Y(X) \right]^2 & \left| CV_k^*(X) - CV_j^*(X) \right| > Y(X) \\ 0 & Otherwise \end{cases} \quad \text{Eq. 3-7}$$

Here the spring constant of the boundary potential K_{b1} is taken to be 1000 $kcal/mol$. The term Y specifies how far the two reduced CVs need to differ for the boundary potential to have an effect. In this study, we choose a Y that depends on the RC in the following form:

$$Y(X) = \eta_1 e^{-\lambda_1 RC(X)} + \eta_2 e^{\lambda_2 RC(X)} \quad \text{Eq. 3-8}$$

with $\eta_1 = 2.367$, $\eta_2 = 1.321 \times 10^{-4}$, $\lambda_1 = \lambda_2 = 9.794$. Figure 3-4 shows Y as a function of the RC. When the RC is around 0 or 1, Y is very large such that the boundary restraint $U_{b1}(X)$ is always zero. When the RC is around 0.5, Y becomes much smaller, and the $U_{b1}(X)$ thus prevents large deviation between the two reduced CVs. Therefore, the boundary restraint here will not affect the equilibrium fluctuations of the CVs in the two metastable states but will enforce the CVs to change together during transitions. In the US [3, 4], for a given transition step, if the RC is defined by multiple CVs (i.e., $N > 1$ in Eq. 2), we apply the boundary potential $U_{b1}(X)$ on every pair of these CVs.

The second type of boundary restraints is applied to the CVs that are not directly involved in defining the RC. However, the deviation of these CVs from the normal ranges could sometimes pose problems in the sampling. These boundary restraints have the following form:

$$U_{\mathrm{b2}}(X) = \begin{cases} \frac{K_{b2}}{2} \; [|CV(X) - cv_0| - \Delta_2 \,]^2 & |CV(X) - cv_0| > \; \Delta_2 \\ 0 & Otherwise \end{cases} \qquad \text{Eq. 3-9}$$

Here cv_0 and Δ_2 specify the midpoint and the width of the range where the CV is not affected by the boundary potential.

The third type of boundary restraint has the following form:

$$U_{b3}(X) = \begin{cases} \frac{K_{b3}}{2} \; [|CV(X) - (\omega \, RC(X) + \beta \,)| - \Delta_3]^2 & |CV(X) - (\omega \, RC(X) + \beta)| > \Delta_3 \\ 0 & Otherwise \end{cases} \qquad \text{Eq. 3-10}$$

It is similar to $U_{\mathrm{b2}}(X)$, except that the constant cv_0 (see Eq. 7) is replaced by a linear function (defined by the slope ω and the intersection β) of the RC. This type of boundary restraint is applied when the CV has substantially different values at the two metastable states and is thus expected to change with the RC during a transition.

3.2.6 Details of Individual Transition Steps

The conformational change and the US[3, 4] for each transition step are described below, with a summary given in Table 3-1..

1) Transition step 1 between the OF state and the intermediate state M_1 involves the sidechain rotation of Trp117 (Figure 3-5A). Therefore, the RC for this step is defined by the two sidechain dihedral angles χ_1 and χ_2 of Trp117.
2) Transition step 2 between the intermediate states M_1 and M_2 mainly involves changes around Leu359, as shown in Figure 3-5B. The Leu359 sidechain points to the exterior and interior of the protein in states M_1 and M2, respectively. A number of degrees of freedom are changed during this transition. We take one torsion-based CV and one distance-based CV to define the RC (Table 3-1) for this step and apply boundary restraints U_{b3} Table 3-5 to ensure proper changes of other CVs.
3) Transition step 3 between the intermediate states M_2 and M_3 mainly involves the conformational change of TMH10. In particular, the Leu366 sidechain on TMH10 crosses the sidechain of Leu113 on TMH3 from one side to the other in this transition. To describe such sidechain crossing, we first introduce an anchor point (denoted as AP here) that is approximately on the line connecting the positions of the L366:C_γ atom in the M_2 and M_3 states. Specifically, the AP is defined as the center of mass for a group of atoms shown inFigure 3-6. All these atoms are on the backbone of transmembrane helices with relatively small fluctuations in the simulations, thus ensuring that the AP can be taken as a rather constant point in the interior of the protein. We further define a CV as the difference between the distances from the AP to the C_γ atoms of Leu366 and Leu113, respectively:

$$CV = \left|\vec{r}\left(\mathrm{L366{:}\,C_\gamma}\right) - \vec{r}(\mathrm{AP})\right| - \left|\vec{r}\left(\mathrm{L113{:}\,C_\gamma}\right) - \vec{r}(\mathrm{AP})\right| \qquad \text{Eq. 3-11}$$

where $\vec{r}()$ represents the coordinate of the given atom or point. This CV could very well distinguish the relative sidechain positions of Leu366 and Leu113 in the metastable states M_2 and M_3 and is therefore used to define the RC for this transition step.
4) Transition step 4 between the intermediate states M_3 and M_4 involves the sidechain rotation of Trp117 (Figure 3-5D) again. This step is opposite to the transition step 1 described earlier and used the same RC as in that step. In the OF→OC transition, this step returns the side chain of Trp117 to its native rotamer in the OF/OC conformation.

Transition step 5 between the intermediate states M_4 and M_5 involves the conformational change of the loop region (Figure 3-5E). In particular, the backbone of the loop adopts distinct conformations in M_4 and M_5. Therefore, we use three backbone ψ torsion angles (Table 3-1)) to define the RC.

Transition step 6 between the intermediate state M_5 and the OC state involves the side chain rotation of Phe355 (Figure 3-5F), with the RC defined by its χ_1 torsion angle.

Whereas the many CVs (Table 1) above are implicated in the complete conformational change, each transition step only involves the changes of a small number of CVs while the others remain roughly constant. In the US [3. 4] simulation of a given transition step, therefore, we apply boundary restraints U_{b2} to the CVs not involved in this step to prevent them from undergoing large spontaneous transitions. These boundary restraints Table 3-2 ensure that each CV has a single status in any metastable state and only makes the transition in the specified step.

In addition, we apply boundary restraints $U_{\mathrm{b2}}(X)$ on several backbones φ and ψ torsion angles (Table 3-3) in all the transition steps. These torsions have similar values in the OF and OC conformations, and their values typically do not deviate much in our unbiased equilibrium simulations. We thus assume that they would stay in the same range during conformational changes between OF and OC as well and apply the boundary restraints to eliminate the possibility of any large transitions in these torsions.

Furthermore, in each individual transition step, some specific degrees of freedom are subject to boundary restraints U_{b2} and U_{b3} to prevent occasional abnormal behaviors. These restraints are given in Tab;e 3-4 and Table 3-5.

3.2.7 Implementation of Umbrella Sampling

To initialize US[3, 4] for an individual transition step between two metastable states, we first carry out a "pulling" simulation in which the umbrella potentials (Eq. 3) are sequentially applied to drive the RC from one end to the other. The same boundary restraints in the corresponding US[3, 4] are also applied in the pulling simulations here. Snapshots from these simulations are then used as the initial coordinates for the subsequent US[3, 4] simulations. As described earlier, we perform two groups of sampling simulations with identical potentials but different initial coordinates. In the *InitOF* group, all the pulling simulations are in the OF→OC direction (Figure 3-3), and the initial coordinates of all the US[3, 4] in this group were thus

originally derived from the OF crystal structure. In contrast, the *InitOC* group has the pulling simulations in the OC→OF direction, and all the initial coordinates for the US simulations originally from the OC crystal structure.

To facilitate equilibration, we employ Hamiltonian replica exchange[33] in all the US[3, 4] simulations. Although each transition step can be separately sampled, in this study, we chose to sample all the transition steps simultaneously in the final production run, which allows the last window of a transition step and the first window of the next transition step to exchange their replicas. The criterion for exchanging any neighboring replicas i and j is based on the change in the total restraint energy due to the exchange: $\Delta U = \left[U_i(X_j) + U_j(X_i)\right] - \left[U_i(X_i) + U_j(X_j)\right]$, with the swap probability given by $\min[\exp(-\Delta U/k_B T), 1]$. This is the standard formulism for Hamiltonian replica exchange[33], regardless of whether the two windows are in the same or different transition step. In the latter case, however, the energy U must include the boundary restraints in addition to the umbrella potential on the RC. In this scheme, a single replica could potentially sample windows across multiple transition steps, thus further facilitating the equilibration of the systems. The exchanges are attempted every 200 time steps (i.e., 0.4 ps). The six transition steps are sampled by a total of 144 umbrella windows Table 3-1. For both the *InitOF* and *InitOC* sets of simulations, the final production run has 150 ns per window, with the last 75 ns used for calculating the free energy profiles by the weighted histogram analysis method[34, 35] (WHAM).

3.2.8 Calculation of Individual Transition Rates

Based on the free energy profiles obtained from the US simulations, we can further calculate the transition rates using an approach we developed previously. In this study, we carried out such kinetics calculations for two of the transition steps, namely, steps 1 and 5, as described below.

We consider each individual transition step between states A and B as an isolated two-state system, assuming that the system is not allowed to visit other metastable states. Based on the free energy profile $G(\alpha)$ as a function of the reaction coordinate (RC) α, the equilibrium probabilities for states A and B are given by

$$P_A = \frac{\int_{-\infty}^{\alpha^*} \exp[-G(\alpha)/k_B T]\, d\alpha}{\int_{-\infty}^{\infty} \exp[-G(\alpha)/k_B T]\, d\alpha}$$

Eq. 3-12

$$P_B = \frac{\int_{\alpha^*}^{\infty} \exp[-G(\alpha)/k_B T]\, d\alpha}{\int_{-\infty}^{\infty} \exp[-G(\alpha)/k_B T]\, d\alpha}$$

where α^* is the location of free energy barrier in $G(\alpha)$.

For transition step 1, where A and B here respectively represent the metastable states OF and M1, we first define a small interval $[\alpha_1, \alpha_2]$, with $[\alpha_1 = 0.45, \alpha_2 = 0.55]$, at the free energy barrier. We then selected ten frames from the US trajectories where the RC is near the barrier. Next, starting from each frame, we performed a simulation of 13 ns, in which the RC is subject to a strong flat-bottom harmonic restraint with a spring constant of 3,750 kcal/mol. The restraint is zero when the RC is inside the interval $[\alpha_1, \alpha_2]$ but acts to pull the RC back when it crosses the boundaries. In addition, all the boundary restraints in the US simulations were also applied here. The velocities and coordinates of the atoms were saved with an interval of 1ns in these simulations, representing an unbiased equilibrium sampling of the microstates when the RC is within $[\alpha_1, \alpha_2]$. We then took the last 12 frames from each simulation, thus obtaining a total of 120 microstates with the RC in the interval $[\alpha_1, \alpha_2]$. Next, for each microstate, we created a conjugate microstate by replicating the coordinates and reverting the direction (i.e., multiplying each component by -1) of the velocities for all the atoms in the system. Now we thus have 120 pairs, each consisting of two microstates that are conjugate of each other. Starting from each of the 240 microstates, we carried out an unbiased simulation under the NVE condition, without any restraint. Importantly, the two simulations in each pair started with identical coordinates but opposite velocities for all the atoms. The simulations were run long enough such that the system commits to either state A or state B.

From each pair of the forward/backward simulations, we calculate a λ value from their trajectories. Specifically, if the forward and backward simulations commit to the same state, λ is set to 0. Otherwise, when the two simulations respectively commit to the two states A and B, they form a transition path (or reactive trajectory) that represents a spontaneous transition between A and B. In this case, λ is set to $1/\tau$, where τ is the total duration that the transition path spends in

the interval $[\alpha_1, \alpha_2]$. If the transition path visits the interval multiple times, τ should be the sum of all the individual durations.

We denote k_0 as the rate of spontaneous transitions from A to B in the equilibrium ensemble of the two-state system. Spontaneous transitions from B to A have the same rate k_0. We previously proved that k_0 can be determined as

$$k_0 = \frac{1}{2} P(\alpha_1 \le \alpha \le \alpha_2) \cdot \langle \lambda(\alpha_1, \alpha_2) \rangle_{\alpha_1 \le \alpha \le \alpha_2} \qquad \text{Eq. 3-13}$$

Here $\langle \lambda(\alpha_1, \alpha_2) \rangle_{\alpha_1 \le \alpha \le \alpha_2}$ can be calculated by the average of λ values over the 120 simulation pairs described earlier. Furthermore, the equilibrium probability for the RC interval $[\alpha_1, \alpha_2]$ can be determined from the free energy profile:

$$P(\alpha_1 \le \alpha \le \alpha_2) = \frac{\int_{\alpha_1}^{\alpha_2} \exp[-G(\alpha)/k_B T]\, d\alpha}{\int_{-\infty}^{\infty} \exp[-G(\alpha)/k_B T]\, d\alpha} \qquad \text{Eq. 3-14}$$

With the obtained k_0, the forward and backward transition rates can be further determined:

$$k_{A\to B} = \frac{k_0}{P_A} = \frac{1}{2} \frac{\int_{\alpha_1}^{\alpha_2} \exp[-G(\alpha)/k_B T]\, d\alpha}{\int_{-\infty}^{\alpha^*} \exp[-G(\alpha)/k_B T]\, d\alpha} \langle \lambda(\alpha_1, \alpha_2) \rangle_{\alpha_1 \le \alpha \le \alpha_2}$$
$$k_{B\to A} = \frac{k_0}{P_B} = \frac{1}{2} \frac{\int_{\alpha_1}^{\alpha_2} \exp[-G(\alpha)/k_B T]\, d\alpha}{\int_{\alpha^*}^{\infty} \exp[-G(\alpha)/k_B T]\, d\alpha} \langle \lambda(\alpha_1, \alpha_2) \rangle_{\alpha_1 \le \alpha \le \alpha_2} \qquad \text{Eq. 3-15}$$

For the transition step 5 between M_4 and M_5, we performed kinetics calculations using the same protocols above, with the following exceptions. First, the chosen interval is $[\alpha_1 = 0.54, \alpha_2 = 0.56]$. Second, in the restrained simulation, the spring constant is 15,000 kcal/mol for the flat-bottom potential on the RC.

