The Energy Landscapes of Metamorphic Proteins

by

© *Bahman Seifi*

Supervisor

Dr. Stefan Wallin

Committee members

Dr. Ivan Saika-Voivod

Dr. Anand Yethiraj

A thesis submitted to the School of Graduate Studies

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Physics and Physical Oceanography

Memorial University of Newfoundland

March 2023

St. John's Newfoundland

Abstract

Most proteins fold into a unique three-dimensional structure called the native state. Recently some examples have been found of so-called metamorphic proteins that undergo reversible large-scale structural transformations between different native states. In this thesis, we develop simulation methods and models to study the thermodynamics of these transformations, both at the coarse-grained and all-atom levels. Because our understanding of the physics fold switching is incomplete, our models utilize in part so-called structure-based or Gō-like potentials, which provide energetic bias towards one, or more, native states. We employ these computational methods to two different fold switch systems: the bacterial protein RfaH and the engineered fold switch system GA/GB. Our models are developed and tested on experimental data for these systems. We study both equilibrium properties, such as stability properties and the characteristics of their energy landscapes, and kinetic properties, such as the mechanism that trigger fold switching and molecular details of the fold switch process. We also study, for the GA/GB system, what role macromolecular crowding effects play for controlling which of the native states is most stable.

General Summary

Proteins are complex molecules that are made up of chains of amino acids and fold to a specific three-dimensional shape called the native state. It is the structure of the protein that determines its function. If not folded properly, proteins may aggregate and cause disease. However, certain proteins can adopt multiple native conformations, and these are known as metamorphic proteins. Metamorphic proteins can switch between different folds in response to particular changes in the environmental conditions. They may be useful as biosensors because of their structural flexibility. Biosensors use biological molecules to identify and measure specific chemicals or biological processes. The first metamorphic protein called Lymphotactin (Ltn) was discovered in 2008, so this field of study is relatively new, but rapidly growing. Biophysical techniques such as nuclear magnetic resonance (NMR), X-ray crystallography, and computational simulations are used to investigate the structural and kinetic properties of these proteins. A thorough understanding of these techniques is necessary to comprehend the molecular mechanisms that govern the behavior of metamorphic proteins. As technology and techniques continue to advance, we can expect further progress in this fascinating field.

Co-authorship Statement

Chapter 3 is based on the following paper:

Bahman Seifi, Adekunle Aina, Stefan Wallin, “Structural fluctuations and mechanical stabilities of the metamorphic protein RfaH”, *Proteins*, 189:289-300, 2021.

I did all simulations and analyses regarding the structural fluctuations (Figs 2-4) in RfaH protein. Dr. Stefan Wallin supervised, guided and contributed to writing the paper.

Chapter 4 is based on the following paper:

Bahman Seifi, and Stefan Wallin, “The C-terminal domain of transcription factor RfaH: Folding, fold switching and energy landscape”, *Biopolymers*, 112:e23420, 2021.

Dr. Stefan Wallin and I developed the hybrid model and designed and performed the computer simulations. I analyzed all the simulations. Dr. Stefan Wallin and I prepared the figures and wrote the manuscript.

Chapter 5 is based on the following preprint paper:

Bahman Seifi, and Stefan Wallin, “Examining the effect of the N-terminal domain of RfaH on domain dissociation and fold switching”. Dr. Stefan Wallin and I designed and performed the computer simulations. I analyzed all the simulations. Dr. Stefan Wallin and I prepared the figures and wrote the manuscript.

Chapter 6 is based on the following preprint paper:

“Simulations of a protein fold switch reveal crowding-induced population shifts driven by disordered regions”. In this project, Saman Bazmi contributed with expertise on macromolecular crowding and I contributed with expertise on fold switching. Overall, Saman Bazmi and I contributed equally. Dr. Stefan Wallin guided and supervised this project and contributed to developing the C_α model. Saman Bazmi and I analyzed all the simulations and prepared all the figures. Dr. Stefan Wallin, Saman Bazmi and I wrote the manuscript.

Acknowledgements

I would like to thank my supervisor Dr. Stefan Wallin, for his invaluable guidance, unwavering support, and profound expertise throughout my Ph.D. program. His mentorship has significantly influenced my academic growth and the quality of my research. I am truly grateful for the opportunity to work under his supervision.

I would also like to thank my committee members, Dr. Ivan Saika-Voivod and Dr. Anand Yethiraj, for their support and helpful comments. Additionally, I extend my gratitude to Saman Bazmi, a member of the computational biophysics lab and my collaborator on the ”Fold-Switching Proteins Under Crowded Conditions” project.

I am thankful to the Memorial University of Newfoundland for the financial support which enabled me to concentrate on my research and writing. I also appreciate everyone at the Department of Physics and Physical Oceanography for creating a pleasant working environment.

I want to express my appreciation to my wife, Nasim, for her cooperation and understanding during the writing of this dissertation. Her love, presence and support have been a constant source of strength and motivation.

I would like to acknowledge and deeply appreciate my father and my family for their continuous support, belief in my abilities, and guidance.

Lastly, I want to take a moment to honor the memory of my late mother. Her love, sacrifices, and unwavering faith in my potential have been a lasting source of inspiration. Although she is no longer with us, her spirit lives on, and I carry her love and wisdom with me every step of the way.

Contents

List of Figures

Chapter 1

Introduction

Gerrit J. Mulder was a Dutch chemist interested in analyzing the chemical composition of cellular material. In the 1830s, he identified a substance composed mainly of carbon, nitrogen, oxygen, and hydrogen. He called this new substance "protein" [1,2]. As we now know, proteins are synthesized in the cell as linear polymer chain molecules composed of amino acids or residues. The amino acids in a protein are attached together by peptide bonds, a strong covalent interaction. For this reason, protein chains are also called polypeptide chains. In a single protein, there are typically more than about 50 amino acids that are covalently bonded to each other, but they can reach thousands of amino acids.

The three-dimensional structure of a protein is determined by the interactions between its amino acids and the solvent environment. In the process of folding, the protein chain folds into its so-called native structure through a series of intermediate conformations. The energy landscape theory of protein folding describes how this process can be completed in a biologically relevant timescale despite the astronomically large number of possible conformations [3, 4]. This process is driven by various interactions between amino acid residues, as well as hydrophobic and hydrophilic interactions between amino acids and the solvent environment. Hydrophobic interactions refer to the tendency of non-polar amino acids to aggregate together, away from water molecules. In contrast, hydrophilic interactions describe the interaction between polar or charged amino acids and water molecules. Together, these two types of interactions play important roles in protein folding, stability, and function. The native structure of a protein is critical for its function, as it determines the location and orientation of functional groups, and determines the protein's ability to bind to other molecules, such as ligands or substrates.

Several exceptions to the above classic picture of proteins have been found over the past decades. One such exception is intrinsically disordered proteins, which are a type of protein that lack a native structure, yet play important roles in biological processes due to their flexibility and ability to interact with a variety of binding partners [5–7]. Another more recent exception is metamorphic proteins. Metamorphic proteins are a unique class of proteins that exhibit the ability to switch between different folds and functions in response to specific stimuli. This dynamic behavior would make them ideal biosensors, as they can modulate their activity in response to changes in the cellular environment [8]. By using computational simulations to study the energy landscape of metamorphic proteins, researchers can gain a deeper understanding of the molecular mechanisms that drive fold switching and how it regulates protein function. This information can help us understand the role of metamorphic proteins in controlling specific activities within cells and how they contribute to biological processes and diseases. Additionally, this study can also advance our understanding of the basic principles of protein folding and function and may lead to the development of new therapies for a range of medical conditions.

Computational simulations are helpful and necessary in this work for several reasons. Firstly, they enable high-throughput atomistically detailed analysis that would be impractical or impossible using experimental methods alone. Secondly, they provide deep insights into the molecular mechanisms of fold switching and how it regulates protein function. Thirdly, they can make predictions about the behavior of metamorphic proteins under different conditions, allowing researchers to test these predictions experimentally. Fourthly, they are generally cost-effective and time-saving compared to experimental methods. Lastly, they complement and extend the insights gained from experimental studies, leading to a more complete understanding of the systems under study.

While AlphaFold2 has significantly improved protein structure prediction [9, 10], it struggles with fold-switching proteins [11], which have multiple stable conformations. The methodology of AlphaFold2 relies heavily on pattern recognition rather than biophysics. As a result, when applied to metamorphic sequences, the predicted structures are biased towards one conformation and fail to capture the other native states. Even in the case that AlphaFold2 could predict both structures, it does not provide insights into the molecular mechanism of fold switching. Understanding how sequence encodes structure is tackled through computer modelling, which highlights the complexity of protein structure and stability, dependent on various interdependent factors. A combination of structure analyses, molecular dynamics simulations, and a novel method for calculating free energy differences is used to explore the reasons for different structural preferences [12].

The work in this thesis addresses the challenge of understanding how the amino acid sequence encodes structure through computer modeling, allowing for the characterization of the native structure of each sequence and how they are modeled onto the fold of the alternate sequence. We use a combination of structure-based model (SBM) with an all-atom model to explore the reasons for the different structural preferences of metamorphic sequences. The results highlight that protein structure and stability are complex and depend on various interdependent factors, with different levels of analysis providing different predictions about the favored sequence-structure combination [13].

1.1 Examples of fold switching proteins

Because the initial and the final state in the fold switching process are folded states, the protein has a major shift in secondary structure and biological functionality.

There are only a few metamorphic proteins that have been studied experimentally in detail so far, and it is still unclear how the fold-switching process happens at the molecular level [14–17]. Fig. (1.1) shows four examples of metamorphic proteins and their structural transformations. The work in this thesis is focused on the metamorphic protein RfaH, and G_A/G_B fold switch system.

The structures of protein sequences at the interface between different folds are

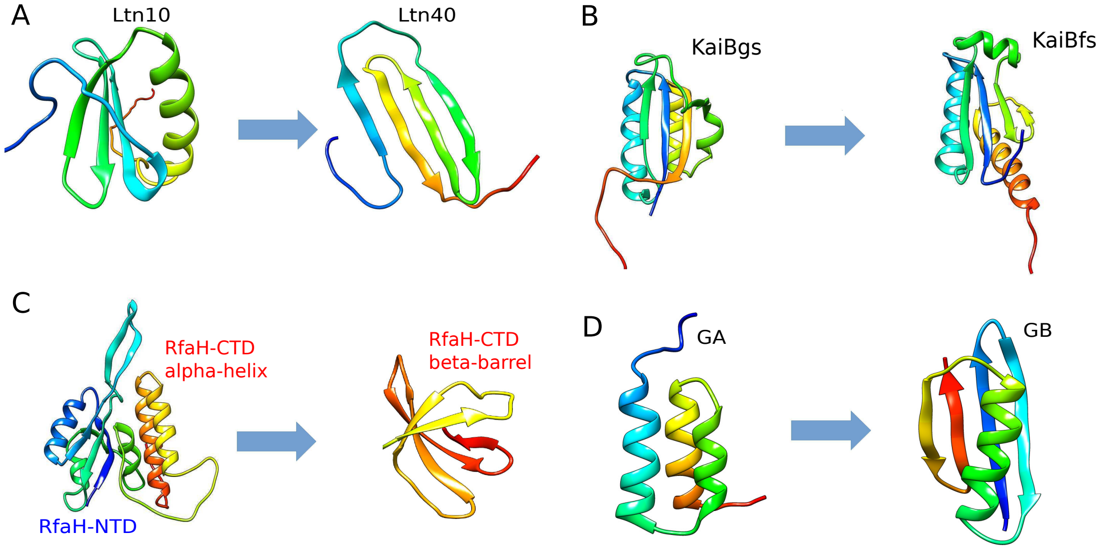

Figure 1.1: Examples of metamorphic proteins; (A) Lymphotactin (Ltn) is a signalling protein important for the immune system. It fold switches upon dimerization; Ltn10 in its α/β fold PDB ID 1J8I and Ltn40 in its all-β dimeric fold PDB ID 2JP1 [18]. (B) The KaiB protein is one of the three proteins in the circadian clock in cyanobacteria. A 24-hour cycle emerges from their interactions. KaiB has two different native states: KaiBgs (a tetramer of asymmetric dimers) and KaiBfs (a monomer). During the day, KaiB is in the KaiBgs fold and switches to the KaiBfs fold at night [16]. KaiBgs PDB ID 2QKE and Kaibfs PDB ID 5JYT. (C) The RfaH protein with two native states and 162 amino acids. RfaH consists of one NTD in a mixed α/β conformation, which remains structurally constant in the two native states, and a CTD with two completely different folds. When the CTD is spatially close to the NTD it is folded into a α-helical hairpin (PDB ID: 5ond). When the CTD is isolated, i.e., spatially far from NTD, it folds into a 5-stranded β-barrel CTD PDB ID: 2lcl). (D)Protein G contains two binding domains, each consisting of 56 amino acids. These domains are engineered fold-switching proteins and differ from each other in only three amino acids. GA95 (PDB ID: 2kdl) and GB95 (PDB ID: 2ldm) [19].

challenging to predict using current computational methods, revealing an incomplete knowledge about protein folding physics. Studies on metamorphic proteins have shown that a limited number of key residues and interactions can result in a change from one fold to another. Mutations can sometimes weaken the original fold and promote new folds and functions, which may account for why alternative folds sometimes have lower predicted energy values in structure prediction algorithms. This adds to the complexity of protein folding [20]. Improved prediction algorithms could result from closer collaboration between computational biophysicists and experimentalists in the field of fold switching, as data on the mutational paths between different folds could be used to refine these algorithms.

1.1.1 RfaH protein

RfaH is a metamorphic protein found in some bacteria, specifically in several *Proteobacteria*, including several pathogenic Gram-negative bacteria in *Escherichia coli* [21]. The protein has two structural domains, an N-terminal domain (NTD) and a C-terminal domain (CTD) [22]. The conformational changes of this protein have been extensively examined experimentally [23–26] and computationally [27–34]. The NTD is stable, but the CTD undergoes a dramatic fold switch when RfaH binds to the RNA polymerase (RNAP) that is paused at an operon. During this switch, the two domains separate and the CTD transforms from a helical hairpin fold to a 5-stranded beta-barrel structure, which is similar to the CTDs of other proteins known as NusG proteins [23, 24, 35–37]. The flexible linker connecting the domains enables the transformed CTD to recruit ribosomal protein S10, forming a physical bridge between RNAP and ribosome. This combination of fold switching and domain separation allows RfaH to regulate transcription, the production of RNA from DNA, and enhance translation, the process by which RNA is used to make protein [25,36]. The transformation of RfaH's CTD into a beta-barrel structure through fold switching is considered a clear example of a fold-switching protein (see Fig. (1.1.C)).

1.1.2 $\mathbf{G_A/G_B}$ fold switch system

The G_A/G_B system is a designed fold-switching protein system where a small number of mutations in the protein sequence can lead to the protein adopting a completely different fold (see Fig (1.1.D)). This fold switching can occur due to the thermodynamic linkage between folding and binding, where the interaction energy of binding can stabilize conformations that would otherwise be overwhelmed by the standard conformation [19]. Understanding this system is important for understanding the evolution of new protein structures and functions and may lead to new approaches for interpreting genetic polymorphisms and other disease-related events [38]. Additionally, the design of globular protein switches that can flip between folds and functions may have applications in the development of more specific targeted drug-delivery systems and the design of tunable nano-scale devices [19].

1.2 Macromolecular crowding

Proteins carry out their functions within cells, which are highly crowded spaces. Macromolecules such as small proteins, carbohydrates, ribosomes, and nucleic acids occupy between 10 to 40 percent of the total volume inside cells, and are often referred to as crowders [39]. In typical experiments and computer simulations, proteins are studied under dilute conditions despite their environment being highly crowded.

Crowding effects within the cell can have significant impacts on various protein processes. For instance, the conformational preferences of intrinsically disordered proteins (IDPs) can be influenced by crowders [40]. Additionally, crowders can affect the stability of proteins. One type of crowding effect is the excluded-volume effect, which is always present and can impact protein folding [41]. In fact, the excluded volume effect due to crowders is expected to stabilize the folded state of proteins because crowder particles have more entropy when the protein is in the folded state than in the unfolded state [42]. However, there have been no published studies on how crowding affects fold switching. Work that is part of this thesis closes this gap in the literature [43].

1.3 Computational methods

Biochemistry experiments carried out on proteins often do not provide direct information on microscopic details. Rather, experimental signals are values averaged over many molecules in bulk solutions. Computer simulations, which in principle provide complete insight into the molecular details of a system, are therefore highly complementary to experiments. However, it is essential that simulations can also provide thermodynamics averages under different conditions in order to verify the underlying model in comparison with the experimental data.

Monte Carlo (MC) and molecular dynamics (MD) are the two most common simulation techniques for simulating molecules on the computer.

Under conditions of constant temperature T, volume, and number of molecules, the probability $p_{\mathrm{B}}(i)$ of finding the system in a microscopic state i is given by the canonical or Boltzmann distribution;

$$p_{\mathrm{B}}(i) = \frac{e^{-\beta E_i}}{\mathcal{Z}}, \tag{1.1}$$

where E_i is the energy of the system in microstate i, $\beta = 1/k_{\mathrm{B}}T$, and $\mathcal{Z} = \sum_i e^{-\beta E_i}$ is the partition function that describes the statistical properties of the system in thermodynamic equilibrium. The thermodynamic average of a physical quantity Q is given by

$$\langle Q \rangle = \frac{\sum_i Q_i e^{-\beta E_i}}{\sum_i e^{-\beta E_i}} = \sum_i p_{\mathrm{B}}(i) Q_i. \tag{1.2}$$

For complex systems, identifying all the states is not easily accessible, and it is limited just to small systems. For large systems, we have to use only a countable set of conformations representing the whole conformation space. If the states are selected with a probability given by Eq.(1.1) the estimate for the thermal average becomes:

$$\langle Q \rangle_{\text{estimate}} = \frac{\sum_i^M Q_i}{M}, \tag{1.3}$$

where M is the number of states and Q_i is the value of Q at the ith generated state [44, 45].

1.3.1 Molecular dynamics

The basic idea of molecular dynamics simulations is to numerically integrate the equations of motion of the particles of a system, and thereby determine the time evolution of the system in microscopic detail. If the system is followed for a sufficiently long time, an MD simulation can be used to find thermodynamic averages of any observable.

Consider a system with N particles with positions $\vec{r}_1, \vec{r}_2, ..., \vec{r}_N$, and a potential energy function described by $E(\vec{r}_1, \vec{r}_2, ..., \vec{r}_N)$. For example, the N particles could be the atoms of a protein molecule. In the absence of chemical reactions, protein motions can be obtained by the numerically solving the classical equations of motion for this many-body system, i.e.,

$$\begin{aligned} \frac{d\vec{r}_i}{dt} &= \frac{\vec{p}_i}{m_i}, \\ \frac{d\vec{p}_i}{dt} &= \vec{F}_i, \end{aligned} \tag{1.4}$$

where $\vec{p}_i$ and m_i are the momentum and mass of atom i, respectively, and $\vec{F}_i = -\nabla_{\vec{r}_i} E(\vec{r}_1, \vec{r}_2, ..., \vec{r}_N)$ is the conformational force on atom i. According to Newton's equation of motion (1.4), it is possible to simulate the evolution of the particles of a system given its initial configuration and an accurate description of the interaction forces. Because of the fast vibrational motion of particles, the time step δt of integration must be small (1-2 fs) and thermal equilibrium is hard to achieve. In the next section, we describe the Monte Carlo simulation method, which can overcome this problem under some circumstances.

1.3.2 Markov chain Monte Carlo

The Monte Carlo (MC) method is a collection of computational techniques that use random numbers to obtain approximate solutions of mathematical problems such as integration and optimization. Markov chain MC simulation is a technique that generates states from a general probability distribution. Hence, it can be employed to find an ensemble of states consistent with a specific thermodynamic condition. States are generated by applying random perturbations to the system called updates or moves, which are either accepted or rejected based on a specific criterion. For an accurate estimate of thermodynamic averages, a proper sampling of phase space is required. An advantage of MC is that it allows large moves, which helps sampling. But this means time evolution is not provided, as MD.

1.3.3 Metropolis algorithm

The basic idea of the Metropolis algorithm is to generate a sequence of states that are biased according to the Boltzmann distribution $p_{\mathrm{B}}(i)$.

The fundamental element in the MC method is the concept of statistical equilibrium, which is expressed as a detailed balance condition. According to the detailed balance condition, for a statistical system in equilibrium, the transition rate between any two physical micro states should be equal. It can be shown that if detailed balance is fulfilled, statistical equilibrium is obtained. If $W(i \to j)$ is the transition probability from state i to j, the detailed balance condition can be written as:

$$p(i)W(i \to j) = p(j)W(j \to i). \tag{1.5}$$

For the Boltzmann distribution, $p(i) = p_{\mathrm{B}}(i)$ (see Eq 1.1) and the detail balance condition implies;

$$\frac{W(i \to j)}{W(j \to i)} = \frac{p(j)}{p(i)} = e^{-\beta \Delta E_{ij}}, \tag{1.6}$$

where $\Delta E_{ij} = E_j - E_i$. Eq. (1.6) suggests an algorithm for generating states i that are biased according to the Boltzmann distribution. It works as follows:

1. Consider the system in an initial state i with energy E_i.
2. Attempt a trial move ($i \to j$); the new state j has energy E_j.
3. Accept or reject the trial state j with the probability

$$P_{\text{acc}} = \begin{cases} e^{-\beta\Delta E_{ij}} & \text{if } \Delta E_{ij} \text{ is positive} \\ 1 & \text{otherwise} \end{cases} . \tag{1.7}$$

4. Repeat from 2.

It can be proven that this algorithm satisfies detailed balance and hence it leads to $p(i) \to p_{\mathrm{B}}(i)$ in the limit of many generated states. Due to detailed balance condition, the transition from one conformation to another in phase space is possible only if the inverse transition is also possible. The order of moves in simulation does not influence the canonical equilibrium as long as the condition of detailed balance is satisfied [46].

1.4 Interactions in biomolecular systems

There are various forces and effects that drive the dynamics of biomolecular processes, such as protein folding, including hydrogen bonds, van der Waals interactions, electrostatics, e.g. ionic bonds (or salt bridges), and effective hydrophobic interactions. Although much of the physics of these various interactions are well understood, research into how they can be best described within (classical) explicit-water molecular dynamics force-fields is ongoing [47].

Hydrogen bonding is an interaction between a donor group DH, which is a hydrogen H attached to an electronegative atom D, and an electronegative acceptor atom A. The interaction is well described by the electrostatics between the positively partially charged H and the negatively partially charged D and A. However, a complete description requires quantum mechanics [48]. An example in proteins is α-helical structures, which are stabilized by hydrogen bonds between a carbonyl group $\mathrm{C{=}O}$ in residue i (acceptor) and an amide group $\mathrm{N{-}H}$ (donor) in residue $i+4$ on the protein backbone. Arranging D, H, and A along a straight line is energetically favorable. Consequently, the α-helical axis is relatively straight, and all donors and acceptors participate in a hydrogen bond (except at the ends of a helix). Also, β-sheets are stabilized by backbone-backbone hydrogen bonds [49].

The hydrophobic effect refers to the tendency of non-polar groups to cluster together in an aqueous environment [50, 51]. While non-polar groups can interact favorably via van der Waals forces [52], this force is quite weak and is generally not the reason for their clustering. Rather, hydrophobic attractions are driven by entropic effects involving the water. Because the water molecules are more ordered near the hydrophobic surface than in the bulk, decreasing the hydrophobic surface exposed to water increases the multiplicity of water molecules. This consequently increases the entropy and decreases the free energy of the mixture. [51].

Proteins typically have both hydrophobic (non-polar) and polar amino acids. For a soluble protein in an aqueous environment such as the inside of a cell, the side chains of non-polar amino acids are usually found inside the protein structure to form a hydrophobic core and polar amino acids are mostly found on the surface [47, 53, 54].

1.5 Outline

In this thesis, we develop a Monte Carlo based simulation method for proteins and apply it to fold switching. In chapter 2, we review the various computational methods that are used in this thesis. In chapter 3, we apply the original PROFASI (Protein Folding and Aggregation Simulator) model, an MC simulation package to characterize the size of structural fluctuations of the two different states of RfaH [34]. In chapter 4, we then describe a new model based on the PROFASI package and augment it with a dual-basin structure-based potential energy term to study the fold switching of CTD of RfaH [55]. In chapter 5, we employ the method developed in chapter 3 to study the mechanism of domain separation RfaH protein [56]. In chapter 6, we developed a coarse-grained C_α model with a dual-basin SBM to study the excluded volume effect arising from macromolecular crowder particles. We apply this model to study the fold switching in the GA95-GB95 proteins [43].

The methods that are developed in this thesis are applied to the RfaH (in sections 3,4 and 5) and GA95-GB95 (in section 6) proteins, but the approach we take is general in that the model can, in principle, be applied to any fold switching protein.

Bibliography

[1] C. Tanford, and J. Reynolds. Nature's robots: a history of proteins. *OUP Oxford*, 2003.

[2] F. H. Portugal, and J. S. Cohen. A Century of DNA: A history of the discovery of the structure and function of the genetic substance. *Mit Press*, 1977.

[3] J. N. Onuchic and P. G. Wolynes. Theory of protein folding. *Curr Opin Struct Biol*, 14:70–5, 2004.

[4] J. Karanicolas and C. L. Brooks. Improved Gō-like models demonstrate the robustness of protein folding mechanisms towards non-native interactions. *J Mol Biol*, 334:309–325, 2003.

[5] P. Tompa, A. Fersht. Structure and function of intrinsically disordered proteins. *CRC press*, 2009.

[6] A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, and et al. Intrinsically disordered protein. *J Mol Graphics Modell*, 19:26–59, 2001.

[7] P. E. Wright, and H. D. Dyson. Linking folding and binding. *Curr Opin Struct Biol*, 19:31–38, 2009.

[8] P. N. Bryan and J. Orban. Proteins that switch folds. *Curr Opin Struct Biol*, 20:482–488, 2010.

[9] J. Jumper, R. Evans, A. Pritzel, and et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589, 2021.

[10] A. N. Lupas, J. Pereira, V. Alva, F. Merino, M. Coles, M. D. Hartmann. The breakthrough in protein structure prediction. *Biochem*, 478:1885–1890, 2021.

[11] D. Chakravarty, and L. L. Porter. AlphaFold2 fails to predict protein fold switching. *Protein Sci*, 31:e4353, 2022.

[12] D. Chakravarty, L. L. Porter. AlphaFold2 fails to predict protein fold switching. *Protein Sci*, 31:e4353, 2022.

[13] J. R. Allison, M. Bergeler, N. Hansen, and W. N. van Gunsteren. Current computer modeling cannot explain why two highly similar sequences fold into different structures. *Biochemistry*, 50:10965–10973, 2011.

[14] R. L. Tuinstra, F. C. Peterson, S. Kutlesa, E. S. Elgin, M. A. Kron, and B. F. Volkman. Interconversion between two unrelated protein folds in the lymphotactin native state. *Proc Natl Acad Sci USA*, 105:5057–5062, 2008.

[15] J. L. Markley, J. H. Kim, Z. Dai, and et al. Metamorphic protein IscU alternates conformations in the course of its role as the scaffold protein for iron-sulfur cluster biosynthesis and delivery. *FEBS Lett*, 587:1172–1179, 2013.

[16] Y. G. Chang, S. E. Cohen, C. Phong, and et al. A protein fold switch joins the circadian oscillator to clock output in cyanobacteria. *Science*, 349:324–328, 2015.

[17] X. Luo, and H. Yu. Protein metamorphosis: the two-state behavior of Mad2. *Structure*, 16:1616–1625, 2008.

[18] A. G. Murzin. Metamorphic proteins. *Science*, 320:1725–1726, 2008.

[19] P. A. Alexander, Y. He, Y. Chen, J. Orban, and P. N. Bryan. A minimal sequence code for switching protein structure and function. *Proc Natl Acad Sci USA*, 106:21149–21154, 2009.

[20] K. Madhurima,and B. Nandi, and A. Sekhar. Metamorphic proteins: the Janus proteins of structural biology. *Open Biol*, 11:210012, 2021.

[21] B. Wang, V. M. Gumerov, E. P. Andrianova, I. B. Zhulin, and I. Artsimovitch. 2020. Origins and molecular evolution of the NusG paralog RfaH. *MBio*, 11:e02717-20, 2020.

[22] B. M. Burmann, S. H. Knauer, A. Sevostyanova, K. Schweimer, R. A. Mooney, R. Landick, I. Artsimovitch, and P. Rösch. An α helix to β barrel domain switch transforms the transcription factor RfaH into a translation factor. *Cell*, 150:291–303, 2012.

[23] G. A. Belogurov, M. N. Vassylyeva, V. Svetlov, S. Klyuyev, N. V. Grishin, D. G. Vassylyev, and I. Artsimovitch. Structural basis for converting a general transcription factor into an operon-specific virulence regulator. *Mol Cell*, 26:117–129, 2007.

[24] D. Shi, D. Svetlov, R. Abagyan, and I. Artsimovitch. Flipping states: a few key residues decide the winning conformation of the only universally conserved transcription factor. *Nucleic Acids Res*, 45:8835–8843, 2017.

[25] P. K. Zuber, K. Schweimer, P. Rösch, I. Artsimovitch, and S. H. Knauer. Reversible fold-switching controls the functional cycle of the antitermination factor RfaH. *Nat Commun*, 10:702, 2019.

[26] P. Galaz-Davison, J. A. Molina, S. Silletti, E. A. Komives, S. H. Knauer, I. Artsimovitch, and C. A. Ramírez-Sarmiento. Differential local stability governs the metamorphic fold switch of bacterial virulence factor RfaH. *Biophys J*, 118:96–104, 2020.

[27] S. Li, B. Xiong, Y. Xu, T. Lu, X. Luo, C. Luo, J. Shen, K. Chen, M. Zheng, and H. Jiang. Mechanism of the all-α to all-β conformational transition of RfaH-CTD: Molecular dynamics simulation and Markov State model. *J Chem Theory Comput*, 10:2255–2264, 2014.

[28] C. A. Ramirez-Sarmiento, J. K. Noel, S. L. Valenzuela, and I. Artsimovitch. Interdomain contacts control native state switching of RfaH on a dual-funneled landscape. *PLOS Comput Biol*, 11:e1004379, 2015.

[29] J. B. GC, Y. R. Bhandari, B. S. Gerstman, and P. P. Chapagain. Molecular dynamics investigations of the α-helix to β-barrel conformational transformation in the RfaH transcription factor. *J Phys Chem B*, 118:5101–5108, 2014.

[30] J. B. GC, B. S. Gerstman, and P. P. Chapagain. The role of the interdomain interactions on RfaH dynamics and conformational transformation. *J Phys Chem B*, 119:12750–12759, 2015.

[31] L. Xiong and Z. Liu. Molecular dynamics study on folding and allostery in RfaH. *Proteins*, 83:1582–1592, 2015.

[32] S. Xun, F. Jiang, and Y. D. Wu. Intrinsically disordered regions stabilize the helical form of the C-terminal domain of RfaH: A molecular dynamics study. *Bioorg Med Chem*, 24:4970–4977, 2016.

[33] J. A. Joseph, D. Chakraborty, and D. J. Wales. Energy landscape for fold-switching in regulatory protein RfaH. *J Chem Theory Comput*, 15:731–742, 2019.

[34] B. Seifi and A. Aina and S. Wallin. Structural fluctuations and mechanical stabilities of the metamorphic protein RfaH, *Proteins*, 89:289–300, 2021.

[35] J. Y. Kang, R. A. Mooney, Y. Nedialkov, J. Saba, T. V. Mishanina, I. Artsimovitch, R. Landick, and S. A. Darst. Structural basis for transcript elongation control by NusG family universal regulators. *Cell*, 173:1650–1662, 2018.

[36] P. K. Zuber, I. Artsimovitch, M. NandyMazumdar, Z. Liu, Y. Nedialkov, K. Schweimer, P. Rösch, and S. H. Knauer. The universally-conserved transcription factor RfaH is recruited to a hairpin structure of the non-template DNA strand. *Elife*, 7:e36349, 2018.

[37] B. Wang, V. M. Gumerov, E. P. Andrianova, I. B. Zhulin, and I. Artsimovitch. Origins and Molecular Evolution of the NusG Paralog RfaH. *mBio*, 11:e02717–20, 2020.

[38] T. L. Solomon, Y. He, N. Sari, Y. Chen, D. T. Gallagher, P. N. Bryan, and J. Orban. Reversible switching between two common protein folds in a designed system using only temperature. *Proc Natl Acad Sci USA*, 120:e2215418120, 2023.

[39] S. B. Zimmerman, S. O. Trach. Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of Escherichia coli. *J Mol Biol*, 222:599–620, 1991.

[40] V. Nguemaha, S. Qin, and H. X. Zhou. Atomistic modeling of intrinsically disordered proteins under polyethylene glycol crowding: quantitative comparison with experimental data and implication of protein–crowder attraction. *J Phys Chem B*, 122:11262–11270, 2028.

[41] J. Mettle, and R. Best. Dependence of protein folding stability and dynamics on the density and composition of macromolecular crowders. *Biophys J*, 98:315–320, 2010.

[42] A. Christiansen, Q. Wang, M. S. Cheung, P. Wittung-Stafshede. Effects of macromolecular crowding agents on protein folding in vitro and in silico. *Biophys Rev*, 5:137–45, 2013.

[43] S. Bazmi, B. Seifi, S. Wallin. Simulations of a protein fold switch reveal crowding-induced population shifts driven by disordered regions. *preprint*, 2023.

[44] V. Muñoz, and W. A. Eaton. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc Natl Acad Sci USA*, 96:11311–11316, 1999.

[45] A. Kolinski, M. Milik, and J. Skolnick. Static and dynamic properties of a new lattice model of polypeptide chains. *J Chem Phys*, 94:3978–3985, 1991.

[46] J. Skolnick, and A. Kolinski. Dynamic Monte Carlo simulations of a new lattice model of globular protein folding, structure and dynamics. *J Mol Biol*, 221:499–

531, 1991.