3.2.9 Calculation of Overall Transition Rates

In this study, we model the conformational change as a series of transition steps. From the kinetic rates of each transition step, we can further obtain the overall rate of the complete conformational change. Specifically, here we use S_0 and S_M to denote the two end states, and S_1, S_2, …, S_{M-1} to represent the intermediate states in a sequential chain. For the transition step between states i and $i+1$, the forward and backward transition rates are denoted as $k_{i\to i+1}$ and $k_{i+1\to i}$, respectively, and can be obtained using the approach described in the previous section. We now aim to calculate the overall rate $k_{0\to M}$. This rate can be defined in a stationary condition where S_0 is the source with a constant population of N and S_M is the drain with a population of 0, such that the constant flux from S_0 to S_M through all the intermediate states is $N \cdot k_{0\to M}$. From the stationary (time-independent) solution of the master equations for this single-chain system, we have the expression for $k_{0\to M}$:

$$k_{0\to M} = K / \sum_{i=0}^{M-1} Q_i \qquad \text{Eq. 3-16}$$

Where

$$K \equiv \prod_{i=0}^{M-1} k_{i\to i+1}$$
$$Q_i \equiv \left(\prod_{j=0}^{i-1} k_{j+1\to j}\right)\left(\prod_{j=i+1}^{M-1} k_{j\to j+1}\right) \qquad \text{Eq. 3-17}$$

These equations can be simplified as follows. At equilibrium, the probability for each state S_i is denoted as P_i. Since $P_{i+1}/P_i = k_{i\to i+1}/k_{i+1\to i}$, we have

$$\frac{P_i}{P_0} = \prod_{j=0}^{i-1} \frac{k_{j\to j+1}}{k_{j+1\to j}} \qquad \text{Eq. 3-18}$$

Therefore,

$$\frac{Q_i}{K} = \frac{\prod_{j=0}^{i-1} k_{j+1\to j}}{\prod_{j=0}^{i} k_{j\to j+1}} = \frac{1}{k_{i\to i+1}} \frac{P_0}{P_i} \qquad \text{Eq. 3-19}$$

We thus have

$$\frac{1}{P_0 k_{0\to M}} = \sum_{i=0}^{M-1} \frac{1}{P_i k_{i\to i+1}} \qquad \text{Eq. 3-20}$$

If we take $1/P_0 k_{0\to M}$ as the "resistance" for transitions between S_0 to S_M, the equation above indicates that this total resistance is simply the sum of the resistances from individual transition steps, analogous to resistors in a series electrical circuit. For a transition step between S_i and S_{i+1}, specifically, the inverse of its resistance is equal to $P_i k_{i\to i+1}$, the number of spontaneous transitions in unit time from S_i to S_{i+1} in the equilibrium ensemble for this multi-state system. The $P_i k_{i\to i+1}$ term is also equal to the rate ($P_{i+1} k_{i+1\to i}$) of spontaneous transitions from S_{i+1} to S_i in the equilibrium ensemble. Therefore, transition steps with smaller $P_i k_{i\to i+1}$ values will have larger contributions to the overall rate $k_{0\to M}$. In particular, a "rate-limiting" step would have a $P_i k_{i\to i+1}$ value that is much smaller than the values from all the other steps. Assuming that the step between states S_l and S_{l+1} is such a rate-limiting step, we may ignore the resistances of the other steps and take only the $P_l k_{l\to l+1}$ to approximately estimate the overall rates as:

$$k_{0\to M} \approx \frac{P_l}{P_0} k_{l\to l+1}$$
$$k_{M\to 0} \approx \frac{P_{l+1}}{P_M} k_{l+1\to l} \qquad \text{Eq. 3-21}$$

3.3 Result

As a case study to demonstrate our computational approaches, we aim to elucidate the conformational changes involving the thin gate of the Mhp1 membrane transporter between the OF and OC states.

We first examine the behaviors of the OF and OC conformations in 100-ns unbiased simulations starting from the corresponding crystal structures. Figure 3-7 shows that both conformations are quite stable during the simulation time, as the Cα atoms in the thin gate remain close to the starting crystal structure and far from the other structure. As described in the Introduction, there are major differences in the local structures, such as backbone torsions and sidechain positions, between the thin gates of OF and OC. Those local structures also remain stable in our unbiased simulations without undergoing any considerable spontaneous conformational transition. Therefore, we conclude that both OF and OC are genuine metastable states, and spontaneous transitions between the two would be rare in typical MD time scales.

3.3.1 Monitoring and Alleviating Convergence Problems

We attempted to reveal the thermodynamics of the conformational change by computing a free energy profile from US simulations between OF and OC. Our major concern in such calculations is the convergence since hysteresis is a common problem that plagues many enhanced samplings of protein conformations. To closely monitor hysteresis, we adopted a strategy of performing each US twice using two different sets of initial coordinates, respectively. Specifically, as described in Methods, we have two groups of simulations, *InitOF* and *InitOC*, each originally starting from the OF and OC crystal structures, respectively. A comparison of the free energy profiles calculated from the *InitOF* and *InitOC* groups thus quantifies the hysteresis in the sampling. Indeed, in some of our preliminary calculations, the free energy profiles from *InitOF* and *InitOC* differed by tens of kcal/mol, thus indicating significant convergence problems. In such cases, we would compare the trajectories in detail to identify those degrees of freedom that exhibit major differences in the *InitOF* and *InitOC* simulations and are thus potentially responsible for the observed hysteresis. Early identification of hysteresis problems allowed us to address them promptly. Through repeated trials and errors, we gained an increasingly better understanding of the important degrees of freedom involved in the conformational change here and reduced the hysteresis in the sampling to an acceptable level. Some important considerations and strategies in this process are described below.

As mentioned earlier, the OF and OC conformations differ in many degrees of freedom. We found that if all these degrees of freedom change simultaneously in a single transition, the resulting free energy barrier would be very high. Therefore, we instead adopted a multi-step

scheme in which the complete conformational change consists of a series of transition steps, each only involving a small number of degrees of freedom. Specifically, our scheme introduces five metastable intermediate states M_1-M_5 between the OF and OC states, with six transition steps connecting all these states into a chain. Each transition step between two adjacent metastable states can be considered a separate two-state system with its own thermodynamics and kinetics. In particular, the free energy profile for each transition step is described by a distinct RC that only incorporates the degrees of freedom that undergo major changes in that step.

For Mhp1 here, although the sidechain of Trp117 has a similar orientation (described by the $\chi1$ and $\chi2$ torsions) in both OF and OC, we found that the sidechain in such position would pose a steric hindrance for the movement of other residues in the region. In our scheme, therefore, when going from OF to OC, the Trp117 side chain undergoes a rotation to an alternative position in transition step 1 to clear the space for other residues to complete their changes in transition steps 2 and 3. The Trp117 sidechain then rotates back to its original position in transition step 4. Consequently, in the intermediate states M_2 and M3, the Trp117 sidechain is in a different position compared to that in the OF and OC.

When running US simulations, the common problem is that some degrees of freedom would make random spontaneous transitions. Such degrees of freedom are not the RC and therefore not subject to the umbrella potentials, but they may nonetheless affect the distribution of the RC and thus the free energy profile. Ideally, the simulations should be long enough such that these orthogonal degrees of freedom are fully equilibrated, with all the values properly sampled. In practice, however, the affordable simulation times are often not sufficient when the orthogonal degrees of freedom make rare transitions with slow kinetics. Our strategy in such cases is to apply boundary restraints on the problematic degrees of freedom. Such restraints effectively impose a boundary for the values that the concerned degree of freedom may take, thus preventing it from undergoing undesired transitions while not affecting its normal dynamics within the allowed range. As detailed in Methods, we applied several types of boundary restraints for different situations. These boundary restraints significantly simplify the sampling by confining it within the desired conformational space and avoiding difficult regions.

By properly choosing the transition steps along with the RC and the boundary restraints for each step, we managed to achieve acceptable convergence, as gauged by comparing the free energy

profiles from the *InitOF* and *InitOC* simulation groups. In the following, we describe the conformational sampling of the Mhp1 thin gate in more detail.

3.3.2 Conformational Thermodynamics of Mhp1 Thin Gate

We carried out US simulations with Hamiltonian replica exchange to calculate the free energy profiles for each transition step. In the final production run, all the transition steps were sampled altogether using a total of 144 umbrella windows. Furthermore, the Hamiltonian replica exchange was not only between windows in the same transition step but also between the last window of a transition step and the first window of the next step. Therefore, a replica may sample multiple transition steps during the US simulations, thereby further facilitating the equilibration of the umbrella windows. The exchange rates between neighboring windows are in the range of 20-40% in our simulations. Furthermore, as shown in Figure 3-8, most of the replicas indeed visited a substantial range of windows, with some covering multiple transition steps.

To analyze the US simulations, we first treat each transition step as a separate two-state system with metastable conformations A and B and calculate a separate free energy profile using the standard WHAM. From each profile, the free energies for metastable states A and B are calculated as

$$G_A = -k_B T \ln \int_{-\infty}^{\alpha^*} \exp\left[-\frac{G(\alpha)}{k_B T}\right] d\alpha$$

$$G_B = -k_B T \ln \int_{\alpha^*}^{\infty} \exp\left[-\frac{G(\alpha)}{k_B T}\right] d\alpha$$

Eq. 3-22

where k_B is the Boltzmann constant. The free energy difference $\Delta G \equiv G_B - G_A$ thus predicts the ratio of the probabilities for the two states in the equilibrium ensemble. The ΔG values for each transition step are provided in Table 3-9. As mentioned earlier, a comparison of the free energies calculated from the *InitOF* and *InitOC* simulations offers an estimate for the hysteresis in the sampling and the convergence of the results. For the individual transition steps here, the maximum deviation of ΔG between *InitOF* and *InitOC* is ~2.1 kcal/mol (Table 3-9), which we consider an acceptable convergence.

By cumulating the stepwise ΔG values above, we obtain the free energies of all the metastable states relative to the first state, OF, as shown in Figure 3-9B. These free energies predict the probabilities of each state in the equilibrium ensemble. Based on this, we plot all the free energy profiles together (shown in Figure 3-9A) by vertically shifting them such that the free energy of each metastable state matches its corresponding value in Figure 3-9B. Thus, these combined profiles reveal the complete thermodynamics of the multistate system here, providing both the free energy levels for the conformations and the barriers between them.

Because boundary restraints were applied in our US simulations as described earlier, their effect should be properly accounted for when constructing the free energy diagram in Figure 3-9, especially since a shared intermediate state may be subject to different boundary restraints when being sampled in different (neighboring) transition steps. To evaluate such effects, we performed equilibrium simulations of 100 ns for all the intermediate states (M_1-M_5) in addition to those for the OF and OC described earlier. From these equilibrium trajectories, we apply the free energy perturbation formulism $\Delta G_b = -k_B T \ln\langle e^{-U_b/k_B T}\rangle$ to calculate the increase of free energy ΔG_b when the boundary restraints U_b are present. Because our boundary restraints were designed to have minimal effects on the metastable states, it turns out that the equilibrium simulations here would only rarely hit any boundary. Consequently, the calculated ΔG_b is $\sim$0.1 kcal/mol or lower for all the metastable states here, which are negligibly small compared to other uncertainties. Therefore, we ignored the correction for the boundary restraints when presenting the free energies in Figure 3-9.

The OF state has two H-bonds formed by local backbone atoms in residues 355-360 (Figure 3-10.1). After the Trp117 sidechain is rotated in step 1, it makes an additional H-bond with the Gln42 side chain in state M_1 (Figure 3-10.2). Next, the rearrangement of Leu359 at step2 makes three new local H-bonds in state M_2 (Figure 3-10.3) while losing both the H-bonds in OF. Subsequently, the rearrangement of TM10 in step 3 makes two additional sidechain H-bonds with residues on TM1 and TM6 in state M_3 (Figure 3-10.4). Then the rotation of the Trp117 sidechain back to its native rotamer in step 4 breaks its H-bond with Gln42 in state M_4 (Figure 3-10.5). In step 5 to state M_5, the loop region changes to an extended conformation, thus breaking most of the local backbone H-bonds. As the last step (i.e., step 6) does not change the H-bonds, and states M_5 and OC (Figure 3-10.6 and Figure 3-10.7) thus share similar H-bonds.

The intermediate state M_3 has the lowest free energy (Figure 3-9B) among all the metastable states (including OF and OC) and is thus predicted to be the most populated conformation at equilibrium. In M_3, the C-terminal half of the loop adopts a distorted helical conformation similar to the OF, whereas the contacts between TM10 and TM6 are similar to the OC. Interestingly, M_3 also has the greatest number of H-bonds (Figure 3-10 and Table 3-6) in comparison to the other states, which might contribute to the favorable free energy for M_3.

3.3.3 Kinetics of Mhp1 Thin Gate

In addition to thermodynamics, kinetic quantities are the other major characteristics of conformational changes. Based on the obtained free energies, we further attempt to calculate the kinetic rates for the transitions of the Mhp1 thin gate. In addition to the end states OF and OC, the state M_3 here is also of major significance because it has the lowest free energy among all the metastable states. Therefore, we calculate the transition rates between these states, as described below.

In our scheme, transitions from the OF to the M_3 states need to take transition steps 1-3, hopping over states M_1 and M2. The kinetic rates between OF and M_3 can thus be determined by the individual transition rates of steps 1-3. In this way, the three transition steps can be lumped together to provide the effective rates between OF and M3, hiding the details such as the intermediates M_1 and M2. Specifically, as elaborated in Methods, each transition step contributes a resistance to the total resistance between OF and M3, and the rate-limiting step would have a predominant resistance compared to other steps. In particular, the resistance of an individual transition step is inversely proportional to the rate of spontaneous transitions for this step in the equilibrium ensemble of the multistate system. Furthermore, a major determinant of such spontaneous transition rate is the equilibrium probability at the barrier top. In the multistate free energy profiles (Figure 3-9A), transition step 1 has the highest free energy barrier and thus the lowest equilibrium probability at the barrier, which suggests it would be the rate-limiting step for transitions between OF and M3. Under this assumption, by ignoring the resistance of the other steps, we only calculate the transition rates for step 1 and use them to estimate the kinetic rates for the OF-M_3 transitions. Similarly, for the M3-OC transitions that consist of steps 4-6, we identify step 5 as the rate-limiting step and only calculate the transition rates for this step.

To calculate the rates in a single transition step, we prepare a number of systems at the free energy barrier top and release each system in two unbiased simulations with opposite initial velocities, as described in Methods. From the time evolution of these unbiased simulations along with the equilibrium probability of the barrier top in the two-state system, the forward and backward transition rates can be obtained. Specifically, if the two simulations starting from the same initial coordinates commit to different states, they form a transition path (reactive trajectory) between the two metastable states, and the durations these transition paths stay at the barrier top are used to calculate the kinetic factor. For transition step 1, more than 15% of the 120 simulation pairs evolve into transition paths, thus allowing an estimate of the transition kinetics. For transition step 5, without any boundary restraints in the 240 unbiased simulations at the barrier top, some backbone torsions in the loop region made spontaneous transitions to values different from those in M_4 or M5. In such cases, the loop would settle into a metastable state with different secondary structures, thus failing to commit to either M_4 or M_5 during the simulation time of 1 ns. All such non-committing simulations were treated as not forming transition paths and contributing zero to the transition rate. Nonetheless, more than 10% of the 120 simulation pairs did form a transition path, thus indicating that spontaneous transitions between M_4 and M_5 would indeed occur with finite rates. Details of the kinetic calculation for the two individual transition steps here are summarized in Table 3-7.