[47] B. M. Jackson, and others. Molecular and cellular biophysics. *Cambridge University Press*, 2006.

[48] P. Atkins, P. W. Atkins, and J. de Paula. Atkins' physical chemistry. *Oxford university press*, 2014.

[49] W. Kauzmann. Some Factors in the Interpretation of Protein Denaturation. *Adv Protein Chem Struct Biol*, 14:1–63, 1959.

[50] C. Tanford. The hydrophobic effect: formation of micelles and biological membranes 2d ed. *J. Wiley.* 1980.

[51] R. S. Spolar, and J. H. Ha, and M. T. Record. Hydrophobic effect in protein folding and other noncovalent processes involving proteins. *Proc. Natl. Acad. Sci. U.S.A.*, 86:8382–8385, 1989.

[52] C. M. Roth, B. L. Neal, and A. M. Lenhoff. Van der Waals interactions involving proteins. *Biophys. J.*, 70:977–987, 1996.

[53] A. K. Dill. Dominant forces in protein folding. *Biochemistry*, 29:7133–7155, 1990.

[54] B. H. Honig, W. L. Hubbell,and R. F. Flewelling. Electrostatic interactions in membranes and proteins. *Annu. Rev. Biophys. Biophys. Chem.*, 15:163–193, 1986.

[55] B. Seifi and S. Wallin. The C-terminal domain of transcription factor RfaH: Folding, fold switching and energy landscape. *Biopolymers*, 112:e23420, 2021.

[56] B. Seifi, and S. Wallin, Examining the effect of the N-terminal domain of RfaH on domain dissociation and fold switching *Preprint*, 2023.

Chapter 2

Methods

In this chapter, we introduce and provide details on some of the methods and models used in the later chapters in the thesis.

2.1 All-atom physics based protein model

Two computational methods, Markov chain Monte Carlo (MCMC) and molecular dynamics (MD), dominate the field of classic molecular simulation. The significant distinction between these methods lies in how the system is updated in each iteration of the system. MCMC is a statistical mechanics technique aiming to generate samples of states associated with some target probability distribution, e.g. the Boltzmann distribution. In contrast, the MD method involves iterating between calculating the forces applied on each particle in the system and using Newton's equations of motion to update their positions and momenta. MD has typically been considered as best-suited for studying dense molecular systems such as the native ensemble of protein system, while MCMC methods are more suitable for highly simplified lattice model calculations. However, MCMC methods are also valid for atomistic models. Moreover, MCMC methods can be efficient for models with an implicit solvent representation for which large-scale chain updates can be performed, Such updates work especially well for systems involving large structural rearrangements such as protein folding and fold switching, which require enormous computer resources to simulate using MD.

For all-atom physics based protein model, we use PROFASI package [1], which is a C++ program package [1] for Monte Carlo simulations of protein folding and aggregation. The model implemented in this package contains all atoms of the protein chains but no explicit water molecules. The bonded interaction potential in this model is given by four sequence-dependent terms,

$$E = E_{\text{loc}} + E_{\text{ev}} + E_{\text{hb}} + E_{\text{sc}}, \tag{2.1}$$

where E_{loc} represents electrostatic interactions between neighbouring amino acids along the protein chain. The E_{ev} is an excluded volume therm with $1/r_{ij}^{12}$ repulsion between pairs of atoms i and j that are in distance r_{ij}. E_{hb} and E_{sc} are effective hydrogen bonding and sidechain-sidechain interactions, respectively, including electrostatic and effectively hydrophobic attractions [1].

2.2 Structure-based models or Gō models

The Gō-like potential was originally pioneered by Gō and coworkers as an approach to protein folding modeling [2]. By construction, the Gō-like potential E_{SBM} is lower for the native structure than for all other conformations [3]. Specifically, any two residues that are in spatial contact in the native state will be attractive. For this reason, Gō-like potentials are also called structure-based models (SBMs). The SBM or Gō-models have been used in both coarse-grained and all-atom simulations. Despite the rather crude approach, SBMs have been successfully applied to the folding of many proteins [3,4]. For example, a coarse-grained C_α model with a structure-based potential reproduced experimentally measured folding rates for a set of small proteins [5].

This thesis uses SBM potentials in two different ways; First, we develop and use an SBM in hybrid with PROFASI package to study the stability, folding, and fold switching of the RfaH protein. Second, we develop a structure-based coarse grain C_α model with single- and dual-basin SBM to study the effect of excluded volume arising from macromolecular crowders in stability, folding, and fold switching of proteins.

2.2.1 C_α model

To study the impact of excluded volume coming from macromolecular crowders on the protein folding and fold switching, we use a structure based coarse-grained C_α model. Each amino acid is represented by a single bead located at the C_α atom. The structure based energy function is given by [5]

$$\begin{aligned} E = & \sum^{\text{bonds}} K_{\text{b}}(b_i - b_i^0)^2 + \sum^{\text{angles}} K_\theta(\theta_i - \theta_i^0)^2 \\ & + \sum^{\text{dihedrals}} K_\phi^{(1)}[1 - \cos(\phi_i - \phi_i^0)] + K_\phi^{(3)}[1 - \cos 3(\phi_i - \phi_i^0)]) \\ & + \sum_{i<j-3}^{\text{native}} \epsilon \left[5\left(\frac{{r_{ij}}^n}{r_{ij}}\right)^{12} - 6\left(\frac{{r_{ij}}^n}{r_{ij}}\right)^{10} \right] + \sum_{i<j-3}^{\text{non-native}} \epsilon \left(\frac{\sigma}{r_{ij}}\right)^{12}, \end{aligned} \tag{2.2}$$

where b_i, θ_i, and ϕ_i are bond lengths, bond angles and torsion angles, respectively, and b_i^0, θ_i^0, and ϕ_i^0 are their values in the native conformation. Also, the strength of interactions are respectively given by K_{b}, K_θ, $K_\phi^{(1)}$, $K_\phi^{(3)}$ and ϵ. The fourth term is the Lennard-Jones attraction between native contacts, and the last term represents repulsions between bead pairs ij that do not form contact in the native structure. The distance r_{ij} is taken between C_α atoms, and the repulsion range is set to $\sigma = 4$ Å [5,6].

2.2.2 Single- and dual-basin structure-based model

A single basin structure-based model is a mathematical model that represents the energy landscape of a biological system as a single energy basin. This model is typically used to study the thermodynamics and kinetics of biomolecular interactions, such as protein folding. The energy landscape of a biological system is typically represented as a free energy surface, which is a function of the coordinates of the system (such as the positions of atoms or residues in a protein). The free energy surface has many local minima and maxima, which correspond to different conformations or states of the system. Due to these minima, it is hard to find the actual minima in simulation.

In this study, an energy term derived from the contact map of folded proteins is used to build a single basin structure-based model, which biases the protein towards the actual energy minimum by creating a deep funnel in the free energy surface [7].

The key advantage of single basin structure-based models is that they are computationally simple and easy to analyze. They can be used to study the thermodynamics and kinetics of biomolecular interactions using techniques such as Monte Carlo simulations or molecular dynamics simulations.

To study metamorphic proteins like RfaH, which have two native states, a single basin SBM would need to be generalized to incorporate two basins or conformational states. This can be achieved by incorporating two energy terms that bias the protein towards different configurations, and thus towards different native states [7–9]. Energy terms are derived from reference structures. These structures can be obtained from experiments or simulation/modeling.

2.3 Langevin Dynamics

In Langevin dynamics, the solvent molecules are not represented explicitly in the simulation. Instead, their effects are incorporated into the equations of motion through a friction term. This method, known as Langevin dynamics, has been widely used to study the kinetics of protein folding [10–12]. The equation of motion for this method is the Langevin equation, which is a stochastic differential equation that describes the time evolution of a system with a subset of the degrees of freedom. The equation for a given particle coordinate i in the system is as follows:

$$m\dot{v}_i(t) = -m\gamma v_i(t) + F_i^{\mathrm{c}}(t) + \Gamma_i(t), \tag{2.3}$$

where m is the mass of the particle, $F_i^{\mathrm{c}}(t)$ is the conformational force, γ_i is the friction coefficient, and $\Gamma_i(t)$ is the random force term. The conformational force $F_i^{\mathrm{c}}(t)$ is equal to the negative gradient of the total potential energy of the system, and drives the system towards the minimum energy state. The friction coefficient γ is related to the solvent viscosity.

The random force term $\Gamma_i(t)$ is a stochastic force that models the effect of the thermal motion of solvent on molecules. The magnitude of $\Gamma_i(t)$ control the temperature T in the system. It is assumed the $\Gamma_i(t)$'s at different times are uncorrelated. Therefore, the autocorrelation function for the random force is given by [13]:

$$\langle \Gamma_i(t)\Gamma_i(t') \rangle = 2m\gamma k_{\mathrm{B}} T \delta(t - t'), \tag{2.4}$$

where k_{B} $\delta(t - t')$ is the Dirac delta function. This equation states that the random force $\Gamma_i(t)$ is generated from a Gaussian white noise process with zero mean and a variance that is proportional to the temperature and the friction coefficient.

2.4 Simulated Tempering Method

Simulated tempering method is based on the idea that at high temperatures, the system has a higher entropy and can explore a larger portion of the phase space, while at low temperatures, the system has a lower entropy and is more likely to be trapped in local energy minima. By running the simulation at different temperatures yet remaining at equilibrium, the method allows the system to explore a wider range of energies, and can result in a more thorough sampling of the phase space. Specifically, simulated tempering uses Metropolis Monte Carlo sampling to explore a non-conventional distribution [14];

$$P(i, m) \propto \exp\left(-\beta_m E(i) + g_m\right), \tag{2.5}$$

where m is a temperature index which ranges from 1 to K, and $\beta_m = 1/k_{\mathrm{B}} T_m$. By introducing the variable m, the simulation can explore various temperatures. If high enough temperatures are visited the simulation can escape low-energy states and increase sampling efficiency at lower temperatures. The parameters g_1 to g_K are important in this technique. If they are set to $g_m = \beta F_m$ where F_m is the system free energy at temperature T_m, all temperatures become equally likely to visit. In simulated tempering, there are two types of Markov Chain Monte Carlo updates: conformational updates $i \to i'$, and temperature updates $m \to m'$. The acceptance probability for the conformational update is the same as that of a regular Metropolis simulation at a constant temperature (see Eq. (1.7)). The acceptance probability for the temperature update is

$$P_{\text{acc}}(m' \to m) = \exp\left(-E(i)(\Delta\beta + \Delta g)\right), \tag{2.6}$$

where $\Delta\beta = \beta_{m'} - \beta_m$ and $\Delta g = g_{m'} - g_m$. At the given temperature T_{m} the probability distribution function will be given by

$$P(i|m) \propto \exp\left(-\beta_m E(i)\right). \tag{2.7}$$

Hence, the state i generated at index m will be biased according to a canonical distribution at temperature T_{m}. The simulated tempering method is useful for studying complex systems such as biomolecules and materials, and has been applied to a variety of fields, including computational chemistry, materials science, and protein folding [15, 16].

Replica-exchange simulated tempering [14, 17] (also known as parallel tempering) is a Monte Carlo simulation technique used to sample the configuration space of complex systems. It involves running multiple simulations at different temperatures i.e. different energy scales simultaneously and periodically exchanging the configurations between them.

2.5 Scaled particle theory

The scaled particle theory (SPT) is a theoretical framework for hard sphere fluids developed in the 1960s [18]. It has been used to analyze the effect of crowders on the folding free energy of biomolecules. We have used Eq. (2.8) to estimate the impact of crowding on the free energy of fold switching in metamorphic proteins. According to SPT, the free energy of inserting a hard sphere of radius R in a hard sphere fluid of particle with radius R_c is given by [18]

$$\beta F = (3x + 3x^2 + x^3)\psi + (\frac{9x^2}{2} + 3x^3)\psi^2 + 3x^3\psi^3 - \ln(1 - \phi_\mathrm{c}), \tag{2.8}$$

where $\beta = 1/k_\mathrm{B}T$, $x = \frac{R}{R_\mathrm{c}}$, $\psi = \frac{\phi_\mathrm{c}}{1-\phi_\mathrm{c}}$, and ϕ_c is total volume fraction occupied by the hard spheres of the fluid. To derive Eq.(2.8) note that we can write $\beta F = -\ln\mathcal{A}$ [20], where $\mathcal{A}$ is the accessible volume fraction when inserting a sphere with radius R into the fluid. The reversible work βF, can be calculated by considering two limits for the sphere with radius R (small sphere and large sphere limits) [21]. For calculation, the radius of the sphere is scaled with parameter q ($R \to qR$). For $q \ll 1$, the inserted particle reduces to a point and the accessible volume for this particle is given by

$$V_\mathrm{acc} = V_\mathrm{T} - \frac{4}{3}\pi N_\mathrm{c}(R_\mathrm{c} + qR)^3, \tag{2.9}$$

where V_T is total volume of box and the second term is the volume of N crowders plus the depletion layer. Then the free volume fraction for a point particle is

$$\mathcal{A}(q \ll 1) = 1 - \frac{4}{3}\pi\rho_\mathrm{c}(R_\mathrm{c} + qR)^3, \tag{2.10}$$

where $\rho_\mathrm{c} = N_\mathrm{cr}/V_\mathrm{T}$ is the hard spheres volume fraction. Therefore,

$$\beta F(q \ll 1) = -\ln[1 - \frac{4}{3}\pi\rho_\mathrm{c}(R_\mathrm{c} + qR)^3]. \tag{2.11}$$

On the other limit, for $q \gg 1$, the required work to insert a sphere with radius R in a sea of crowders is approximately the work to create a hole with radius R inside the crowders, which is given by

$$F(q \gg 1) = \frac{4}{3}\pi (qR)^3 \Pi_{\mathrm{c}}, \tag{2.12}$$

where Π_{c} is the osmotic pressure of the dispersion of hard sphere fluid. In SPT, $F(q \ll 1)$ is expanded up to order q^2 and $F(q \gg 1)$ is added as the q^3 term. At the end, we can put $q = 1$ to find the βF which is

$$\beta F = -\ln[1 - \phi_{\mathrm{c}}] + 3x\psi + \frac{1}{2}(6x^2\psi + 9x^2\psi^2) + \frac{4}{3}\pi R^3 \beta \Pi_{\mathrm{c}}, \tag{2.13}$$

where $x = \frac{R}{R_{\mathrm{c}}}$, $\psi = \frac{\phi_{\mathrm{c}}}{1-\phi_{\mathrm{c}}}$, and $\phi_{\mathrm{c}} = (4/3)\pi \rho_{\mathrm{c}} R_{\mathrm{c}}^3$ is fluid volume fraction. For pure hard spheres, the osmotic pressure Π_{c} using the Percus–Yevick equation is given by [21];

$$\frac{\beta \Pi_{\mathrm{c}}}{\rho_{\mathrm{c}}} = \frac{1 + \phi_{\mathrm{c}} + {\phi_{\mathrm{c}}}^2}{(1 - \phi_{\mathrm{c}})^3} = \frac{1}{1 - \phi_{\mathrm{c}}} + \frac{2\phi_{\mathrm{c}}}{(1 - \phi_{\mathrm{c}})^2} + \frac{3{\phi_{\mathrm{c}}}^2}{(1 - \phi_{\mathrm{c}})^3}. \tag{2.14}$$

Finally, inserting the Π_{c} from Eq. (2.14) to Eq. (2.13) leads to Eq.(2.8). If the radius of the unfolded state of a protein is approximated using a Gaussian chain, SPT predicts that excluded volume crowders strongly increase the stability of the folded state. This effect is monotonic in ϕ_{c} and can be described by a two-parameter model, which can be extracted from the Eq. (2.8).

2.6 Two-state model

Conformational change in biomolecules are often cooperative transition between two well-defined states. One example is the folding of small proteins, which often can be described as a transition between an unfolded (U) and folded (F) states. The two states are separated by an energy barrier, and the protein can transition between the two states through thermal fluctuations or other external perturbations. The partition function can then be written as

$$\mathcal{Z} = \mathcal{Z}_{\mathrm{U}} + \mathcal{Z}_{\mathrm{F}} = e^{-\beta F_{\mathrm{U}}} + e^{-\beta F_{\mathrm{F}}}, \tag{2.15}$$

where $\exp(-\beta F_{\mathrm{U}}) = \sum_{i \in U} \exp(-\beta E_i)$ and $\exp(-\beta F_{\mathrm{F}}) = \sum_{i \in F} \exp(-\beta E_i)$. Then the thermal average of observable Q will be

$$\langle Q \rangle = \frac{Q_{\mathrm{U}} + Q_{\mathrm{F}} e^{-\beta \Delta F}}{1 + e^{-\beta \Delta F}}, \tag{2.16}$$

where Q_{U} and Q_{F} are the values of the Q in the unfolded and folded states, respectively, and $\Delta F = F_{\mathrm{F}} - F_{\mathrm{U}}$ is the free energy of folding,

In protein folding, the midpoint temperature (T_{m}) is the temperature at which the protein equally populates its folded and unfolded states. Because $\Delta F = \Delta E - T \Delta S$, where ΔE and ΔS are energy and entropy of folding, and $\Delta F(T_{\mathrm{m}}) = 0$, we have

$$\beta \Delta F = \Delta E (\frac{1}{k_{\mathrm{B}} T} - \frac{1}{k_{\mathrm{B}} T_{\mathrm{m}}}), \tag{2.17}$$

and the thermal average of observable Q in terms of $Q_{\mathrm{U}}, Q_{\mathrm{F}}, \Delta E$, and T_{m} is given by

$$\langle Q \rangle = \frac{Q_{\mathrm{U}} + Q_{\mathrm{F}} e^{-\Delta E (\frac{1}{k_{\mathrm{B}} T} - \frac{1}{k_{\mathrm{B}} T_{\mathrm{m}}})}}{1 + e^{-\Delta E (\frac{1}{k_{\mathrm{B}} T} - \frac{1}{k_{\mathrm{B}} T_{\mathrm{m}}})}}, \tag{2.18}$$

which is known as a two-state equation. If Q_{U} and Q_{F} are known, a folding curve $\langle Q \rangle$ versus T can be fit using ΔE and T_{m} as free parameters. Alternatively, Q_{U}, and Q_{F} can also be left as free parameters. T_{m} is often used in folding studies as a measure of native state stability.

Bibliography

[1] A. Irbäck, and S. Mohanty. PROFASI: a Monte Carlo simulation package for protein folding and aggregation. *J Comput Chem*, 27:1548–1555, 2006.

[2] H. Taketomi, Y. Ueda, and N. Go. Studies on protein folding, unfolding and fluctuations by computer simulation. *Int J Pept Protein Res*, 7:445–459, 1975.

[3] J. N. Onuchic and P. G. Wolynes. Theory of protein folding. *Curr Opin Struct Biol*, 14:70–5, 2004.

[4] J. Karanicolas and C. L. Brooks. Improved Gō-like models demonstrate the robustness of protein folding mechanisms towards non-native interactions. *J Mol Biol*, 334:309–325, 2003.

[5] S. Wallin, and H. S. Chan. Conformational entropic barriers in topology-dependent protein folding: perspectives from a simple native-centric polymer model. *J Phys Condens Matter*, 18:S307,2006.

[6] P. K. Zuber, I. Artsimovitch, M. Nandymazumdar, Z. Liu, Y. Nedialkov, K. Schweimer and S. H. Knauer. The universally-conserved transcription factor

RfaH is recruited to a hairpin structure of the non-template DNA strand. *Elife*, 7:34–36, 2018.

[7] B. Seifi, and S. Wallin. The C-terminal domain of transcription factor RfaH: Folding, fold switching and energy landscape. *Biopolymers*, 112:e23420, 2021.

[8] B. Seifi and S. Wallin. Examining the effect of the N-terminal domain of RfaH on domain dissociation and fold switching. *Preprint*, 2023.

[9] S. Bazmi, B. Seifi and S. Wallin. Effect of crowding on a fold switch protein is determined by its disordered tails. *Preprint*, 2023.

[10] O. F. Lange and H. Grubmüller. Collective Langevin dynamics of conformational motions in proteins. *J Chem Phys*, 124:214903, 2006.

[11] Z. Guo and D. Thirumalai. Kinetics of protein folding: nucleation mechanism, time scales, and pathways. *Biopolymers*, 36:83–102, 1995.

[12] J. D. Honeycutt and D. Thirumalai. The nature of folded states of globular proteins. *Biopolymers*, 332:695–709, 1992.

[13] T. Veitshans, D. Klimov and D. Thirumalai. Protein folding kinetics: timescales, pathways and energy landscapes in terms of sequence-dependent properties. *Fold Des*, 2:1–22, 1997.

[14] E. Marinari, and G. Parisi. Simulated tempering: a new Monte Carlo scheme. *Europhys Lett*, 19:451–458, 1992.

[15] R. C. Bernardi, M. C. R. Melo, K. Schulten. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim Biophys Acta*

Gen, 1850:872–877, 2015.

[16] H. E. U. Hansmann, and Y. Okamoto. New Monte Carlo algorithms for protein folding. *Curr Opin Struct Biol*, 9:177–183, 1999.

[17] A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov, and P. N. Vorontsov-Velyaminov. New approach to Monte Carlo calculation of the free energy: Method of expanded ensembles. *J Chem Phys*, 96:1776–1783, 1992.

[18] J. L. Lebowitz, and J. S. Rowlinson. Thermodynamic properties of mixtures of hard spheres. *J Chem Phys*, 41:133–138, 1964.

[19] A. P. Minton. Models for excluded volume interaction between an unfolded protein and rigid macromolecular cosolutes: macromolecular crowding and protein stability revisited. *Biophys J*, 88:971–985, 2005.

[20] B. Widom. Intermolecular Forces and the Nature of the Liquid State: Liquids reflect in their bulk properties the attractions and repulsions of their constituent molecules. *Science*, 157:375–382, 1967.

[21] J. P. Hansen, and I R. McDonald. Theory of Simple Liquids. With Applications to Soft Matter. *AP*, 2013.

Chapter 3

Structural fluctuations and mechanical stabilities of the metamorphic protein RfaH

Abstract

RfaH is a compact two-domain bacterial transcription factor that functions both as a regulator of transcription and an enhancer of translation. Underpinning the dual functional roles of RfaH is a partial but dramatic fold switch, which completely transforms the approximately 50-amino acid C-terminal domain (CTD) from an all-α state to an all-β state. The fold switch of the CTD occurs when RfaH binds to RNA polymerase (RNAP), however, the details of how this structural transformation is triggered is not well understood. Here we use all-atom Monte Carlo simulations to characterize structural fluctuations and mechanical stability properties of the full-length RfaH and the CTD as an isolated fragment. In agreement with experiments, we find that interdomain contacts are crucial for maintaining a stable, all-α CTD in free RfaH. To probe mechanical properties, we use pulling simulations to measure the work required to inflict local deformations at different positions along the chain. The resulting mechanical stability profile reveals that free RfaH can be divided into a "rigid" part and a "soft" part, with a boundary that nearly coincides with the boundary between the two domains. We discuss the potential role of this feature for how fold switching may be triggered by interaction with RNAP.

3.1 Introduction

Proteins are classically seen to fold spontaneously into an essentially unique three-dimensional structure as determined by their amino acid sequence [1]. Conformational fluctuations occur in the neighborhood of the native structure and are necessary for function [2, 3]. Various exceptions to this rule have emerged over the past two decades, inviting a more dynamic view of proteins. For example, some proteins are triggered to unfold – partly or wholly – upon receiving a signal, such as the binding of a ligand [4], while others are intrinsically disordered, i.e., lack a single stable conformation under native conditions [5]. Still other proteins have been found to transform from one folded structure to another, whereby they undergo major changes to their secondary structure contents, packing of core hydrophobic sidechains and overall shape [6]. Because these proteins essentially switch folds, they have been termed metamorphic [7] or transformer [8] proteins. Only a handful of metamorphic proteins have been studied in detail [6] but many more were recently identified [9].

One of the most dramatic examples of a metamorphic protein is the two-domain bacterial antiterminator RfaH, a member of the NusG family of transcription factors [10]. In its free state, RfaH exists in a domain-closed form with a tight interface between its N-terminal domain (NTD) and its C-terminal domain (CTD) [11], as shown in Fig. (3.1). While the mixed α/β fold of the NTD is shared with other NusG proteins, the α-helical hairpin of the CTD is drastically different. However, upon binding to RNA polymerase (RNAP) paused at an operon that contains a so-called ops site (a 12 nucleotide regulatory segment), the two domains separate and the CTD undergoes a complete transformation into a 5-stranded β-barrel structure that is essentially identical to the CTDs of other NusG proteins [10]. A flexible linker connecting the two domains makes it possible for the structurally transformed CTD to recruit ribosomal protein S10, forming a physical bridge between RNAP and ribosome. Hence, the combination of fold switching and domain separation allows RfaH to both regulate transcription and enhance translation. The development of the fold switching capability of RfaH [12] also represents one of the most clear examples of fold switching as a mechanism of protein structure evolution [13–16].

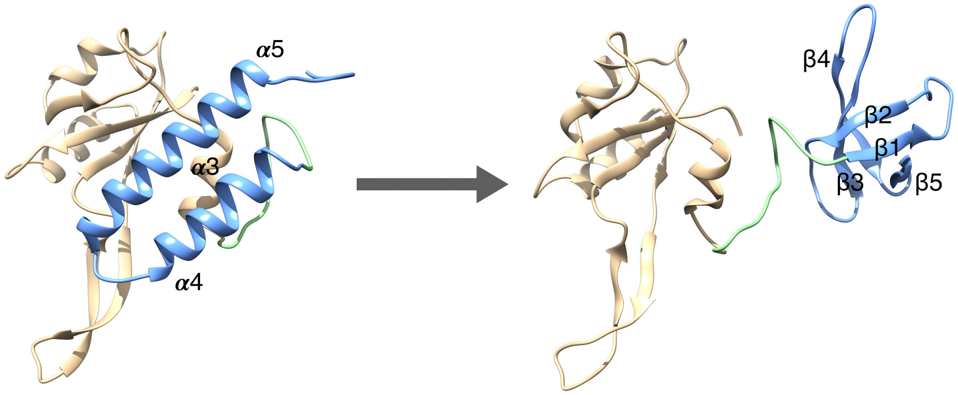

Figure 3.1: Fold switching of the bacterial transcription factor RfaH. (A) In its free state, RfaH adopts a compact three-dimensional structure (PDB id 2oug) consisting of an NTD (beige) in a mixed-α/β fold and a CTD (blue) in an α-helical hairpin fold. The two domains form a tight interface and are connected by a linker (green). (B) The binding of RfaH to RNAP triggers the two domains to separate [53], which causes the CTD to spontaneously re-fold into a 5-stranded β-barrel (PDB id 2lcl) while the NTD retains its fold. Missing segments in the X-ray crystal structure of domain-closed RfaH (2oug), including the linker region, were added using Modeller [41]. Molecular structures were rendered using UCSF Chimera [59].

Several computational studies have focused on elucidating the molecular details of the all-α to all-β transition of the RfaH-CTD [17–25]. Because of the large-scale nature of the structural transition non-traditional conformational sampling methods have been applied, such as targeted molecular dynamics [19], replica-exchange with tunnelling [24] and discrete path sampling [25]. Simulations of the isolated CTD have suggested that the transition proceeds via an unstructured intermediate state with little secondary structure [17, 18, 24, 25]. The transition has also been studied in the context of the full RfaH protein with all-atom and coarse-grained simulations, indicating that contacts between NTD and CTD may persist through at least part of the structural transition [19, 22]. Xun *et al.* [23] studied the overall stability to the free state of full length RfaH and found that the disordered linker region (see Fig. (3.1)) may contribute to stability by making energetically favorable interactions with ordered regions in both CTD and NTD.

Like most other known metamorphic proteins [26–31] (but not all [32]), RfaH undergoes a partial fold switch, i.e., a transformation that takes part of the protein to a different fold while the remaining part is left unchanged. We reasoned therefore that it is pertinent to understand how various structural and dynamical properties of metamorphic proteins vary along the sequence. Of particular interest are the differences between the structurally variable and unchanged parts. Here we set out to probe local structural fluctuations and mechanical stabilities of RfaH. To this end, we use Monte Carlo simulations in combination with a computationally efficient all-atom protein model [33]. This model has previously been applied to a range of protein processes, including folding [34], conformational fluctuations of disordered proteins [35, 36], and protein-peptide binding [37]. To probe local mechanical stability properties we apply the technique of Das *et al.* [38], which was developed to mechanically characterize proteins that are prone to misfolding [39]. The basic idea [38] is to simulate single-molecule pulling experiments that inflict local structural deformations at various sites on the protein surface. The work required to inflict such a deformation as a function of sequence index constitutes a type of local rigidity profile of the structure. Applied to RfaH, we find that structural rigidities are distinctly lower in the CTD than in the NTD, which may help regulate fold switching by RNAP. In probing both thermal fluctuations and mechanical properties we exploit the computational efficiency of our model, which allows us to carry out multiple independent trajectories for each condition such that heterogeneity between individual runs can be averaged out.

3.2 Materials and Methods

3.2.1 Representative structures

We obtained the X-ray crystal structure of Belogurov *et al.* [11] (PDB id 2oug) of the domain-closed form of RfaH from the Protein Data Bank [40]. The missing N-terminal residue (methionine), NTD-CTD linker 101–114 (single letter code: PKDIVDPAT-PYPGD) and the C-terminal tail 157–162 (TEFRKL) were added using the homology modelling tool MODELLER [41]. We also obtained the NMR structure of Burmann *et al.* [10] (PDB id 2lcl) of the isolated CTD in the all-β form. We retained the ordered part of the structure, i.e., residues Gly113–Leu162.

3.2.2 Computational protein model

Simulations were carried out using the computational protein model described in Ref. [33] and implemented in the software package PROFASI [42]. This model combines an all-atom protein representation with an effective potential energy function (no explicit solvent molecules) with 4 terms: $E = E_{\text{loc}} + E_{\text{ev}} + E_{\text{hb}} + E_{\text{sc}}$. The local term E_{loc} includes interactions between mainchain partial charges and provides a good local description of the protein chain, and the term E_{ev} implements excluded-volume effects between all atom pairs. Structure formation is driven mainly by the remaining two terms, E_{hb} and E_{sc}, which represent hydrogen bonding and sidechain-sidechain interactions, respectively. Hydrogen bonding is implemented with directionally-dependent explicit attractions between donor and acceptor groups. The term E_{sc} includes both pairwise attractions between sidechain charges and effective hydrophobic interactions. Parametrization of the effective energy function was carried out on the folding of 17 different proteins with diverse secondary structure contents, such that the global free energy minimum of each sequence represented the experimentally determined native structure.

3.2.3 Initial model conformations

All simulations were started from two initial model conformations derived from the experimental structures of the domain-closed (2oug) and domain-separated (2lcl) forms of RfaH through a “regularization” process. This process is necessary because, in our model, bond lengths and bond angles are held fixed at standard values [33]. As a consequence, there are geometric constraints imposed on model conformations that are in general not satisfied by experimental protein structures. A model conformation is instead specified by a set of dihedral backbone angles, ϕ and ψ, and a set of dihedral sidechain angles, χ. Finding a model conformations that regularizes an experimental structure is a two-step procedure: First, a model conformation is found that minimizes the root-mean-square deviation (RMSD) taken with respect to the experimental structure, regardless of the value of the potential energy E; thereafter, the minimum-RMSD conformation is taken as a starting point to minimize E while not increasing the RMSD too much. Overall, the process identifies a good structural approximation of the experimental structure, such that the potential energy function E is at a local minimum and all geometric constraints are satisfied. The two initial model conformations obtained this way for the experimental structures 2oug and 2lcl have RMSD values of 1.8 Å and 1.6 Å, respectively.

3.2.4 Monte Carlo updates

All simulations are carried out with two different types of small-step MC updates. Sidechains are updated by selecting and turning a single sidechain torsional angle, χ. Rotamers, i.e., statistically frequent sidechain configurations, are not imposed but allowed to arise from local interactions in the model. The backbone is updated by applying so-called Biased Gaussian Steps (BGS) [43]. The BGS move works by turning up to 8 consecutive backbone dihedral angles, i.e. the ϕ and ψ of 4 consecutive amino acids, in a coordinated manner such that a quasi-local deformation of the protein backbone is achieved. The BGS move has two free parameters a and b that control the acceptance rate and degree of bias, respectively [43]. Our simulations were carried out with $a = 300$ and $b = 10$. The fractions of sidechain and backbone MC updates were set to 70% and 30%, respectively. Hence, we assume here that the sidechains dynamics are fast compared to the backbone dynamics.

3.2.5 Fixed-temperature simulations

Thermal fluctuations are assessed through fixed-temperature Metropolis-Hastings [44, 45] MC simulations started from one of the initial model conformations (see above). For each starting conformation and temperature T, we carried out 30 independent simulations of 10^5 MC cycles each. An MC cycle corresponds to n elementary MC updates, where n is the number of degrees of freedom in the protein chain, i.e., the number of turnable ϕ, ψ and χ angles. For full-length RfaH, n = 740, and for the isolated CTD, n = 239.

3.2.6 Pulling simulations

To probe local structural rigidity of domain-closed RfaH with respect to mechanical forces, we use pulling simulations carried out in the following way. Forces were applied to two tethering points. We take one of the tethering points to be the residue closest to the center of mass of the protein (generally Phe56). The other point is the residue for which we want to quantify local structural rigidity. Mechanical forces are implemented using the energy term $E_{\text{pull}} = \frac{1}{2}k(\Delta d - vt)^2$, which is added to the PROFASI base energy function E. In E_{pull}, k is the spring constant, v the pulling speed, t the time since the pulling motion started, and $\Delta d = d - d_0$, where d is the distance between the tethering points at time t and d_0 the distance at $t = 0$. Our pulling simulations are carried out at 273 K with $k = 1.0$ e.u./Å^2 and $v = 10^{-3}$ Å/MC cycle, and are terminated when the extension $vt = 5$ Å. For each pair of tethering points, 10 independent pulling runs are carried out.

The energy (and temperature) scale of the PROFASI model has been set previously by mapping the experimentally determined melting temperature of the protein Trp cage (315 K) [46] to the folding temperature determined in internal model units from equilibrium simulations [33]. The energy scale found this way is 1eu= 5.615 kJ/mol [33], where eu is the internal energy unit in PROFASI. With this energy scale, the spring constant used in the present work becomes $k = 5.6$ kJ/mol/Å^2.

In order to assign physical units to the speed v in our pulling simulations, we must estimate the timescale of the small-step MC kinetics used in present work. We do that here in a 3-step procedure by again using Trp cage as a reference protein: (1) 100 independent folding simulations are carried out at 279 K. All simulations are started from the same extended conformation but supplied with different random seeds. The MC move set is the same as described above. (2) The mean-first-passage time (MFPT) to reach the native state in these simulations is determined, where the native state is considered reached when RMSD $<$ 3 Å and RMSD is determined with respect to the experimental structure of Trp cage [46]. (3) Finally, the observed MFPT = 159,000 MC cycles is mapped to the experimentally determined Trp cage folding time, i.e., $4\mu s$ [47]. This procedure gives a timescale of 1 MC cycle = 2.5×10^{-11}s, meaning that our pulling speed v is approximately 4.0 mm/s. This pulling speed is comparable to the 2.5 mm/s used in Ref. [39] and at the lower end of typical speeds used in atomistic computational pulling simulations (1 mm/s to 1 m/s) [48].

An alternative way to set the timescale in our simulations would have been to follow Habibi et al. [49], who relied on autocorrelation times in the relaxation to the native state following mechanical perturbations of the structure of interest. The approach taken here allow us to maintain a single reference protein (Trp cage) for determining both the energy and time scales of our model.

3.3 Results

3.3.1 Structural fluctuations in domain-closed RfaH

To probe structural fluctuations in the full-length domain-closed form of RfaH, we carry out small-step Monte Carlo simulations initiated from the structure in Fig. (3.1) (left) at the 6 different temperatures 273, 300, 310, 320, 330, and 340 K, as described in Methods. In quantifying the structural divergence from the starting point of these simulations, we focus on two types of quantities: (1) root-mean-square deviations (RMSD), taken with respect to the experimental structure of free RfaH, and (2) secondary structure contents (α and β), taken over different positions along the RfaH chain. We find substantial variations in the MC time evolution of these quantities even among simulation runs carried out at the same T (see Fig. S1). This is not unexpected given the heterogeneity between trajectories typically seen in molecular dynamics unfolding simulations of small proteins [50]. In order to focus on robust trends, we therefore construct an "average trajectory" using 30 independent runs at each T. For example, Fig. 3.2A shows $\langle\text{RMSD}\rangle$ as a function of MC time at $T = 273$ K, where $\langle\rangle$ indicates an average over the 30 runs. We find that $\langle\text{RMSD}\rangle$

gradually increases to around 6 Å over the first half of the simulation time and then remains flat throughout the rest of the simulations, indicating that the domain-closed form of RfaH is reasonably stable in our model at $T = 273$ K.

At $T \geq 310$ K, we find instead a gradual unfolding of RfaH, as illustrated in Fig. 3.2B for $T = 310$ K. Importantly, the loss of structure is more pronounced for the CTD than for the NTD; $\langle$RMSD$\rangle$ determined over the NTD region remains below 5 Å while $\langle$RMSD$\rangle$ determined over the CTD region reaches 8 Å at the end of the simulation time. Another way to quantify this difference is to consider the Pearson correlation coefficients, ρ, formed between full-length RMSD and NTD RMSD (ρ_{NTD}) or between full-length RMSD and CTD RMSD (ρ_{CTD}). We determined ρ_{NTD} and ρ_{CTD} for each trajectory using conformations taken from the second half. At $T = 310$ K, we find $\langle\rho_{\mathrm{NTD}}\rangle = 0.30$ and $\langle\rho_{\mathrm{CTD}}\rangle = 0.64$, indicating that structural fluctuations in the CTD indeed dominate over those in the NTD. The trend is even clearer at higher temperatures, e.g., $\langle\rho_{\mathrm{NTD}}\rangle = 0.17$ and $\langle\rho_{\mathrm{CTD}}\rangle = 0.73$ at $T = 340$ K, and only slightly weaker at the lowest studied temperature, i.e., $\langle\rho_{\mathrm{NTD}}\rangle = 0.39$ and $\langle\rho_{\mathrm{CTD}}\rangle = 0.60$ at $T = 273$ K. Taken together, from the perspective of RMSD, we conclude that domain-closed RfaH is overall stable over the length of our simulations at $T = 273$ K while, at higher Ts, the protein unfolds gradually (on average) driven primarily by loss of structure in the CTD.

This view is re-enforced by the results from our secondary structure analysis. In this analysis, we consider all conformations taken from the second half of the trajectories and use them to calculate average α- and β-structure contents, $\langle\alpha\rangle$ and $\langle\beta\rangle$. We find that, overall, the loss of secondary structure with increasing temperature

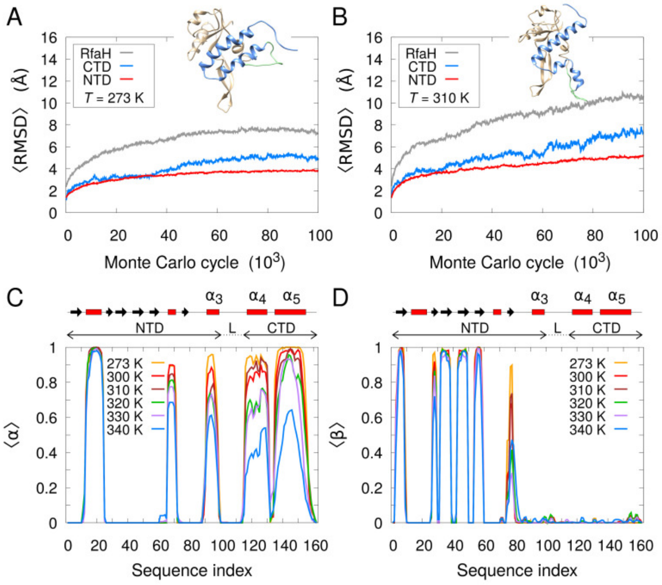

Figure 3.2: Thermal stability properties of domain-closed RfaH. Shown is the average root-mean-square deviation, ⟨RMSD⟩, as a function of MC time, for simulations carried out at temperatures (A) T = 273 K and (B) T = 310 K. The RMSD is determined with respect to the experimental structure of domain-closed RfaH (PDB id 2oug; see Fig. 3.1), taken over residues 1-162 (RfaH), 1-100 (NTD) or 115-156 (CTD). Brackets ⟨⟩ indicate an average over 30 independent simulations carried out at each T. All runs were initialized from a regularized version of 2oug (see Methods). Note that full-length RMSD is calculated over all residues in RfaH, including the domain linker (101-114) and C-terminal tail (157-162), while the NTD and CTD RMSDs are calculated only over structurally ordered regions in 2oug. This tend to make the full-length RMSD larger than either of the two "domain" RMSDs, despite the normalization by chain length in the RMSD measure. The representative structures at (A) 273 K and (B) 310 K are shown in ribbon representation with colors as in Fig. (3.1) and taken at MC step 50×10^3. The structures have RMSD = 5.7 and 9.3 Å, respectively, where RMSD is determined over residues 1-162. Shown is also the average (C) α-helix and (D) β-sheet contents, as determined over the second half of the simulations, as functions of the sequence index. Indicated above the graphs in (C) and (D) are the locations of the CTD, NTD and linker between the domains (L), as well as the α-helices (red solid boxes) and β-strands (black solid arrows) in 2oug. Assignment of secondary structures were carried using STRIDE [60].

is more pronounced for the CTD than for the NTD. For NTD (residues 1-100), we find $\langle\alpha\rangle = 0.25$ and $\langle\beta\rangle = 0.31$ at $T = 273$ K, which decrease to $\langle\alpha\rangle = 0.19$ and $\langle\beta\rangle = 0.26$ at $T = 340$ K. For CTD (residues 115-162), the α-helix content decreases from $\langle\alpha\rangle = 0.75$ at $T = 273$ K to $\langle\alpha\rangle = 0.39$ at $T = 340$ K. To examine individual secondary structure elements, we determine also $\langle\alpha\rangle$ and $\langle\beta\rangle$ as functions of sequence position, as shown in Figs. 3.2C and 3.2D. We see that loss of secondary structure in the NTD is due mainly to the short β-strand at positions 75-78 and helix α_3. In the CTD, the largest loss of secondary structure occurs for α_4, especially the N-terminal half of α_4. Helix α_5 also loses structure but mainly at the highest studied temperature ($T = 340$ K). A poor stability of the N-terminal part of α_4, especially the segment Val116–Gly121, has been noted before [10,23] and a dynamic network analysis found relatively large structural fluctuations of the α_3 helix [19].

3.3.2 Isolated CTD: all-α and all-β forms

The CTD segment can adopt two different folds, as shown in Fig. 3.1. We investigate structural fluctuations in the basins of both structural forms in a similar way as for full-length RfaH, i.e., we carry out small-step MC simulations of the isolated CTD started from either the all-α state or the all-β state. From the experiments of Burmann *et al.*, we know that the all-β state is the most thermally stable given the complete absence of NMR signals from the all-α form for the isolated CTD [10].

The all-β form of the CTD is indeed quite stable in our simulations, as shown in Figs. 3.3A and 3.3B. At $T = 310$ K, $\langle\text{RMSD}\rangle$ remains below 4 Å for most of the simulation time. By contrast, at the same T, the all-α state simulations lead to $\langle \text{RMSD} \rangle > 6$ Å after only a few MC cycles. The origin of this stark difference is apparent from the secondary structure profiles in Fig. 3.4. In the absence of a nearby NTD, both α_4 and α_5 lose much of their structures (see Fig. 3.4A). The tendency is, however, stronger for α_4 than for α_5. A substantial loss of helicity of α_4 occurs at all temperatures, meaning that α_4 is inherently unstable in our model. These results suggest that the NTD-CTD inter-domain interactions are necessary to stabilize the all-α CTD in RfaH, especially α_4, in line with experiments [10] and previous computational work [22]. The simulations started from the all-β state retain much more of the secondary structure (see Figs. 3.4C and 3.4D). At the highest studied T (340 K), we find an overall β-sheet content $\langle \beta \rangle = 0.43$ compared to 0.52 for the starting structure, 2lcl, a reduction by 17%. The corresponding reduction in the overall α-helix content for the simulations started in the all-α form is 63% ($\langle \alpha \rangle = 0.27$ at $T = 340$ K and 0.73 for the starting structure, 2oug). A closer look at Fig. 3.4D reveals that the smallest loss of structure in the all-β CTD simulations is generally found around the β-hairpin formed by β3 and β4 (Arg138 - Lys155), while the largest loss occurs in β5 (Phe159–Lys161).

Given the experimentally observed higher stability of the all-β state of the isolated CTD chain one should, in principle, expect simulations started in the all-α state to eventually re-fold into the all-β form, unless the temperature is too high. In our simulations, however, we do not observe a spontaneous fold switch into the β-barrel structure at any T. A possible explanation is that the trajectories carried out are too short to observe the transition. It is also possible that our computational model does not fully capture the free energy landscape of the isolated CTD, which should

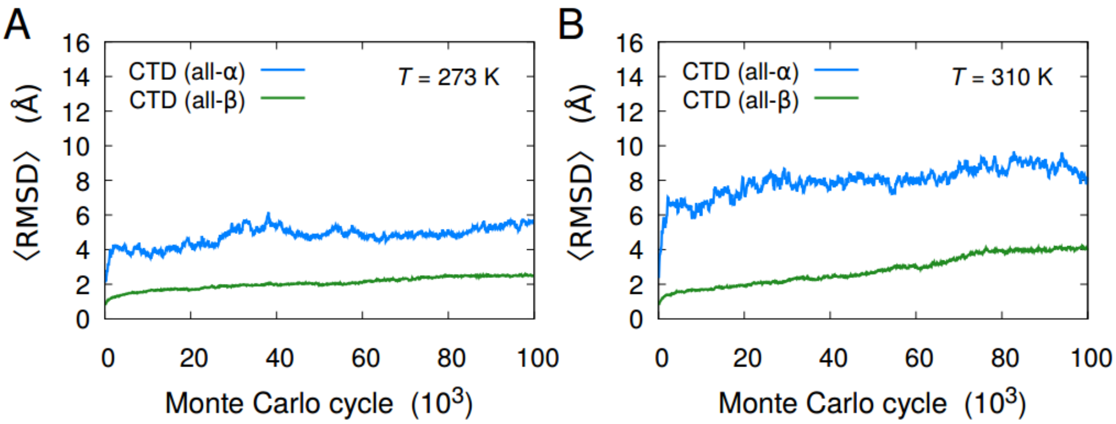

Figure 3.3: Stability of the all-α and all-β forms of isolated RfaH-CTD. Shown is ⟨RMSD⟩ as a function of MC time for simulations carried out at (A) $T = 273$ K or (B) $T = 310$ K. Simulations of the all-α state (blue) were initialized from a regularized version of the structure 2oug (residues 114-162) and RMSD is calculated with respect to 2oug. Simulations of the all-β state (green) were initialized from a regularized version of the structure 2lcl (residues 112-162) and RMSD is calculated with respect to 2lcl. Brackets ⟨⟩ indicate an average over 30 independent simulations.

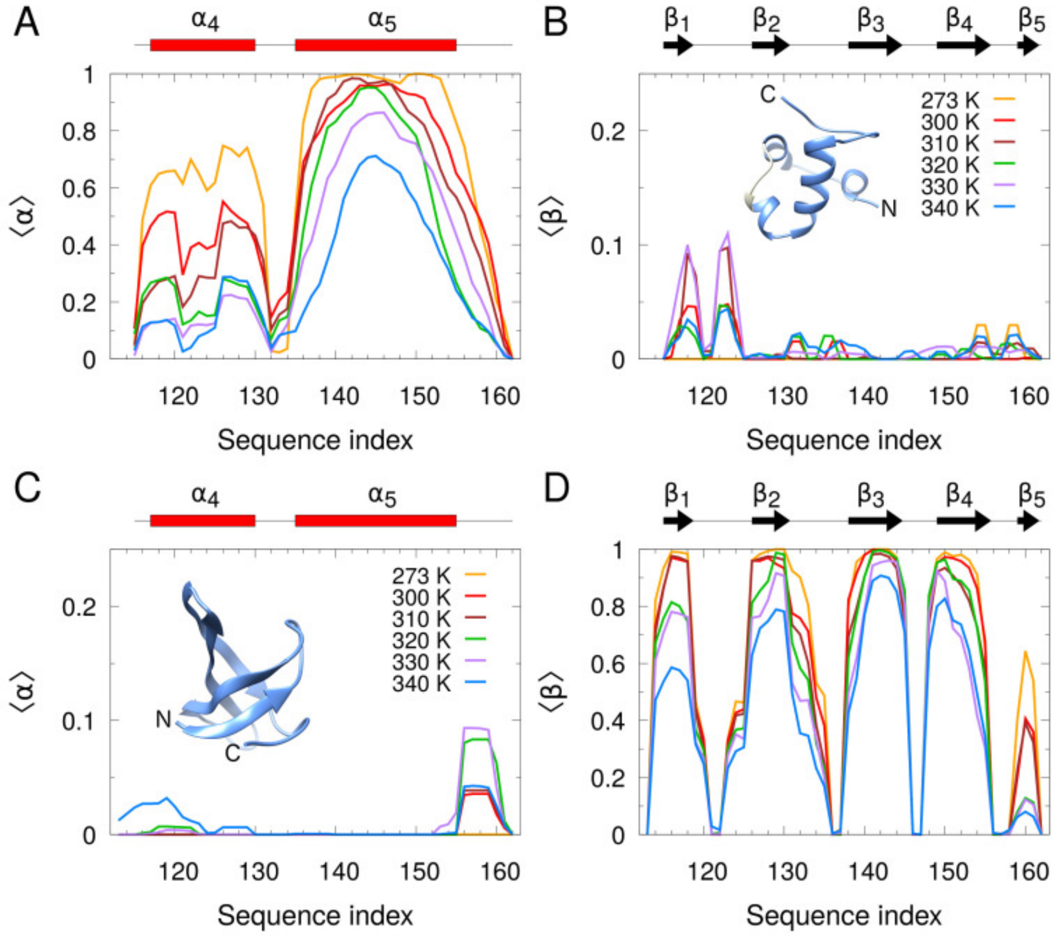

Figure 3.4: Stabilities of secondary structure elements in the two different forms of the isolated CTD. Shown are the average α-helix ($\langle\alpha\rangle$) and β-sheet ($\langle\beta\rangle$) contents as a function of sequence position for simulations started in the all-α (A, B) and the all-β (C, D) forms of the CTD, as determined over the second half of the simulations. Brackets $\langle\rangle$ indicate an average over 30 independent simulations. Results are given for simulations carried out at 6 different temperatures, 273 (orange), 300 (red), 310 (brown), 320 (green), 330 (purple) and 340 (blue) K. Color labels are the same in A-D but only shown in B and C. Sequence indices corresponds to those of the full-length RfaH. Shown are also two representative structures from simulations at 310 K of (B) the all-α and (C) the all-β forms. Both structures are taken from the middle of a simulation run and colored as in Fig. 3.1, except in (B) where the α-helical hairpin loop (residues 132-134) that is present in the starting structure (2oug) is shown in beige. N- and C-termini are indicated.

include a (dominant) funnel directed towards the β-barrel fold and thereby guarantee folding from any starting point for long enough simulations. Capturing the effects that control the subtle free energy balances in metamorphic proteins have, however, been recognized as a particularly challenging task in molecular simulations [51]. In this regard, it is encouraging that some tendencies for β-sheet formation, e.g., in the β1-β2 hairpin region, are seen at intermediate temperatures (see Fig. 3.4B).

3.3.3 Mechanical stabilities of domain-closed RfaH

We turn now to an analysis of mechanical stabilities of the domain-closed, free RfaH structure. The basic idea of this analysis [38] is to quantify the resistance to force-induced local deformations at different positions along the protein chain. The resistance to deformation at a given site provides a measure of the local structural rigidity. The local deformations can be obtained by simulating the effect of an atomic force microscopy (AFM) cantilever pulling on various C_α atom sites on an immobilized protein. To this end, we carry out MC simulations at $T = 273$ K with stretching forces applied to two tethering points. One of the tethering points is a centrally located position in the protein, playing the role of immobilizing the protein, and the other tethering point determines the sequence position that is being probed for its local mechanical stability.

A few typical pulling trajectories carried out on RfaH are given in Fig. 3.5A. They show that the pulling force exhibits large fluctuations and it is therefore not suitable as measure of mechanical stability. However, as was pointed out in Ref. [38], the cumulative work W performed by the external force versus the pulling extension is a relatively smooth function. The work W required to reach a given pulling extension threshold is therefore a more robust measure of local structural rigidity. Here we note additionally that even though the cumulative work curves from individual pulling runs are smooth, they can still differ significantly due to the underlying force fluctuations, as illustrated in Fig. 3.5B. Therefore, as a measure of local mechanical stability, we take here the work W required to reach a pulling extension of 5 Å, averaged over 10 independent pulling trajectories.

The work W required to inflict these local deformations as a function of sequence index constitutes a type of mechanical stability profile [39]. For domain-closed RfaH, we find a stability profile with some interesting properties, as shown in Fig. 3.6A. A clear division of the structure into two parts can be made: a structurally rigid part (residues ≈1-80) and a structurally soft part (residues ≈80-162). The division thus nearly coincides with the NTD-CTD division. The divisions do not coincide exactly, however. Helix α_3, which belongs to NTD but is located close to the domain interface, exhibits mechanical stabilities that are as low as for the two CTD helices, α_4 and α_5. The mechanical stability of the long, extended β-hairpin loop of the NTD (residues 31-52) may seem surprisingly high (see Fig 3.6B), given that previous simulations have found this segment to be highly dynamic [19]. The high mechanical stability of this hairpin loop in our analysis is likely an artifact of the pulling direction used. For residues in this loop, the line connecting the two tethering points run almost parallel to the two strands in this β-hairpin, giving a high resistance to elongation. However, structural rigidities in regions with secondary structure, with the exception of α_3, are generally higher in the NTD than in the CTD.

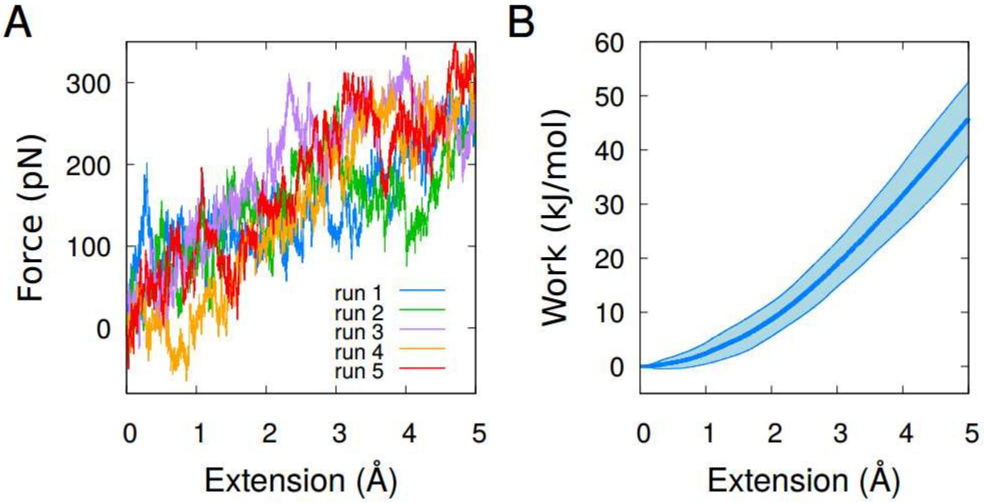

Figure 3.5: Probing local mechanical rigidities for the free, domain-closed RfaH structure through pulling simulations. Simulations are carried out with a constant pulling speed, v, and two tethering points: one centrally located residue (generally Phe56) and one for which local structural rigidity is probed. The pulling force is given by $F = -k(\Delta d - vt)$, where Δd is the change in distance between the tethering points since the start of pulling at time $t = 0$, and vt is the pulling extension. Ten independent simulations are carried out for each amino acid position that is probed for stability. Shown are examples of pulling runs obtained for the two tethering points Phe56 and Met1. (A) The pulling force F as a function of the extension for 5 of the 10 independent runs. (B) The cumulative work W as a function of pulling extension averaged over all 10 runs (thick dark blue line) and the standard deviation of W determined over the 10 runs (shaded light blue areas).

Within the CTD, the highest local mechanical stabilities are found in the 128-146 region, which includes parts of helices α_4 and α_5, and the loop between them. As a result, this "tip" of the α-helical hairpin may be important for triggering fold switching because it requires separation of CTD from the rest of RfaH. Indeed, several interactions close to this region have been found to impact the population balance between the two RfaH structural states. These include hydrophobic inter-domain interactions involving, e.g., Ile129, Phe130 and Leu141, and the buried inter-domain Glu48-Arg138 salt bridge [12, 19, 22]. Disrupting some of these interactions by point mutations, e.g., F130V, have been found to destabilize the domain interface and lead to an increased population of the domain-separated state [12].

As a final probe of the mechanical properties of domain-closed RfaH, we carry out 10 independent pulling simulations in which the tethering points are the N- and C-terminal amino acids of RfaH. These simulations are terminated when the extension reaches 200 Å, as illustrated for one of the runs in Fig. 3.7A. We find that mechanically extending RfaH via its termini leads to an initial unraveling of the CTD while, in contrast, the NTD remains much more unchanged. The work required to reach the full extension 200 Å is, on average, 542 ± 8 kJ/mol, as shown in Fig. 3.7B. As a point of comparison, we find that the work required to reach the same extension in pulling simulations of the NTD as an isolated domain (i.e., residues 1-100 excised from 2oug) is slightly higher, 605 ± 6 kJ/mol (data not shown). A closer inspection of the 10 pulling simulations of domain-closed RfaH reveals that 6 of the 10 runs exhibit a single dominant peak in the force followed by a rapid decrease. In each of the 6 cases, the rapid decrease in the force coincides with the detachment of helix α_4, as illustrated in Figs. 3.7C and 3.7D. Moreover, inter-domain contacts in the maximum-force conformations involve several hydrophobic amino acids in α_4, including Phe126, Ile129 and Phe130. Interestingly, in each of the 6 maximum-force conformations, Arg138 has previously detached from the NTD (see Fig. 3.7C), suggesting that force resistance in our pulling simulations is not dominated by the Glu48-Arg138 salt bridge but rather by hydrophobic interactions.

3.4 Discussion

We have investigated chain fluctuations and mechanical properties of the two structural forms of the bacterial transcription factor RfaH. A computationally efficient all-atom model [33] and MC sampling [43] allowed us to average over many independent runs, alleviating the statistical uncertainty deriving from the heterogeneity between trajectories. We have found that although the free (domain-closed) RfaH form is stable in our simulations at the lowest studied T, the all-α CTD, i.e., the part of RfaH undergoing fold switching, exhibits the largest structural fluctuations. These fluctuations in the CTD drive the loss of structure at higher Ts. A reduced thermal stability has been found for other metamorphic proteins [29], and may indeed be a hallmark of this class of proteins [6]. We also found that the all-α CTD as an isolated fragment does not retain its overall structure in our simulations, even at the lowest studied T. Hence, maintaining the all-α fold of the CTD in free RfaH depends critically on energetically favorable interdomain contacts. This observation is consistent with experiments [10,12] and previous computational work [19,22,23], validating our computational approach.

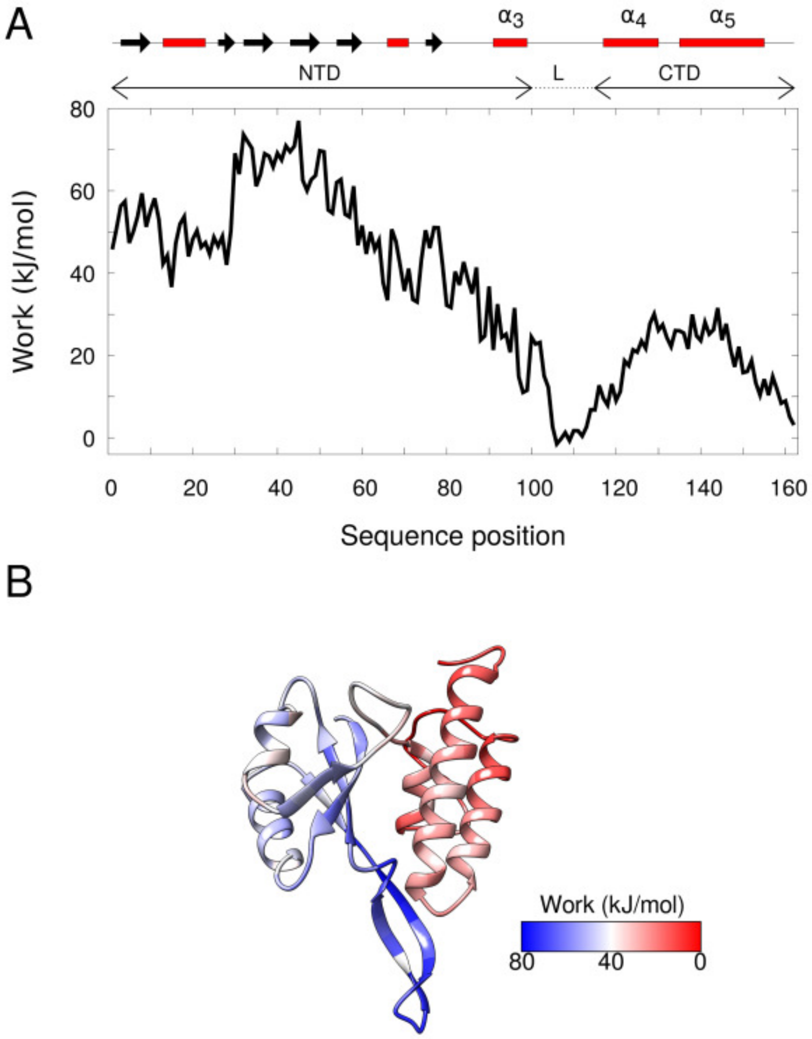

Figure 3.6: Mechanical stability profile of domain-closed RfaH. (A) Average work required to induce a small local deformation in the RfaH domain-closed structure in pulling simulations (see Fig. 3.5) as a function of sequence index. One of the tethering points in the pulling simulations is a centrally located residue (generally Phe56), chosen because it is the residue closest to the center-of-mass. The other tethering point is the residue for which the mechanical rigidity is probed. Upon probing residue positions 45–65, Leu7 is used as a central tethering point instead of Phe56 in order to avoid excessive work deriving from chain connectivity. (B) Heat map of the average pulling work in (A) projected onto a ribbon representation of the domain-closed RfaH structure (PDB id 2oug).

Comparing in more detail the two α-helices of the CTD we find that α_4, in particular, gains structure from its interactions with NTD. In our simulations of the isolated CTD, we find that α_4 loses almost all of its α-helix structure within the simulation time, while α_5 is much more resistant to structure loss. The coarse-grained simulations of Xiong et al. [21] showed comparable melting temperatures for α_4 and α_5, implying similar stabilities. However, other atomistic simulations [17, 18, 20, 23] have reached conclusions similar to ours, namely that α_4 is inherently less stable than α_5. The low stability of α_4 may impact its behavior within the full-length RfaH protein. Indeed, chemical shifts for C_α and CO groups in the segment Val116-Gly121, i.e., the N-terminal part of α_4, are close to those expected for a random coil state [10], indicating that α_4 may be shorter in solution than in the crystal state. Hydrogen-deuterium exchange experiments showed that α_4 is also more flexible than the rest of the CTD, including the loop between α_4 and α_5 [52]. Together with previous work, our study suggests that the all-α CTD is characterized by relatively large structural fluctuations, which may derive from a low inherent stability of the α_4 helix.

Our mechanical analysis shows that domain-closed RfaH can be divided into a structurally rigid part and a structurally soft part with a boundary that nearly, but not exactly, coincides with the division between NTD and CTD (see Fig. 3.6). This feature of domain-closed RfaH suggests a potential role for local mechanical properties for how fold switching is triggered by RNAP paused at an ops site. In the domain-closed form of RfaH, the RNAP binding site on the RfaH-NTD is masked by tight interactions with the all-α CTD [11]. It has therefore been suggested that an encounter complex consisting of RNAP, ops and RfaH is transiently formed, leading to domain

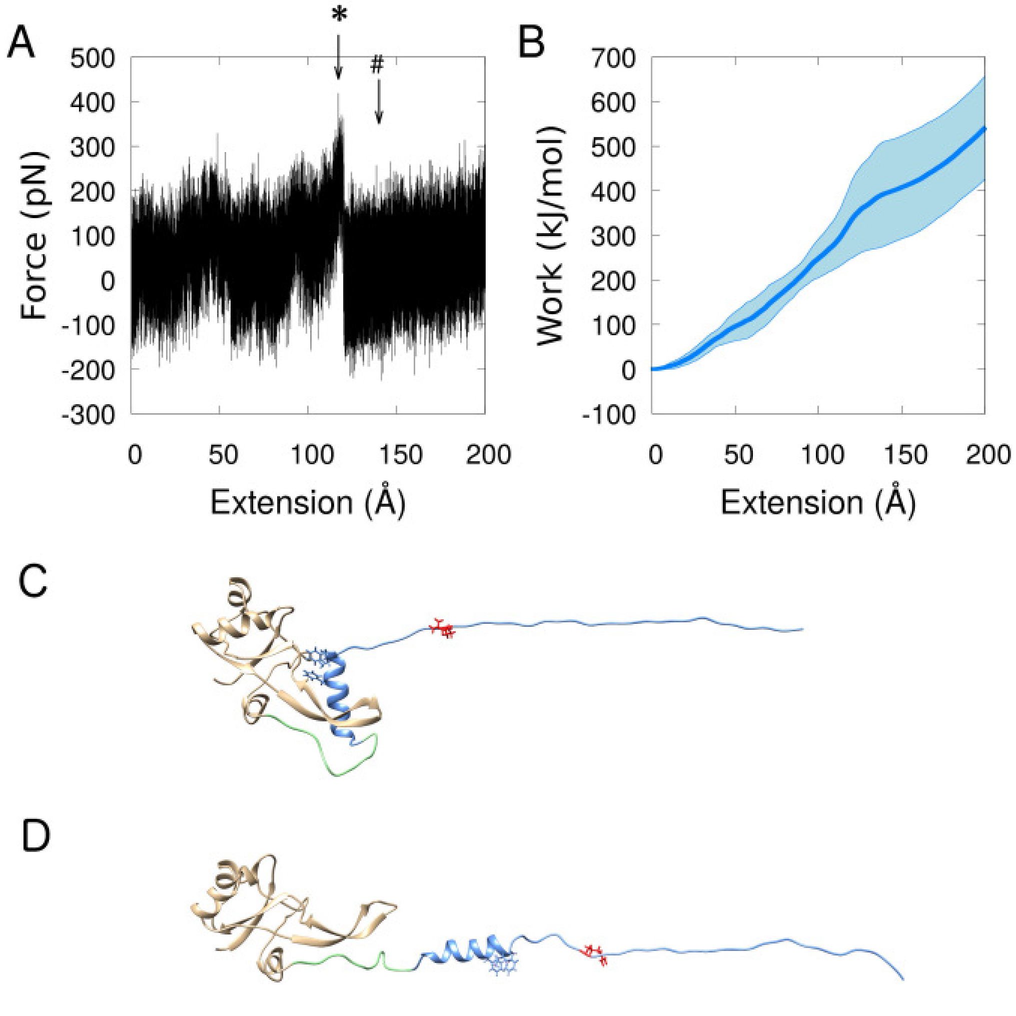

Figure 3.7: Mechanical stretching of domain-closed RfaH (2oug) via its N- and C-termini. (A) The force as function of pulling extension for 1 out of 10 trajectories obtained. (B) The cumulative work W as a function of pulling extension averaged over the 10 runs (thick dark blue line) and the standard deviation of W determined over the 10 runs (shaded light blue areas). (C) Conformation exhibiting the maximum pulling force in the trajectory shown in (A) and marked by "$*$". (D) Conformation marked by "#" in the trajectory in (A). Colors and molecular representation in (C) and (D) are the same as in Fig. (3.1), except for the residues Phe126, Ile129, Phe130 (blue) and Arg138 (red), which are highlighted in stick representation.

separation and unmasking of the RNAP binding site [10]. The structure of this encounter complex and how it triggers domain separation remain unclear, however [53]. A recent cryo-EM structure of the RNAP-ops with RfaH in its domain-separated form shows some important features [54]. For example, the RNAP binding surface on RfaH-NTD binds the RNAP clamp helices β'CH and the gate loop helix βGL, and the NTD helices α_1 and α_2 bind two flipped out ops-bases in the non-template DNA strand, as shown in Fig. 3.8A. Some of the favorable interactions between RfaH-NTD and RNAP-ops appear sterically feasible also in an encounter complex with RfaH in a domain-closed form. For example, the base-specific interactions with the NTD helices α_1 and α_2 may be feasible because the site on the DNA strand with the flipped out bases is relatively exposed. By contrast, interactions involving the RNAP binding surface are clearly infeasible in such an encounter complex, both because the CTD (all-α state) masks the binding surface and because the CTD would sterically clash with RNAP [22].