By taking the rates of transition step 1 and ignoring the resistance of other steps, we obtain (Eq. 2.19) the effective transition rates (with details provided in Table 3-8) between the OF and M_3 states: $k_{OF\to M3} \sim 1.7\times10^{2}$ S^{-1}, and $k_{M3\to OF} \sim 1.2$ S^{-1}. This simplifies the multistate system of OF, M_1, M_2, M_3 into a two-state system of OF and M_3 only. Similarly, using the obtained rates for step 5, we can reduce the multistate system of M_3, M_4, M_5, OC into a two-state system of M_3 and OC, with the effective transition rates of $k_{M_3\to OC} \sim 8\times10^{-3}$ S^{-1} and $k_{OC\to M_3} \sim 1.8\times10^{-1}$ S^{-1}. We thus finally obtain both the thermodynamics and the kinetics concerning the three major states OF, M3, and OC, as summarized in Figure 3-11.

By piecing together two opposite simulations starting from the same point at the barrier top, the transition paths represent spontaneous transitions between the two metastable states, thus providing valuable information about how such transitions would occur.

3.3.4 Free Energy Profiles at Each Step Transition

For both the *InitOF* and *InitOC* sets of simulations, the free energy profiles are calculated by the weighted histogram analysis method (WHAM). We calculated the free energy at temperatures 300K at each transition step along the chosen reaction coordinate (RC). Each transition step involves a two-state transition between adjacent metastable conformations. The CVs are a function of the atomic coordinates of the protein. Moreover, the CVs in this study are defined either by torsion dihedral angle, a bond length, or consisting of a combination of such collective variables (CV). Between the two adjacent intermediate states, each value of the CV is reduced to have a value between ~0.0 and ~1.0. Correspondingly, we measured the average of all the CVs as the selected reaction coordinate. From the US simulations, the free energy profile $G(\alpha)$ as a function of the reaction coordinate (RC) α is calculated separately at each transition step. In a two-state system, the free energy difference is given by $G_A = -k_B T \ln \int_{-\infty}^{\alpha^*} \exp\left[-\frac{G(\alpha)}{k_B T}\right] d\alpha$ for state A and $G_B = -k_B T \ln \int_{\alpha^*}^{\infty} \exp\left[-\frac{G(\alpha)}{k_B T}\right] d\alpha$ for state B. where α^* is the location of free energy barrier. Table 3-9 shows the energy difference between the two metastable states as $\Delta G = G_B - G_A$

Figure 3-12 shows the free energy profiles at each step transition along with the selected RC separately. At each transition step, the statistical errors were measured by the uncertainties of the mean forces at each window with respect to the closest metastable state to the OF state (i.e., The window with RC = 0)

3.4 Discussion

The direct task in this study is to elucidate the conformational changes between the OF and the OC states of the Mhp1 membrane transporter. In comparison to the transitions between the outward- and inward-facing states, the conformational changes here are relatively localized and of much smaller scales. However, this seemingly simple conformational change still has a significant degree of complexity. In our proposed mechanism, there are several intermediate metastable states between the OF and OC conformations. It is particularly unexpected that the intermediate state M3 has a lower calculated free energy than both the OF and the OC. Although M3 has some resemblance to the thin gate in some other Mhp1 structures (e.g., PDB ID: 2JLO[18]), one would expect the OF conformation described by the crystal structure (PDB ID: 2JLN[18]) to be the most

stable state for the substrate-free Mhp1. In addition to potential computational errors due to the force field[36-38] and the sampling, another contributing factor to the apparent discrepancy here is that our simulation systems do not have the bound Na^+ ion as in the crystal structures, and ion binding[39-41] could likely shift the relative free energies of the OF, OC, and M3 conformations.

More broadly, our case study on Mhp1 here serves to explore the simulation methodology for characterizing protein conformational changes in general. Despite the development of a large variety of enhanced sampling techniques in recent years, protein conformational changes remain a significant challenge for all-atom simulations. Inevitably, most enhanced sampling methods operate under certain assumptions, e.g., that the slow kinetics of the conformational transition can be captured by a collective coordinate and that all relevant orthogonal degrees of freedom can be properly equilibrated within the simulation time. When dealing with protein conformations, however, such assumptions are often not valid. Therefore, instead of developing or improving any particular enhanced sampling technique, we focus on a few practical strategies here to handle the enormous complexity of the conformational space, such that the sampling can still be feasible and reliable.

A key component in our strategies is a stringent method for gauging the convergence of conformational sampling. We first briefly define the meaning of convergence here. In general, the statistical accuracy of the sampling would increase with simulation times. Theoretically, an infinitely long simulation should reproduce the equilibrium ensemble, and the results therein can be considered the "correct" answer as far as the sampling is concerned[42]. Importantly, an infinitely long sampling will be completely independent of the initial coordinates. Regardless of the starting configuration of the simulation, the statistics from any infinite sampling will always be identical and will reproduce the Boltzmann distribution that does not depend on the specific time evolution[43, 44]. In contrast, results from simulations of finite times may depend on their starting configuration (i.e., the history), a phenomenon called hysteresis. The sampling error for a set of finite simulations is the deviation of its result from the correct answer, and an unacceptably large sampling error means that the simulations have not converged. As the correct answer from the ideal (infinitely long) simulations is unknown in practice, so is the exact sampling error. Consequently, all methods for examining the convergence rely on certain criteria to estimate the sampling errors and detect convergence problems. Whereas no criterion guarantees to identify all convergence problems, some are more capable than others.

A common method for checking convergence is to monitor the time evolution of the concerned quantities: if their values are not stabilized over the simulation time, the simulation clearly has not converged yet. However, the reverse is often not true. If the system is trapped in a metastable state without visiting other more probable states, all the concerned quantities would still appear to be fully stabilized during the simulation despite their values being far from the correct answers in true equilibrium. Some improvements in this aspect can be made by comparing multiple simulations. In umbrella sampling, e.g., inconsistency between the histograms from neighboring windows could indicate convergence problems that are otherwise not manifest in the individual trajectories.[34] In this study, our adopted strategy for examining convergence is to compare simulations with different initial coordinates. As discussed before, if the simulation is sufficiently long, the starting configuration should be "forgotten" with a vanishing effect on the resulting statistics. Therefore, a large deviation in the results between simulations with different initial coordinates is a clear evidence of hysteresis, indicating that a satisfactory convergence has not been achieved yet. This approach often captures convergence problems missed by other methods.

In our case study of Mhp1 here, the convergence is examined by comparing two independent groups of simulations originated from the two end conformations, i.e., the OF and the OC crystal structures, respectively. Sampling from the two end states has the following benefit for the free energy calculations. In general, when starting from conformation A and driving it toward conformation B (e.g., in our pulling simulations described in the Methods), the effect of hysteresis in an insufficient sampling tends to be an overestimation of the free energy difference $\Delta G \equiv G_B - G_A$. Reversely, when driving from conformation B toward A, hysteresis tends to bias toward underestimating the ΔG above. Therefore, the values obtained from these two groups of simulations would likely enclose the correct ΔG. Furthermore, such comparison is especially important for evaluating the simulation designs in the early exploratory stage. If the free energies from the two simulation groups persistently exhibit large differences and do not appear to converge toward each other over the simulation time, the simulation design would need adjustment to address the problem. On the other hand, an obvious drawback of this approach is the doubled computational cost, as it runs two groups of simulations for the same free energy profile. In this aspect, if an acceptable convergence appears plausible, one may combine the two groups into one set of simulations in the final production run.

Our strategy of comparing simulations with different initial coordinates should be applicable for testing convergence in all enhanced sampling techniques. In this study, we adopt umbrella sampling due to its conceptual simplicity and its practical convenience for diagnosis. Here, if the free energies from the two simulation groups differ significantly, it is easy to narrow down (by checking the mean forces[34]) the major discrepancy to certain umbrella window(s). Then, by closely comparing the two trajectories in the same window, the degrees of freedom responsible for the discrepancy often manifest themselves and could thus be properly considered to improve the simulation design. Using this strategy, through repeated trials and errors, we came up with the simulation design presented here in which the hysteresis is no longer significant. In our final results, the maximum difference between the free energies from the *InitOF* and *InitOC* groups is ~2.1 kcal/mol, which is acceptable albeit not great. It appears that our *InitOF* and *InitOC* systems have subtle differences in other regions of the protein as well as in specific protein-lipid/water interactions,[45] resulting in small but persistent deviations in the sampling that are difficult to eliminate without running the simulations much longer.

We also performed kinetics calculations by releasing the system at the free energy barrier in pairs of unbiased simulations with opposite initial velocities. In addition to the transition rates, these calculations revealed the quality of the RC[46] that defines the free energy profile. The results show that although our chosen RCs are not perfect, a substantial fraction of the simulation pairs did form a transition path that represents a spontaneous transition between the two metastable states, thereby allowing a rigorous calculation of the rate constants.[46] In equilibrium simulations starting from either metastable state, such transitions would take many orders of magnitude longer to occur. In comparison, therefore, our approach is much more efficient for generating rare events and elucidating all the details in the spontaneous transitions.[46, 47] Furthermore, whereas the free energy profile depends on the chosen reaction coordinate, the transition rates are independent of such artificial choices.[46] These kinetics calculations thus provide the intrinsic properties of the equilibrium ensemble as well as the major experimental observables.

In our strategies for achieving convergence, boundary restraints are powerful tools to simplify the conformational space and avoid certain complications in the sampling. Because the boundary restraints act to restrict some dimensions orthogonal to the RC, they may reduce the accessible microstates and thus increase the free energies in certain sections of the RC. In this study, all our applied boundaries are wide enough on the metastable states such that the free

energies or the equilibrium probabilities of the states are hardly affected, as explained in the Results. In contrast, some of the boundaries are narrow in the transition region (see Figure 3-4) and may increase the free energy barrier, which would in turn result in an underestimation of the transition rates. In this aspect, good boundary restraints should be such that in most of the unbiased spontaneous transitions, the system still stays inside the boundaries during barrier crossings, even without the restraints. In such cases, the boundary restraints will preserve the kinetics while facilitating the free energy sampling. Furthermore, in our kinetics calculations, all the unbiased simulations starting at the barrier top were not subject to any restraint. Therefore, the obtained transition paths indeed represent transitions that spontaneously occur in the equilibrium ensemble. Overall, when sampling highly complicated protein conformations, proper boundary restraints could prove instrumental or even indispensable for achieving convergence.

Our work also highlights the need to extend beyond simple two-state models when exploring protein conformational space. The conformational change of Mhp1 in this study, albeit of relatively small scale, still involves the transitions in many degrees of freedom, both local ones such as torsion angles and global ones such as the position and orientation of protein domains. All these diverse degrees of freedom do not necessarily have to make transitions simultaneously during spontaneous conformational changes. Indeed, as suggested by our results, it would be more plausible for the Mhp1 thin gate to take multiple transition steps in a complete conformational change, each involving just a few degrees of freedom. The sampling of such multi-step transitions also requires distinct RCs for each step, rather than a single collective coordinate for the entire conformational change as in many other simulation studies. Surveys of proteins with multiple known conformations reveal that their structures also differ in many local and global degrees of freedom, thus suggesting that the multi-step scheme could be common for protein conformational changes in general.

One concern for our approaches here is the involvement of many artificial choices. Indeed, our simulation design, including the transition steps, the RC for each step, and all the boundary potentials, is primarily based on our understanding of the mechanism. Whereas there are potentially many metastable states for Mhp1, we only focus on a few as the intermediates for the conformational change here. Similarly, by applying each boundary restraint, we have limited our consideration to the conformational space within the boundary only. In fact, spontaneous conformational changes in the equilibrium ensemble may take place through different routes

(pathways), and we merely characterized one plausible route in this work. Problems of this sort are often tackled in the field of quantum-mechanical calculations for chemical reactions, where multiple hypothetical mechanisms are proposed and tested to identify the most favored one with the lowest energetic barrier. Similarly, for protein conformational changes here, given that an exhaustive search for all the possible routes is clearly infeasible, it is justified to explore plausible routes based on intuitions, examine them through rigorous calculations, and possibly compare multiple alternatives. Thermodynamic consistency demands that all calculations, if correct, should give the same free energy differences between common metastable states, regardless of the routes or the chose RCs. In contrast, different routes may have different transition rates, and the most relevant routes are the ones taken the most frequently by the spontaneous conformational changes in the equilibrium ensemble. Further explorations, therefore, could likely update our knowledge about the mechanism of the given conformational change.

In summary, simulating complex protein conformational changes is a challenging task. The strategies demonstrated in this study are designed to alleviate some of the difficulty arising from the enormous complexity and to facilitate the sampling of protein conformations. By properly designing the simulations and carefully examining their convergence, it should be possible to obtain reliable and reproducible thermodynamic and kinetic quantities, thereby elucidating the molecular mechanism for the reversible conformational change.