To see which part of CTD might be subject to repulsion from clashes with RNAP-ops in the encounter complex, we optimally superimposed the domain-closed form of RfaH onto the domain-separated form of RfaH found in the RNAP-ops elongation complex using the structurally conserved NTD, as shown in Fig. 3.8B. We make two observations: (1) the extended β-hairpin formed by residues 31-52 on the NTD shows different orientations in the RNAP-ops bound state and the free RfaH state; (2) the structural overlap resulting from the superposition of the domain-closed RfaH structure involves mainly α_5 on the RfaH-CTD and the clamp helices β'CH and the gate loop helix βGL, on RNAP-ops, as shown Fig. 3.8C. Interestingly, the overlap-

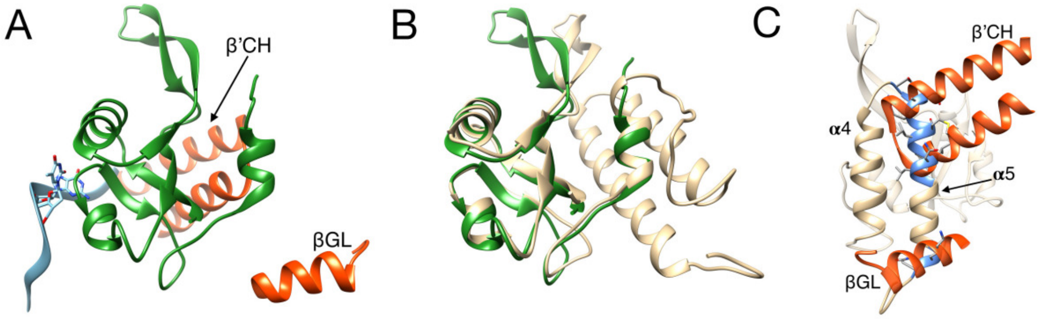

Figure 3.8: Binding to RNAP triggers RfaH domain separation and fold switching. (A) Cryo-EM structure of RfaH bound to RNAP in a paused state (PDB id 6c6s) [54]. Shown are the NTD (green; the CTD is not shown), the nontemplate DNA strand (blue), and the β'CH clamp helices and the βGL helices (red). Two flipped out bases on the nontemplate DNA strand interact with the helices α_1 and α_2 of the RfaH-NTD and are shown in stick representation. (B) The free domain-closed form of RfaH (beige) is optimally superimposed onto the domain-separated form of RfaH bound to RNAP-ops (green; the CTD is not shown). (C) The free domain-closed RfaH, in the same orientation as in (B), together with β'CH and βGL, which clashes severely with parts of the α_5 helix of the CTD (blue and stick representation).

ping residues on α_5 is roughly the region that exhibits the highest local mechanical stabilities within the CTD (cf. Figs. 3.6A and 3.8C). A possibility is therefore that interactions in the encounter complex between domain-closed RfaH and RNAP-ops are characterized by net attractive interactions with the NTD and net repulsive interactions with the CTD. The resulting opposing forces on NTD and CTD, in combination with the peculiar mechanical rigidity profile of domain-closed RfaH (see Fig. 3.6), might help trigger domain separation. A potential way to test this idea experimentally would be to structurally probe the encounter complex between domain-closed RfaH and RNAP-ops. In this regard, it would be extremely interesting to see if contacts could be detected between α_5 and β'CH using NMR spectroscopy, possibly by artificially stabilizing the all-α CTD fold through cross-linking techniques. Probing the structure of encounter complexes have been achieved, e.g., using paramagnetic resonance (PRE) and NMR spectroscopy techniques [55, 56].

Finally, we note more generally that the presence of structural subdomains is an apparent common property of metamorphic proteins [9]. Typically, one subdomain is structurally variable, i.e. it switches fold, while the rest of the protein remains unchanged. This property is exemplified by RfaH, although RfaH might be unique in that fold switching and domain separation occur together. A protein engineering study demonstrated that two radically different conformations of a 3-α-helix bundle protein could be stabilized by different subdomain interactions [16]. Finding novel metamorphic proteins could therefore be aided by features that identifies subdomains that are prone to fold switching. For example, Porter and Looger searched for novel metamorphic proteins based in part on the idea of subdomains are thermodynamically independent folding units [9]. Such identifying features can be combined with sequence-based features, such as the inaccuracy [57] and diversity [58] of predicted secondary structure. Our work here suggests that some subdomains prone to fold switching might be identified by probing the local mechanical stabilities of protein structures, as demonstrated for the free form of RfaH in Fig. 3.6. Local mechanical stabilities are relatively inexpensive to obtain computationally. With potentially as many as 4% of nonredundant proteins in the Protein Data Bank capable of fold switching [9], most still unidentified, future efforts to discover new metamorphic proteins are likely to be fruitful.

3.5 Conclusion

We have used all-atom Monte Carlo simulations to characterize thermal fluctuations and structural rigidities in the basins of the two structural forms of the transcription factor RfaH, in which the CTD is either in an all-α state closely interacting with the NTD or in an all-β state separated from the NTD. In line with previous experiments, we have found that energetically favorable NTD-CTD interactions are critical for stabilizing the domain-closed form, validating our computational approach. By measuring the resistance to local structural deformations by mechanical forces along the sequence, we have found that domain-closed RfaH can be divided into a structurally rigid part and a structurally soft part. Specifically, the CTD, i.e., the segment of RfaH undergoing fold switching, is characterized by particularly low structural rigidities. Mechanical profiling using pulling simulations as carried out here might, therefore, along with other identifying features, help identify metamorphic proteins.

Further, we speculate that the special mechanical features of domain-closed RfaH play a role in triggering domain separation in RfaH upon binding to the ops-paused RNAP complex. Further experimentation, focusing on the structural features of the transient encounter complex of RfaH and RNAP, will be necessary to determine to the precise mechanism underpinning domain separation and fold switching of RfaH.

Acknowledgements

The study was partly funded by the Natural Sciences and Engineering Research Council of Canada (grant RGPIN-2016-05014) and enabled by computational resources provided by Compute Canada.

Bibliography

[1] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.

[2] H. Frauenfelder, G. A. Petsko, and D. Tsernoglou. Temperature-dependent X-ray diffraction as a probe of protein structural dynamics. *Nature*, 280:558–563, 1979.

[3] J. A. McCammon, B. R. Gelin, and M. Karplus. Dynamics of folded proteins. *Nature*, 267:585–590, 1977.

[4] D. M. Mitrea and R. W. Kriwacki. Regulated unfolding of proteins in signaling. *FEBS Lett*, 587:1081–1088, 2013.

[5] J. Habchi, P. Tompa, S. Longhi, and V. N. Uversky. Introducing protein intrinsic disorder. *Chem Rev*, 114:6561–6588, 2014.

[6] P. N. Bryan and J. Orban. Proteins that switch folds. *Curr Opin Struct Biol*, 20:482–488, 2010.

[7] A. G. Murzin. Biochemistry. Metamorphic proteins. *Science*, 320(5884):1725–1726, Jun 2008.

[8] S. H. Knauer, P. Rösch, and I. Artsimovitch. Transformation: the next level of regulation. *RNA Biol*, 9:1418–1423, 2012.

[9] L. L. Porter and L. L. Looger. Extant fold-switching proteins are widespread. *Proc Natl Acad Sci USA*, 115:5968–5973, 2018.

[10] B. M. Burmann, S. H. Knauer, A. Sevostyanova, K. Schweimer, R. A. Mooney, R. Landick, I. Artsimovitch, and P. Rösch. An α helix to β barrel domain switch transforms the transcription factor RfaH into a translation factor. *Cell*, 150:291–303, 2012.

[11] G. A. Belogurov, M. N. Vassylyeva, V. Svetlov, S. Klyuyev, N. V. Grishin, D. G. Vassylyev, and I. Artsimovitch. Structural basis for converting a general transcription factor into an operon-specific virulence regulator. *Mol Cell*, 26:117–129, 2007.

[12] D. Shi, D. Svetlov, R. Abagyan, and I. Artsimovitch. Flipping states: a few key residues decide the winning conformation of the only universally conserved transcription factor. *Nucleic Acids Res*, 45:8835–8843, 2017.

[13] M. H. Cordes, R. E. Burton, N. P. Walsh, C. J. McKnight, and R. T. Sauer. An evolutionary bridge to a new protein fold. *Nat Struct Biol*, 7:1129–1132, 2000.

[14] T. Sikosek, E. Bornberg-Bauer, and H. S. Chan. Evolutionary dynamics on protein bi-stability landscapes can potentially resolve adaptive conflicts. *PLOS Comput Biol*, 8:e1002659, 2012.

[15] C. Holzgräfe and S. Wallin. Smooth functional transition along a mutational pathway with an abrupt protein fold switch. *Biophys J*, 107:1217–1225, 2014.

[16] L. L. Porter, Y. He, Y. Chen, J. Orban, and P. N. Bryan. Subdomain interactions foster the design of two protein pairs with 80% sequence identity but different folds. *Biophys J*, 108:154–162, 2015.

[17] S. Li, B. Xiong, Y. Xu, T. Lu, X. Luo, C. Luo, J. Shen, K. Chen, M. Zheng, and H. Jiang. Mechanism of the all-α to all-β conformational transition of RfaH-CTD: Molecular dynamics simulation and Markov State model. *J Chem Theory Comput*, 10:2255–2264, 2014.

[18] J. B. GC, Y. R. Bhandari, B. S. Gerstman, and P. P. Chapagain. Molecular dynamics investigations of the α-helix to β-barrel conformational transformation in the RfaH transcription factor. *J Phys Chem B*, 118:5101–5108, 2014.

[19] J. B. GC, B. S. Gerstman, and P. P. Chapagain. The role of the interdomain interactions on RfaH dynamics and conformational transformation. *J Phys Chem B*, 119:12750–12759, 2015.

[20] N. Balasco, D. Barone, and L. Vitagliano. Structural conversion of the transformer protein RfaH: new insights derived from protein structure prediction and molecular dynamics simulations. *J Biomol Struct Dyn*, 33:2173–2179, 2015.

[21] L. Xiong and Z. Liu. Molecular dynamics study on folding and allostery in RfaH. *Proteins*, 83:1582–1592, 2015.

[22] C. A. Ramirez-Sarmiento, J. K. Noel, S. L. Valenzuela, and I. Artsimovitch. Interdomain contacts control native state switching of RfaH on a dual-funneled landscape. *PLOS Comput Biol*, 11:e1004379, 2015.

[23] S. Xun, F. Jiang, and Y. D. Wu. Intrinsically disordered regions stabilize the helical form of the C-terminal domain of RfaH: A molecular dynamics study. *Bioorg Med Chem*, 24:4970–4977, 2016.

[24] N. A. Bernhardt and U. H. E. Hansmann. Multifunnel landscape of the fold-switching protein RfaH-CTD. *J Phys Chem B*, 122:1600–1607, 2018.

[25] J. A. Joseph, D. Chakraborty, and D. J. Wales. Energy landscape for fold-switching in regulatory protein RfaH. *J Chem Theory Comput*, 15:731–742, 2019.

[26] M. H. Cordes, N. P. Walsh, C. J. McKnight, and R. T. Sauer. Evolution of a protein fold in vitro. *Science*, 9:325–328, 1999.

[27] D. R. Littler, S. J. Harrop, W. D. Fairlie, L. J. Brown, G. J. Pankhurst, S. Pankhurst, M. Z. DeMaere, T. J. Campbell, A. R. Bauskin, R. Tonini, M. Mazzanti, S. N. Breit, and P. M. Curmi. The intracellular chloride ion channel protein CLIC1 undergoes a redox-controlled structural transition. *J Biol Chem*, 279:9298–9305, 2004.

[28] X. Luo, Z. Tang, G. Xia, K. Wassmann, T. Matsumoto, J. Rizo, and H. Yu. The Mad2 spindle checkpoint protein has two distinct natively folded states. *Nat Struct Mol Biol*, 11:338–345, 2004.

[29] P. A. Alexander, Y. He, Y. Chen, J. Orban, and P. N. Bryan. A minimal sequence code for switching protein structure and function. *Proc Natl Acad Sci USA*, 106:21149–21154, 2009.

[30] M. López-Pelegrín, N. Cerdà-Costa, A. Cintas-Pedrola, F. Herranz-Trillo, P. Bernadó, J. R. Peinado, J. L. Arolas, and F. X. Gomis-Rüth. Multiple stable conformations account for reversible concentration-dependent oligomerization and autoinhibition of a metamorphic metallopeptidase. *Angew Chem Int Ed Engl*, 53:10624–10630, 2014.

[31] Y. G. Chang, S. E. Cohen, C. Phong, W. K. Myers, Y. I. Kim, R. Tseng, J. Lin, L. Zhang, J. S. Boyd, Y. Lee, S. Kang, D. Lee, S. Li, R. D. Britt, M. J. Rust, S. S. Golden, and A. LiWang. A protein fold switch joins the circadian oscillator to clock output in cyanobacteria. *Science*, 349:324–328, 2015.

[32] R. L. Tuinstra, F. C. Peterson, S. Kutlesa, E. S. Elgin, M. A. Kron, and B. F. Volkman. Interconversion between two unrelated protein folds in the lymphotactin native state. *Proc Natl Acad Sci USA*, 105:5057–5062, 2008.

[33] A. Irbäck, S. Mitternacht, and S. Mohanty. An effective all-atom potential for proteins. *PMC Biophysics*, 2:1–24, 2009.

[34] S. Mohanty, J. H. Meinke, O. Zimmermann, and U. H. Hansmann. Simulation of Top7-CFr: a transient helix extension guides folding. *Proc Natl Acad Sci USA*, 105:8004–8007, 2008.

[35] J. Petrlova, A. Bhattacherjee, W. Boomsma, S. Wallin, J. O. Lagerstedt, and A. Irbäck. Conformational and aggregation properties of the 1-93 fragment of apolipoprotein A-I. *Protein Sci*, 23:1559–1571, 2014.

[36] B. M. Coady, J. D. Marshall, L. E. Hattie, A. M. Brannan, M. N. Fitzpatrick, K. E. Hickey, S. Wallin, V. Booth, and R. J. Brown. Characterization of a

peptide containing the major heparin binding domain of human hepatic lipase. *J Pept Sci*, 24:e3123, 2018.

[37] I. Staneva, Y. Huang, Z. Liu, and S. Wallin. Binding of two intrinsically disordered peptides to a multi-specific protein: a combined Monte Carlo and molecular dynamics study. *PLOS Comput Biol*, 8:e1002682, 2012.

[38] A. Das and S. S. Plotkin. Mechanical probes of SOD1 predict systematic trends in metal and dimer affinity of ALS-associated mutants. *J Mol Biol*, 425:850–874, 2013.

[39] A. Das and S. S. Plotkin. SOD1 exhibits allosteric frustration to facilitate metal binding affinity. *Proc Natl Acad Sci USA*, 110:3871–3876, 2013.

[40] H. M. Berman. The Protein Data Bank: a historical perspective. *Acta Crystallogr, A, Found Crystallogr*, 64:88–95, 2008.

[41] A. Sali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, 234:779–815, 1993.

[42] A. Irbäck and S. Mohanty. PROFASI: A Monte Carlo simulation package for protein folding and aggregation. *J Comput Chem*, 27(13):1548–1555, Oct 2006.

[43] G. Favrin, A. Irbäck, and F. Sjunnesson. Monte Carlo update for chain molecules: biased Gaussian steps in torsional space. *J Chem Phys*, 114:8154–8158, 2001.

[44] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, and A. H. Teller. Equation of state calculations by fast computing machines. *J Chem Phys*, 21:1087–1092, 1953.

[45] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[46] J. W. Neidigh, R. M. Fesinmeyer, and N. H. Andersen. Designing a 20-residue protein. *Nat Struct Biol*, 9:425–430, 2002.

[47] L. Qiu, S. A. Pabit, A. E. Roitberg, and S. J. Hagen. Smaller and faster: the 20-residue Trp-cage protein folds in 4 μs. *J Am Chem Soc*, 124:12952–12953, 2002.

[48] S. Sheridan, F. Gräter, and C. Daday. How fast is too fast in force-probe molecular dynamics simulations? *J Phys Chem B*, 123:3658–3664, 2019.

[49] M. Habibi, J. Rottler, and S. S. Plotkin. As simple as possible, but not simpler: Exploring the fidelity of coarse-grained protein models for simulated force spectroscopy. *PLOS Comput Biol*, 12:e1005211, 2016.

[50] R. Day and V. Daggett. Ensemble versus single-molecule protein unfolding. *Proc Natl Acad Sci USA*, 102:13445–13450, 2005.

[51] J. R. Allison, M. Bergeler, N. Hansen, and W. F. van Gunsteren. Current computer modeling cannot explain why two highly similar sequences fold into different structures. *Biochemistry*, 50:10965–10973, 2011.

[52] P. Galaz-Davison, J. A. Molina, S. Silletti, E. A. Komives, S. H. Knauer, I. Artsimovitch, and C. A. Ramírez-Sarmiento. Differential local stability governs the metamorphic fold switch of bacterial virulence factor RfaH. *Biophys J*, 118:96–104, 2020.

[53] P. K. Zuber, K. Schweimer, P. Rösch, I. Artsimovitch, and S. H. Knauer. Reversible fold-switching controls the functional cycle of the antitermination factor RfaH. *Nat Commun*, 10:702, 2019.

[54] J. Y. Kang, R. A. Mooney, Y. Nedialkov, J. Saba, T. V. Mishanina, I. Artsimovitch, R. Landick, and S. A. Darst. Structural basis for transcript elongation control by NusG family universal regulators. *Cell*, 173:1650–1662, 2018.

[55] C. Tang, J. Iwahara, and G. M. Clore. Visualization of transient encounter complexes in protein-protein association. *Nature*, 444:383–386, 2006.

[56] A. N. Volkov, Q. Bashir, J. A. Worrall, G. M. Ullmann, and M. Ubbink. Shifting the equilibrium between the encounter state and the specific form of a protein complex by interfacial point mutations. *J Am Chem Soc*, 132:11487–11495, 2010.

[57] S. Mishra, L. L. Looger, and L. L. Porter. Inaccurate secondary structure predictions often indicate protein fold switching. *Protein Sci*, 28:1487–1493, 2019.

[58] N. Chen, M. Das, A. LiWang, and L.-P. Wang. Sequence-based prediction of metamorphic behavior in proteins. *Biophys J*, 119:1380–1390, 2020.

[59] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin. UCSF Chimera–a visualization system for exploratory research and analysis. *J Comput Chem*, 25:1605–1612, 2004.

[60] D. Frishman and P. Argos. Knowledge-based protein secondary structure assignment. *Proteins*, 23:566–579, 1995.

Chapter 4

The C-terminal domain of transcription factor RfaH: Folding, fold switching and energy landscape

Abstract

We simulate the folding and fold switching of the C-terminal domain (CTD) of the transcription factor RfaH using an all-atom physics-based model augmented with a dual-basin structure-based potential energy term. We show that this hybrid model captures the essential thermodynamic behavior of this metamorphic domain, i.e., a change in the global free energy minimum from an α-helical hairpin to a 5-stranded β-barrel upon the dissociation of the CTD from the rest of the protein. Using Monte Carlo sampling techniques, we then analyze the energy landscape of the CTD in terms of progress variables for folding towards the two folds. We find that, below the folding transition, the energy landscape is characterized by a single, dominant funnel to the native β-barrel structure. The absence of a deep funnel to the α-helical hairpin state reflects a negligible population of this fold for the isolated CTD. We observe, however, a higher α-helix structure content in the unfolded state compared to results from a similar but fold switch-incompetent version of our model. Moreover, in folding simulations started from an extended chain conformation we find transiently formed α-helical structure, occurring early in the process and disappearing as the chain progresses towards the thermally stable β-barrel state.

4.1 Introduction

Proteins are increasingly being discovered with a remarkable ability to switch between folds with widely different structures [1–3]. While it is not uncommon for proteins to undergo large-scale motions after their initial folding, such as domain-swapping [4] or other hinge-like motions [5], fold switching is a distinct phenomenon. It involves a reorganization of the protein at the most basic structural level at play in folding, i.e., secondary structure (α-helices and β-sheets). Despite the remarkable complexity of these molecular transformations, fold switching is reversible and thereby controlled by the system's free energy. In this sense, it can be said that metamorphic proteins adhere to Anfinsen's thermodynamic principle (or hypothesis) of protein folding [6]. Clearly, however, fold switching fundamentally challenges the idea of a unique native conformation, which was a central aspect of the classic view of folding since emerging from the pioneering refolding experiments on ribonuclease A [7]. It is important to note that fold switching typically occurs only when triggered by specific changes to the local environment (or milieu) of the protein, such as salt concentration [8], redox condition [9] or oligomerization state [10]. In the absence of such a trigger, metamorphic proteins typically fold to an apparently unique structure, which masks their fold switching capabilities. As a result, metamorphic proteins often go unrecognized [11].

Metamorphic proteins thus encode two different folds within a single amino acid sequence even though, as mentioned, they typically adopt a single fold for a given (constant) local milieu. It is natural, then, to ask: what impact does this dual encoding have on their folding? This question is in fact related to a classic line of inquiry in protein folding, namely whether the mechanism of folding is conserved among homologous proteins or, more generally, among sequences adopting the same fold [12,13]. It was observed, remarkably, that sequences with low sequence similarity (but still adopting the same native fold) often fold in a very similar manner [14–16] however, this conservation breaks down at very low sequence similarity [17] and does not extend to all fold classes [17–19]. The sequences of metamorphic proteins have diverged from their homologs to the point of nearly adopting a new unique fold. Characterizing the folding of metamorphic proteins may therefore shed light on the extent to which the amino acid sequence can re-shape the energy landscape within a single fold. It may also be practically important in the field of protein metamorphism because any identifying feature, including folding behavior, can be used to discover as of yet unknown fold switching events [10, 11, 20].

Here we study the C-terminal domain (CTD) of the transcriptional antiterminator protein RfaH, a prototypical metamorphic protein [21]. The fold switching exhibited by this protein has been studied extensively both experimentally [22–25] and computationally [26–33]. The RfaH CTD, on its own, i.e., dissociated or excised from the rest of the protein, folds spontaneously into a β-barrel-like fold with 5 strands, as shown in Fig. (4.1). This structure is virtually identical to the CTD structure of some other members of the NusG/Spt5 family of transcription factors to which RfaH belongs [34]. However, as part of the full-length RfaH, the CTD adopts instead an α-helical hairpin [22]. This entirely different folded state is stabilized by favorable interactions with the RfaH N-terminal domain (NTD) [23]. The switch in structure from all-α to all-β is triggered when RfaH binds to RNA polymerase in a paused state, which underpins RfaH's regulatory function in transcription and translation [24, 35]. From an evolutionary perspective, recent phylogenetics analyses revealed that RfaH likely evolved from NusG in Proteobacteria through gene duplication and functional divergence [36].

Simulating the structural transitions of metamorphic proteins on the computer has the potential for detailed insight but is technically challenging. Some studies have

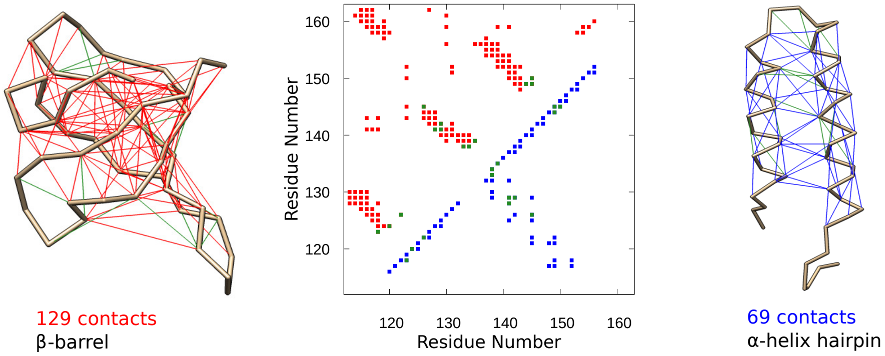

Figure 4.1: The two different folds of the C-terminal domain (CTD) of RfaH. Native contacts in the two folds, determined using the shadow map method [56], are shown as thin lines in the C_α-traces (left and right) and as solid squares in the contact maps (center). The two contact sets, denoted C^β (β-barrel) and C^α (α-helical hairpin), have 13 common contacts (green).

focused on alleviating the formidable problem of achieving sufficient conformational sampling for these large-scale transitions by various enhanced sampling techniques [26, 28, 32, 37, 38]. This approach assumes, however, an underlying model that accurately captures the delicate free energy balance between folds. This is not guaranteed even for modern explicit-water molecular dynamics force fields [39], although it has been achieved within the framework of a coarse-grained 3-letter protein model [40–42]. Other studies have relied on so-called structure-based models (SBM). The basic idea of SBMs is to make the contacts present in a given target (typically native) structure artificially attractive, while non-native interactions are left neutral or even made repulsive [43]. SBMs can be constructed with several basin of attractions, each one representing a different target structure. Single-basin SBMs have a long tradition in protein folding studies [44] and have had notable successes in reproducing detailed experimental data on specific proteins, such as folding rates [45, 46] and transition-state structures [47], despite the *ad hoc* nature of the approach. Dual- or multi-basin SBMs are the logical extension to metamorphic proteins, and such models have been applied to both the CTD of RfaH [27, 30] and the mutation-induced fold switch between the GA and GB proteins [48, 49].

To study the folding and fold switching of the RfaH CTD, we take a hybrid approach that combines a physics-based model for protein folding with a dual-basin SBM. We develop this hybrid sequence-structure based model based on its basic thermodynamic behavior, which we determine using extensive Monte Carlo (MC) simulations. Hence, we require that this model captures the switch in global free energy minimum between the two folds of RfaH CTD, depending on whether the CTD is part of the full-length RfaH chain or an isolated fragment. We also consider a version of our model with a single-basin SBM that folds the RfaH CTD but does not exhibit proper fold switching. By comparing closely with this model, we are able to probe the impact of encoding for two different folds on various features of the energy landscape of RfaH CTD.

4.2 Materials and Methods

4.2.1 Physics-based computational protein model

All simulations were carried out using the software package PROFASI [50]. The physics-based model implemented in this package is described in Ref. [51]. Briefly, the model combines an all-atom protein representation with an effective potential energy function with 4 terms: $E^{(0)} = E_{\text{loc}} + E_{\text{ev}} + E_{\text{hb}} + E_{\text{sc}}$. Geometrically, there are some constraints on allowed protein conformations, e.g., fixed bond lengths and bond angles, such that $E^{(0)}$ is a function only of the Ramachandran (torsional) angles, ϕ and ψ, determining the backbone conformation, and various sidechain torsional angles, χ, determining the sidechain conformations. Solvent effects are implicitly taken into account by the energy function. The term E_{ev} implements excluded-volume between all atom pairs using $1/r^{12}$-repulsions. The term E_{loc} includes interactions between atoms close along the chain, e.g., between partial charges on neighboring peptide planes, which help provide a good local chain description. The remaining two terms, E_{hb} and E_{sc}, represent hydrogen bonding and sidechain-sidechain interactions, respectively. Hydrogen bonding is implemented through orientationally dependent attractions between donor and acceptor groups. The term E_{sc} includes both pairwise interactions between partial charges on sidechains and effective hydrophobic attractions.

4.2.2 Equilibrium Monte Carlo simulations

To determine the equilibrium (thermodynamic) behavior of the CTD of RfaH, either as an isolated fragment or as part of full-length RfaH, we used simulated tempering MC. Simulated tempering is a generalized ensemble MC method that works by allowing a random walk both in conformational space and in temperature space, such that equilibrium sampling at a set of selected temperatures is obtained in a single run. We performed sampling in conformational space using three different types of moves: (1) a pivot move that updates a single Ramachandran ϕ- or ψ-angle; (2) Biased Gaussian Steps (BGS) that work by updating up to 8 consecutive ϕ, ψ-angles such that an approximately local chain deformation is obtained [52]; and (3) a sidechain move that updates a single sidechain torsional angle, χ. While (1) gives global changes in conformation, (2) and (3) give local (or small-step) changes. In all our simulations, the fraction of sidechain moves was held fixed at 59%. In our equilibrium simulations, the remaining 41 % of moves were divided between pivot and BGS. The BGS move was not used at the highest simulated temperatures, where the pivot move is highly effective at enhancing conformational sampling [53].

For the isolated CTD, the thermodynamic behavior (for each strength λ of the structure-based potential) was determined using 5 or 10 independent simulated tempering runs of each 3×10^7 MC cycles, where a cycle is 239 elementary MC steps (the number of turnable ϕ, ψ or χ angles in the protein chain). An exception is $\lambda = 0.30$, for which 20 independent runs were carried out. For the full-length RfaH, the thermodynamic behavior was determined using 20 independent runs of at least 1×10^7 MC cycles, where a cycle is 740 elementary steps. In the full-length simulations, the backbone chain corresponding to positions 1-100 (i.e., the ordered region of the NTD) was held fixed in its initial (native) conformation by disallowing BGS and pivots in this region. All sidechains were allowed to move.

4.2.3 Small-step "kinetic" Monte Carlo simulations

Our "kinetic" MC runs of the isolated CTD differed from the simulated tempering runs in two respects: (1) global (i.e. pivot) moves were turned off; and (2) the temperature was held fixed. Two different starting conformations were used. Our folding simulations were started from a rigid extended conformation, obtained by setting all backbone dihedral angles to $\phi = 180°$ and $\psi = 180°$, followed by a few MC cycles of relaxation at high temperature to remove any atom-atom clashes. This starting structure was basically a highly extended, rod-like conformation. Our fold switching simulations were started from the regularized (see section 2.4) α-helical hairpin structure (PDB id 2oug).

4.2.4 Representative structures

As representative structures of the two RfaH folds, we used the X-ray structure of the full-length RfaH with PDB id 2oug [22] and the NMR derived structure of the isolated CTD with PDB id 2lcl [21]. The missing residues in 2oug, including the linker region (101-114) and flexible C-terminal tail (157-162), were added back using a homology modeling tool [54], as described previously [33]. The 2lcl structure of the isolated CTD was truncated to retain the ordered part (residues 113-162).

Following our previous work [33], we subjected our representative structures to PROFASI "regularization", which is a procedure to find an energetically relaxed model conformation that closely approximates a given experimental structure. Regularization is necessary because model conformations in our physics-based model [51] are subject to some geometrical constraints, e.g., fixed bond lengths and bond angles, (see Section 2.1) that in general are not satisfied by experimental structures. The regularized model conformations we obtained for 2oug and 2lcl have RMSD values of 1.8 Å and 1.6 Å, respectively, taken over all non-hydrogen atoms in the protein chain.

4.2.5 Native contact maps

We obtained native contact sets, $C = \{ij|$ if residues i and j are in contact$\}$, for the two regularized model structures (derived from 2oug and 2lcl) by submitting them to the SMOG webserver [55] with the coarse-graining option "All-atom Calpha" and otherwise default parameters. At this level of coarse-graining, two residues i and j are considered to form a contact ij if any atom-atom contact is present between i and j according to the shadow map algorithm [56]. For the β-barrel structure (2lcl), we obtained this way a set of 129 native contacts, C^{β}. For the α-helical hairpin structure (2oug), retaining only contacts for which both i and j belong to the segment 115-156, we obtained a set of 69 native contacts, C^{α}. The C^{α} and C^{β} contact maps are shown in Fig. (4.1). In the simulations of the full-length RfaH using the dual-basin SBM, C^{α} was replaced with the contact map $C^{\alpha}_{\mathrm{RfaH}}$, shown in Fig. (4.4.B), with 157 native contacts. $C^{\alpha}_{\mathrm{RfaH}}$ includes all contacts in C^{α} and, in addition, all the NTD-CTD inter-domain native contacts, i.e., with residue i belonging to the segment 1-100 and residue j belonging to the segment 113-156.

4.2.6 Observables

The progress variables, Q_{α} and Q_{β}, are the fraction of the native contacts formed in C^{α} and C^{β}, respectively (see Fig. (4.1)). In determining Q_{α} and Q_{β}, a contact between residues i and j is considered formed if $r_{ij} < 1.2r^{0}_{ij}$, where the distances r_{ij} and r^{0}_{ij} are defined in the text following Eq. (4.2) in section 3.1. The root-mean-square deviations, RMSD_{α} and RMSD_{β}, are taken with respect to two representatives (experimental) structures of the all-α and all-β CTD folds, respectively, and determined over C_α atoms. Secondary structure assignments, used for the calculation of α-helix content, were performed using STRIDE [57].