3.5 Figures and Tables

Table 3-1. Information of US[3, 4] for each of the six transition steps. The column of $CV(X)$ lists the CVs that define the RC for each transition step. The values of cv_A and cv_B are used to convert each CV to its reduced form (Eq. 3-3). The column of K provides the spring constant for the umbrella potential on the RC (Eq. 3-4). The values of α_A and α_B determine the range of the umbrella windows, as explained in the text. The last column gives the parameter Δ involved only in the potentials (Eq. 3-6) for the two end windows. The number of windows given in the table has included the two end windows, thus corresponding to $M + 2$ in the text.

$$CV^*(X) \equiv \frac{CV(X) - cv_A}{cv_B - cv_A}$$

Step	$CV(X)$	cv_A	state	cv_B	state	$K\ \frac{kcal}{mol}$	Number of US windows	α_A	α_B	Δ
1	Dihedral W117 sidechain χ_1	290.0°	OF	195.0°	M_1	750	15	0.11	0.89	0.10
	Dihedral W117 sidechain χ_2	−60.0°		65.0°						
2	Bond Q121:C_α - L359: C_γ	15.7 Å	M_1	9.5 Å	M_2	2000	31	0.02	1.0	0.055
	Dihedral P352:C_α - L359:N - L359:C_α - L359:C_β	295.0°		200.0°						
3	$\lvert\vec{r}(\text{L366:C}_\gamma) - \vec{r}(\text{AP})\rvert - \lvert\vec{r}(\text{L113:C}_\gamma) - \vec{r}(\text{AP})\rvert$	5.0 Å	M_2	−5.0 Å	M_3	4000	33	0.05	0.89	0.10
4	Dihedral W117 sidechain χ_1	195.0°	M_3	290.0°	M_4	850	16	0.11	0.89	0.10
	Dihedral W117 sidechain χ_2	65.0°		−60.0°						
5	Dihedral V358 backbone ψ	20.0°	M_4	115.0°	M_5	2000	32	0.02	0.96	0.055
	Dihedral G357 backbone ψ	320.0°		180.0°						
	Dihedral F355 backbone ψ	10.0°		135.0°						
6	Dihedral F355 sidechain χ_1	310.0°	M_5	175.0°	OC	610	17	0.15	0.85	0.15

Table 3-2. Boundary restraints on the reduced CVs involved in defining any of the RCs Table 3-1). All the restraints have Δ_2= 0.4 and $K_{b2} = 1000\ kcal/mol$. The parameter cv_0^* is given in the entries for the restraint on each CV in each transition step. Each reduced CV is defined using the corresponding values of α_A and α_B in Table 3-1. For the first CV (Dihedral W117 sidechain χ1) here, its reduced form CV^* is defined using the α_A and α_B in the transition step 1 (instead of step 4) in Table 3-1.

$$U_{b2}(X) = \begin{cases} \frac{K_{b2}}{2}\ [|CV^*(X) - cv_0^*| - \Delta_2\ ]^2 & |CV^*(X) - cv_0^*| > \Delta_2 \\ 0 & Otherwise \end{cases}$$

$\boldsymbol{CV^*(X)}$	**Step 1**	**Step 2**	**Step 3**	**Step 4**	**Step 5**	**Step 6**
Dihedral W117 sidechain χ_1	-	1.0	1.0	-	0.0	0.0
Bond Q121:C_α - L359: C_γ	0.0	-	1.0	1.0	1.0	1.0
Dihedral P352:C_α - L359:N - L359:C_α - L359:C_β	0.0	-	1.0	1.0	1.0	1.0
$\lvert\vec{r}(\text{L366:C}_\gamma) - \vec{r}(\text{AP})\rvert - \lvert\vec{r}(\text{L113:C}_\gamma) - \vec{r}(\text{AP})\rvert$	0.0	0.0	-	1.0	1.0	1.0
Dihedral V358 backbone ψ	0.0	0.0	0.0	0.0	-	1.0
Dihedral G357 backbone ψ	0.0	0.0	0.0	0.0	-	1.0
Dihedral F355 backbone ψ	0.0	0.0	0.0	0.0	-	1.0
Dihedral F355 sidechain χ_1	0.0	0.0	0.0	0.0	0.0	-

Table 3-3. Common boundary restraints on some backbone φ and ψ torsion angles for all the transition steps

$$U_{b2}(X) = \begin{cases} \frac{K_{b2}}{2} \ [|CV(X) - cv_0| - \Delta_2 \ ]^2 & |CV(X) - cv_0| > \ \Delta_2 \\ 0 & Otherwise \end{cases}$$

$CV(X)$	cv_0	Δ_2	K_{b2} $(\frac{kcal}{\text{mol rad}^2})$
Dihedral Q354 backbone φ	$-75.0°$	60.0°	100
Dihedral Q354 backbone ψ	0.0°	60.0°	100
Dihedral F355 backbone φ	$-87.0°$	80.0°	100
Dihedral A356 backbone φ	$-82.0°$	80.0°	100
Dihedral A356 backbone ψ	$-20.0°$	60.0°	100
Dihedral G357 backbone φ	$-75.0°$	80.0°	100
Dihedral V358 backbone φ	$-82.0°$	80.0°	100
Dihedral L359 backbone φ	$-80.0°$	80.0°	100
Dihedral L359 backbone ψ	$-40.0°$	80.0°	100

Table 3-4. Additional boundary restraints U_{b2} for each individual transition step.

$$U_{b2}(X) = \begin{cases} \frac{K_{b2}}{2}\ [|CV(X) - cv_0| - \Delta_2\ ]^2 & |CV(X) - cv_0| > \Delta_2 \\ 0 & Otherwise \end{cases}$$

step	$CV(X)$	$cv_0(X)$	Δ_2	K_{b2}
1	Dihedral N360:C_α - L359:C_α - V358:C_α - G357:C_α	$-75.0°$	75.0°	$100.0(\frac{kcal}{\text{mol rad}^2})$
	Dihedral N360:N - L359:C - L359:C_α - L359:C_β	145.0°	50.0°	$100.0(\frac{kcal}{\text{mol rad}^2})$
	Bond V358:C - F362:N	6.5Å	1.5Å	$25.0\frac{kcal}{mol\ Å^2}$
2	Angle P352:C_α - L359:C_α - L359:C_β	70.0Å	50.0Å	$100.0\frac{kcal}{mol\ rad^2}$
	Bond A369:C_α - L373:N	6.25Å	1.75Å	$25.0\frac{kcal}{mol\ Å^2}$
	Bond L113:C_α - L366:C_γ	8.25Å	8.0Å	$25.0\frac{kcal}{mol\ Å^2}$
3	Bond V358:C - F362:N	4.00Å	1.00Å	$25.0\frac{kcal}{mol\ Å^2}$
	Bond A369:C_α - L373:N	6.25Å	1.75Å	$25.0\frac{kcal}{mol\ Å^2}$
	Bond S343:C_α - A367:C_α	15.0Å	8.0Å	$25.0\frac{kcal}{mol\ Å^2}$
	Bond G347:C_α - L363:C_α	12.0Å	5.7Å	$25.0\frac{kcal}{mol\ Å^2}$
4	Dihedral N360:N - L359:C - L359:C_α - L359:C_β	90.0°	45.0°	$100.0(\frac{kcal}{\text{mol rad}^2})$
	Dihedral L359:N - V358:C - V358:C_α - V358:C_β	0.0°	50.0°	$100.0(\frac{kcal}{\text{mol rad}^2})$
	Bond V358:C - F362:N	4.0Å	1.5Å	$25.0\frac{kcal}{mol\ Å^2}$
5	Dihedral N360:C_α - L359:C_α - V358:C_α - G357:C_α	185.0°	95.0°	$100.0(\frac{kcal}{\text{mol rad}^2})$
	Dihedral N360:N - L359:C - L359:C_α - L359:C_β	90.0°	45.0°	$100.0(\frac{kcal}{\text{mol rad}^2})$
	Bond Q121:C_α - L359:C_γ	10.5Å	2.5Å	$25.0\frac{kcal}{mol\ Å^2}$
6	Bond V358:O - F362:N	2.0Å	3.5Å	$25.0\frac{kcal}{mol\ Å^2}$
	Bond V358:C_α - T361:C_α	3.0Å	4.5Å	$25.0\frac{kcal}{mol\ Å^2}$

Table 3-5. Additional boundary restraints U_{b3} (with the boundary changing linearly with the RC) for individual transition steps.

$$U_{b3}(X) = \begin{cases} \frac{K_{b3}}{2} \ [|CV(X) - (\omega \ RC(X) + \beta)| - \Delta_3]^2 & |CV(X) - (\omega \ RC(X) + \beta)| > \Delta_3 \\ 0 & Otherwise \end{cases}$$

step	$CV(X)$	ω	β	Δ_3	K_{b3}
2	Dihedral N360:C_α - L359:C_α - V358:C_α - G357:C_α	$-70.0°$	105.0°	40.0°	$50.0(\frac{kcal}{\text{mol rad}^2})$
	Dihedral N360:N - L359:C - L359:C_α - L359:C_β	60.0°	75.0°	45.0°	$50.0(\frac{kcal}{\text{mol rad}^2})$
	Bond L366:C_α - L370:C_α	0.36Å	6.75Å	1.10Å	$25.0\frac{kcal}{mol\ Å^2}$
	Bond L365:C_α - A369:C_α	-0.20Å	6.60Å	1.00Å	$25.0\frac{kcal}{mol\ Å^2}$
	Bond N364:C_α - S368:C_α	0.02Å	6.06Å	0.55Å	$25.0\frac{kcal}{mol\ Å^2}$
	Bond T361:C_α - L365:C_α	0.07Å	6.23Å	0.65Å	$25.0\frac{kcal}{mol\ Å^2}$
	Bond A356:C_α - N360:C_α	-2.03Å	7.87Å	1.00Å	$25.0\frac{kcal}{mol\ Å^2}$
	Bond L359:C_α - L363:C_α	-0.02Å	6.10Å	0.60Å	$25.0\frac{kcal}{mol\ Å^2}$
	Bond G357:C_α - T361:C_α	-0.63Å	9.04Å	1.20Å	$25.0\frac{kcal}{mol\ Å^2}$
	Bond A356:C - N360:N	-1.70Å	5.95Å	0.90Å	$25.0\frac{kcal}{mol\ Å^2}$
	Bond V358:C - F362:N	2.25Å	4.11Å	1.30Å	$25.0\frac{kcal}{mol\ Å^2}$
5	Bond L124:C_α - A356:C_α	3.92Å	15.25Å	2.50Å	$25.0\frac{kcal}{mol\ Å^2}$
	Bond L118:C_α - A356:C_α	-3.58Å	17.17Å	2.00Å	$25.0\frac{kcal}{mol\ Å^2}$

Table 3-6. 1 and 0 represent the type of H-bonds present and absent in the related conformational state, respectively. The H-bonds form by carbonyl C=O and amide N-H either at the backbone or side chain of relevant residues.

Donor-acceptor distance < 4.0 Å, Donor-acceptor angle > 140°							
Hydrogen Bond	**OF**	**M_1**	**M_2**	**M_3**	**M_4**	**M_5**	**OC**
(L359 backbone NH) → (F355 backbone O)	1	1	0	0	0	0	0
(N360 Backbone NH) → (G357 Backbone O)	1	1	0	0	0	0	0
(W117 Sidechain NH) ↔ (Q42 Sidechain O, NH)	0	1	1	1	0	0	0
(V358 Backbone NH) → (F355 Backbone O)	0	0	1	1	1	0	0
(L359 Backbone NH) → (A356 Backbone O)	0	0	1	1	1	0	0
(F362 Backbone NH) → (V358 Backbone O)	0	0	1	1	1	1	1
(Q51 Sidechain NH) ↔ (N360 Sidechain O, NH)	0	0	0	1	1	1	1
(S422 Sidechain OH) ↔ (S368 Sidechain OH)	0	0	0	1	1	1	1

Table 3-7. Spontaneous transition rate k_0 calculated by 120 unbiased simulations at steps 1 and 5 for *InitOF* and *InitOC*, separately. At step1, the chosen interval is $[\alpha_1 = 0.45, \alpha_2 = 0.55]$ and at step 5 the chosen interval is $[\alpha_1 = 0.54, \alpha_2 = 0.56]$. The transition rates' unit is s^{-1}.

Step1 Trp117	N_{sim}	N_{TP}	$< \lambda(\alpha 1, \alpha 2) > s^{-1}$	$p(\alpha 1 < \alpha < \alpha 2)$	k_0	$k_{OF \to M1}$	$k_{M1 \to OF}$
InitOF	120	19	$2.4 * 10^{10}$	$1.1 * 10^{-8}$	131.93	159.8	756.9
InitOC	120	21	$3.7 * 10^{10}$	$0.9 * 10^{-8}$	161.8	180.2	1583.6
Step5 Loop	N_{sim}	N_{TP}	$< \lambda(\alpha 1, \alpha 2) > s^{-1}$	$p(\alpha 1 < \alpha < \alpha 2)$	k_0	$k_{M4 \to M5}$	$k_{M5 \to M4}$
InitOF	120	15	$2.69 * 10^{9}$	$3.06 * 10^{-11}$	0.041	0.042	4.313
InitOC	120	14	$2.08 * 10^{9}$	$1.98 * 10^{-11}$	0.021	0.021	0.961

Table 3-8. Calculation of the overall transition rates between OF and M_3 with OF-M_1 being the rate-limiting step and between M_3 and OC with M_4-M_5 being the rate-limiting step. The overall rates were obtained according to Eq. 3-21 in Method, with the kinetics of the rate-limiting steps taken from Table 3-7. The unit of the transition rates is s^{-1}.

OF-M3	P_{OF}	P_{M1}	P_{M2}	P_{M3}	$k_{OF\rightarrow M3}$ Forward	$k_{M3\rightarrow OF}$ Backward
InitOF	0.0084	0.0018	0.0005	0.9893	159.8	1.3772
InitOC	0.0061	0.0007	0.0067	0.9865	180.2	1.1237
M3-OC	P_{M3}	P_{M4}	P_{M5}	P_{OC}	$k_{M3\rightarrow OC}$ Forward	$k_{OC\rightarrow M3}$ Backward
InitOF	0.7731	0.1903	0.0018	0.0348	0.0103	0.2231
InitOC	0.7434	0.2217	0.0048	0.0301	0.0063	0.1532

Table 3-9. Energy difference between two metastable states of each transition step. The energy difference between the state with RC=1 and RC = 0. The energy unit for both ΔG and the estimated error is kcal/mol.

		InitOF		***InitOC***	
Step	State	ΔG	Estimated Error	ΔG	Estimated Error
1 (W117)	OF-M_1	0.93	0.15	1.30	0.08
2 (L359)	M_1-M_2	0.76	0.16	-1.35	0.19
3 (TMH10)	M_2-M_3	-4.53	0.34	-2.98	0.43
4 (W117)	M_3-M_4	0.84	0.15	0.72	0.09
5 (Loop)	M_4-M_5	2.76	0.18	2.28	0.19
6 (F355)	M_5-OC	-1.78	0.12	-1.09	0.15

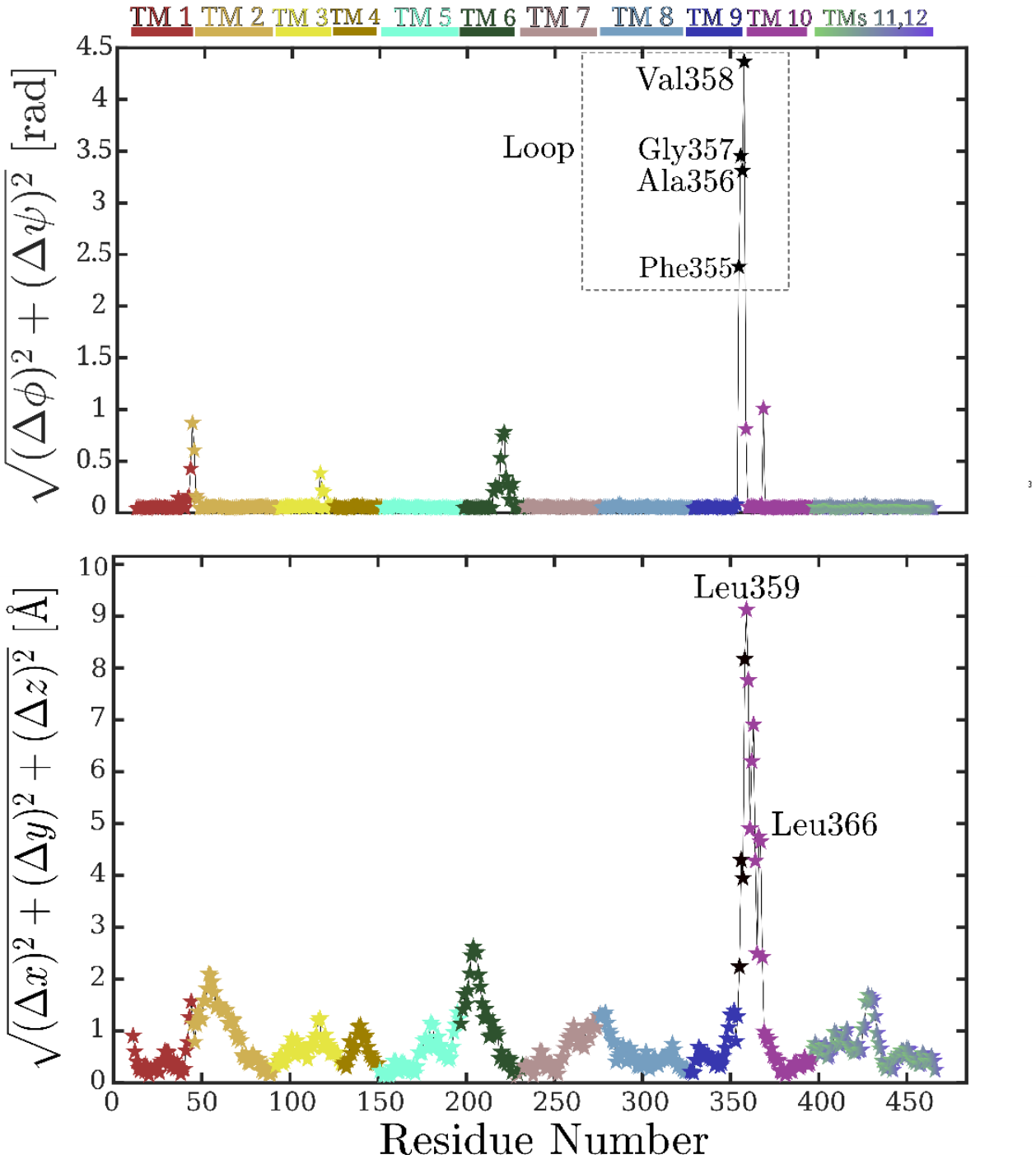


Figure 3-1. Top) The difference in backbone torsion angles for each residue between the crystal structures of OF and OC. bottom) The difference for each C_α atom between its positions in the OF and OC structures after alignment.

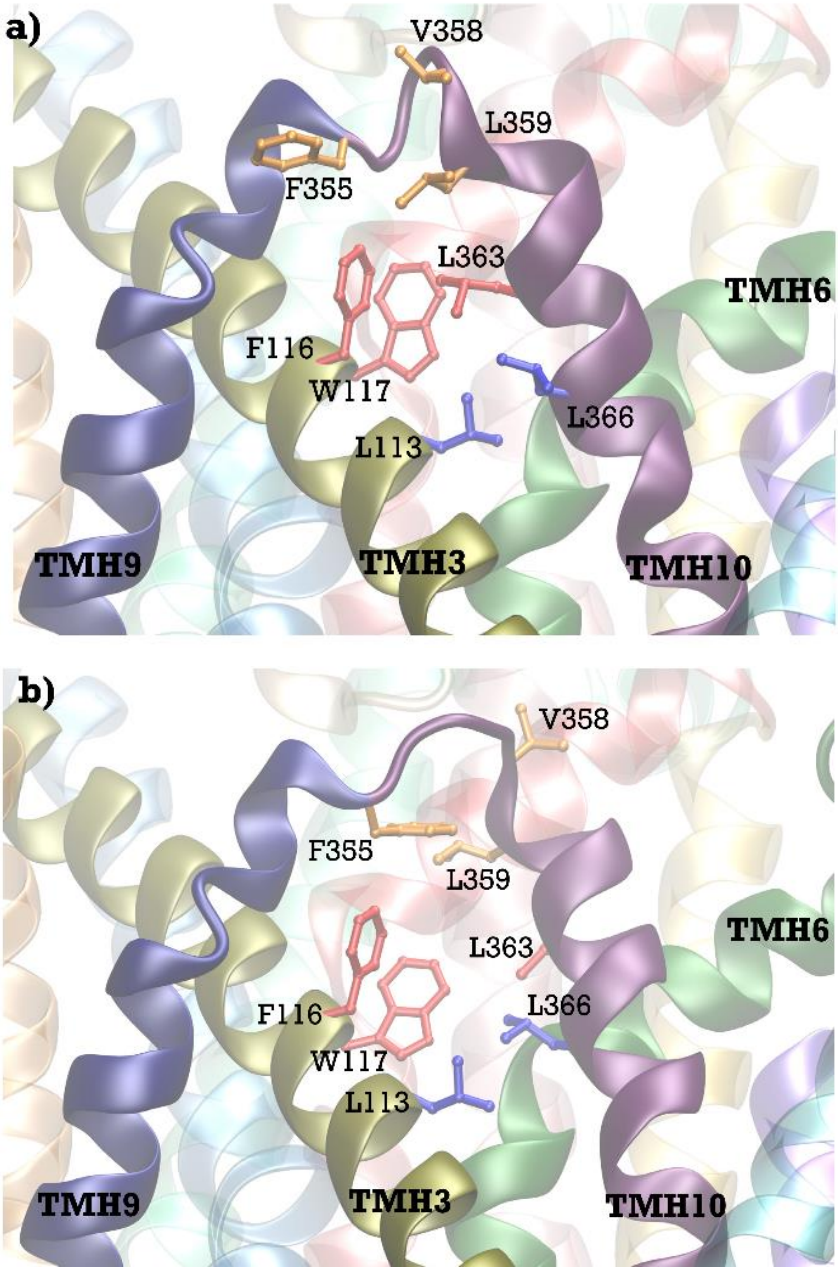


Figure 3-2. Protein structures around the thin gate in the OF (a) and OC (b) states. Relevant residues that highlight the difference between the two structures are shown and labeled.

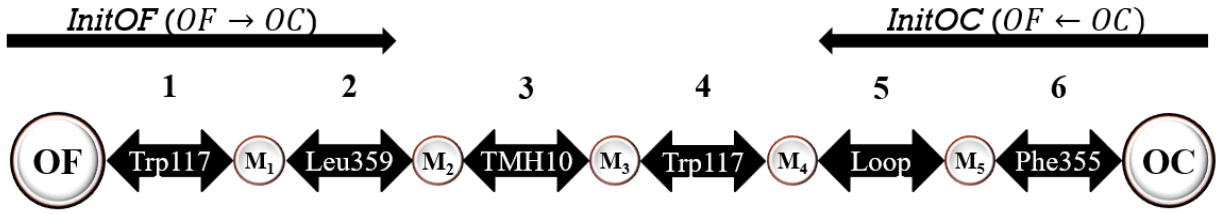


Figure 3-3. Stepwise transition model. The model consists of six transitions (1-6) with five metastable states (M_1-M_5) between the OF and OC states at the two ends. Two groups of simulations, *InitOF*, and *InitOC*, starting from the OF and OC crystal structures, respectively, were performed in this study.

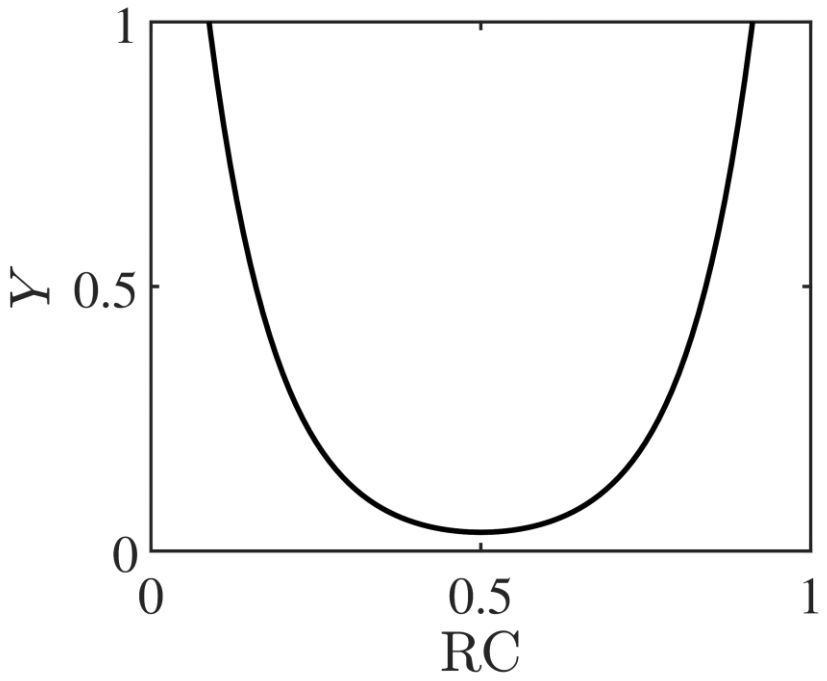


Figure 3-4. The dependence of the parameter Y (in Eq. 5) on the RC. If the difference between the two reduced CVs exceeds Y, a harmonic potential will act to reduce the difference (see Eq. 5).

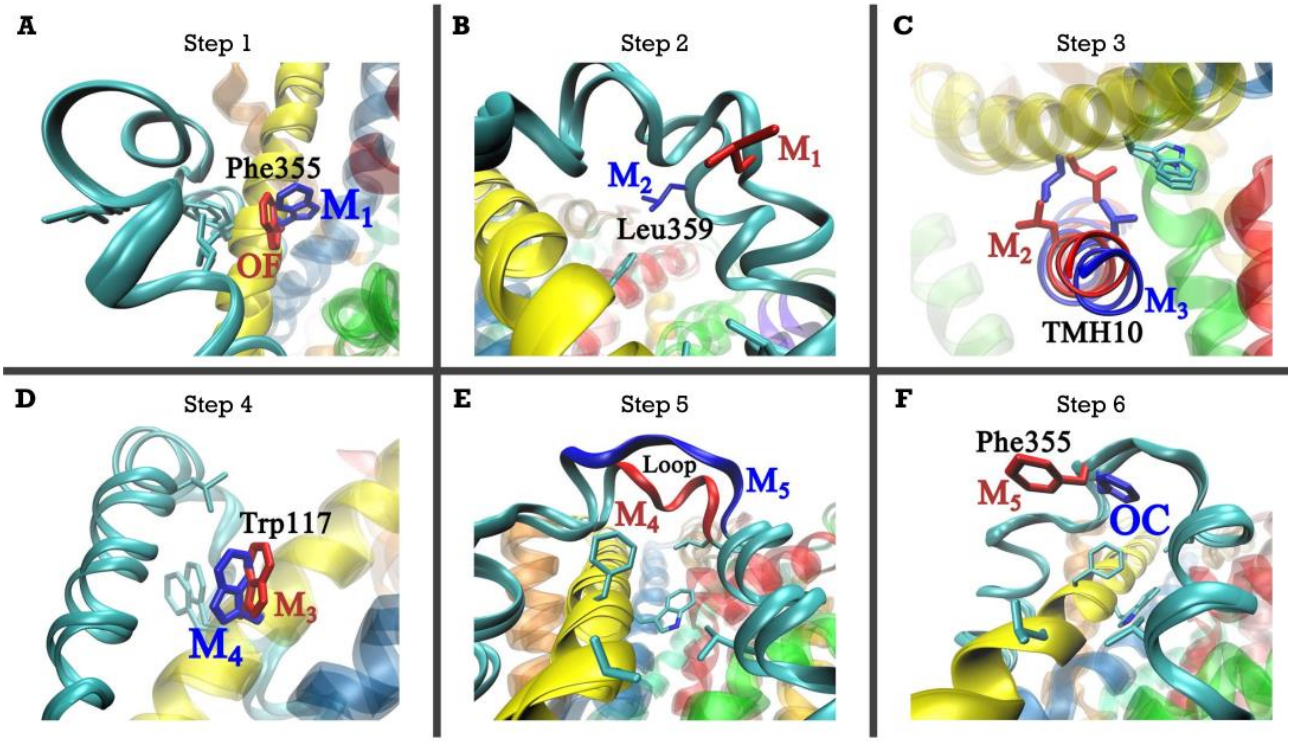


Figure 3-5. Protein conformational change with the sequence of transition steps between the OF and OC states of Mhp1. From step1 (A) to step6 (F), the transition starts with the protein segment in red, and the transition ends in blue color. These sequences undergo the protein from OF to OC state. Reversely, for backward transition, the protein segment starts with blue and ends with red at each step transition. Therefore, the sequence from step6 (F) to step1 (A) causes the protein to go through OC to OF state.