4.2.7 Correction term for the dual-basin SBM

In our dual-basin SBM, attractions are included between any pair of residues that form a contact either in the all-α fold (C^α) or in the all-β fold (C^β). However, there are 13 common contacts in C^α and C^β, as shown in Fig. (4.1). To avoid the unreasonable situation that these 13 contacts have strengths up to twice that of all other contacts, we use the convention that, for each of these 13 contacts, only the most favorable of the two possible contact energies contributes towards the total energy. This is achieved by a correction term, included in the potential energy function $E^{(2)}$ of the dual-basin SBM (see section 3.3), given by

$$E_{\mathrm{corr}}(\lambda^\alpha, C^\alpha; \lambda^\beta, C^\beta) = -\epsilon \sum_{ij} \min[\lambda^\alpha g(r_{ij}, r_{ij}^{\alpha,0}), \lambda^\beta g(r_{ij}, r_{ij}^{\beta,0})] \, , \qquad (4.1)$$

where the sum goes over the 13 contacts ij that are present in both C^α and C^β. In this equation, the distance r_{ij} and the function g are defined in the text following Eq. (4.2) in section 3.1, and the strengths, λ^α and λ^β, are defined in the text following Eq. (4.5) in section 3.3. The reference distances $r_{ij}^{\alpha,0}$ and $r_{ij}^{\beta,0}$ are the values of r_{ij} determined for the all-α and all-β (regularized) experimental structures of the CTD, respectively.

4.3 Results and Discussion

4.3.1 Hybrid sequence-structure-based simulation approach

An ideal computational model for simulating metamorphic proteins would be able to capture the large-scale structural transitions of these proteins and, at the same time, be computationally tractable. However, as mentioned in Introduction, constructing such a model that relies entirely on sequence information (i.e., a physics-based model) is challenging at present. We therefore pursue here a hybrid approach in which an implicit solvent physics-based model is augmented with an SBM. Similar approaches have been tested before [58–60], although not for the system under study here. Specifically, we construct the potential energy function of the hybrid model as a linear combination of the physics-based and the structure-based potentials. Our aim is to pick the relative strength of the SBM as small as possible, while requiring that the hybrid model as a whole exhibits a thermodynamic behavior in agreement with available experimental data.

As our starting point, we use the protein model developed in Ref. [51]. This model is based on an effective all-atom (solvent free) potential energy function, which includes terms for the major driving forces of protein structure formation, including hydrophobic and electrostatic attractions and hydrogen bonding. Parametrization of this model was done by requiring that a set of 17 different amino acid sequences exhibit global free energy minima corresponding to their respective experimentally determined native structures. Interestingly, this “top-down” approach to parameterization also leads to thermodynamic behaviors, such as melting temperatures, that for many of the sequences were in quantitative agreement with experimental data [51].

With this model in mind as a baseline, we formulate a structure-based, or Gō-like [44], potential, which provides an energetic bias towards a single native structure encoded as the set of residue-residue contacts present in the structure, C (i.e., the contact map of the structure). We pick this energy term to have the form

$$E_{\mathrm{SB}}(C) = -\epsilon \sum_{ij \in C} g(r_{ij}, r_{ij}^0) \,, \tag{4.2}$$

where the sum goes over all contacts ij in C, ϵ is the energy unit of our baseline model [51], and r_{ij} is the C_β-C_β distance between the residues at positions i and j. In the case of a glycine residue at position i or position j, r_{ij} is instead taken to be the C_α-C_β distance (if glycine at i only), C_β-C_α distance (if glycine at j only) or C_α-C_α distance (if glycines at both i and j). The reference distance r_{ij}^0 is found by determining r_{ij} for the (native) structure used to obtain the contact map, C. The quantity $g(r_{ij}, r_{ij}^0)$, where

$$g(r, r^0) = e^{-(r-r^0)^2/2\xi^2} \,, \tag{4.3}$$

measures the extent to which the native contact ij is correctly formed. The width of the Gaussian function g is controlled by the parameter ξ, which we set to $\xi = 1$ Å. Previous work [58–60] that have similarly combined physics-based and SBMs, have exclusively formulated their SBMs using C_α-C_α distances. Here we primarily use C_β-atoms for quantifying native contact formation because we expect that C_β-C_β distances, through the function $g(r_{ij}, r_{ij}^0)$, to provide a higher specificity towards the native structure.

In the following, we apply Eq. 4.2 to the two different folds of the RfaH CTD. This gives two structure-based potentials, $E_{\mathrm{SB}}(C^\alpha)$ and $E_{\mathrm{SB}}(C^\beta)$, where C^α and C^β are the contact maps of the two experimentally determined structures displayed in Fig. (4.1). In section 4.3.2, we combine our physics-based model with each term separately to create two hybrid models in which the SBM has a single basin of attraction. In section 4.3.3, we combine these potentials into a hybrid model with a dual-basin SBM.

4.3.2 Single-basin SBM

Combining the SBM defined by Eq. 4.2 with our physics-based model results in a potential energy function of the form,

$$E^{(1)} = E^{(0)} + \lambda E_{\mathrm{SB}}(C)\,, \tag{4.4}$$

where $E^{(0)}$ is the energy function defined in Ref [51] and λ is the strength of the structure-based term, which should be seen as a free parameter in this approach.

We apply Eq. (4.4) to the CTD of RfaH, with either $C = C^{\alpha}$ or $C = C^{\beta}$ (see Fig. (4.1)). Quite generally, we expect that increasing the relative strength of the SBM should increase the "nativeness" of the generated conformational ensemble. In other words, increasing λ should increase Q_α in the case $C = C^{\alpha}$ and increase Q_β in the case $C = C^{\beta}$, where Q_α and Q_β are the fraction of native contacts formed with regards to the α-helical hairpin and β-barrel, respectively (see Methods). We indeed observe such a trend, as shown in Fig. (4.2). However, beyond this general trend, we find stark qualitative differences. The term $E_{\mathrm{SB}}(C^{\beta})$ has a large effect on the stability of the β-barrel fold. For large enough λ, the melting curves become characterized by a sharp decrease in Q_β with increasing temperature, indicating a cooperative transition. Indeed, the height of the heat capacity peak ($C_{\mathrm{v}}^{\mathrm{max}}$) increases with λ (see Fig. (4.2.C)). Moreover, increasing λ also leads to a shift in the folding transition to higher temperatures. As seen from Fig. (4.3.A), this shift is similar, but not identical, whether the transition is characterized by the midpoint temperature, T_{m}, found from fitting the Q_{β} melting curves to a two-state model (see Fig. (4.2)), or by the temperature T^{*}, defined by $C_{\mathrm{v}}^{\mathrm{max}}$. The single-basin SBM applied to the α-helical hairpin fold ($C = C^{\alpha}$) has a much weaker effect on the thermodynamic behavior. The fraction of native contacts, Q_{α}, does not reach 0.5, even at much lower temperatures. Strikingly, $C_{\mathrm{v}}^{\mathrm{max}}$ decreases with increasing λ in this case, highlighting the lack of a folding transition in the $C = C^{\alpha}$ case.

The degree of cooperativity in the folding of the β-barrel can be roughly quantified by the difference $\Delta T = T_{\mathrm{m}} - T^{*}$. For a strongly cooperative (all-or-none) folding transition, the transition midpoint should be basically independent of which precise structural aspect is used to characterize the transition, i.e., ΔT should be small. Indeed, we find that ΔT is small for large λ, as shown in Fig. (4.3.B). Substantial deviations appear only at $\lambda \lesssim 0.25$. Based on these results, we pick $\lambda = 0.30$ as a reasonable strength of the structure-based term, making the assumption that the folding of the CTD is cooperative. Practically, we designate the simulated temperature closest to the heat capacity peak as the folding temperature, T_{f}. Because, for $\lambda = 0.30$, ΔT is very small we have $T_{\mathrm{f}} \approx T_{\mathrm{m}} \approx T^{*}$ (see Fig. (4.2.C)).

It must be noted that the potential energy function of the hybrid model, $E^{(1)}$, does not preserve the energy scale of the underlying physics-based model [51]. As seen from Eq. 4.4, the scale of $E^{(1)}$ is controlled by the strength λ and the number of contacts in the contact map, C. Therefore, the physical units for temperature and energy established in Ref. [51] no longer holds (for $\lambda > 0$). A link between real and

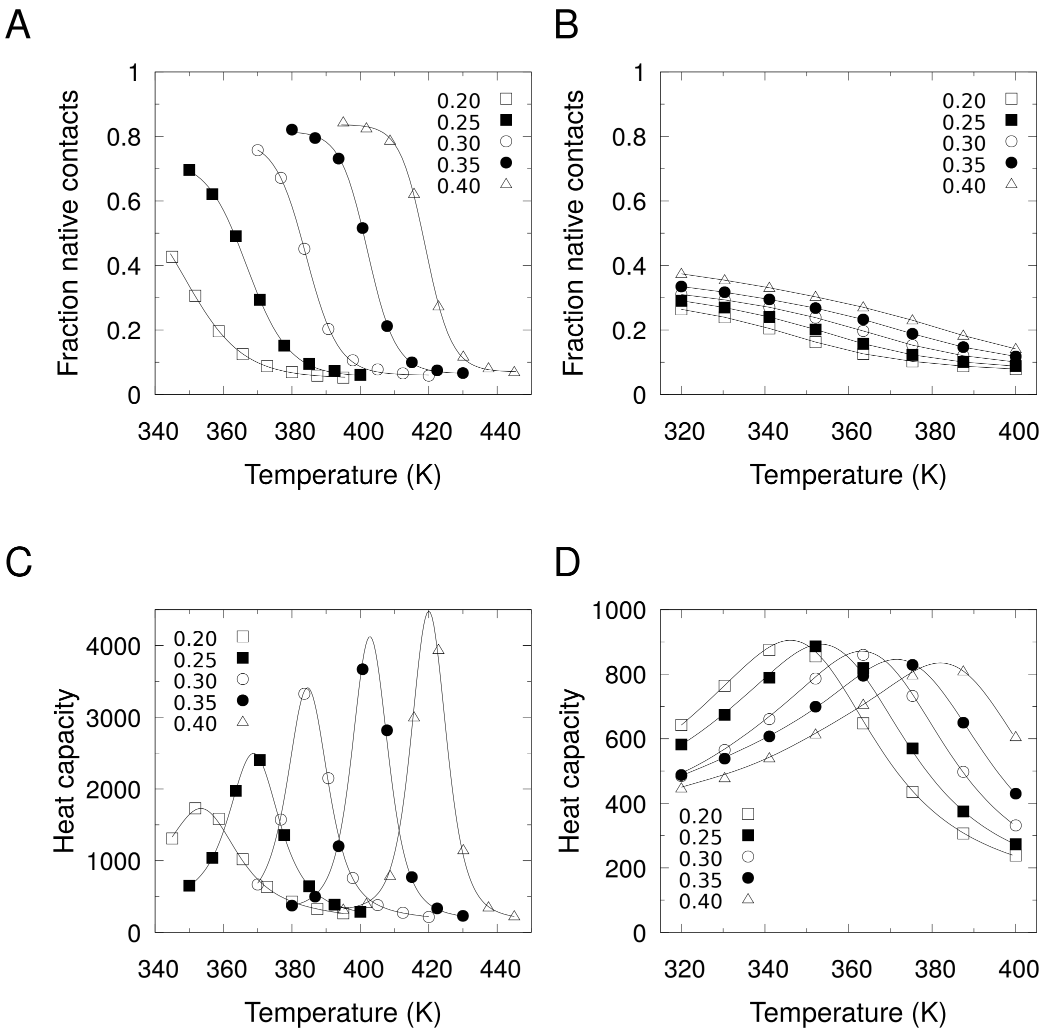

Figure 4.2: Equilibrium behavior of the hybrid model with a single-basin SBM. Shown are results obtained with the energy function in Eq. 4.4 with (A, C) a β-barrel SBM ($C = C^{\beta}$), (B, D) an α-helical hairpin SBM ($C = C^{\alpha}$) and a range of strengths $\lambda = 0.20$–0.40. As functions of the temperature: (A) Q_{β}, (B) Q_{α} and (C, D) heat capacity $\mathrm{C_v}/k_\mathrm{B}$, where $\mathrm{C_v} = \left(\left\langle E^2 \right\rangle - \left\langle E \right\rangle^2\right)/k_\mathrm{B}T^2$, E is the energy from Eq. 4.4, and k_B is Boltzmann's constant. Solid curves in (A) are fits to the two-state equation $\langle Q \rangle = (Q^\mathrm{U} + Q^\mathrm{N}K)/(1+K)$, where $K = \exp\left(-\Delta E(1/k_\mathrm{B}T - 1/k_\mathrm{B}T_\mathrm{m}\right)$ and Q^U, Q^N, ΔE and T_m (midpoint temperature) are fit parameters. Solid lines between points in (B) are drawn to guide the eye. Solid curves in (C) and (D) are obtained using multiple-histogram reweighting [61].

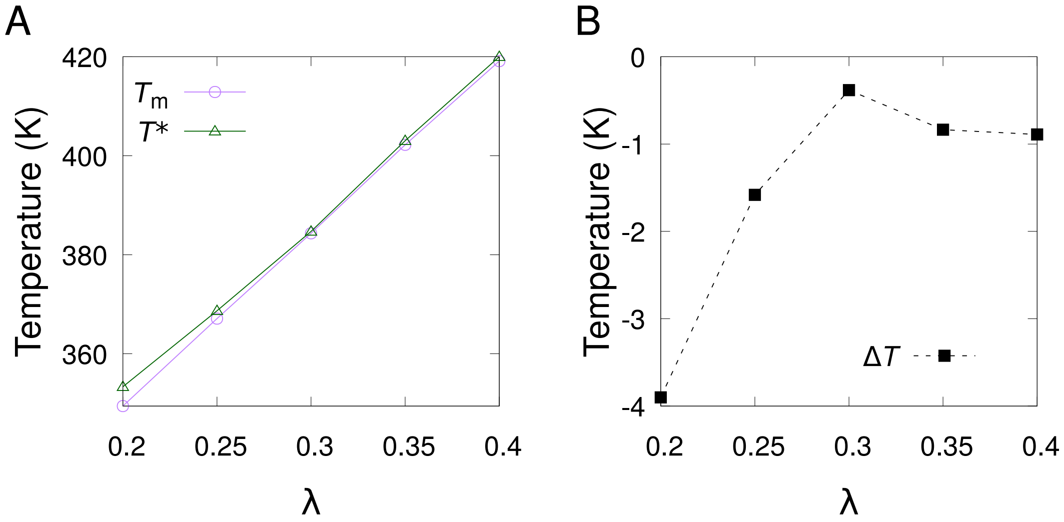

Figure 4.3: Impact of the SBM strength λ on the folding transition. Shown are results for our hybrid model with the single-basin β-barrel SBM ($C = C^{\beta}$). (A) T_m and T^* as functions of λ, where T_m is the midpoint temperature, obtained from the two-state fits in Fig. (4.2.A), and T^* is the C_v maximum temperature, obtained from Fig. (4.2.C). (B) The difference $\Delta T = T_\mathrm{m} - T^*$ as function of λ. Dashed lines between points are drawn to guide the eye.

model units can, in principle, be re-established separately for each specific protein to which $E^{(1)}$ is applied and the particular value of λ used (e.g., by matching the folding temperatures found in simulations and experiments). In this work, however, we use the temperature scale arising in our hybrid model, i.e., without re-scaling, while at the same time making sure that all comparisons with experiments are performed under at least roughly equivalent conditions (e.g. folding or unfolding conditions).

4.3.3 Dual-basin SBM

We now combine the two structure-based terms, $E_{\text{SB}}(C^{\alpha})$ and $E_{\text{SB}}(C^{\beta})$, into a dual-basin SBM. Following our general approach, the energy function of the resulting hybrid model becomes

$$E^{(2)} = E^{(0)} + \lambda^{\alpha} E_{\text{SB}}(C^{\alpha}) + \lambda^{\beta} E_{\text{SB}}(C^{\beta}) - E_{\text{corr}}(\lambda^{\alpha}, C^{\alpha}; \lambda^{\beta}, C^{\beta})\,, \tag{4.5}$$

where the strengths of the two structure-based terms, λ^{α} and λ^{β}, are selected to be $\lambda^{\alpha} = \lambda^{\beta} = \lambda = 0.30$ based on the results of the previous section. The last term, E_{corr}, is a correction necessary to avoid double counting common contacts in C^{α} and C^{β} (see Fig. (4.1)), which receive contributions from both $E_{\text{SB}}(C^{\alpha})$ and $E_{\text{SB}}(C^{\beta})$. The role of E_{corr} is, for each contact present in both C^{α} and C^{β}, to eliminate the weaker of the two interactions (see Methods).

We apply Eq. 4.5 to the CTD as an isolated fragment and to the full-length RfaH. For the isolated CTD, we find that Q_{β} increases with decreasing temperature, as shown in Fig. (4.4). In other words, despite the dual-basin nature of the SBM in our hybrid model, the isolated CTD folds into a stable β-barrel at low T. As a comparison, we show also in Fig. (4.4.A) the results for the single-basin SBM, which, interestingly, gives a higher stability compared to the dual-basin SBM. The simulations of full-length RfaH are carried out in the following way. First, the contact map C^{α} in Eq. 4.5 is replaced with $C^{\alpha}_{\mathrm{RfaH}}$, which in addition to all contacts in C^{α} also includes NTD-CTD interdomain contacts, as shown in Fig. (4.4.B). Second, for computational reasons, we keep the NTD backbone fixed in its native conformation while the linker and CTD regions are free to move. All sidechains are also left free. From these simulations we find that, within the context of the full-length RfaH, the CTD switches into an α-helix rich state, as shown by the increasing Q_{α} with decreasing temperature in Fig. (4.4.A).

The temperature dependence of Q_{α} for the full-length RfaH, like the temperature dependence of Q_{β} for the isolated CTD, is basically sigmoidal in shape. However, at low T, Q_{α} for full-length RfaH converges to a lower fraction of native contacts than Q_{β} for isolated CTD (cf. Fig. (4.4.A) (left) and Fig. (4.4.A) (right)), suggesting an incompletely formed α-helical hairpin. To determine if this is the case, we plot the α-helix content as a function of sequence position at a relatively low T (see Fig. (4.4.C)). We find that helix α_4 (residues 116—130) is significantly less structured than helix α_5 (residues 134—156) in our model. According to the NMR experiments of Burmann et al. [21] on free RfaH, the segment Val116-Gly121 displays chemical shifts that are more in line with a random coil state than an α-helix, suggesting that the 6 N-terminal residues of α_4 might be mainly disordered in the solution state. This segment is indeed particularly poorly structured in our model (see Fig. (4.4.C)). To correct for this partial lack of structure, we construct an alternative nativeness measure, $Q_{\alpha}^{(49)}$, which is the same as Q_{α} except that all contacts involving 116-121 are ignored. Indeed, $Q_\alpha^{(49)}$ is consistently higher than Q_α across all temperatures (at most 23% higher, seen at the lowest studied T) indicating that the α-helical hairpin is formed at low T, with the exception of the N-terminal part of α_4, which, in line with experiments, remain largely unstructured. We note also that poorly formed α-structure found in the 116-121 region as well as in the C-terminal end of $\alpha 5$ (residues 150-156), is consistent with the hydrogen/deuterium exchange mass spectrometry (HDXMS) experiments of Galaz-Davison et al. [25].

In our simulations of full-length RfaH, we used the 2oug structure of free RfaH to represent the NTD. Subsequent structural analysis of the RfaH-DNA complex (PDB id 5ond; PMID: 29741479) [35] has revealed a one-residue misplacement of the last 17 NTD residues, i.e., the region 84-100 (Irina Artsimovitch, personal communication). The 84-100 region wholly includes helix $\alpha 3$, which is located at the NTD-CTD interface and interacts with both $\alpha 4$ and $\alpha 5$. The potential misplacement in 2oug could be corrected for by switching to 5ond as the representative structure of the NTD. The impact on our results, if any, would presumably be to increase the propensity for α-helix structure in $\alpha 4$ and $\alpha 5$, due to more favorable energetic interactions with $\alpha 3$ captured by our physics-based model. Additional simulations would be needed to confirm this, however. We emphasize that the choice of structural representation of the NTD has no effect on our simulations of the isolated CTD.

4.3.4 Energy landscape with a single dominant funnel

Having showed that our hybrid model with a dual-basin SBM captures the basic thermodynamic behavior of the RfaH CTD, i.e., its all-α-to-all-β switch in global

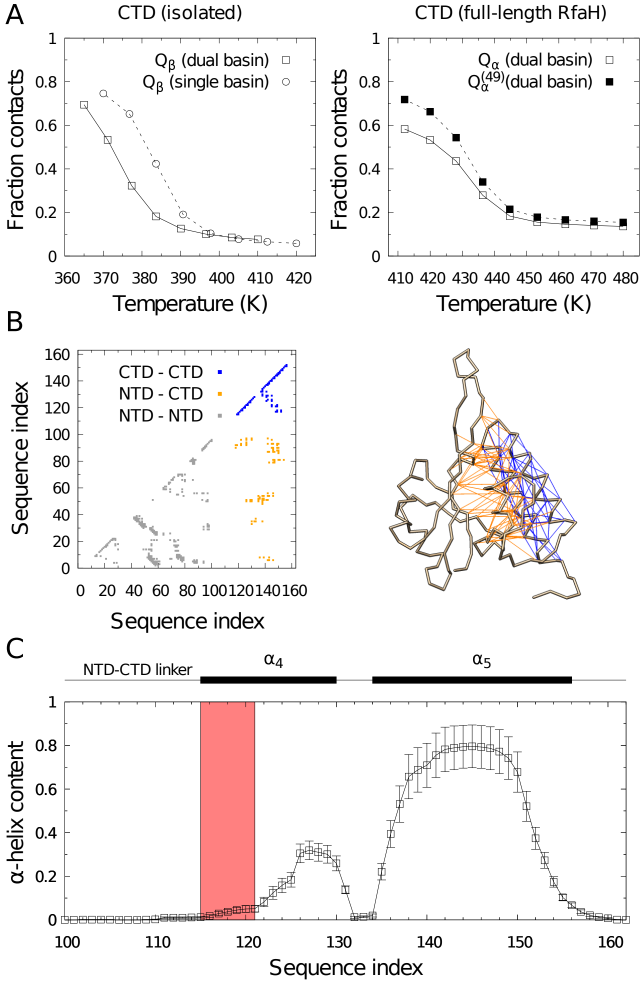

Figure 4.4: Equilibrium behavior of the hybrid model with a dual-basin SBM. (A) Temperature dependence of the fraction of native contacts (Q_β or Q_α) obtained from simulations of the isolated CTD (left) or full-length RfaH (right). For comparison, data for the single-basin β-barrel SBM is re-drawn from Fig. (4.2.A). (B) Native contact map (left) and structure in cartoon representation (right) of full-length RfaH (PDB id 2oug). Intra-CTD, inter-domain and other native contacts are shown in blue, yellow and gray, respectively. Blue and yellow contacts make up the contact set, $C^{\alpha}_{\mathrm{RfaH}}$. (C) The average α-helix content as function of sequence position for full-length RfaH, taken at 420 K (the NTD region, fixed in simulations, is not shown). $Q_\alpha^{(49)}$ is determined as Q_α but taken over a reduced set of 49 contacts that excludes contacts in the segment Val116–Gly121 (orange shaded region). All results are obtained with $\lambda^\alpha = \lambda^\beta = 0.30$.

free energy minimum upon excision from the rest of the protein, we turn to the folding of this domain. We first determine the free energy surface $F(Q_\alpha, Q_\beta)$ at the folding temperature, T_f, as shown in Fig. (4.5.A). At this temperature, one might expect three distinct free energy minima corresponding to the unfolded state, which must be populated at T_f, and two ordered states, the all-α fold and all-β folds. However, there are only two major minima in the Q_α-Q_β plane: (1) a low-Q_α, low-Q_β minimum, i.e., the unfolded state; and (2) a low-Q_α, high-Q_β, minimum, i.e., the β-barrel state. Although centered around $Q_\alpha \approx$ 0.2-0.4, the basin of attraction (1) includes some states with free energies within $\approx$ 4-5$k_\mathrm{B}T$ of the basin minimum and relatively high native contact fractions, $Q_\alpha \approx 0.6$, and higher still in terms of $Q_\alpha^{(49)}$ (see Fig. (4.5.A), inset). These states do not represent a separate funnel towards the all-α state, however, but are rather characteristics of the unfolded state ensemble (see section 3.5). The absence of low-energy states competing with the β-barrel fold is clear from the free energy surface $F(E, \mathrm{RMSD}_\beta)$ in Fig. (4.5.B). Competing states also do not appear under folding conditions (see Fig. S2). At $T < T_\mathrm{f}$, the population of the unfolded state decreases as ordered states with lower energies are increasingly favored. As a result, the energy landscape becomes dominated by a single funnel toward the β-barrel, as can be seen from the free energy profiles $F(Q_\alpha)$ and $F(Q_\beta)$ in Fig. (4.6).

Recent computational studies of the RfaH CTD have found energy landscapes with clear two-funnel character [32, 38], in apparent contradiction with our results. These studies relied on molecular dynamics simulations with either implicit or explicit water in combination with advanced sampling and analysis techniques. For example,

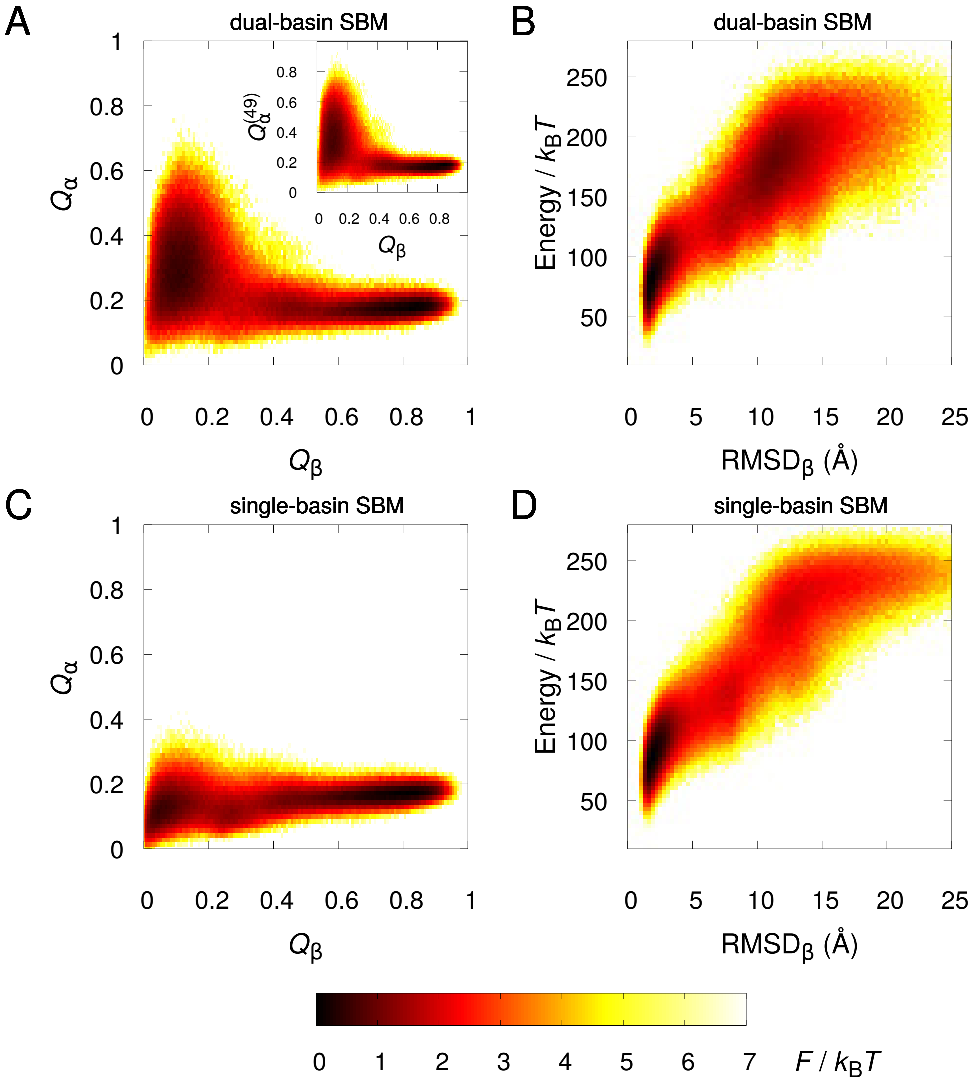

Figure 4.5: Free energy surfaces at the folding transition: single-basin SBM vs dual-basin SBM. Free energy surfaces $F(X_1, X_2) = -k_BT \ln P(X_1, X_2)$, with (A, B) $X_1 = Q_\alpha$ and $X_2 = Q_\beta$ or (B, D) X_1 equal to the total energy and $X_2 = \mathrm{RMSD}_\beta$. The probability distributions $P(X_1, X_2)$ are obtained at the respective T_fs of the models, i.e., (nominally) 377 K for the dual-basin SBM and 384 K for the single-basin (β-barrel) SBM.

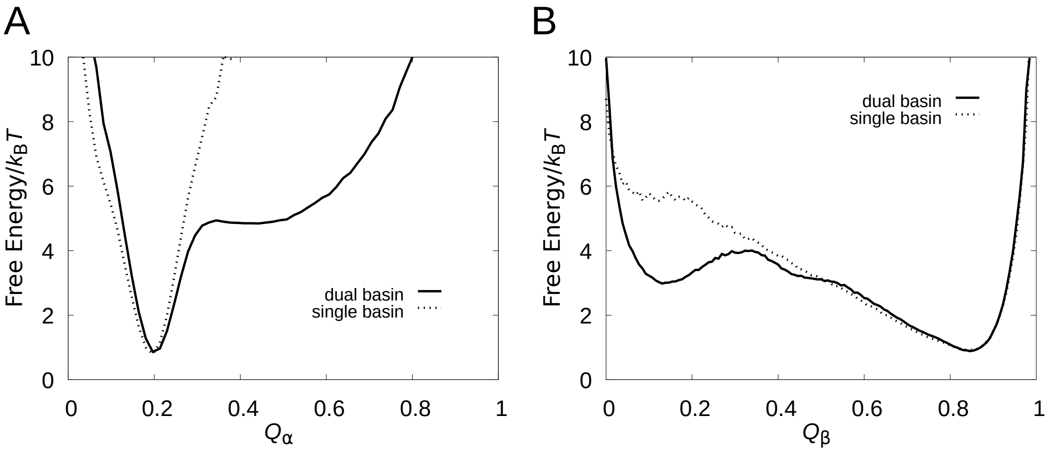

Figure 4.6: Free energy profiles under folding conditions. Free energy as function of (A) Q_α and (B) Q_β for our hybrid models with dual-basin (solid curves) or single-basin (dashed curves) SBMs, taken at $T \approx 0.96 T_\mathrm{f}$.

Bernhardt et al. [38] achieved enhanced sampling using a Hamilton replica-exchange method that couples their physical model to a Gō potential, which seeds the sampling of conformational space. It is possible that these types of techniques are able to detect fine features of the energy landscape that are not apparent in our calculations. However, the lack of a major competing minimum in the energy landscape of the isolated CTD is consistent with [^{1}H,^{15}N]-HSQC NMR data [21]. Upon protease-induced cleavage of the NTD-CTD linker, thus releasing the CTD into the solution, Burmann et al. [21] observed the emergence of the spectra from the β-barrel fold and a complete disappearance of the spectra from the α-helical hairpin. Although the α-helical hairpin is clearly encoded in the amino acid sequence of the CTD, these experimental results limit the possible “depth” of this funnel and hence its significance in comparison with the large number of other local minima that are present in any protein energy landscape.

4.3.5 Impact of encoding for two folds on the unfolded state

To delineate the impact of the fold switching capability of the CTD on its folding, we compare the results we obtained using the dual-basin SBM with those using the single-basin SBM (applied to the β-barrel contact map, i.e., $C = C^{\beta}$). Both hybrid models fold the CTD into a stable β-barrel (see Fig. (4.4.A)), but only the dual-basin SBM exhibits proper fold switching. We therefore reason that differences between the two folding energy landscapes can be attributed to the unique fold switching capability of the CTD.

While the differences between the two free energy surfaces $F(Q_{\alpha}, Q_{\beta})$ at T_{f} (see Fig. (4.5.A) and (4.5.C), obtained for the single- and dual-basin SBMs, are not large, it is interesting to examine them in detail. We find that, for the single-basin SBM, the unfolded state is characterized by a rather narrow minimum, especially in the Q_{α} direction. For the dual-basin SBM, the unfolded state is much broader in the this direction, indicating relatively large fluctuations in Q_{α}. Because many of the native contacts in C^{α} are local, including i, i+4-contacts (see Fig. (4.1)), these fluctuations suggest the presence of α-helix structure in the unfolded state. Indeed, we find that the α-helix content in the unfolded state, as defined by $Q_{\beta} < 0.5$, is significantly higher for the dual-basin SBM (12.3 ± 0.7 %) than for the single-basin SBM (1.4 ± 0.1 %), as quantified by STRIDE [57]. Hence, the main effect of encoding for the α-helical hairpin within the hybrid model with a dual-basin SBM, is the appearance of residual α-helical structure in the unfolded state, U.

In states other than U, the difference between the two hybrid models is much smaller, which is most easily seen from the free energy profiles in Fig. (4.6). At large Q_β, the shapes of the two $F(Q_\beta)$ curves are remarkably similar. This similarity is likely a consequence of the high structural dissimilarity of the two folds; contacts unique to the α-helical fold are either highly unstable or otherwise impossible to form due to topological constraints, once folding has progressed towards the β-barrel beyond a certain point (roughly $Q_\beta \approx 0.5$).