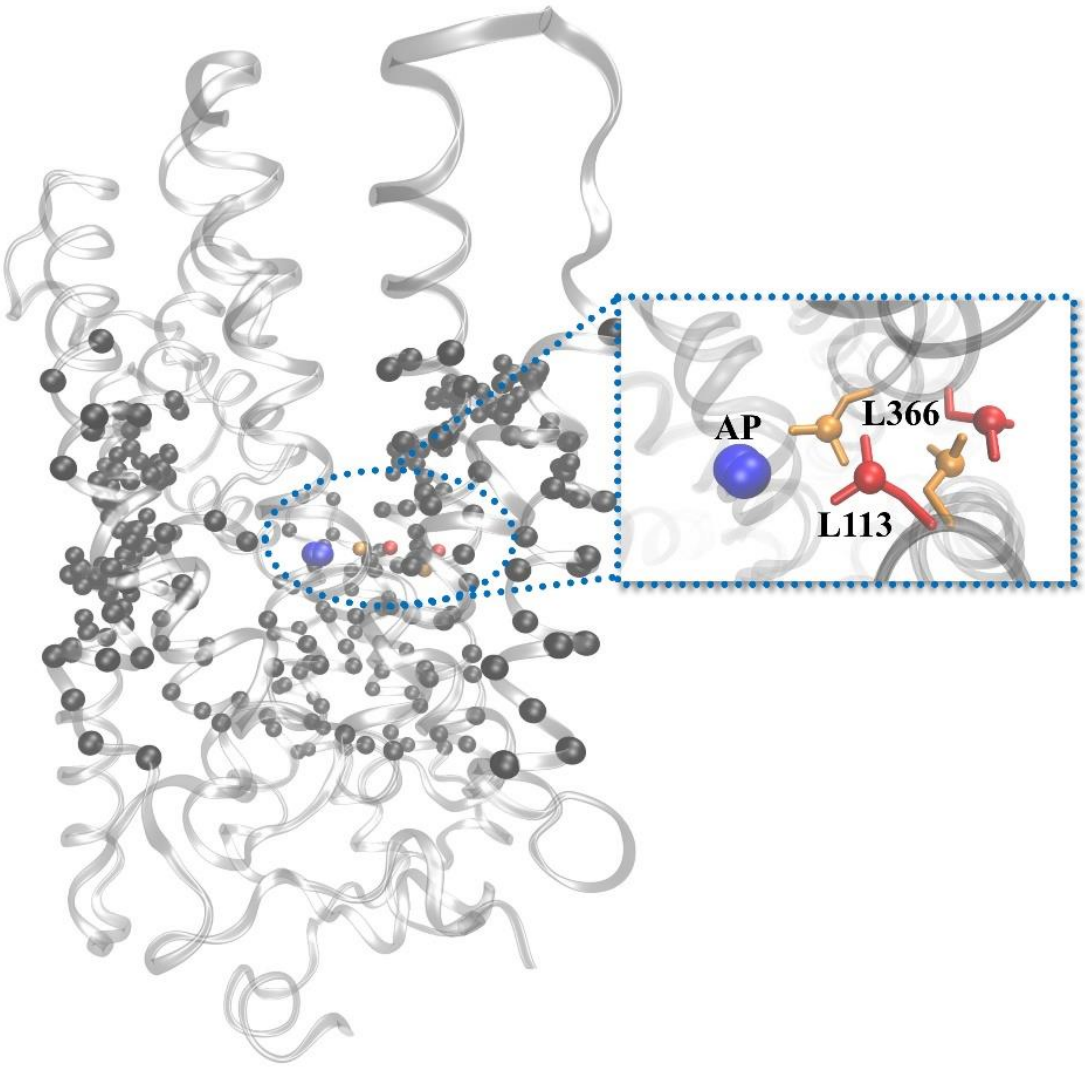


Figure 3-6. The anchor point (AP) is used to define the RC for transition step 3 between conformations M_2 and M3. The AP (blue spheres) is defined as the center of mass for the atoms shown in black spheres. The sidechains of L366 and L113 at the M_2 and M_3 states are shown in red and orange, respectively.

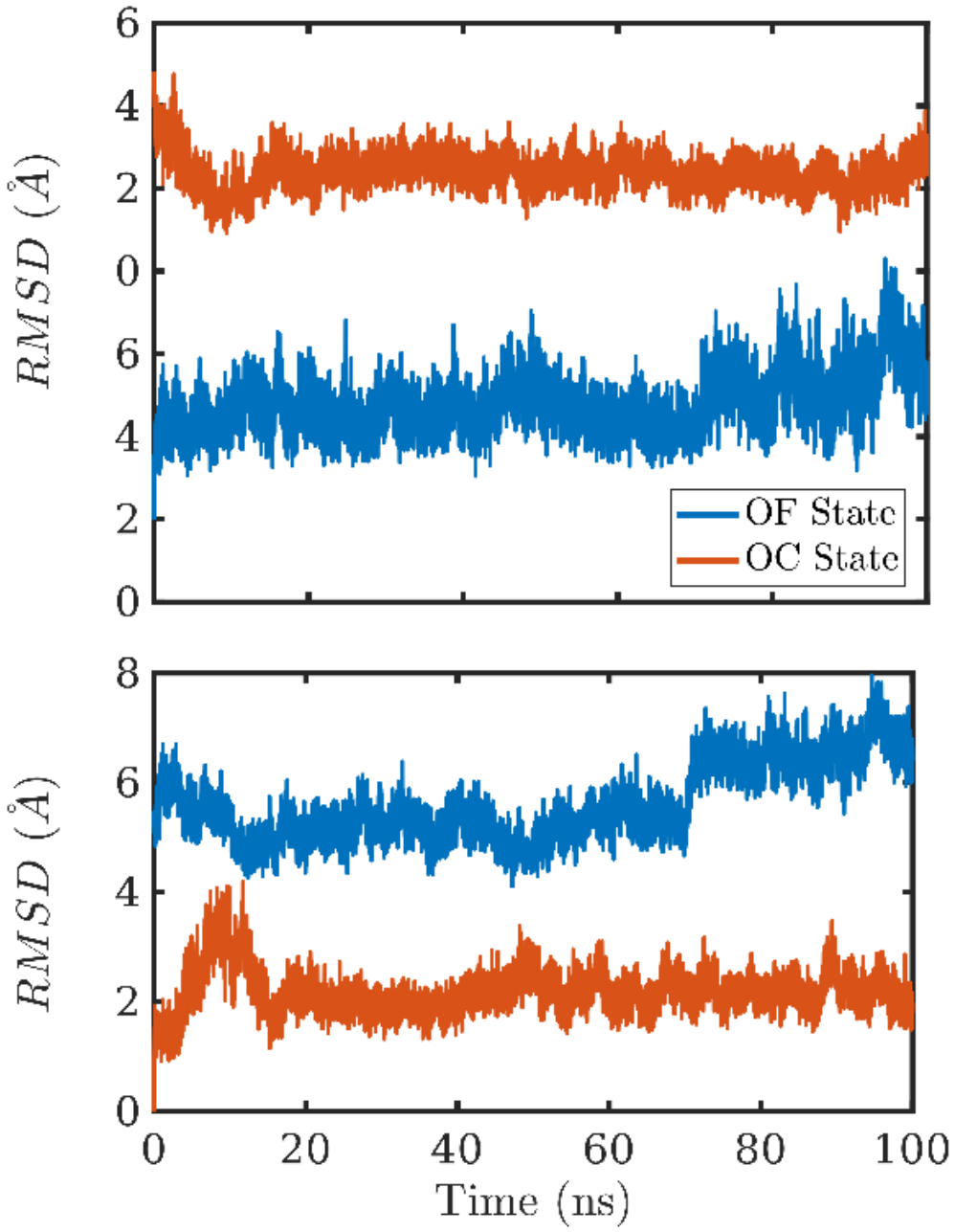


Figure 3-7. Root mean square deviation (RMSD) of Cα atoms of residues from 355 to 368. The two unbiased simulations' trajectories, first aligned by the entire Cα atoms of the protein crystal structure as the reference conformation. Top) The OF crystal structure was used as the reference confirmation. Bottom) The reference coordinate is OC crystal structure. The two 100 ns unbiased simulations show no significant conformational changes.

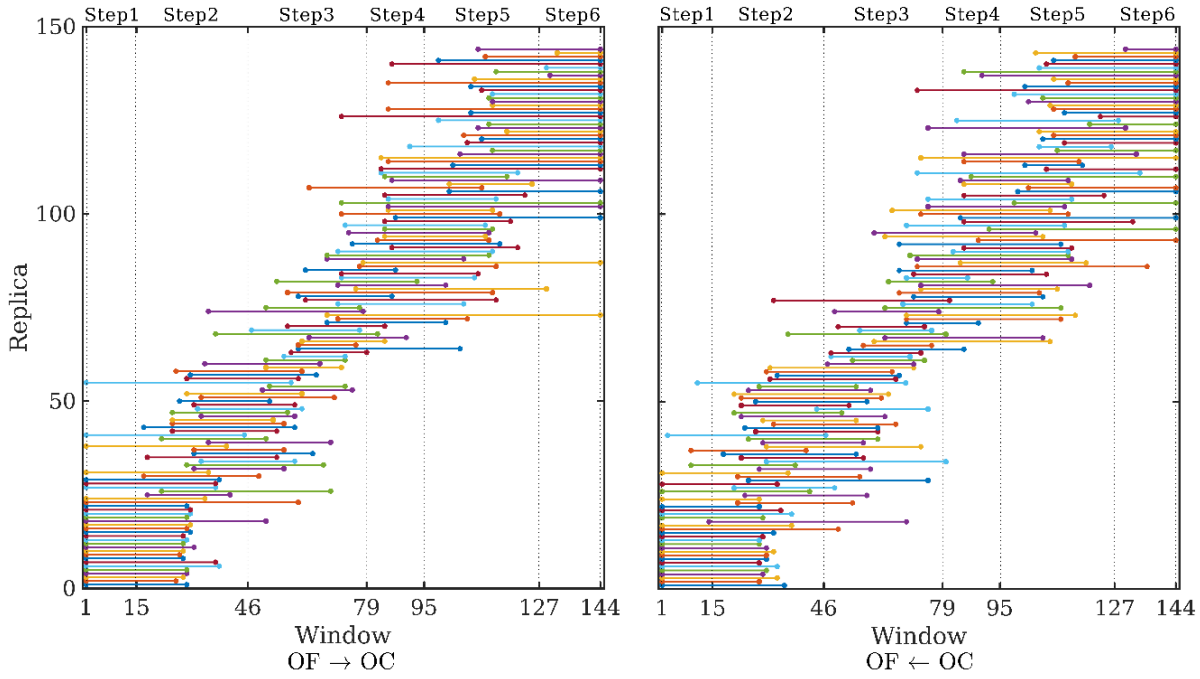


Figure 3-8. Umbrella windows that each replica sampled during Hamiltonian replica exchange MD. The left figure shows the *InitOF* transition (OF→OC), while the right represents the *InitOC* transition (OF←OC).

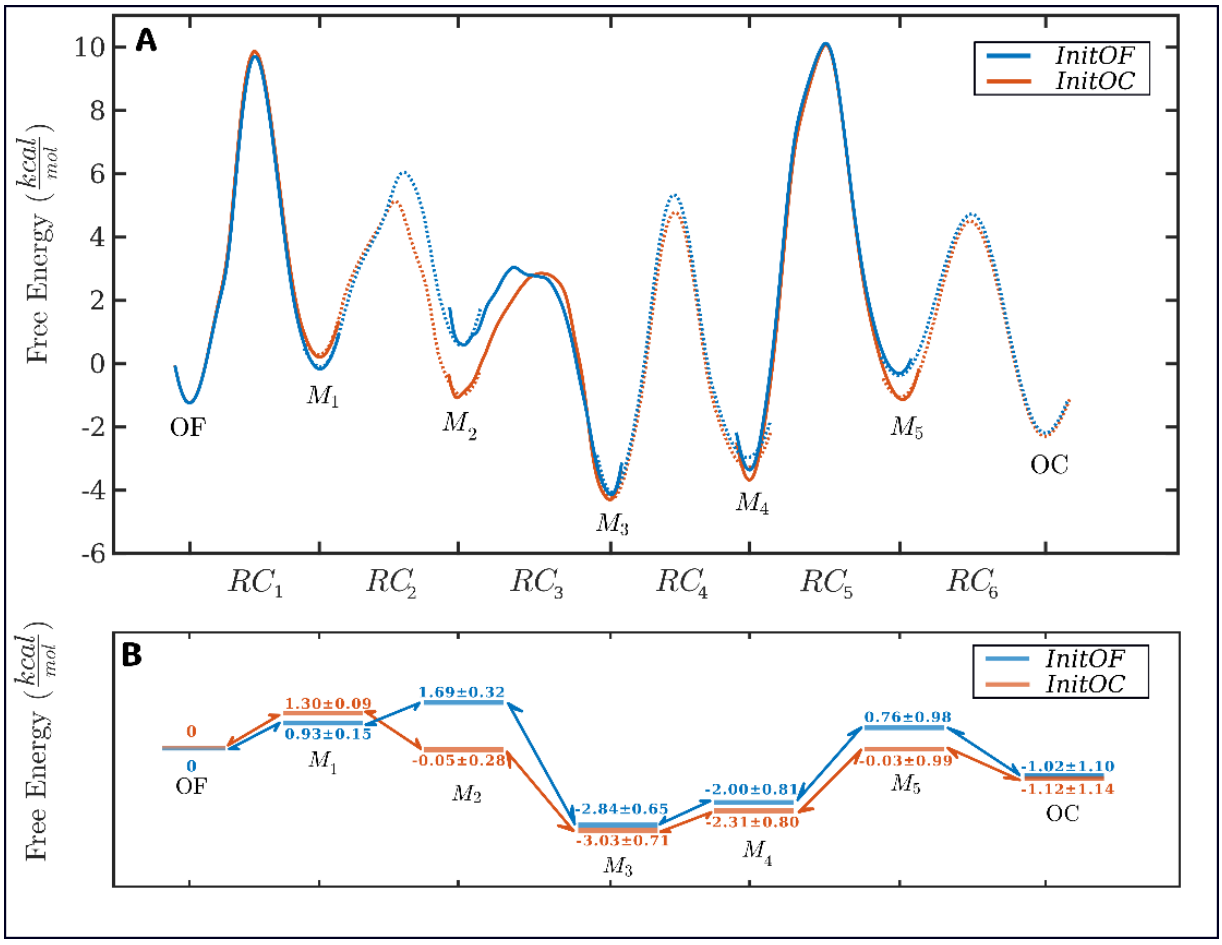


Figure 3-9 A) Free energy profile of MHP1 between the outward-facing open and outward-facing occluded state. B) The value of the G_A and G_B at different metastable state. At each step transition, we measured the statistical errors from the uncertainties of the mean forces at each window with respect to the first umbrella window at OF state.

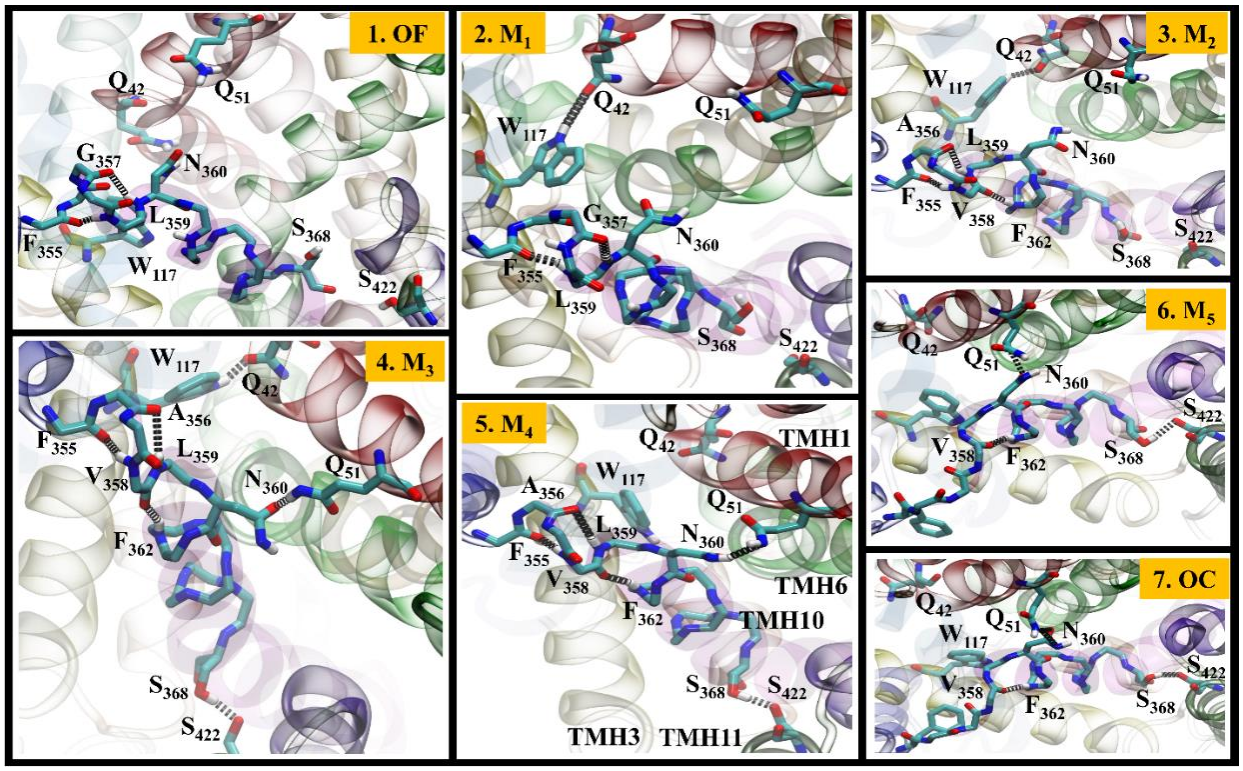


Figure 3-10. H-bonds at seven conformational states, OF, M_1-M_5, and OC, are shown by black rings between Nitrogen atoms in blue, Hydrogen in white, and Oxygen in red color. All H-bonds are specified by donor-acceptor distance to be smaller than 4.0 Å, and the donor-acceptor angle to be larger than 140°. The name of the transmembrane helices is shown only on panel 5.M_4, which can be found at the other metastable state with the same color.

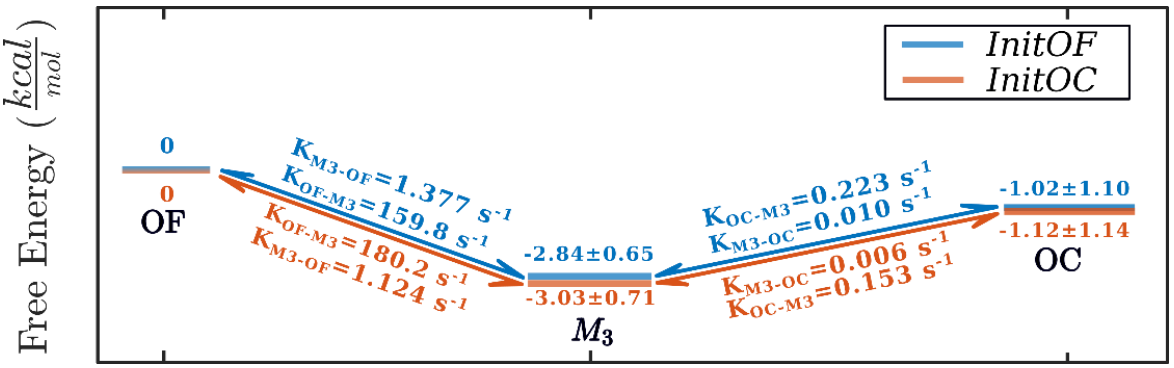


Figure 3-11. The multistate system of OF, M_1, M_2, M_3 are presented as a two-state system of OF and M3. Similarly, M3, M4, M5, OC are presented as a two-state system of M_3 and OC. The kinetic rate for both forward and backward transitions between these three states is shown in blue for *InitOF* and red for *InitOC* transitions.

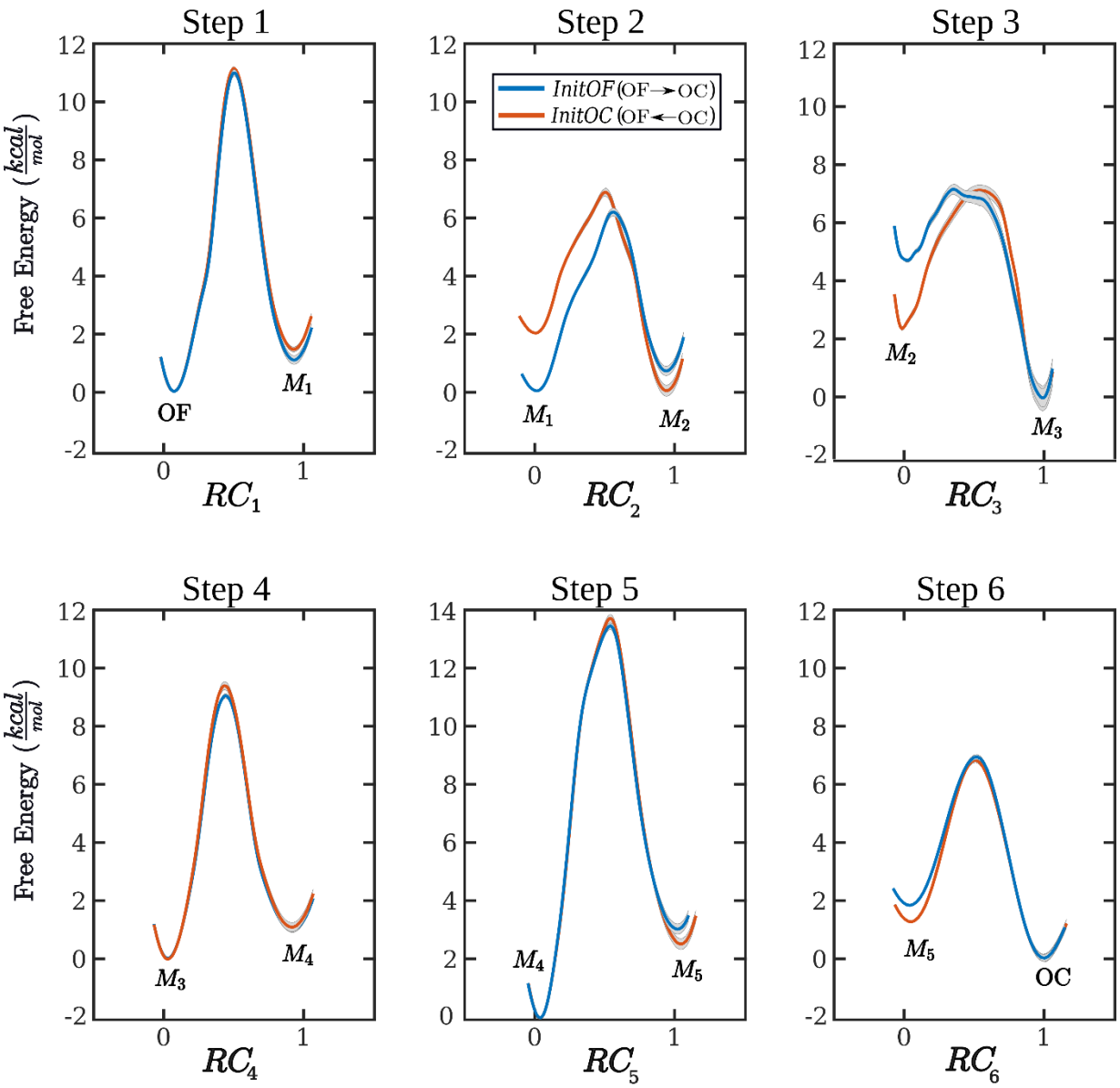


Figure 3-12. Free energy profile along the selected reaction coordinate at each transition step. The *InitOF* transition, which is from $OF \rightarrow OC$ state is shown in blue, and *InitOC* transition ($OF \leftarrow OC$) is shown in red. At each transition step, the statistical errors were measured by the uncertainties of the mean forces at each window with respect to the window with $RC = 0$.

3.6 References

1. Shaw, D.E., et al., *Atomic-level characterization of the structural dynamics of proteins.* Science, 2010. **330**(6002): p. 341-6.

2. Lindorff-Larsen, K., et al., *How fast-folding proteins fold.* Science, 2011. **334**(6055): p. 517-20.

3. Kästner, J., *Umbrella sampling.* Wiley Interdisciplinary Reviews: Computational Molecular Science, 2011. **1**(6): p. 932-942.

4. Shea, J.-E. and C.L. Brooks III, *From folding theories to folding proteins: a review and assessment of simulation studies of protein folding and unfolding.* Annual review of physical chemistry, 2001. **52**(1): p. 499-535.

5. Bolhuis, P.G., et al., *Transition path sampling: throwing ropes over rough mountain passes, in the dark.* Annu Rev Phys Chem, 2002. **53**: p. 291-318.

6. Leone, V., et al., *Targeting biomolecular flexibility with metadynamics.* Curr Opin Struct Biol, 2010. **20**(2): p. 148-54.

7. Hamelberg, D., J. Mongan, and J.A. McCammon, *Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules.* Journal of Chemical Physics, 2004. **120**(24): p. 11919-29.

8. Darve, E., D. Rodriguez-Gomez, and A. Pohorille, *Adaptive biasing force method for scalar and vector free energy calculations.* J Chem Phys, 2008. **128**(14): p. 144120.

9. Faradjian, A.K. and R. Elber, *Computing time scales from reaction coordinates by milestoning.* J Chem Phys, 2004. **120**(23): p. 10880-9.

10. Perilla, J.R., et al., *Computing ensembles of transitions from stable states: Dynamic importance sampling.* J Comput Chem, 2011. **32**(2): p. 196-209.

11. Huber, G.A. and S. Kim, *Weighted-ensemble Brownian dynamics simulations for protein association reactions.* Biophys J, 1996. **70**(1): p. 97-110.

12. Isralewitz, B., M. Gao, and K. Schulten, *Steered molecular dynamics and mechanical functions of proteins.* Curr Opin Struct Biol, 2001. **11**(2): p. 224-30.

13. E, W., W. Ren, and E. Vanden-Eijnden, *String method for the study of rare events.* Physical Review B, 2002. **66**(5): p. 52301.

14. Maragliano, L., et al., *String method in collective variables: minimum free energy paths and isocommittor surfaces.* J Chem Phys, 2006. **125**(2): p. 24106.

15. E, W., W. Ren, and E. Vanden-Eijnden, *Finite temperature string method for the study of rare events.* J Phys Chem B, 2005. **109**(14): p. 6688-93.

16. Vanden-Eijnden, E. and M. Venturoli, *Revisiting the finite temperature string method for the calculation of reaction tubes and free energies.* J Chem Phys, 2009. **130**(19): p. 194103.

17. Ovchinnikov, V., M. Karplus, and E. Vanden-Eijnden, *Free energy of conformational transition paths in biomolecules: the string method and its application to myosin VI.* J Chem Phys, 2011. **134**(8): p. 085103.

18. Weyand, S., et al., *Structure and molecular mechanism of a nucleobase–cation–symport-1 family transporter.* Science, 2008. **322**(5902): p. 709-713.

19. Simmons, K.J., et al., *Molecular mechanism of ligand recognition by membrane transport protein, Mhp1.* The EMBO journal, 2014. **33**(16): p. 1831-1844.

20. Suzuki, S.i. and P.J. Henderson, *The hydantoin transport protein from Microbacterium liquefaciens.* Journal of bacteriology, 2006. **188**(9): p. 3329-3336.

21. Humphrey, W., A. Dalke, and K. Schulten, *VMD: visual molecular dynamics.* Journal of molecular graphics, 1996. **14**(1): p. 33-38.

22. Phillips, J.C., et al., *Scalable molecular dynamics with NAMD.* Journal of computational chemistry, 2005. **26**(16): p. 1781-1802.

23. Phillips, J.C., et al., *Scalable molecular dynamics on CPU and GPU architectures with NAMD.* The Journal of chemical physics, 2020. **153**(4): p. 044130.

24. MacKerell Jr, A.D., et al., *All-atom empirical potential for molecular modeling and dynamics studies of proteins.* The journal of physical chemistry B, 1998. **102**(18): p. 3586-3616.

25. Best, R.B., et al., *Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi1$ and $\chi2$ dihedral angles.* Journal of chemical theory and computation, 2012. **8**(9): p. 3257-3273.

26. MacKerell Jr, A.D., M. Feig, and C.L. Brooks, *Improved treatment of the protein backbone in empirical force fields.* Journal of the American Chemical Society, 2004. **126**(3): p. 698-699.

27. Klauda, J.B., et al., *Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types.* The journal of physical chemistry B, 2010. **114**(23): p. 7830-7843.

28. Jorgensen, W.L., et al., *Comparison of simple potential functions for simulating liquid water.* The Journal of chemical physics, 1983. **79**(2): p. 926-935.

29. Ryckaert, J.-P., G. Ciccotti, and H.J. Berendsen, *Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes.* Journal of computational physics, 1977. **23**(3): p. 327-341.

30. Miyamoto, S. and P.A. Kollman, *Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models.* Journal of computational chemistry, 1992. **13**(8): p. 952-962.

31. Darden, T., D. York, and L. Pedersen, *Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems.* The Journal of chemical physics, 1993. **98**(12): p. 10089-10092.

32. Feller, S.E., et al., *Constant pressure molecular dynamics simulation: the Langevin piston method.* The Journal of chemical physics, 1995. **103**(11): p. 4613-4621.

33. Fukunishi, H., O. Watanabe, and S. Takada, *On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction.* The Journal of chemical physics, 2002. **116**(20): p. 9058-9067.

34. Zhu, F. and G. Hummer, *Convergence and error estimation in free energy calculations using the weighted histogram analysis method.* Journal of computational chemistry, 2012. **33**(4): p. 453-465.

35. Kumar, S., et al., *The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method.* Journal of computational chemistry, 1992. **13**(8): p. 1011-1021.

36. Piana, S., K. Lindorff-Larsen, and D.E. Shaw, *How robust are protein folding simulations with respect to force field parameterization?* Biophysical journal, 2011. **100**(9): p. L47-L49.

37. Henriques, J., C. Cragnell, and M. Skepö, *Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment.* Journal of chemical theory and computation, 2015. **11**(7): p. 3420-3431.