4.3.6 Is there an activation barrier to the all-α-to-all-β fold switch?

Although we find in our hybrid approach, with a dual-basin SBM, a dominant funnel to the β-barrel fold, conformations sufficiently close to the α-helical hairpin will, by construction, be energetically biased towards this attractor. Therefore, a conformation prepared in this all-α state might still need to overcome an activation barrier before it can proceed downhill the energy landscape towards the β-barrel native state. Such a situation might occur *in vivo* when RfaH binds to RNA polymerase, which triggers the CTD to dissociate from the NTD [24].

In order to explore the potential barrier of fold switching in the all-α-to-all-β direction, we carry out small-step "kinetic" MC simulations (see Methods). We consider two different starting points: (1) the α-helical hairpin and (2) an extended, open conformation. The temperature is held fixed at $T = 365$ K, at which the β-barrel is the thermodynamically dominant state (see Fig. (4.4.A)). Hence, both starting points should eventually transform into the β-barrel fold, although the timescales of the two transformations could differ. These two sets of simulations started from points (1) and (2) probe the fold switching and folding of the RfaH CTD, respectively.

In Fig. (4.7.A) and B, we show the relaxation of the α-helix content, $\langle\alpha\rangle$, and the fraction of native contacts, $\langle Q_\beta\rangle$, where $\langle\rangle$ indicates an average taken over 100 independent runs, towards their respective equilibrium values at this T , i.e., $\langle\alpha\rangle_{\text{equil}} = 0.02 \pm 0.01$ and $\langle Q_\beta\rangle_{\text{equil}} = 0.71 \pm 0.01$. In terms of Q_β, the folding simulations, which start at $Q_\beta = 0$, rapidly "overtake" the fold switching simulations. This behaviour indicates some degree of kinetic trapping early in the fold switching process. However, the associated barrier cannot be large because the two sets of simulations become statistically indistinguishable at around 1-2 million MC cycles, when $\langle Q_\beta\rangle$ is still far from its equilibrium value. A convergence of the two sets of simulations occurs also in terms of $\langle\alpha\rangle$ at roughly the same point.

Interestingly, and perhaps surprisingly, the folding simulations exhibit an initial increase in α-helix structure content to a maximum of $\langle\alpha\rangle \approx 0.20$, and thereafter a much slower decrease following closely the trend of the fold switching simulations. This gradual decrease comes from the conversion of more and more of the trajectories into the β-barrel fold. Taken together, these simulations suggest a fold-switching process from the α-helical hairpin state (N_α) to the β-barrel state (N_β) proceeding as $\text{N}_\alpha \rightarrow \text{U} \rightarrow \text{N}_\beta$, where U is the unfolded state.

This basic scheme is confirmed when examining in more structural detail individual fold switching events. Figure (4.8) shows RMSD_α and RMSD_β as functions of MC time for a typical fold switching trajectory, where RMSD_α and RMSD_β are the root-mean-square deviations taken with respect to the representative all-α and all-β CTD structures, respectively. After only a few MC cycles, RMSD_α increases rapidly, and the chain settles into an intermediate state with large conformational fluctuations.

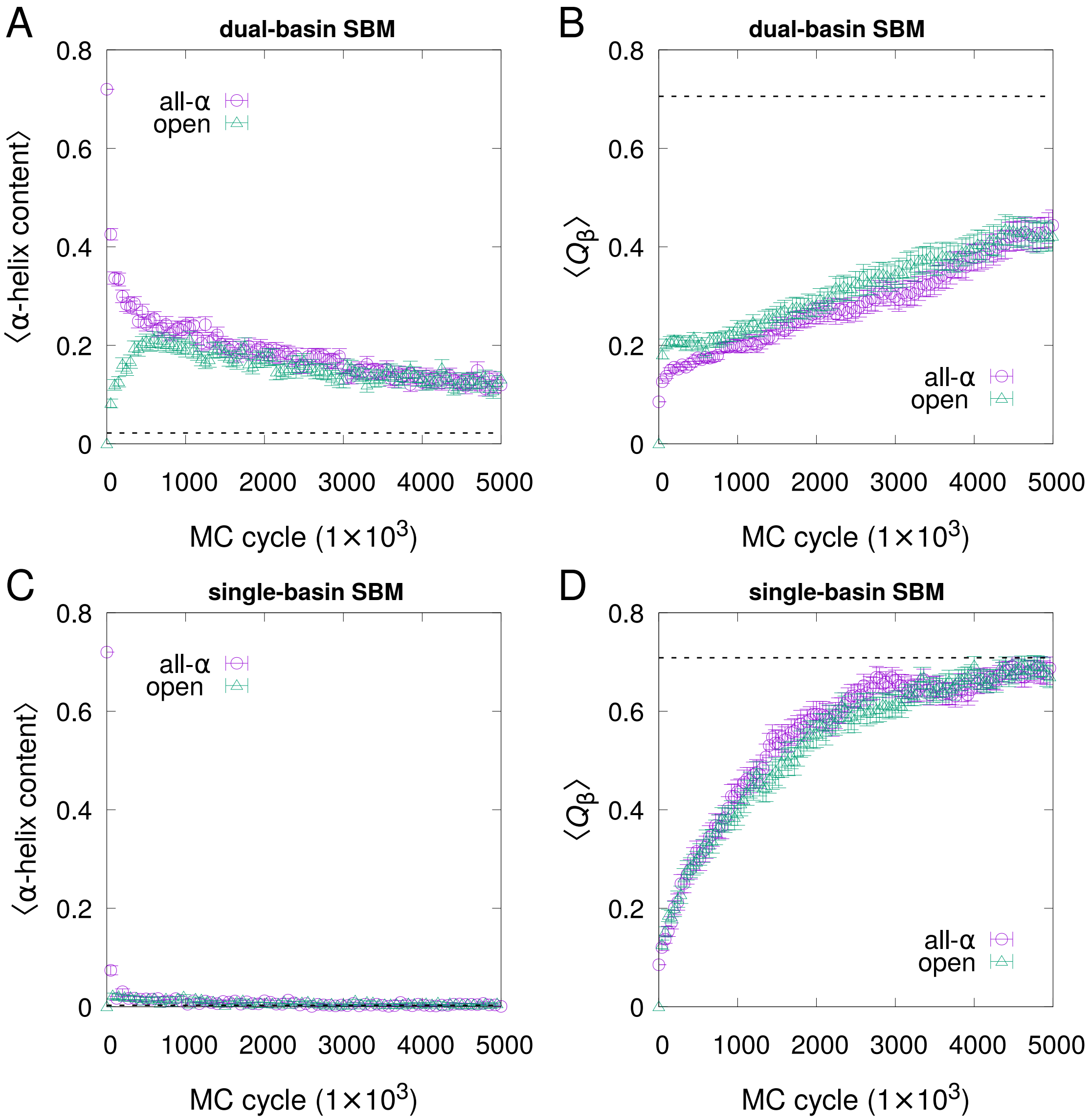

Figure 4.7: Folding and fold switching of the RfaH CTD. Evolution of the α-helix content, $\langle\alpha\rangle$, and the fraction of native contacts, $\langle Q_\beta\rangle$, in small-step MC simulations started in either the α-helical hairpin fold (purple circles) or in an open (rod-like) chain conformation (green triangles). The average $\langle\rangle$ is taken over 100 independent trajectories. Results are shown for our hybrid model (A and B) with the dual-basin SBM at $T = 365$ K and (C and D) with the single-basin SBM at $T = 375$ K. Equilibrium values for the respective models and temperatures are indicated with dashed horizontal lines. Error bars show standard errors calculated over the 100 trajectories.

Residual α-helical structure are found upon inspection of conformations in this state, as shown in Fig. (4.8). Eventually, the chain switches abruptly to an all-β state, characterized by low RMSD_β values and much smaller fluctuations. Overall, we find that, of all the trajectories that switch folds within the simulation time (47 out of 100), none proceeds directly from the all-α state to the all-β state but instead proceed via an intermediate state. Because of the presence of residual α-helix structure and large structural fluctuations, we identify this intermediate state with the unfolded state, U.

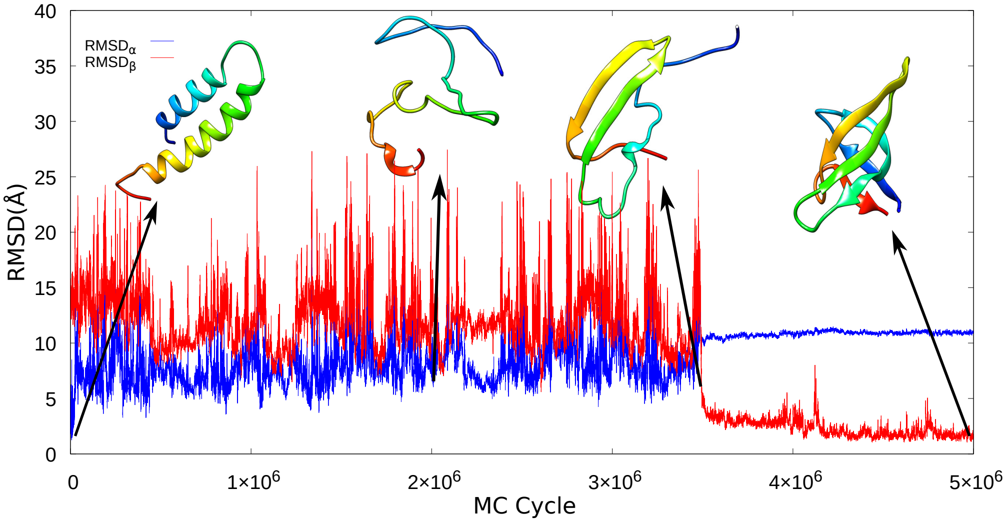

Figure 4.8: Example of a fold switching trajectory. Evolution of the root-mean-square deviation determined with respect to the α-helical hairpin structure (RMSD_α, PDB id 2oug) or the β-barrel structure (RMSD_β; PDB id 2lcl), for one of the 100 dual-basin SBM fold switching runs in Fig. (4.7).

We also examine the question of what underpins the formation of transient α-helix structure during the folding of the CTD. Although our hybrid model is based in part on structure, ultimately it is (in Nature) the sequence of the CTD that determines its conformational behavior. In this regard, it is interesting that the CTD of RfaH and the CTD of RfaH's fold-switch incapable paralog NusG have a low sequence identity (16%, based on the sequence alignment given Shi et al. [23]). To explore the extent to which our physics-based model, on its own, is able to capture the sequence-encoded conformational preferences of the RfaH CTD, we apply this model (i.e., our hybrid model with $\lambda = 0$) to the chain segments corresponding to helix $\alpha4$ (residues 116–130; VIITEGAFEGFQAIF) and helix $\alpha5$ (residues 135–155; GEARSMLLLNLINKEIKHSVK) of the CTD all-α structure. We find α-helical structure formation for the $\alpha5$ segment but not for the $\alpha4$ segment (see Fig. S3). At 25°C, the average α-helical contents are 0.03 ± 0.01 for $\alpha4$ and 0.43 ± 0.03 for $\alpha5$. Parts of the $\alpha5$ segment must therefore be converted to β-structure through cooperative effects during folding. These results for these segments could be tested by biophysical characterization using, e.g., circular dichroism. However, in our hybrid model, at the chosen strength $\lambda = 0.30$, the SBM part of the energy function likely plays a major role in determining conformational preferences. This can be seen from Fig. (4.7.C), which shows that folding simulations of the CTD obtained with the single-basin SBM (i.e., no structure-based bias towards the all-α fold) exhibit very low amounts of α-helical structure.

Finally, we note an interesting difference in the relaxation behavior between the single- and dual-basin SBMs (cf. Fig. (4.7.B) and Fig. (4.7.D)), namely the slower relaxation exhibited by the dual-basin SBM. This difference indicates a higher degree of "roughness" in the energy landscape of the dual-basin SBM, presumably as a consequence of the conflicting conformational preferences built into this model.

4.4 Conclusions

Metamorphic proteins are often discussed in terms of an energy landscape with two separate and “coexisting” funnels in order to rationalize their ability to adopt two different folds. However, most naturally occurring metamorphic proteins adopt a unique fold under a given constant local environment and switch to a different fold only upon a change in the environment. To examine the energy landscape of the CTD of RfaH, we developed and tested a hybrid all-atom model that combines a physics-based model with a dual-basin structure-based potential (dual-basin SBM). We showed that this model captures the required change in global free energy minimum upon the excision of the CTD from RfaH. Applying this model to the isolated CTD, we found an energy landscape that is characterized by a single dominant funnel towards the β-barrel fold, with no sign of a second funnel towards the α-helical hairpin fold. Our model thus suggests that a multifunneled energy landscape cannot be assumed for metamorphic proteins. Further, we found a relatively high α-helix content in the unfolded state of the CTD. Such α-helical structure was largely absent in our hybrid model with a single-basin SBM. Biophysical characterizations, e.g., using circular dichroism, of the CTDs of RfaH and other members of the general NusG family of transcription factors under weakly unfolding conditions, would provide an interesting experimental test of the computational results of this work.

4.5 Author contributions

Stefan Wallin and Bahman Seifi developed the hybrid model, and designed and performed the computer simulations. Bahman Seifi analyzed the simulations. Stefan Wallin and Bahman Seifi prepared the figures and wrote the manuscript.

4.6 Data availability

Data available on request from the authors. Patch to the PROFASI software package for adding structure-based energy terms available at:

https://github.com/jswallin/PROFASI_plugin_SBM.

Acknowledgements

This work was supported by grants from Memorial University and Natural Sciences and Engineering Research Council of Canada (NSERC), and by computational resources provided by Compute Canada.

Conflict of interest

None to declare.

Bibliography

[1] P. N. Bryan and J. Orban. Proteins that switch folds. *Curr Opin Struct Biol*, 20:482–488, 2010.

[2] A. F. Dishman and B. F. Volkman. Unfolding the mysteries of protein metamorphosis. *ACS Chem Biol*, 13:1438–1446, 2018.

[3] M. Lella and R. Mahalakshmi. Metamorphic proteins: emergence of dual protein folds from one primary sequence. *Biochemistry*, 56:2971–2984, 2017.

[4] Y. Liu and D. Eisenberg. 3D domain swapping: as domains continue to swap. *Protein Sci*, 11:1285–1299, 2002.

[5] S. Hayward. Structural principles governing domain motions in proteins. *Proteins*, 36:425–435, 1999.

[6] J. A. Vila. Metamorphic proteins in light of Anfinsen's dogma. *J Phys Chem Lett*, 11:4998–4999, 2020.

[7] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.

[8] R. L. Tuinstra, F. C. Peterson, S. Kutlesa, E. S. Elgin, M. A. Kron, and B. F. Volkman. Interconversion between two unrelated protein folds in the lymphotactin native state. *Proc Natl Acad Sci USA*, 105:5057–5062, 2008.

[9] D. R. Littler, S. J. Harrop, W. D. Fairlie, L. J. Brown, G. J. Pankhurst, S. Pankhurst, M. Z. DeMaere, T. J. Campbell, A. R. Bauskin, R. Tonini, M. Mazzanti, S. N. Breit, and P. M. Curmi. The intracellular chloride ion channel protein CLIC1 undergoes a redox-controlled structural transition. *J Biol Chem*, 279:9298–9305, 2004.

[10] Y. G. Chang, S. E. Cohen, C. Phong, W. K. Myers, Y. I. Kim, R. Tseng, J. Lin, L. Zhang, J. S. Boyd, Y. Lee, S. Kang, D. Lee, S. Li, R. D. Britt, M. J. Rust, S. S. Golden, and A. LiWang. A protein fold switch joins the circadian oscillator to clock output in cyanobacteria. *Science*, 349:324–328, 2015.

[11] L. L. Porter and L. L. Looger. Extant fold-switching proteins are widespread. *Proc Natl Acad Sci USA*, 115:5968–5973, 2018.

[12] A. Zarrine-Afsar, S. M. Larson, and A. R. Davidson. The family feud: do proteins with similar structures fold via the same pathway? *Curr Opin Struct Biol*, 15:42–49, 2005.

[13] A. A. Nickson and J. Clarke. What lessons can be learned from studying the folding of homologous proteins? *Methods*, 52:38–50, 2010.

[14] D. S. Riddle, V. P. Grantcharova, J. V. Santiago, E. Alm, I. Ruczinski, and D. Baker. Experiment and theory highlight role of native state topology in SH3 folding. *Nat Struct Biol*, 6:1016–1024, 1999.

[15] F. Chiti, N. Taddei, P. M. White, M. Bucciantini, F. Magherini, M. Stefani, and C. M. Dobson. Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nat Struct Mol Biol*, 6:1005–1009, 1999.

[16] J. C. Martinez and L. Serrano. The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nat Struct Biol*, 6:1010–1016, 1999.

[17] D. Trotter and S. Wallin. Effects of topology and sequence in protein folding linked via conformational fluctuations. *Biophys J*, 118:1370–1380, 2020.

[18] S. S. Cho, Y. Levy, and P. G. Wolynes. Quantitative criteria for native energetic heterogeneity influences in the prediction of protein folding kinetics. *Proc Natl Acad Sci USA*, 106:434–439, 2009.

[19] A. Kluber, T. A. Burt, and C. Clementi. Size and topology modulate the effects of frustration in protein folding. *Proc Natl Acad Sci USA*, 115:9234–9239, 2018.

[20] S. Mishra, L. L. Looger, and L. L. Porter. Inaccurate secondary structure predictions often indicate protein fold switching. *Protein Sci*, 28:1487–1493, 2019.

[21] B. M. Burmann, S. H. Knauer, A. Sevostyanova, K. Schweimer, R. A. Mooney, R. Landick, I. Artsimovitch, and P. Rösch. An α helix to β barrel domain switch transforms the transcription factor RfaH into a translation factor. *Cell*, 150:291–303, 2012.

[22] G. A. Belogurov, M. N. Vassylyeva, V. Svetlov, S. Klyuyev, N. V. Grishin, D. G. Vassylyev, and I. Artsimovitch. Structural basis for converting a general

transcription factor into an operon-specific virulence regulator. *Mol Cell*, 26:117–129, 2007.

[23] D. Shi, D. Svetlov, R. Abagyan, and I. Artsimovitch. Flipping states: a few key residues decide the winning conformation of the only universally conserved transcription factor. *Nucleic Acids Res*, 45:8835–8843, 2017.

[24] P. K. Zuber, K. Schweimer, P. Rösch, I. Artsimovitch, and S. H. Knauer. Reversible fold-switching controls the functional cycle of the antitermination factor RfaH. *Nat Commun*, 10:702, 2019.

[25] P. Galaz-Davison, J. A. Molina, S. Silletti, E. A. Komives, S. H. Knauer, I. Artsimovitch, and C. A. Ramírez-Sarmiento. Differential local stability governs the metamorphic fold switch of bacterial virulence factor RfaH. *Biophys J*, 118:96–104, 2020.

[26] S. Li, B. Xiong, Y. Xu, T. Lu, X. Luo, C. Luo, J. Shen, K. Chen, M. Zheng, and H. Jiang. Mechanism of the all-α to all-β conformational transition of RfaH-CTD: Molecular dynamics simulation and Markov State model. *J Chem Theory Comput*, 10:2255–2264, 2014.

[27] C. A. Ramirez-Sarmiento, J. K. Noel, S. L. Valenzuela, and I. Artsimovitch. Interdomain contacts control native state switching of RfaH on a dual-funneled landscape. *PLOS Comput Biol*, 11:e1004379, 2015.

[28] J. B. GC, Y. R. Bhandari, B. S. Gerstman, and P. P. Chapagain. Molecular dynamics investigations of the α-helix to β-barrel conformational transformation in the RfaH transcription factor. *J Phys Chem B*, 118:5101–5108, 2014.

[29] J. B. GC, B. S. Gerstman, and P. P. Chapagain. The role of the interdomain interactions on RfaH dynamics and conformational transformation. *J Phys Chem B*, 119:12750–12759, 2015.

[30] L. Xiong and Z. Liu. Molecular dynamics study on folding and allostery in RfaH. *Proteins*, 83:1582–1592, 2015.

[31] S. Xun, F. Jiang, and Y. D. Wu. Intrinsically disordered regions stabilize the helical form of the C-terminal domain of RfaH: A molecular dynamics study. *Bioorg Med Chem*, 24:4970–4977, 2016.

[32] J. A. Joseph, D. Chakraborty, and D. J. Wales. Energy landscape for fold-switching in regulatory protein RfaH. *J Chem Theory Comput*, 15:731–742, 2019.

[33] B. Seifi and A. Aina and S. Wallin. Structural fluctuations and mechanical stabilities of the metamorphic protein RfaH. *Proteins*, 89:289–300, 2021.

[34] J. Y. Kang, R. A. Mooney, Y. Nedialkov, J. Saba, T. V. Mishanina, I. Artsimovitch, R. Landick, and S. A. Darst. Structural basis for transcript elongation control by NusG family universal regulators. *Cell*, 173:1650–1662, 2018.

[35] P. K. Zuber, I. Artsimovitch, M. NandyMazumdar, Z. Liu, Y. Nedialkov, K. Schweimer, P. Rösch, and S. H. Knauer. The universally-conserved transcription factor RfaH is recruited to a hairpin structure of the non-template DNA strand. *Elife*, 7:e36349 2018.

[36] B. Wang, V. M. Gumerov, E. P. Andrianova, I. B. Zhulin, and I. Artsimovitch. Origins and Molecular Evolution of the NusG Paralog RfaH. *mBio*, 11:e02717–20

2020.

[37] A. Roy, A. Perez, K. A. Dill, and J. L. Maccallum. Computing the relative stabilities and the per-residue components in protein conformational changes. *Structure*, 22:168–175, 2014.

[38] N. A. Bernhardt and U. H. E. Hansmann. Multifunnel landscape of the fold-switching protein RfaH-CTD. *J Phys Chem B*, 122:1600–1607, 2018.

[39] J. R. Allison, M. Bergeler, N. Hansen, and W. F. van Gunsteren. Current computer modeling cannot explain why two highly similar sequences fold into different structures. *Biochemistry*, 50:10965–10973, 2011.

[40] C. Holzgräfe and S. Wallin. Smooth functional transition along a mutational pathway with an abrupt protein fold switch. *Biophys J*, 107:1217–1225, 2014.

[41] C. Holzgräfe and S. Wallin. Local versus global fold switching in protein evolution: insight from a three-letter continuous model. *Phys Biol*, 12:026002, 2015.

[42] A. Aina and S. Wallin. Multisequence algorithm for coarse-grained biomolecular simulations: Exploring the sequence-structure relationship of proteins. *J Chem Phys*, 147:095102, 2017.

[43] R. D. Hills and C. L. Brooks. Insights from coarse-grained Gō models for protein folding and dynamics. *Int J Mol Sci*, 10:889–905, 2009.

[44] N. Gō and H. Taketomi. Respective roles of short- and long-ranged interactions in protein folding. *Proc Natl Acad Sci USA*, 75:559–563, 1978.

[45] S. Wallin and H. S. Chan. Conformational entropic barriers in topology-dependent protein folding: perspectives from a simple native-centric polymer model. *J Phys Condens Matter*, 18:S307, 2006.

[46] L. L. Chavez, J. N. Onuchic, and C. Clementi. Quantifying the roughness on the free energy landscape: entropic bottlenecks and protein folding rates. *J Am Chem Soc*, 126:8426–8432, 2004.

[47] J. S. Yang, S. Wallin, and E. I. Shakhnovich. Universality and diversity of folding mechanics for three-helix bundle proteins. *Proc Natl Acad Sci USA*, 105:895–900, 2008.

[48] L. Sutto and C. Camilloni. From A to B: a ride in the free energy surfaces of protein G domains suggests how new folds arise. *J Chem Phys*, 136:185101, 2012.

[49] M. Kouza and U. H. Hansmann. Folding simulations of the A and B domains of protein G. *J Phys Chem B*, 116:6645–6653, 2012.

[50] A. Irbäck and S. Mohanty. PROFASI: A Monte Carlo simulation package for protein folding and aggregation. *J Comput Chem*, 27:1548–1555, 2006.

[51] A. Irbäck, S. Mitternacht, and S. Mohanty. An effective all-atom potential for proteins. *PMC Biophysics*, 2:1–24, 2009.

[52] G. Favrin, A. Irbäck, and F. Sjunnesson. Monte Carlo update for chain molecules: biased Gaussian steps in torsional space. *J Chem Phys*, 114:8154–8158, 2001.

[53] N. Madras and A. D. Sokal. The pivot algorithm: A highly efficient Monte Carlo method for the self-avoiding walk. *J Stat Phys*, 50:109–186, 1988.

[54] A. Sali, E. Shakhnovich, and M. Karplus. How does a protein fold? *Nature*, 369:248–251, 1994.

[55] J. K. Noel, M. Levi, M. Raghunathan, H. Lammert, R. L. Hayes, J. N. Onuchic, and P. C. Whitford. SMOG 2: A versatile software package for generating structure-based models. *PLOS Comput Biol*, 12:e1004794, 2016.

[56] J. K Noel, P. C. Whitford, and J. N. Onuchic. The shadow map: a general contact definition for capturing the dynamics of biomolecular folding and function. *J Phys Chem B*, 116:8692–8702, 2012.

[57] M. Heinig and D. Frishman. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res*, 32:W500–502, 2004.

[58] J. H. Meinke and U. H. E. Hansmann. Protein simulations combining an all-atom force field with a Go term. *J Phys Condens Matter*, 19:285215, 2007.

[59] T. Sikosek, H. Krobath, and H. S. Chan. Theoretical insights into the biophysics of protein bi-stability and evolutionary switches. *PLOS Comput Biol*, 12:e1004960, 2016.

[60] Y. Wang, P. Tian, W. Boomsma, and K. Lindorff-Larsen. Monte Carlo sampling of protein folding by combining an all-atom physics-based model with a native state bias. *J Phys Chem B*, 122:11174–11185, 2018.

[61] A. M. Ferrenberg and R. H. Swendsen. Optimized Monte Carlo data analysis. *Phys Rev Lett*, 63:1195–1198, 1989.

Chapter 5

Examining the effect of the N-terminal domain of RfaH on domain dissociation and fold switching

Abstract

RfaH is a two-domain metamorphic protein involved in both gene regulation and enhancement of translation. Its dual functions rely on two key characteristics: domain separation and fold switching of its C-terminal domain (CTD). In the free state, the CTD is in an all-α state; when RfaH binds to RNA polymerase (RNAP), the CTD completely transforms into an all-β state separated from the NTD. However, the mechanism of domain separation in the RfaH protein is unknown. Here we hypothesize that a change in the relative orientation of the extended hairpin in the NTD (β3-β4), which occurs upon binding to RNAP, can trigger the CTD to dissociate from NTD. To test this hypothesis, we build a RfaH structure with remodeled extended hairpin using a homology modelling tool (structure H_1). We use an all-atom physics-based model enhanced with a dual-basin structure-based potential to simulate domain separation driven by unfolding of the CTD under condition with NTD fixed in its folded state. We apply the model to both free RfaH and H_1. In line with our hypothesis, we find that the CTD has a reduced stability for model H_1 compared to free RfaH. We also studied the reverse fold switching; we use fixed NTD with all-β CTD and analyze the secondary structure of the CTD during the reverse fold switching from the all-β state to the all-α state.

5.1 Introduction

As increasing number of proteins are being discovered with an ability to switch from one native state to another [1–4]. One of the best-studied examples is RfaH, which is a protein involved in the regulation of transcription and translation in *Escherichia coli* [4–6]. In isolation, this protein adopts a fold in which the C-terminal domain (CTD) adopts a helical hairpin that is tightly packed against the N-terminal domain (NTD). It has been shown that when the NTD and CTD dissociate, the CTD spontaneously transforms into a β-barrel fold [6]. This can be achieved in vitro, e.g, by cutting the flexible linker that connects the NTD and CTD with an enzyme. It is known that in the *E. coli* cell, the two domains dissociate when the RfaH molecule interacts with RNA polymerase (RNAP). However, the mechanism that triggers domain separation is unknown [6, 7].

Three dimensional structures of RfaH have been determined experimentally in the protein's free state (PDB id 5ond) using X-ray crystallography [5] and in its fold-switched state bound to RNAP (PDB id 6c6s) using cryo-electron microscopy [8]. No structural information exists so far for the bound state of RfaH before fold switching has occurred. It has been suggested that RfaH forms an encounter complex upon initial binding to RNAP, which finalizes domain dissociation. [9]. Interestingly, a structural comparison of these two structures (5ondA and 6c6s) reveals changes to the NTD as we and others have noted previously [6, 7, 10]. One of these differences is the orientation of the extended β3-β4 hairpin (approx residues 31–52) of the RfaH NTD, as shown in Fig. (5.1.A). Fold switching of the CTD has been heavily studied experimentally [6,11–13] and through simulations [14–21]. However, most simulation studies have focused on the isolated CTD.

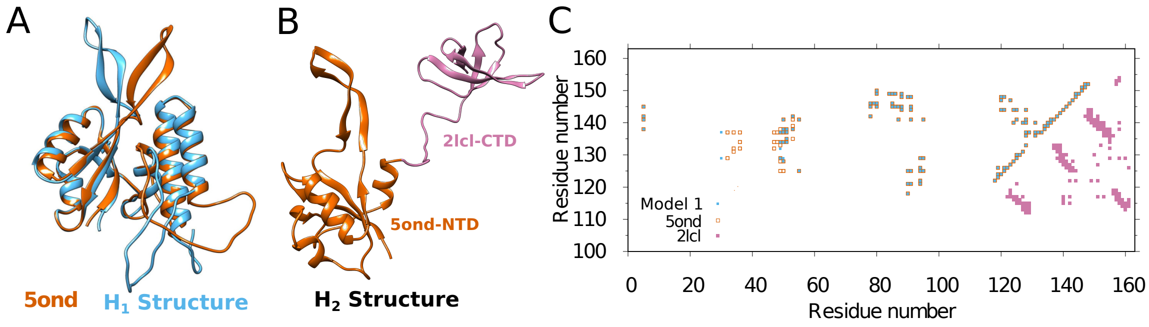

Figure 5.1: RfaH states and contact maps. (A) Two different structures of the α-state of RfaH: the experimentally determined structure of free RfaH (pdb id 5ond in vermilion) and a hypothetical structure of RfaH in the RfaH/RNAP encounter complex (H_1 in blue). (B) Model structure H_2 of the β-state of RfaH, in which the β-barrel structure of the isolated CTD (pdb id 2lcl) is combined with the free RfaH structure (5ond) from the free form of RfaH. H_1 and H_2 are created using homology modeling as described in Methods. (C) Native residue-residue contacts within the CTD (residues 113-162) for the structures 5ond, H_1 (above the diagonal) and H_2 (below the diagonal), and between the NTD and CTD for 5ond and H_1. Contacts within the NTD are identical for all three structures and not shown. In this work we apply our all-atom hybrid model for fold switching [10] to the 5ond and H_1 structures, creating computational models for the free RfaH state and for RfaH in the state H_1 ("model H_1").

Here we focus on two potential effects of the NTD for RfaH fold switching; (1) the dissociation of the two RfaH domains, and (2) the reverse fold switch, i.e., the transformation of the CTD from the all-β to the all-α state. In particular, we focus on the mechanism that triggers fold switching. One recent study by Galaz-Davison et al, explains the key role of the NTD in stabilizing the transformation of RfaH from an α-hairpin to a β-barrel structure upon binding with RNAP using memory potential terms [22].

Based on the structural changes of the NTD, our working hypothesis is that the change in the orientation of the extended hairpin upon binding to RNAP plays a role in controlling the "affinity" between the CTD and NTD. To test this hypothesis, we construct the structural model H_1, a chimera between 5ond and 6c6s (see Fig. 5.1,A). This structure is almost identical to the free form of RfaH (5ond) in all regions of the chain except the β3-β4 extended hairpin, which has an orientation taken from the final bound form (6c6s). The idea is that structure H_1 might represent some structural aspects of the RfaH encounter complex. We compared simulations of the free RfaH and the H_1 in terms of the stability of the CTD.

To study the reverse fold switching, we construct another homology modeling structure H_2, with NTD taken from the 5ond structure [5] and CTD from the 2lcl structure [7]. The structure H_2 represents the state of RfaH just after the dissociation from RNAP. During the subsequent β-to-α fold switch, CTD must eventually contact the NTD in order to form the close interface present in the α-state of RfaH. We examine, in particular, the role of the NTD for this process.

5.2 Materials and Methods

5.2.1 Homology modeling

The H_1 structure is created from the two experimental structures 6c6s and 5ond in the following way. First, 6c6s and 5ond are optimally superimposed, i.e. the root mean square deviation (RMSD) is minimized over all rigid-body translations and rotations of one of the chains. In this superposition, RMSD is determined only over the segments 2-29 and 54-99, which are selected because they are the most conserved regions between the structures. A PDB file is then created by combining fragments from the two structures: segments 1-29, 54-99, and 115-162 are taken from 5ond, while 31-52 extended hairpin region is taken from 6c6s. Finally, these structural fragments are used to create a complete full-length structure H_1 using the homology modelling tool MODELLER [24] (see Fig. (5.1.A)). The structure H_2 is created similarly by combining residues 1-99 from 5ond and 113-162 from 2lcl.

5.2.2 Computational protein model

All simulations were carried out using the hybrid model developed in Ref [10]. This model is a hybrid model in the sense that it combines the physics based model in the software package PROFASI [25] and a dual-basin structure-based model (SBM) or Gō-like model [26]. The dual-basin nature of the model means it includes bias towards two different reference structures.

In our model for the free RfaH state, the two reference structures are 5ond and H_2. The model for the RfaH state with an altered extended hairpin orientation is constructed using the two reference structures H_1 and H_2. We refer to these as our model for the free RfaH state and model H_1.