38. Hazel, A.J., et al., *Folding free energy landscapes of beta-sheets with non-polarizable and polarizable CHARMM force fields.* J Chem Phys, 2018. **149**(7): p. 072317.

39. Misra, V.K. and D.E. Draper, *On the role of magnesium ions in RNA stability.* Biopolymers: Original Research on Biomolecules, 1998. **48**(2-3): p. 113-135.

40. Boudker, O., et al., *Coupling substrate and ion binding to extracellular gate of a sodium-dependent aspartate transporter.* Nature, 2007. **445**(7126): p. 387-393.

41. Khafizov, K., et al., *Investigation of the sodium-binding sites in the sodium-coupled betaine transporter BetP.* Proceedings of the National Academy of Sciences, 2012. **109**(44): p. E3035-E3044.

42. Chipot, C. and A. Pohorille, *Free energy calculations.* Springer series in chemical physics, 2007. **86**: p. 159-184.

43. Swope, W.C., J.W. Pitera, and F. Suits, *Describing protein folding kinetics by molecular dynamics simulations. 1. Theory.* The Journal of Physical Chemistry B, 2004. **108**(21): p. 6571-6581.

44. Rhee, Y.M. and V.S. Pande, *Multiplexed-replica exchange molecular dynamics method for protein folding simulation.* Biophysical journal, 2003. **84**(2): p. 775-786.

45. Immadisetty, K., J. Hettige, and M. Moradi, *Lipid-Dependent Alternating Access Mechanism of a Bacterial Multidrug ABC Exporter.* ACS Cent Sci, 2019. **5**(1): p. 43-56.

46. Zhu, F., *Calculating transition rates from durations of transition paths.* The Journal of chemical physics, 2017. **146**(12): p. 124128.

47. Zhou, X. and F. Zhu, *Calculating Single-Channel Permeability and Conductance from Transition Paths.* J Chem Inf Model, 2019. **59**(2): p. 777-785.

CHAPTER 4. SUMMARY AND CONCLUSION

The primary purpose of this thesis was to characterize the protein conformational changes between different stable states of proteins. The standard Molecular Dynamics simulation was implemented by US with HREMD to generate the trajectory between the conformational changes of protein states. The CHARMM force field [1] with the NAMD [2, 3] program was utilized in the MD simulations. We measured the free energy profile by using the WHAM equations [4].

4.1 Protein Folding

A well-defined reaction coordinate should precisely modify all the conformations of a molecular system. Many proteins may adopt multiple conformations that can be reversibly converted from one to another. Therefore, a "good" reaction coordinate allows us to obtain the dynamics along a chosen coordinate, but the projected dynamics will be affected with long-time memory for poor choices. Moreover, a well-defined reaction coordinate should distinguish all the possible intermediate states. It is essential since we would be able to measure several physical properties of our biological system. However, because we mainly deal with large-scale systems, it is impossible to cross the energetic barrier and sample all over the conformational space by an unbiased simulation. Even if we cross, the inadequate sampling at the top of the energetic regions may cause a high error. Therefore, enhanced sampling methods allow us to apply a biased potential to cross the energetic barriers and measure our system's physical properties.

Spontaneous transitions between native and non-native protein conformations usually are rare events that hardly occur in typical unbiased molecular dynamics simulations. It was demonstrated that the thermodynamics of protein folding could be well described by reaction coordinate: $Q(X) = \frac{1}{N}\sum \frac{1}{1+\exp[-\beta(r_{ij}-\lambda\, r_{ij}^0)]}$. This function represents the collective fraction of the native contacts between the protein atoms. The function's range is between two values, 0 and 1, and it can describe a two-state system. In 2013 Best et al. [5] based on microsecond unbiased simulations of small proteins, parameterized $\beta, \lambda, and\ r_{ij}^0$. We used this function for the reaction coordinate of the umbrella sampling simulation to investigate whether this function can measure the conformational states similar to Best et al. [5] except with an enhanced sampling approach. C++ programming code was generated to speed up the processing time, and because NAMD [2,

3] uses the TCL/TK environment, the SWIG (simplify wrapper and interface generator) package [6] was applied for TCL/TK to interface with C/C++.

We implemented US [7] with HREMD [8], using Q as the reaction coordinate to model Trp-Cage [9] and BBA [10] protein folding. The CHARMM force field [1] was used in molecular dynamics simulations. Our results from simulations showed a satisfactory convergence along with Q. Besides the native structure, multiple folded states can be observed in the reconstructed equilibrium ensemble (Figure 2-1). We tested protein Trp-cage simulation at three different temperatures (270 K, 280 K, and 290 K). $T = 290K$ was reported as the melting temperature, where both folded and unfolded states are equally populated at the equilibrium probability. We expected, along with the reaction coordinate in the US simulation by lowering the temperature, the population shifts from non-native state toward the native state and by increasing the temperature the populated states shifts from the native state to the non-native state. The cumulative distribution function (CDF) for Trp-Cage at the three temperatures shows that the equilibrium populations of the native and the non-native states are barely comparable, and manipulating the temperature does not shift the equilibrium probability toward the native or non-native states (Figure 2-2).

Our result indicated that even without native contacts, some protein structures are stabilized by hydrogen bonds not present in the native state. Overall, our result showed that although Q is a reasonably reliable parameter to analyze the simulations, it is not necessarily the best reaction coordinate for US simulation. In particular, the folding of the α-helix is a slow degree of freedom for Trp-Cage. The reaction coordinate may probably be improved by incorporating parameters that describe the α-helix conformation as well. Even though Q function can adequately describe our system's physical properties, the reaction coordinate is not entirely efficient in distinguishing between different intermediate states of our targeted proteins. For example, there are function domains that return multiple conformations. Technically, the equilibrium ensemble shows that the α-helix transition is almost orthogonal to the reaction coordinate Q.

4.2 Toward Convergence in Free Energy Calculation by Stepwise Model

As we mentioned in the previous section, our results indicated that the reaction coordinate Q characterizes several orthogonal conformations. To overcome the orthogonality behavior or hysteresis of a reaction coordinate, we developed another research to measure the transition from state A to state B with a better choice of the reaction coordinate. However, it is impossible to

define the entire transition of a large-scale biological system by only one reaction coordinate without having orthogonal conformations. Therefore, the hypothesis of the project was to implement a stepwise transition. In the way that we start from metastable A, then we have a transition to another intermediate state, m_1, and from the intermediate state m_1, we go to m_2, and so on. Therefore, by N transitions, we can get to the metastable state B. m_1, m_2, and m_{N-1} are the intermediate states between the two metastable states A and B. With this strategy, along with a favorable reaction coordinate, we can avoid orthogonal conformations. Because all the degrees of freedom have a sequential transition separately. Thus, we expect to get higher accuracy in our thermodynamics measurements.

The intermediate states can be obtained by measuring the structural differences of the two stable states A and B. For example, assume there are only two dihedral torsion angles, ψ_1 and ψ_2 significantly different at both stable states A and B. At state A, we have ψ_1^A and ψ_2^A, besides at state B, the torsion angles' microstates are ψ_1^B and ψ_2^B. As a result, these two torsion angles can be considered as the two degrees of freedom of the entire transition. We start from metastable state A with ψ_1^A and ψ_2^A. The ψ_1 transition involves the conformational change of the first torsion angle, resulting in reaching the intermediate state m with ψ_1^B and ψ_2^A. During the ψ_1 transition, we apply a boundary potential over ψ_2^A to keep this degree of freedom unchanged. Next, from the intermediate state m_1, there is a transition to metastable state B with ψ_1^B and ψ_2^B. However, this time, we apply boundary potential over ψ_1^B to keep this degree of freedom unchanged during the ψ_2 transition. The schematic of the transition can be shown as $A \overset{\psi_1}{\leftrightarrow} M \overset{\psi_2}{\leftrightarrow}$ B. As a case study, we used the transmembrane protein MHP1. The two stable states of this protein were selected as the Outward-Facing Open (OF) and Outward-Facing Occluded (OC) state with the crystal structure in the protein data bank (PDB) as 2JLN [11] and 4D1B [11, 12], respectively. For each step transition, the reaction coordinate was defined by a simple dihedral torsion angle, angle distance, and bond length.

We could obtain six transition steps with five intermediate states as M_1-M_5 that connect the two OF and OC stable states Figure 3-3. The detail of individual steps transition in the US can be found in Table 3-1Each step transition consists of one or multiple collective variables (CV) that are a function of the atomic coordinate of the protein. These CVs are a subset of degrees of freedom that describe the RC of each step transition. We also applied several restraints on some other degrees of freedom which in this thesis are known as "boundary restraints" to facilitate proper

sampling. Therefore, the boundary restraints mainly serve in the US to prevent the protein undergo undesired spontaneous transitions. Additionally, the boundary restraints used here do not affect the normal dynamics of the system. Also, three types of boundary potentials were used in the US simulation, explained in detail from Eq 3-7 to Eq 3-10.

We measured the free energy profile of each step transition by using umbrella sampling US along with the HREMD method at $T = 300K$. We performed two independent sampling simulations with different initial structures: the transition initiates from OF state ($OF \rightarrow OC$) indicated as the *InitOF* transition and the transition initiates from the OC state indicated as *InitOC* transition ($OF \leftarrow OC$). By comparing the two obtained free energy profiles with the stepwise model, we will then imply the extent of convergence in our calculations. By the defined reaction coordinates, we would be able to measure any physical properties of our targeted system, such as the gate's kinetics of MHP1 outward-facing conformations.

Figure 3-9 shows the free energy profiles as a function of the reaction coordinate for the overall transition between OF and OC states obtained from the US simulation trajectories. At each step transition, the free energy profile is calculated from the WHAM equations. Our results indicate that at intermediate states M_3 and M_4 compared to other states, the protein structures are stabilized by a more significant number of hydrogen bonds in protein MHP1. The energy difference between OF and OC states in our study is $\Delta G = -1.02 \pm 1.10$ kcal/mol and $\Delta G = -1.12 \pm 1.14$ kcal/mol for forward ($OF \rightarrow OC$) and backward ($OF \leftarrow OC$) transition, respectively. In addition to thermodynamics, kinetic quantities were measured based on the conformational changes. Based on the obtained free energies, we calculated the kinetic rates for the transitions of the Mhp1 thin gate. In addition to the end states OF and OC, the state M_3 has the lowest free energy among all the metastable states, so we calculate the transition rates between these three states. we obtained the effective transition rates between the OF and M_3 states: $k_{OF \rightarrow M3}$ ~ 1.7×10^{2} S^{-1}, and $k_{M3 \rightarrow OF}$ ~ 1.2 S^{-1}. With the effective transition rates between M_3 and OC states: of $k_{M_3 \rightarrow OC}$ ~ 8×10^{-3} S^{-1} and $k_{OC \rightarrow M_3}$ ~ 1.8×10^{-1} S^{-1}. The thermodynamics and the kinetics concerning the three major states OF, M3, and OC, are presented in Figure 3-11 3-11.

4.3 Future Research Direction

Compared to other similar studies, in our approach, the entire conformational space is not described by only one reaction coordinate, and the complete conformational change involves some sequential transition steps. In which the individual transition step connects two metastable states. Additionally, each step transition requires certain distinct degrees of freedom described by a unique reaction coordinate. This sequence in our research is as follows and is also represented by Figure 3-3.

$$OF \leftrightarrow M_1 \leftrightarrow M_2 \leftrightarrow M_3 \leftrightarrow M_4 \leftrightarrow M_5 \leftrightarrow OC$$

From OF to OC state, the sequence of the intermediate states (M_1-M_5) is one distinct selected pathway presented in our study. Any combination of intermediate states from M_1 to M_5 has the potential to be a possible transition pathway. Thus, a future research direction is to obtain the free energy profile of any possible transition pathways and kinetics rate measurements. If the energetic barrier is the lowest, then the related transition pathway would be the most favorable for the protein.

Even though the suggested future research needs extensive MD simulations, because the related degrees of freedom with RCs are already obtained in our study, it would be straightforward research. Moreover, we introduced boundary restraints perpendicular to each RC to prevent the protein from being trapped in unfavorable conformations or scape from a desirable conformation. For sure, for any combinations, a new set of boundary restraints is required to be determined.

The bacterial hydantoin transporter Mhp1 is a case study for our proposed methodology. Because of the high-resolution crystal structures available for all the stable states of protein Mhp1, this membrane transporter was selected as an excellent protein to study the alternating access model in atomic details. By now, between the two OF and OC states, no free energy difference experimentally has been reported in the literature. Therefore, another future research direction is to select a case study protein with available free energy differences from the experiment to test the introduced methodology more precisely.

4.4 References

1. Best, R.B., et al., *Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi1$ and $\chi2$ dihedral angles.* Journal of chemical theory and computation, 2012. **8**(9): p. 3257-3273.

2. Phillips, J.C., et al., *Scalable molecular dynamics with NAMD.* Journal of computational chemistry, 2005. **26**: p. 1781-1802.

3. Phillips, J.C., et al., *Scalable molecular dynamics on CPU and GPU architectures with NAMD.* The Journal of chemical physics, 2020. **153**(4): p. 044130.

4. Zhu, F. and G. Hummer, *Convergence and error estimation in free energy calculations using the weighted histogram analysis method.* Journal of computational chemistry, 2012. **33**(4): p. 453-465.

5. Best, R.B., G. Hummer, and W.A. Eaton, *Native contacts determine protein folding mechanisms in atomistic simulations.* Proceedings of the National Academy of Sciences, 2013. **110**(44): p. 17874-17879.

6. Beazley, D.M. *SWIG: An Easy to Use Tool for Integrating Scripting Languages with C and C++*. in *Tcl/Tk Workshop*. 1996.

7. Kästner, J., *Umbrella sampling.* Wiley Interdisciplinary Reviews: Computational Molecular Science, 2011. **1**(6): p. 932-942.

8. Fukunishi, H., O. Watanabe, and S. Takada, *On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction.* The Journal of chemical physics, 2002. **116**(20): p. 9058-9067.

9. Barua, B., et al., *The Trp-cage: optimizing the stability of a globular miniprotein.* Protein Engineering Design and Selection, 2008. **21**: p. 171-185.

10. Sarisky, C.A. and S.L. Mayo, *The ββα fold: explorations in sequence space.* Journal of molecular biology, 2001. **307**(5): p. 1411-1418.

11. Weyand, S., et al., *Structure and molecular mechanism of a nucleobase–cation–symport-1 family transporter.* Science, 2008. **322**(5902): p. 709-713.

12. Simmons, K.J., et al., *Molecular mechanism of ligand recognition by membrane transport protein, Mhp1.* The EMBO journal, 2014. **33**(16): p. 1831-1844.