The PROFASI package is a simulation model for all-atom protein system with four energy terms: $E^{(0)} = E_{\text{loc}} + E_{\text{ev}} + E_{\text{hb}} + E_{\text{sc}}$. Conformationally, there are some restrictions on protein structures e.g. fixed bond lengths and bond angles, and $E^{(0)}$ is a function only of the backbone torsional angles, ϕ and ψ, and various sidechain torsional angles, χ. The local term E_{loc} includes interactions between atoms close along the mainchain with partial charges and provides a proper local description of the protein chain. The E_{ev} is the excluded-volume energy term arising from repulsion energy ($1/r^{12}$) between all atom pairs. The remaining two terms, E_{hb} and E_{sc}, which, respectively, represent hydrogen bonding and sidechain-sidechain interactions, underpin the main structure formation. Hydrogen bonding is implemented through directionally dependent explicit attractions between donor and acceptor groups. The term E_{sc} includes both effective hydrophobic attractions and pairwise interactions between sidechain charges. In this way, solvent effects are implicitly considered by the energy function.

The dual basin SBM provides two energetic biases towards to the two native structures with all-α CTD and all-β CTD encoded as sets of residue-residue contacts present in each structure as shown in Fig. (5.1.A and B). The structure-based potential can be written

$$E_{\text{SB}} = \lambda^{\alpha} E_{\text{SB}}(C^{\alpha}) + \lambda^{\beta} E_{\text{SB}}(C^{\beta}) - E_{\text{corr}}(\lambda^{\alpha}, C^{\alpha}; \lambda^{\beta}, C^{\beta}) \,, \tag{5.1}$$

where the C^{α} and C^{β} are contact map sets for native states all-α CTD and all-β CTD, respectively (see Fig. 5.1.C). λ^{α} and λ^{β} are the strengths of the two structure-based terms ($\lambda^{\alpha} = \lambda^{\beta} = 0.3$). The last term, E_{corr}, is a correction term to avoid double counting the energy of common contacts in C^{α} and C^{β}. The hybrid model combines the physics-based model and the dual basin SBM such that simulations are carried out with the energy $E = E^{(0)} + E_{\text{SB}}$ [10].

5.2.3 Equilibrium Monte Carlo simulations

To characterize the equilibrium behavior of the CTD of RfaH as part of full-length RfaH, we used ordinary fixed temperature Metropolis MC. We performed sampling by allowing a random walk in conformational space using three different types of moves: (1) a pivot move that updates a single Ramachandran ϕ- or ψ-angle; (2) Biased Gaussian Steps (BGS) that work by updating up to 8 consecutive ϕ, ψ-angles such that an approximately local chain deformation is obtained [27]; and (3) a sidechain move that updates a single sidechain torsional angle, χ. While (1) gives global changes in conformation, (2) and (3) give local (or small-step) changes. In all our simulations, the fraction of sidechain moves was held fixed at 58%. In our equilibrium simulations, the remaining 42 % of moves were divided between pivot and BGS [28].

For both free RfaH and the H_1 structures, the thermodynamic behavior was determined using at least 10 independent runs of each 1×10^7 MC cycles, where a cycle is 560 elementary MC steps (the number of turnable ϕ, ψ or χ angles in the protein chain). For both structures, the backbone chain corresponding to positions 1-100 (i.e., the ordered region of the NTD) was held fixed in its initial (native) conformation by disallowing BGS and pivots in this region. All sidechains were allowed to move.

5.2.4 Small-step "kinetic" Monte Carlo simulations

Our "kinetic" MC runs of the chimeric structure H_2 differed from the equilibrium simulations runs, in that the global (i.e. pivot) moves were turned off. The small-step "kinetic" simulation is suitable for fold switching simulations. Our fold switching simulations started from the regularized H_2 structure, Fig. (5.1.B)

5.2.5 Native contact maps

The native contact sets $C = \{ij|$ if residues i and j are in contact $\}$, are obtained by submitting three regularised model structures (derived from free RfaH, H_1, and H_2) to the SMOG webserver [31] with the coarse-graining option "Calpha" and otherwise default parameters. In these contact sets, if there is atom-atom contact between two residues i and j according to the shadow contact map algorithm [32], two residues i and j are considered to form a contact ij.

For the 5ond structure, we obtain a set of 137 native contacts with 58 contact in α-helical hairpin in CTD (C^{α}) and 79 NTD-CTD inter-domain contacts (contacts with residue i in segment 1-100 and residue j in the segment 113-156). For the H_1 structure, there are 120 native contacts, and the difference between 5ond and H_1 contacts is in inter domain contacts that are between extended hairpin in NTD (residues 30 to 53) and CTD. Thus, there are 58 contacts in α-helical hairpin in CTD (C^{α}) and 62 NTD-CTD inter-domain contacts. NTD-CTD inter-domain native contacts, i.e., with residue i belonging to the segment 1-100 and residue j belonging to the segment 113-156. For the β-barrel structure (2lcl), we obtained this way a set of 130 native contacts, C^{β}. The contact maps are shown in Fig (5.1.C).

5.2.6 Observables

The variable Q_{α} is the fraction of 46 contacts within residues 122-162 of C^{α} (native contacts α-helical CTD for both 5ond and H_1 structures), and Q_{β} is the fraction of C^{β} (native contacts of β-barrel CTD for 2lcl structure). The Q_{α} and Q_{β} measure how much the structure folds toward the all-α and all-β native states, respectively (see Fig. (5.1.C)). Q_α excludes some native contacts in the NTD part of α_4, which is known to be disordered in solution [10]. To calculate Q_α and Q_β we consider a contact between residues i and j in CTD as a formed contact if $r_{ij} < 1.2r^0_{ij}$, where r^0_{ij} and r_{ij} are distances in the native state and extracted configuration in a given MC cycle, respectively.

The root-mean-square deviations, RMSD$_\alpha$ and RMSD$_\beta$, are measured over C$_\alpha$ atoms and are taken with respect to two representative structures of the all-α (5ond-CTD) and all-β CTD (2lcl) folds, respectively. Secondary structure assignments, used for the calculation of α-helix and β-barrel content, were generated using STRIDE [33].

Domain-domain distance measures the distance between the C-alpha atom of residues Phe56 and Gly135 that, respectively, are close to the center of mass of 5ond-NTD and 2lcl-CTD in their folded states.

5.3 Results and discussion

5.3.1 Role of the NTD extended hairpin for CTD stability

We start by comparing the thermodynamic stability of the CTD in the context of two different NTD conformations represented by the free form structure (5ond) and our H$_1$ structure. To do this, we carry out equilibrium simulations using our hybrid all-atom model in which the NTD backbone is held fixed. The NTD conformation is taken to be either in the 5ond or H$_1$ conformation. Hence, in these simulations, CTD is able to fold and unfold while the NTD remains folded. We can determine a midpoint temperature for this folding-unfolding transition. From a physical perspective, these folding simulations correspond to thermal unfolding experiments of free RfaH in which the NTD is kept stable with covalent cross-links.

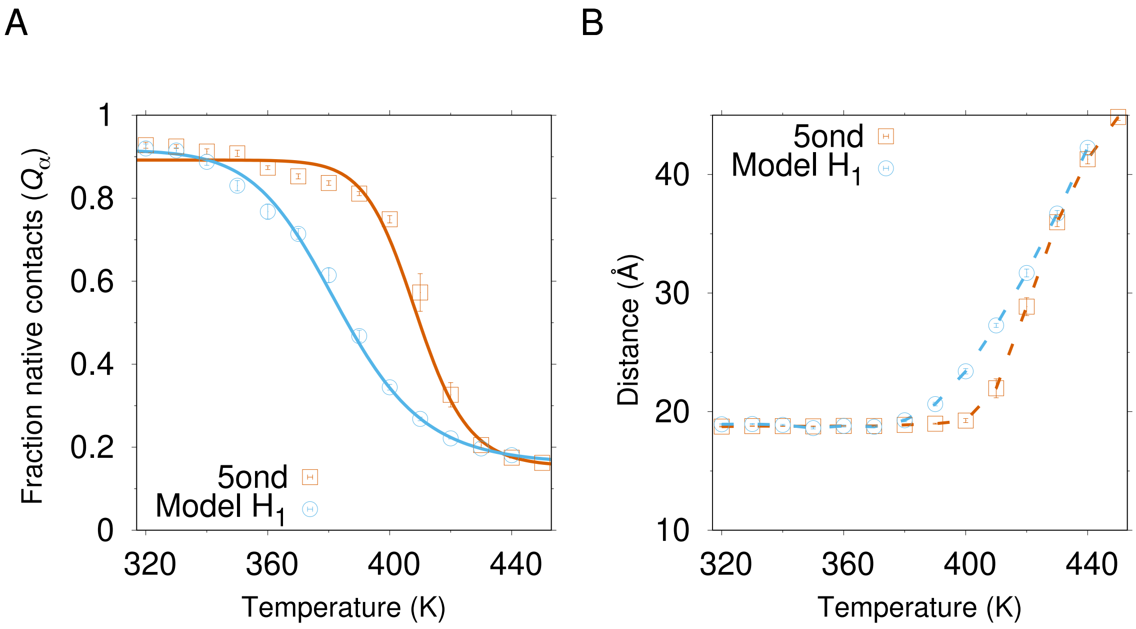

Figure 5.2: Impact of the NTD structure on the stability of the all-α CTD state. Shown is the temperature dependence of the fraction of CTD native contacts, obtained from equilibrium simulations of our model for the free RfaH state (squares; vermilion) and our model for the H_1 state (circles; blue). Solid curves in are fits to the two-state equation $\langle Q \rangle = (Q^{\mathrm{U}} + Q^{\mathrm{N}}K)/(1+K)$, where $K = \exp\left(-\Delta E(1/k_{\mathrm{B}}T - 1/k_{\mathrm{B}}T_{\mathrm{m}}\right)$ and Q^{U}, Q^{N}, ΔE and T_{m} (midpoint temperature) are fit parameters. According to the two state model, midpoint temperature for free RfaH and H_1 are $410K$, and $380K$, respectively.

Burmann et al. [7] utilized NMR experiments on unbound RfaH and discovered that the Val116-Gly121 section had chemical shifts indicating it was more similar to a random coil than an α-helix. This suggested that the first six residues of helix α_4 were mostly disordered in the solution phase. Similarly, the same segment was also poorly structured in the model presented in this study (see Fig. S3). To address this, we used C^{α} contacts except that all contacts involving residues 116-121 were excluded in Q_{α} to measure the nativeness of α-helical state. This measure consistently demonstrated higher values than the fraction of all native α contacts at all temperatures, indicating that the α-helical hairpin is formed at low temperatures, with the exception of the N-terminal portion of α_4, which, in agreement with the experiments, remains largely unstructured. The formation of a weak α-structure in the 116-121 region, as well as in the C-terminal end of $\alpha 5$ (residues 150-156), is consistent with the hydrogen/deuterium exchange mass spectrometry (HDXMS) experiments conducted by Galaz-Davison et al. [13].

According to the temperature dependence of the fraction of native contacts, the midpoint temperatures for 5ond and H_1, respectively, are 410K and 380K, as shown in Fig.(5.2,A). These results show that when the extended hairpin in the RfaH protein adopt a new bent conformation, the CTD domain becomes unstable in comparison with the 5ond structure. It should be mentioned that the midpoint temperature that we obtain for H_1 is close to the isolated β-strand CTD, which is 375K [21].

In Fig.(5.2,B), we measure the temperature dependence of domain-domain distance between the NTD and CTD, which shows that NTD-CTD detachment occurs at a lower temperature for H_1 than free RfaH. Overall, the above research is in line with our hypothesis that the orientation of the extended hairpin plays a significant role in triggering domain dissociation.

5.3.2 Energy landscape for free RfaH and H_1

Having shown, in line with our hypothesis, that the structural change from 5ond to model H_1 leads to reduced stability of the CTD, we turn to study the free energy landscape in the two different states.

We first determine the free energy profile $F(Q_\alpha)$ at the midpoint temperature T_m as shown in Fig. (5.3). For free RfaH, at T_m, there are two minima corresponding to the unfolded state (low-Q_α) and the all-α fold (high-Q_α), with a free energy barrier separating the states. However, for the H_1 model, there is one minimum stretching over a range of Q_α. Significantly, the small free energy barrier between unfolded state and all-α fold in free RfaH disappears in the H_1 model.

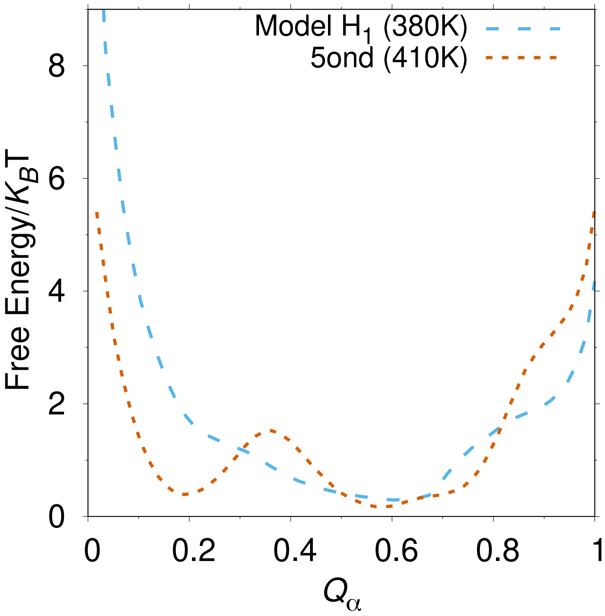

Figure 5.3: Free energy profiles in midpoint temperature. Free energy as function of Q_α for our hybrid models of free RfaH (vermilion) and H_1 (blue) with dual basin SBM, taken at their respective midpoint temperatures, $T = 410K$ and $T = 380K$.

Additional insight can be gained from the free energy surface $F(Q_\alpha, Q_\beta)$ at T_m. As shown in Fig. (5.4.A), for free RfaH, there are two free energy minima at T_m, which respectively correspond to the unfolded state (low-Q_α, low-Q_β), and all-α fold (high-Q_α, low-Q_β). For the H_1 model, there is one minimum in free energy, which stretches from low-Q_α to high-Q_α (see Fig. (5.4.B)). There is no minima at $Q_\beta \approx 0.7 - 0.9$, which are the values of Q_β that correspond to a fully formed β-barrel state [10]. Hence, no complete fold switching in the α-to-β direction occurs in these simulations. However, a shallow free energy minimum at $Q_\beta \approx 0.3 - 0.5$ appears for model H_1 indicating some tendency towards fold switching for this model. Interestingly, the midpoint temperature $T_\mathrm{m} = 380$ K for model H_1 is close to the folding temperature for the isolated CTD (375 K) [10]. Some refolding into the β-barrel state would therefore be expected in our model, if the CTD chain were fully detached from NTD at low enough temperature. A transition of the CTD into the β-barrel state may therefore be hindered by interactions with the hydrophobic binding surface on NTD, which is also the binding site for RNAP. Indeed, the domain-domain distance is only slightly larger at 380 K than at lower T where the folded helical hairpin state dominates, indicating an incomplete separation between the domains even at $T \approx T_\mathrm{m}$.

Taken together, these results indicate that the change in orientation of the extended hairpin alone is not sufficient to trigger a transformation into the β-barrel fold. It may also be required that the NTD binding surface be engaged with another partner molecule, such that this large hydrophobic surface can be hidden. Such an engagement could be achieved in vivo by the binding of the NTD binding surface to the tip of the two coiled-coil helices in the β clamp domain of RNAP, as in the final bound state of

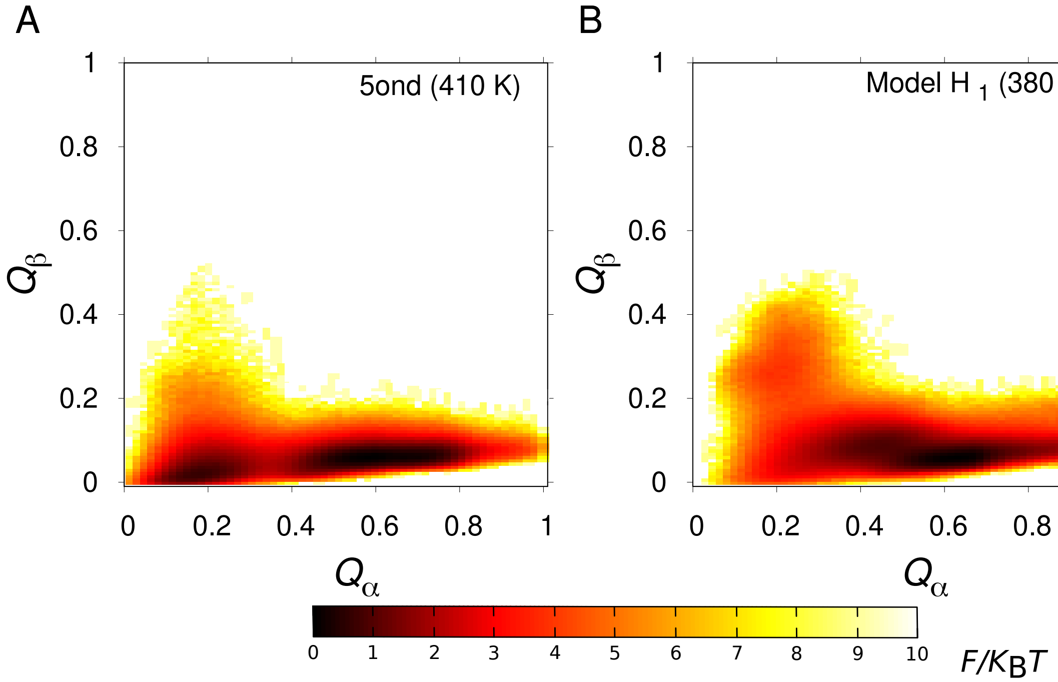

Figure 5.4: Free energy surfaces for free RfaH and the model H_1 at midpoint temperature. Free energy surfaces $F(X_1, X_2) = -k_B T \ln P(X_1, X_2)$, with $X_1 = Q_\alpha$ and $X_2 = Q_\beta$ for (A) the free RfaH model at $T = 410K$ and (B) the H_1 model at $T = 380K$.

RfaH [15]. In some simulations starting from the H_1 we observed fold switching from all-α state to partially folded β state (see Fig. (5.5)).

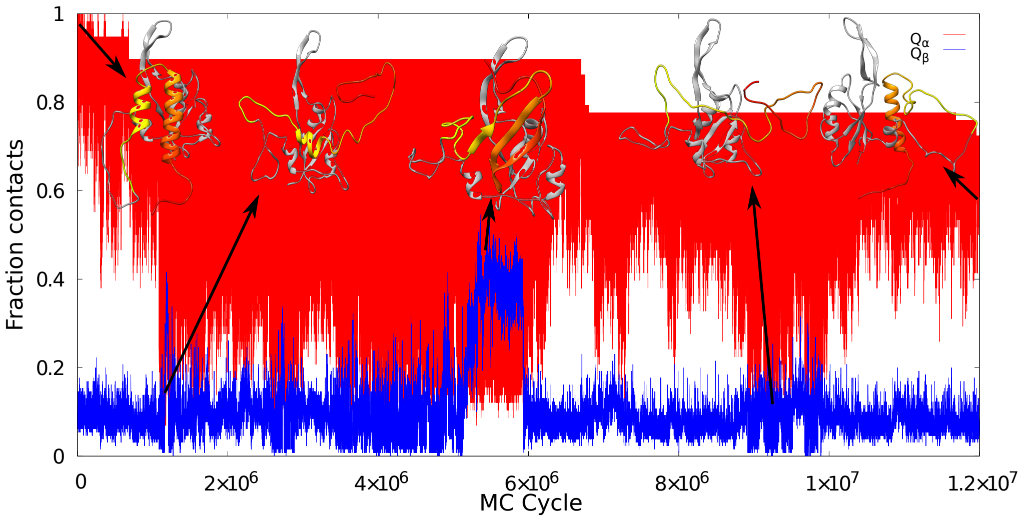

Figure 5.5: Example of a fold switching trajectory in the model H_1 with a partial transition to the all-β CTD state. The red and blue lines are fractions of α and β contacts, respectively.

5.3.3 Reverse fold switching

For studying the reverse fold switching, we initialize the system in the β state and carry out simulations using the free RfaH model at a temperature $T < T_{\rm m}$, i.e., under conditions where the α state is thermodynamically stable. We use small-step "kinetic" Monte Carlo simulations in order to mimic the time dependence of this process. The idea is to study how the CTD switches its fold from the all-β state to the all-α state in the presence of folded NTD.

Fig (5.6.A) shows RMSD_α and RMSD_β as functions of MC time for a typical reverse fold switching trajectory where RMSD_α and RMSD_β are, respectively, the root-mean-square deviations taken with respect to the representative all-α and all-β CTD structures. After 3.5×10^6 MC cycles, there is a jump in RMSD$_\beta$ from ≈2 Å to a state with RMSD$_\beta$ ≈10-12 Å and at 4.2×10^6 MC cycles there is another jump to RMSD$_\beta$ ≈15-30 Å. The last transition is a transition from a partially folded state to an intermediate state, I. Because of the large fluctuations in both RMSD$_\alpha$ and RMSD$_\beta$, it is clear that this intermediate state is highly disordered. Finally, after 7.5×10^6 MC cycles, the CTD transitions from I into the all-α state.

For more detail, we can see the time evolution of secondary structures in the reverse fold switching. As shown in Fig.(5.6.C), the first jump in RMSD$_\beta$ belongs to the unfolding of β1 and β5, which happen at the same time. Structurally, this means an opening of the β-barrel structure, which aligns with the results from Galaz-Davison et al. [**?**,22]. As seen in Fig.(5.6.B), after this transition, the domain-domain distance exhibits a sudden decrease in the size of fluctuations at 3.6×10^6 MC cycle, suggesting the CTD enters a bound state with NTD, although with brief visits to larger domain-domain distances. Without having a significant change in RMSD and secondary structure between 3.6×10^6 to 3.9×10^6 MC cycles, the CTD becomes closer to the NTD and sticks to the NTD at 3.9×10^6 MC cycle. In this intermediate state, the CTD is close to the NTD, and there are brief formation of various local secondary structure elements, including in the α_4 and α_5 regions. Finally, the CTD folds to the all-α states with low RMSD$_\alpha$, a well ordered α5, and even smaller domain-domain distances.

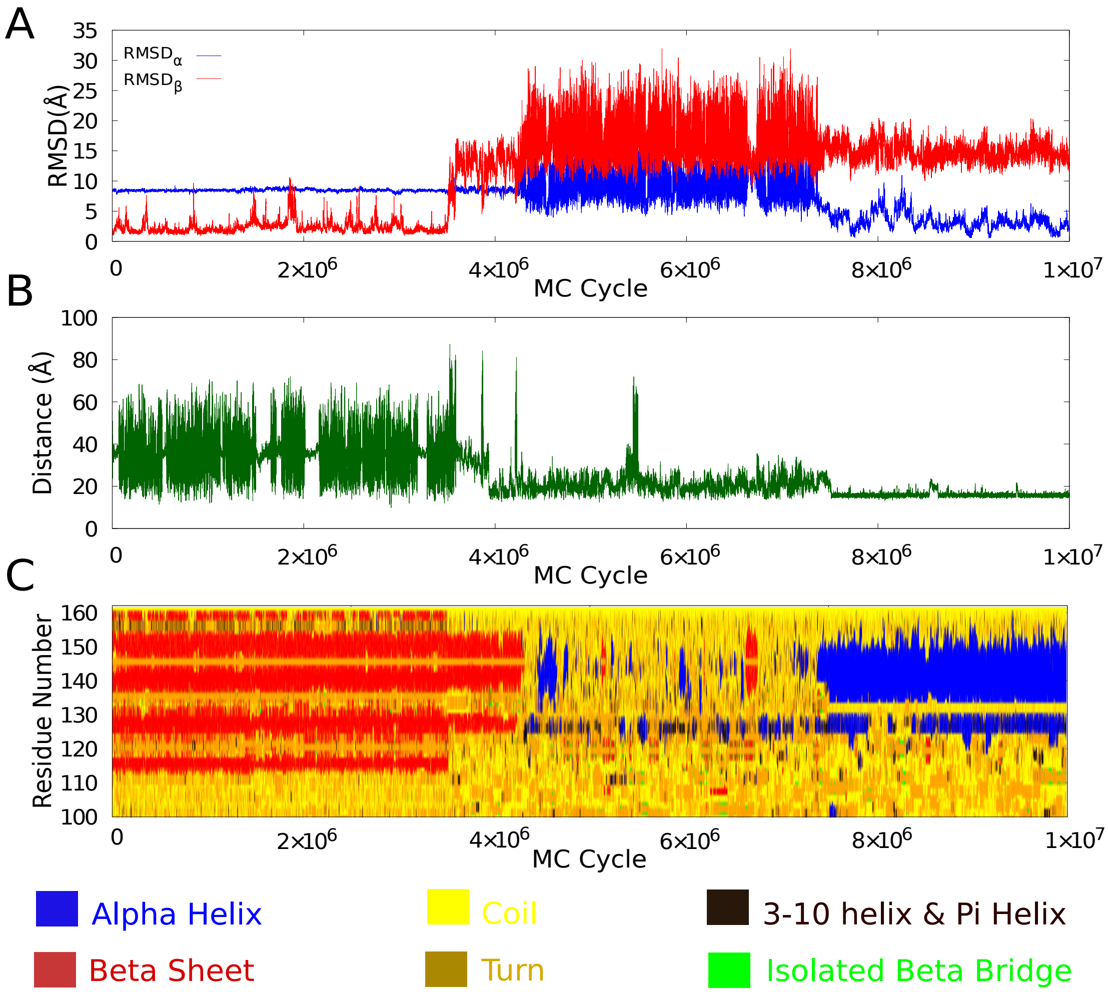

Figure 5.6: Example of a reverse fold switching trajectory. Time evolution of (A) the root-mean-square deviation determined for the CTD and taken with respect to the α-helical hairpin structure (RMSD$_\alpha$, PDB id 5ond) or the β-barrel structure (RMSD$_\beta$; PDB id 2lcl), (B) the domain-domain distance (see Methods), and (C) secondary structure elements at different positions along the chain, including the linker region (residues 100-112) and the CTD (residues 113-162).

5.4 Conclusion

We have used Monte Carlo simulations and an all-atom hybrid model [10] to study domain dissociation and β to α fold switching in the transcription factor RfaH. The RfaH protein adopts a unique fold under a given constant local environment, and its CTD switches its fold upon a change in the environment from an all-α state closely interacting with the NTD to an all-β state separated from the NTD. We have a hypothesis that changing the relative orientation of the extended $\beta 3$-$\beta 4$ hairpin upon binding to RNAP could sufficiently change the environment for CTD to switch its fold. To test this hypothesis, we built a chimera structure H_1 and compared its characteristics with free RfaH. Applying our computational model to both the H_1 structure and free RfaH, we found the midpoint temperature for H_1 is much smaller in our model than free RfaH, and it is close to the midpoint temperature of isolated CTD. At the midpoint temperature, the H_1 simulations exhibit a higher β content than the free RfaH simulations. Further, we studied the reverse fold switching of the CTD in a condition that the NTD is fixed in the free RfaH state. These simulations suggest a specific order of unfolding of the secondary structure elements in CTD, before refolding into the α state takes place.

Bibliography

[1] L. Looger, A. K. Majumdar and L. Porter. Identification and Prediction of Fold-Switching Proteins. *Biophys. J*, 118:480a, 2020.

[2] P. N. Bryan and J. Orban. Proteins that switch folds. *Curr Opin Struct Biol*, 20:482–488, 2010.

[3] A. F. Dishman and B. F. Volkman. Unfolding the mysteries of protein metamorphosis. *ACS Chem Biol*, 13:1438–1446, 2018.

[4] M. Lella and R. Mahalakshmi. Metamorphic proteins: emergence of dual protein folds from one primary sequence. *Biochemistry*, 56:2971–2984, 2017.

[5] P. K. Zuber, I. Artsimovitch, M. NandyMazumdar, Z. Liu, Y. Nedialkov, K. Schweimer, P. Rösch, and S. H. Knauer. The universally-conserved transcription factor RfaH is recruited to a hairpin structure of the non-template DNA strand. *Elife*, 7:e36349 2018.

[6] P. K. Zuber, K. Schweimer, P. Rösch, I. Artsimovitch, and S. H. Knauer. Reversible fold-switching controls the functional cycle of the antitermination factor RfaH. *Nat Commun*, 10:702, 2019.

[7] B. M. Burmann, S. H. Knauer, A. Sevostyanova, K. Schweimer, R. A. Mooney, R. Landick, I. Artsimovitch, and P. Rösch. An α helix to β barrel domain switch transforms the transcription factor RfaH into a translation factor. *Cell*, 150:291–303, 2012.

[8] J. Y. Kang, R. A. Mooney, Y. Nedialkov, J. Saba, T. V. Mishanina, I. Artsimovitch, R. Landick, and S. A. Darst. Structural basis for transcript elongation control by NusG family universal regulators. *Cell*, 173:1650–1662, 2018.

[9] P. K. Zuber, T. Daviter, R. Heißmann, U. Persau, K. Schweimer, and S. H. Knauer. Structural and thermodynamic analyses of the β-to-α transformation in RfaH reveal principles of fold-switching proteins . *Elife*, 11:e76630, 2022.

[10] B. Seifi, and S. Wallin. The C-terminal domain of transcription factor RfaH: Folding, fold switching and energy landscape. *Biopolymers*, 112:e23420, 2021.

[11] G. A. Belogurov, M. N. Vassylyeva, V. Svetlov, S. Klyuyev, N. V. Grishin, D. G. Vassylyev, and I. Artsimovitch. Structural basis for converting a general transcription factor into an operon-specific virulence regulator. *Mol Cell*, 26:117–129, 2007.

[12] D. Shi, D. Svetlov, R. Abagyan, and I. Artsimovitch. Flipping states: a few key residues decide the winning conformation of the only universally conserved transcription factor. *Nucleic Acids Res*, 45:8835–8843, 2017.

[13] P. Galaz-Davison, J. A. Molina, S. Silletti, E. A. Komives, S. H. Knauer, I. Artsimovitch, and C. A. Ramírez-Sarmiento. Differential local stability governs the

metamorphic fold switch of bacterial virulence factor RfaH. *Biophys J*, 118:96–104, 2020.

[14] S. Li, B. Xiong, Y. Xu, T. Lu, X. Luo, C. Luo, J. Shen, K. Chen, M. Zheng, and H. Jiang. Mechanism of the all-α to all-β conformational transition of RfaH-CTD: Molecular dynamics simulation and Markov State model. *J Chem Theory Comput*, 10:2255–2264, 2014.

[15] C. A. Ramirez-Sarmiento, J. K. Noel, S. L. Valenzuela, and I. Artsimovitch. Interdomain contacts control native state switching of RfaH on a dual-funneled landscape. *PLOS Comput Biol*, 11:e1004379, 2015.

[16] J. B. GC, Y. R. Bhandari, B. S. Gerstman, and P. P. Chapagain. Molecular dynamics investigations of the α-helix to β-barrel conformational transformation in the RfaH transcription factor. *J Phys Chem B*, 118:5101–5108, 2014.

[17] J. B. GC, B. S. Gerstman, and P. P. Chapagain. The role of the interdomain interactions on RfaH dynamics and conformational transformation. *J Phys Chem B*, 119:12750–12759, 2015.

[18] L. Xiong and Z. Liu. Molecular dynamics study on folding and allostery in RfaH. *Proteins*, 83:1582–1592, 2015.

[19] S. Xun, F. Jiang, and Y. D. Wu. Intrinsically disordered regions stabilize the helical form of the C-terminal domain of RfaH: A molecular dynamics study. *Bioorg Med Chem*, 24:4970–4977, 2016.

[20] J. A. Joseph, D. Chakraborty, and D. J. Wales. Energy landscape for fold-switching in regulatory protein RfaH. *J Chem Theory Comput*, 15:731–742,

2019.

[21] B. Seifi and A. Aina and S. Wallin. Structural fluctuations and mechanical stabilities of the metamorphic protein RfaH. *Proteins*, 89:289–300, 2021.

[22] P. Galaz-Davison, and E. A. Román, and C. A. Ramírez-Sarmiento. The N-terminal domain of RfaH plays an active role in protein fold-switching. *PLoS Comput Biol*, 17:e1008882, 2021.

[23] Y. Liu and D. Eisenberg. 3D domain swapping: as domains continue to swap. *Protein Sci*, 11:1285–1299, 2002.

[24] A. Sali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, 234:779–815, 1993.

[25] A. Irbäck and S. Mohanty. PROFASI: A Monte Carlo simulation package for protein folding and aggregation. *J Comput Chem*, 27:1548–1555, 2006.

[26] N. Gō and H. Taketomi. Respective roles of short- and long-ranged interactions in protein folding. *Proc Natl Acad Sci USA*, 75:559–563, 1978.

[27] G. Favrin, A. Irbäck, and F. Sjunnesson. Monte Carlo update for chain molecules: biased Gaussian steps in torsional space. *J Chem Phys*, 114:8154–8158, 2001.

[28] N. Madras and A. D. Sokal. The pivot algorithm: A highly efficient Monte Carlo method for the self-avoiding walk. *J Stat Phys*, 50:109–186, 1988.

[29] A. Sali, E. Shakhnovich, and M. Karplus. How does a protein fold? *Nature*, 369:248–251, 1994.

[30] A. Irbäck, S. Mitternacht, and S. Mohanty. An effective all-atom potential for proteins. *PMC Biophysics*, 2:1–24, 2009.

[31] J. K. Noel, M. Levi, M. Raghunathan, H. Lammert, R. L. Hayes, J. N. Onuchic, and P. C. Whitford. SMOG 2: A versatile software package for generating structure-based models. *PLOS Comput Biol*, 12:e1004794, 2016.

[32] J. K Noel, P. C. Whitford, and J. N. Onuchic. The shadow map: a general contact definition for capturing the dynamics of biomolecular folding and function. *J Phys Chem B*, 116:8692–8702, 2012.

[33] M. Heinig and D. Frishman. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res*, 32:W500–502, 2004.

[34] J. H. Meinke and U. H. E. Hansmann. Protein simulations combining an all-atom force field with a Go term. *J Phys Condens Matter*, 19:285215, 2007.

Chapter 6

Simulations of a protein fold switch reveal crowding-induced population shifts driven by disordered regions

Abstract

Macromolecular crowding effects on globular proteins, which typically adopt a single stable fold, and intrinsically disordered proteins, which lack a stable fold, have been widely studied. However, much less is known about crowding effects on fold-switching proteins, a class of proteins characterized by their ability to reversibly switch between distinct folds. Here we study the mutationally driven switch between the folds of GA and GB, the two 56-amino acid binding domains of *Streptococcal* Protein G, using a structure-based dual-basin model and Langevin dynamics sampling. We show first that, in the absence of crowders, the fold populations P_{A} and P_{B} are controlled by the strengths of native contacts in the two folds, κ_{A} and κ_{B}. A population balance, $P_{\mathrm{A}} \approx P_{\mathrm{B}}$, is obtained for $\kappa_{\mathrm{B}}/\kappa_{\mathrm{A}} = 0.92$. The resulting model protein, which we denote G^{*}_{AB}, is then subject to crowded conditions with different packing fractions, ϕ_{c}. We find that the presence of crowders promotes the GB population and reduces the GA population, reaching $P_{\mathrm{B}}/P_{\mathrm{A}} \approx 4$ at $\phi_{\mathrm{c}} = 0.39$. We analyze the ϕ_{c}-dependence of the crowding-induced GA-to-GB fold switch using scaled particle theory (SPT). SPT provides a qualitative, but not quantitative, fit of our data, suggesting effects beyond a spherical description of the folds. We prove that the terminal regions of the chain, which are intrinsically disordered only in the GA fold, play a dominant role in determining the response of the fold switch to crowding effects.

6.1 Introduction

Most globular proteins rely on a single fold to carry out their function. However, recently proteins have been discovered with an ability to switch between different folds [1–4], a phenomenon called fold switching. By adopting an alternative structure, these fold-switching proteins (also termed metamorphic [5] or transformer [6] proteins) gain the ability to carry out an additional unrelated function. For example, a switch from a helical hairpin to a β-barrel transforms the *Escherichia coli* protein RfaH from a transcription factor to a translational activator [7]. Consistent with this view, fold switching is often regulated [8]. A range of cellular signals has been associated with fold switching, such as changes in salt concentration [9], redox conditions [10], and oligomerization [11]. Fold switching also underpins evolutionary changes in protein structure [12–14], in which case fold switching is driven by mutations.

In this work, we investigate the effects of macromolecular crowding on fold switching. To this end, we focus on the binding domains of Protein G, GA and GB, which form one of the most well-characterized fold switch systems [15] (see Fig. 6.1a). It was demonstrated that a set of substitution mutations can be found which drastically increases the sequence identity of GA and GB, while still retaining their respective native structures and binding partners [15]. For example, the variants GA95 and GB95 differ in only 3 amino acid positions. Hence, three additional substitutions (L20A, I30F and L45Y) applied to GA95 cause an abrupt switch from the 3α fold of GA to the $4\beta + \alpha$ fold of GB. Later it was shown that there are multiple ways in which a single substitution can tip the balance from one fold to the other, e.g., L20A applied to the variant GB98-T25I [16]. These experiments on GA and GB were,

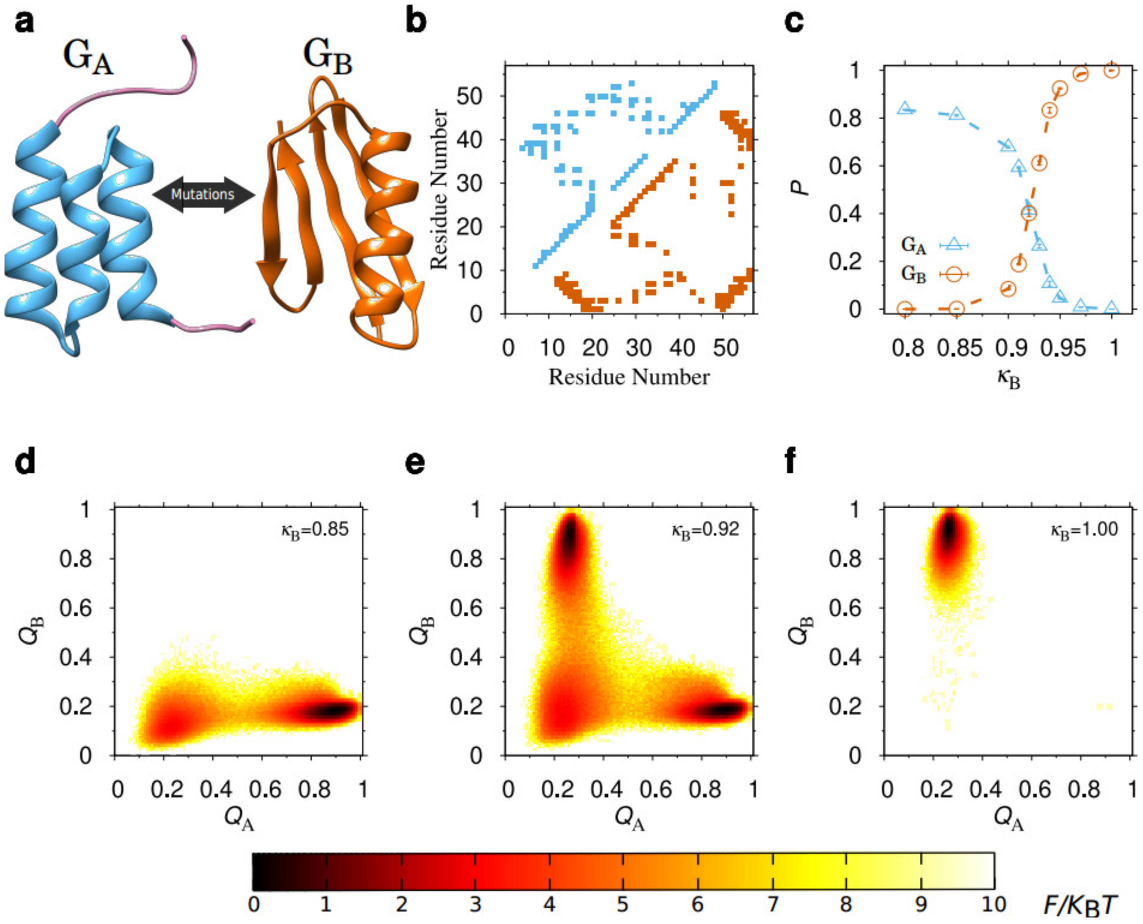

Figure 6.1: The GA/GB fold-switch system. (a) Representative experimental structures of the GA and GB folds shown in ribbon: GA95 (PDB id 2KDL; blue) and GB95 (PDB id 2KDM; orange). In GA95, residue positions 1-7 and 53-56 are intrinsically disordered (purple). (b) Contact maps of the GA95 (above diagonal) and GB95 (below diagonal) structures. (c) Populations of the GA (triangles) and GB (circles) folds as functions of the GB contacts strengths, $\kappa_{\rm B}$. (d-e) Free energy surface $F(Q_{\rm A}, Q_{\rm B}) = -k_{\rm B}T \ln P(Q_{\rm A}, Q_{\rm B})$, where $Q_{\rm A}$ and $Q_{\rm B}$ are the fractions of GA and GB contacts, respectively, T is the temperature, $k_{\rm B}$ Boltzmann's constant, and $P(Q_{\rm A}, Q_{\rm B})$ is a probability distribution, obtained at three different values of $\kappa_{\rm B}$. Error bars in (c) and all other figures represent 1σ errors estimated from results of independent simulations.

however, carried out in dilute protein solutions and therefore in the absence of any crowding effects.

We carry out our simulations with a coarse-grained structure-based model, which we develop and test on the GA/GB fold switch in the absence of crowders (see Methods). In its original form, the structure-based approach involves a potential energy landscape with a single basin of attraction constructed by making native contacts attractive and non-native contacts repulsive. This type of modeling has provided important insights into several aspects of protein folding [17–20]. The natural extension to fold switching is a potential with dual basins of attraction corresponding to the two alternative folds [21–26]. Our dual-basin model for GA/GB fold switching permits us to mimic the progression of mutations along a pathway from one fold to the other by tuning the relative interaction strengths of residue-residue contacts in the GA and GB folds (see Fig. 6.1b). To understand the effect of crowding, we focus on the point along the mutational pathway where the GA and GB folds exhibit roughly equal fold propensities, which we reasoned should be especially susceptible to crowding effects.

6.2 Results

6.2.1 Mimicking the mutational pathway between the GA and GB folds

We first simulate the GA/GB system in the absence of crowders at a fixed temperature T sufficiently low for low-energy folded conformations to dominate over those in the unfolded state (U). By varying the strength κ_{B} of GB contacts, keeping the strength of GA contacts fixed ($\kappa_{\mathrm{A}} = 1$), we can control the relative population of the two folds in our model, as shown in Fig. 6.1c. While GB is the dominant state at high κ_{B} ($\gtrsim 0.97$) GA dominates at low κ_{B} ($\lesssim 0.85$), where there is also a non-zero population of U. At an intermediate value, $\kappa_{\mathrm{B}} = \kappa^{*} = 0.92$, the populations of GA (P_{A}) and GB (P_{B}) are almost equal, $P_{\mathrm{A}} \approx P_{\mathrm{B}} \approx 0.39 - 0.42$. The drastic population shifts between states GA, GB, and U, can be seen from the free energy surfaces $F(Q_{\mathrm{A}}, Q_{\mathrm{B}})$, where Q_{A} and Q_{B} are the fractions of formed GA and GB contacts, respectively, taken at different κ_{B} values (see Fig. 6.1d-f).

The sharp structural transition around $\kappa_{\mathrm{B}} \approx \kappa^{*}$ is reminiscent of experiments showing that very few mutational steps (or a single step) is sufficient to tip the balance from GA to GB, or vice versa, for carefully selected mutational pathways [15]. Moreover, the minimum in the total folded population $P_{\mathrm{tot}} = P_{\mathrm{A}} + P_{\mathrm{B}}$ at $\kappa_{\mathrm{B}} \approx \kappa^{*}$ (see Fig. 6.2a) is in line with the partial loss of stability seen for GA and GB sequences close to the transition point, e.g., GA98 and GB98, [15] as well as for other fold switching proteins [1, 27]. These results allow us to interpret κ_{B} as a continuous parameter mimicking the number of steps taken along a mutational pathway connecting the GA and GB folds. The point $\kappa_{\mathrm{B}} = \kappa^{*}$ thus represents a sequence located on the border between GA and GB. Although a sequence with a perfect GA and GB population balance was not reported, it has been found for other fold switching systems, e.g., the E48S variant of RfaH [7] and the N11L mutant of the Switch Arc protein [28]. We denote our $\kappa_{\mathrm{B}} = \kappa^{*}$ model protein with GAB*.

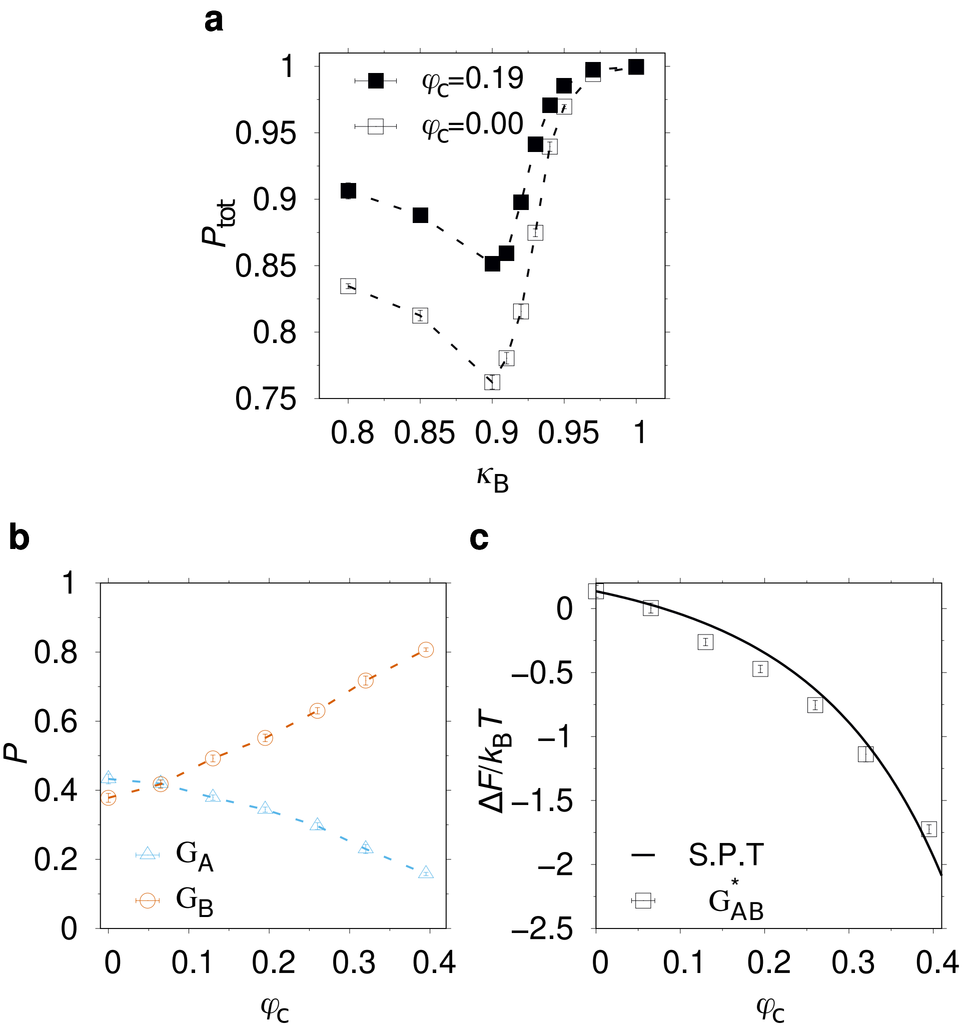

Figure 6.2: Crowding effects on absolute and relative fold populations in the GA/GB fold switch. (a) The total native population $P_{\rm tot} = P_{\rm A} + P_{\rm B}$ as function of the contact strength $\kappa_{\rm B}$ in the absence (open squares) and presence of crowders at packing fraction $\phi_{\rm c} = 0.19$ (filled squares). (b) GA ($P_{\rm A}$; triangles) and GB ($P_{\rm B}$; circles) fold populations as functions of $\phi_{\rm c}$. (c) Free energy of fold switching $\Delta F_{\rm fs} = -k_{\rm B}T \ln P_{\rm B}/P_{\rm A}$ (squares) as function of $\phi_{\rm c}$, fitted to Eq. 6.4 with δ as a single free parameter (solid curve). Dashed lines between points are drawn to guide the eye.

6.2.2 Macromolecular crowding effects on the GA/GB fold switch

Next we introduce spherical crowder particles of radius $R_{\mathrm{cr}} = 12$ Å into our simulations and study their effects on the GA/GB fold switch. Because of hard-core steric repulsions, the protein chain must at all times avoid the space occupied by these particles. For single-fold proteins, such loss of free volume typically stabilizes the native state because any extended conformation in U becomes entropically disfavored relative to compact, folded conformations [29]. The same argument can be applied to each fold of a metamorphic protein. Hence, the overall stability of all folded states should increase. Indeed, as shown in Fig. 6.2a, the addition of crowders increases the total population $P_{\mathrm{tot}} = P_{\mathrm{A}} + P_{\mathrm{B}}$ across all values of κ_{B}. Interestingly, a reduced native state stability is a common feature of fold-switching proteins [1]. One example is the poor stability of sequences on either side of the GA/GB switch point [15]. Hence, crowding effects, if indeed providing an overall stabilization, might alleviate the partial loss of stability suffered by bridge sequences in evolutionary fold-switch transitions [30].

To investigate how the relative population of the GA and GB folds is affected by crowders we focus on GAB*. Fig. 6.2b shows that, as ϕ_{c} increases, the population balance exhibited by GAB* at $\phi_{\mathrm{c}} = 0$ swings towards GB at the expense of GA, i.e., P_{B} increases while P_{A} decreases. The effect on GAB* is not small. For example, $P_{\mathrm{A}}/P_{\mathrm{B}} \approx 4$ at $\phi_{\mathrm{c}} = 0.39$ as compared to ≈ 1 at $\phi_{\mathrm{c}} = 0$. Hence, the effect of steric repulsions between crowders and protein is to favor to GB over GA.

To quantitatively analyze this population shift we apply scaled particle theory (SPT) [31]. In this theory, the free energy cost of inserting a hard sphere of radius R into a fluid of hard spheres of radii R_{cr} with packing fraction ϕ_{c} can be analytically expressed (see Methods). SPT has been used to model crowding-induced changes to the unfolding free energy $\Delta F_{\mathrm{unf}} = F_{\mathrm{U}} - F_{\mathrm{N}}$, where F_{U} and F_{N} are the free energies of U and N, respectively [29, 32, 33]. Here we adapt SPT to fold switching by treating the GA and GB folds as spheres of radii R_{A} and R_{B}. With the parametrization $R_{\mathrm{A}} = R_0 + \delta$ and $R_{\mathrm{B}} = R_0 - \delta$, where R_0 and δ are two parameters, the free energy difference can be written

$$\beta\Delta F_{\mathrm{SPT}} = 6\left[(a + 6ab + ab^2 + \frac{a^3}{3})\rho + (3ab + 3ab^2 + a^3)\rho^2 + (3ab^2 + a^3)\rho^3\right], \quad (6.1)$$

where $a = \delta/R_{\mathrm{cr}}$, $b = R_0/R_{\mathrm{cr}}$, $\rho = \phi_{\mathrm{c}}/(1-\phi_{\mathrm{c}})$ and $\beta = 1/k_{\mathrm{B}}T$. We fit the measured crowding induced changes in free energy of fold switching, $\Delta F_{\mathrm{switch}} = F_{\mathrm{B}} - F_{\mathrm{A}} = -k_{\mathrm{B}}T \ln[P_{\mathrm{B}}/P_{\mathrm{A}}]$ to Eq. 6.1 with δ as a single free parameter, fixing $R_0 = 10.9$ Å to the average radius of gyration of the GA95 and GB95 native structures (see Fig. 6.1a). As shown in Fig. 6.2c, the fits gives $\delta = -0.75 \pm 0.04$ Å, which is reasonable because the radius of gyration of the GA95 and GB95 native structures are $R_{\mathrm{g}}^{\mathrm{A}} = 11.4$ Å and $R_{\mathrm{g}}^{\mathrm{B}} = 10.5$ Å, respectively. The quality of the fit ($\chi^2/(n-1) = 13.5$, sample size $n = 7$) indicates, however, that SPT does not completely describe the observed crowding effects on $\Delta F_{\mathrm{switch}}$.

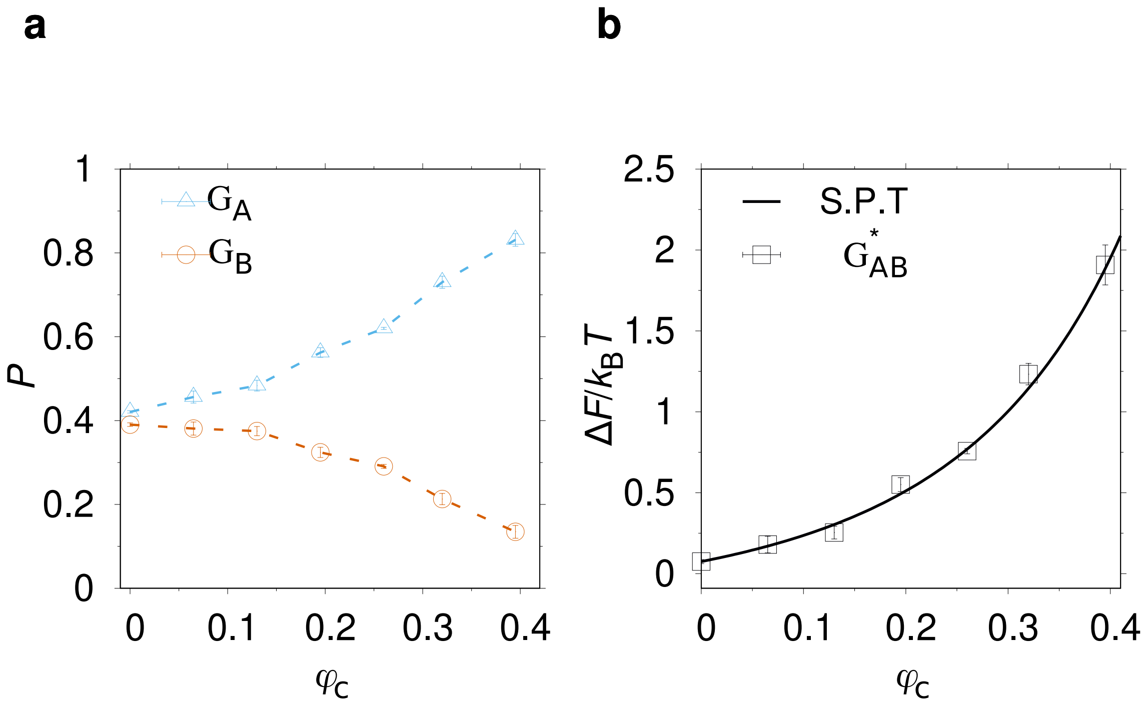

Figure 6.3: Disordered tail segments control crowding effect on fold switch. Results are shown for simulations with a modified potential energy function that ignores hard-core steric repulsions between any crowder and beads in chain segments 1-7 and 53-56 (see text). (a) GA (triangles) and GB (circles) fold populations as function of ϕ_c. (b) Fit of $\Delta F_{\mathrm{switch}}$ to scale particle theory (solid curve).

6.2.3 Disordered tails control the crowding effect on the fold switch

The two terminal segments of the GA95 structure, residues 1-7 and 53-56, are intrinsically disordered (see Fig. 6.1a). Hence, the GA-to-GB fold switch involves a disorder-order transition of these tail regions. Given their flexible nature, it is likely that the tails contribute substantially to the volume excluded by the protein when occupying the GA fold. Indeed, if the terminal segments are ignored, the radius of gyration of GA95 is reduced by almost 30 %, $R_{\mathrm{g}}^{\mathrm{A},7\text{-}50} = 8.2$ Å. Together with the poor fit with SPT (see Fig. 6.2c), these results suggest a major role for the tail segments in how the GA/GB fold switch is impacted by crowding.

To show that this is indeed the case, we carry out crowding simulations with a modified potential energy function $E_{\mathrm{mod}}^{(\mathrm{db})}$, in which all crowder-protein interactions have been turned off for residues in the 1-7 and 53-56 regions. Hence, in these simulations, these N- and C-terminal segments become invisible to the crowders, which thus freely overlap with these residues. Although unphysical, this computational experiment logically tests the role of the tail regions in our model under crowded conditions. Note that crowders can overlap with the tails regardless of which state is populated by the protein. Moreover, at $\phi_{\mathrm{c}} = 0$, there is no change because intra-chain interactions are unaffected. The results are shown in Fig. 6.3. Strikingly, with the modified potential $E_{\mathrm{mod}}^{(\mathrm{db})}$, the impact of crowding reverses such that the GA fold becomes increasingly favored over GB with increasing ϕ_{c}. Hence, our computational experiment shows that the volume excluded by the disordered tails in the GA fold is the dominant factor affecting the balance between the folds in the presence of crowders.

6.2.4 Comparing with crowding effects on single-fold proteins

Above we have shown that the crowders induce a population shift in GAB*, which is due to the presence of disordered tails. For single-fold (monomorphic) proteins, purely repulsive crowders typically enhance the stability of the native state [34]. Naively, one may therefore expect that the native state of monomorphic GB ($\kappa_{\mathrm{B}} > \kappa^{*}$) would be more strongly stabilized by the crowders than monomorphic GA ($\kappa_{\mathrm{B}} < \kappa^{*}$). To test this idea, we determine the folding midpoint temperature, T_{m}, for the model proteins with $\kappa_{\mathrm{B}} = 0.85$, which adopts the single fold GA, and $\kappa_{\mathrm{B}} = 1.00$, which adopts the single fold GB (see Fig 6.1), over a range of ϕ_{c}. As seen in Fig. 6.4a-c, both proteins exhibit a monotonic increase in T_{m} with increasing ϕ_{c}, indicating stabilization. The relative increase in T_{m} for monomorphic GB is indeed somewhat larger than for monomorphic GA. The difference is relatively small, however. We also perform similar simulations using the single-basin energy functions $E^{(\mathrm{A})}$ and $E^{(\mathrm{B})}$, (see Methods) which lack entirely a bias towards the alternative fold. For these models, the crowding-induced increases in T_{m} are almost identical (see Fig. 6.4d). Taken together, these results suggest that determining the crowding response of a fold switching with two "co-existing" folds may not be easily obtained from experiments on two different single-fold proteins representing the folds.

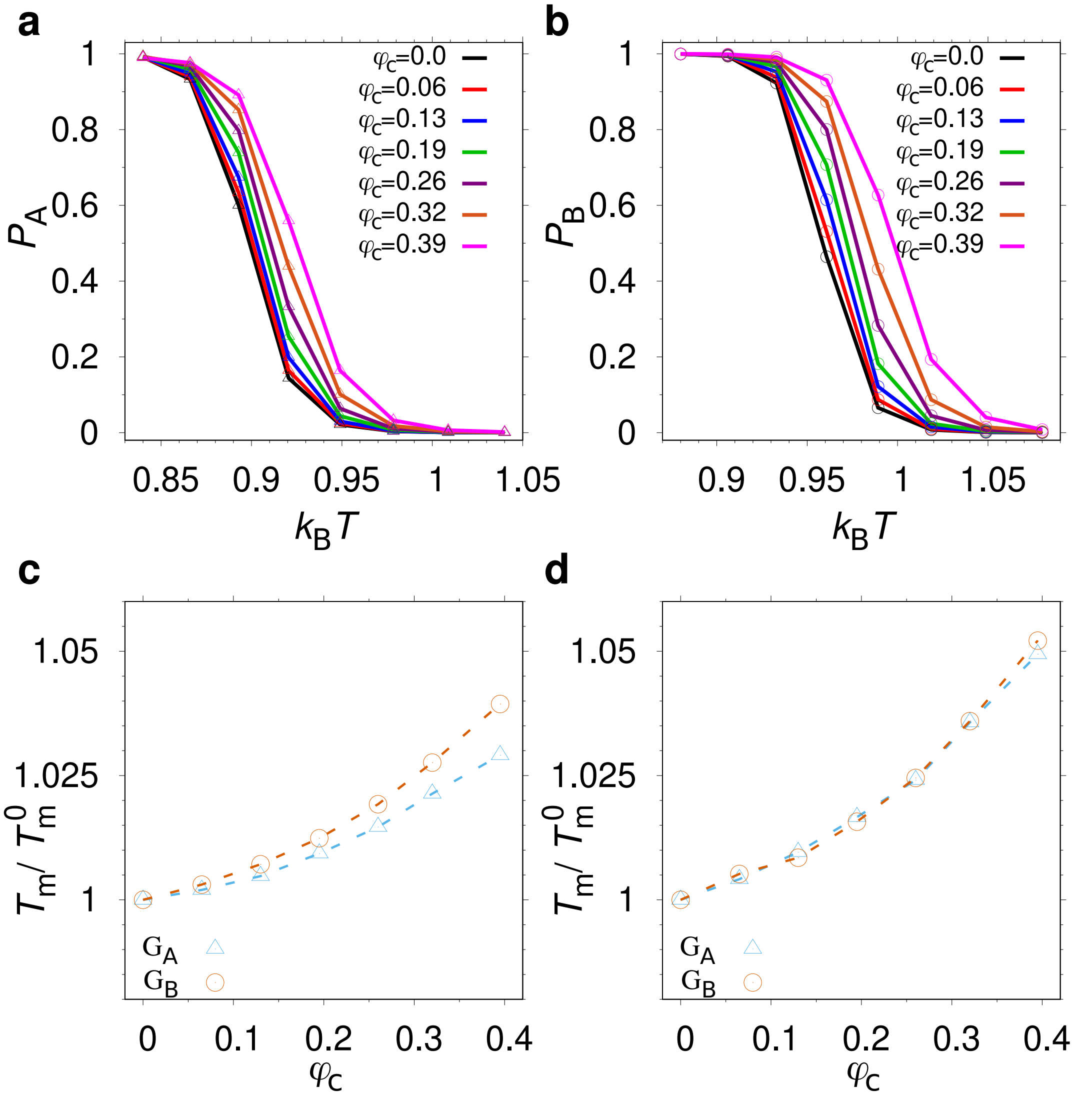

Figure 6.4: Crowding effects on GA and GB single-fold proteins. (a) GA fold population obtained with our dual-basin structure-based model with weak GB contacts ($\kappa_B = 0.85$), as function of temperature. (b) GB fold population obtained with the same model but with strong GB contacts ($\kappa_B = 1.00$). (c) Midpoint temperature, T_m, as function of ϕ_c. T_m is obtained by fitting the folding curves in (a) and (b) to a two-state model. (d) T_m as function of ϕ_c, obtained with single-basin structure-based models for GA and GB. In both (c) and (d), T_m^0 is the value of T_m at $\phi_c = 0$.

6.2.5 The unfolded state changes character across the fold switch

The results in Fig. 6.4 are at first surprising because ΔF_{switch} for a fold switching protein can be obtained from the relation

$$\Delta F_{\text{switch}} = \Delta F^{\text{A}}_{\text{unf}} - \Delta F^{\text{B}}_{\text{unf}}\,, \tag{6.2}$$

where $\Delta F^{\text{A}}_{\text{unf}} = F_{\text{U}} - F_{\text{A}}$ and $\Delta F^{\text{B}}_{\text{unf}} = F_{\text{U}} - F_{\text{B}}$ are defined in direct analogy with the unfolding free energy of single fold proteins. Equation 6.2 expresses that a decrease in ΔF_{switch} results when the crowding-induced stabilization of fold GB relative to U is stronger than the stabilization of fold GA. However, Eq. 6.2 is only guaranteed to hold when ΔF_{switch}, $\Delta F^{\text{A}}_{\text{unf}}$ and $\Delta F^{\text{B}}_{\text{unf}}$ are determined for the same protein for which U provides a common reference. We therefore examine if the drastic structural shift for low energy (folded) conformations in the GA/GB fold switch is accompanied by changes in U.

We first characterize U across the fold switch in the absence of crowders, i.e., upon changing the contact strength κ_{B}, as shown in Fig. 6.5a and b. With increasing κ_{B}, and therefore increasing GB population, the unfolded state radius of gyration $R_{\text{g}}^{(\text{U})}$ decreases. Additionally, U becomes more "GB-like" as shown by the increase in $Q_{\text{B}}^{(\text{U})}$, i.e., the fraction of formed GB contacts in U. These results are in line with simulations of single-fold proteins showing that native contacts in β-proteins tend to promote chain collapse during folding more efficiently than α-proteins [35].

In the GA to GB fold switch driven by crowding we similarly find a compaction of U (see Fig. 6.5c and d). For $\phi_{\text{c}} > 0.20$, $R_{\text{g}}^{(\text{U})}$ becomes smaller than for any value of κ_B in the case of no crowders. Moreover, $Q_\mathrm{A}^{(\mathrm{U})}$ and $Q_\mathrm{B}^{(\mathrm{U})}$ both increase with ϕ_c. Hence, fold switching driven either by mutation or crowding substantially impacts the structural characteristics of U. Both compaction and the formation of residual structure due to crowding have been observed for various single-fold proteins [36,37].

6.3 Discussion

Fold switching in proteins involves major structural changes, including in shape and amino acid composition of surface regions. As a result, fold switching should be inherently susceptible to macromolecular crowding effects. Here we tested this idea by applying a dual-basin structure-based protein model and purely repulsive crowders to the GA/GB fold switch. We found that the addition of crowders indeed alters the free energy balance between the two folds. The effect increases monotonically with ϕ_c. At $\phi_\mathrm{c} = 0.39$, the change in ΔF_switch is $\approx 2\ k_\mathrm{B}T$ in magnitude. While no experiment probing crowding effects on fold switching is available for comparison, a recent study demonstrated a key role for molecular shape in crowding by exploiting alternative dimer forms of two almost identical sequences [38].

Our results show that the response to crowding is determined by chain segments at the N- and C-terminal ends, which are intrinsically disordered only in the GA fold. The volume excluded by these disordered segments leads to an entropic stabilization of the GB fold relative GA. Interestingly, order-disorder transitions occur frequently in protein fold switching [1]. An example besides GA/GB is the human chemokine XCL1, which switches fold upon dimerization. In its monomeric (chemokine) fold, XCL1 adopts an α-helix in its C-terminal region, which becomes disordered when the protein transforms to its dimeric fold-switched state [12]. It should be pointed out that crowder interactions other than hard-core steric repulsions may modify the effect of the crowders, as is the case for single-fold proteins. Non-specific attractive (soft) interactions between protein and crowders generally counteract the stabilizing effect of volume exclusion [39], sometimes even leading to a net destabilization [40].

Most studies on fold switching have quite naturally focused on the structure and dynamics of the different folded states and their interconversions. However, our simulations of the GA/GB switch reveal that fold switching may be accompanied by substantial changes in U (see Fig. 6.5) even in the absence of crowders. Under conditions favoring GA, we find that U is rather expanded and dominated by local contacts while becoming more compact and forming more non-local contacts as the conditions shift to favor GB. In previous simulations of the metamorphic RfaH [25], we showed that the isolated C-terminal domain (CTD), which adopts a stable β-barrel in isolation, exhibits a propensity for α-helical structure in U. This helical propensity was demonstrated experimentally by Zuber et. al. [27], who suggested further that the presence of residual helical structure may help initiate the reverse fold switching of RfaH, i.e., the transformation from the β-barrel to its alternative all-α fold. Taken together, the above considerations suggest that an improved understanding of U may give further insights into fold switching mechanisms as well as effects from crowding.

In addition to changes to the relative population of the two folds, we have found that the presence of crowders increase the total population of the GA and GB folds relative to U. An overall stabilization of ordered states might be especially beneficial to fold-switching proteins, which often exhibit reduced stabilities [1]. Poor stabilities of

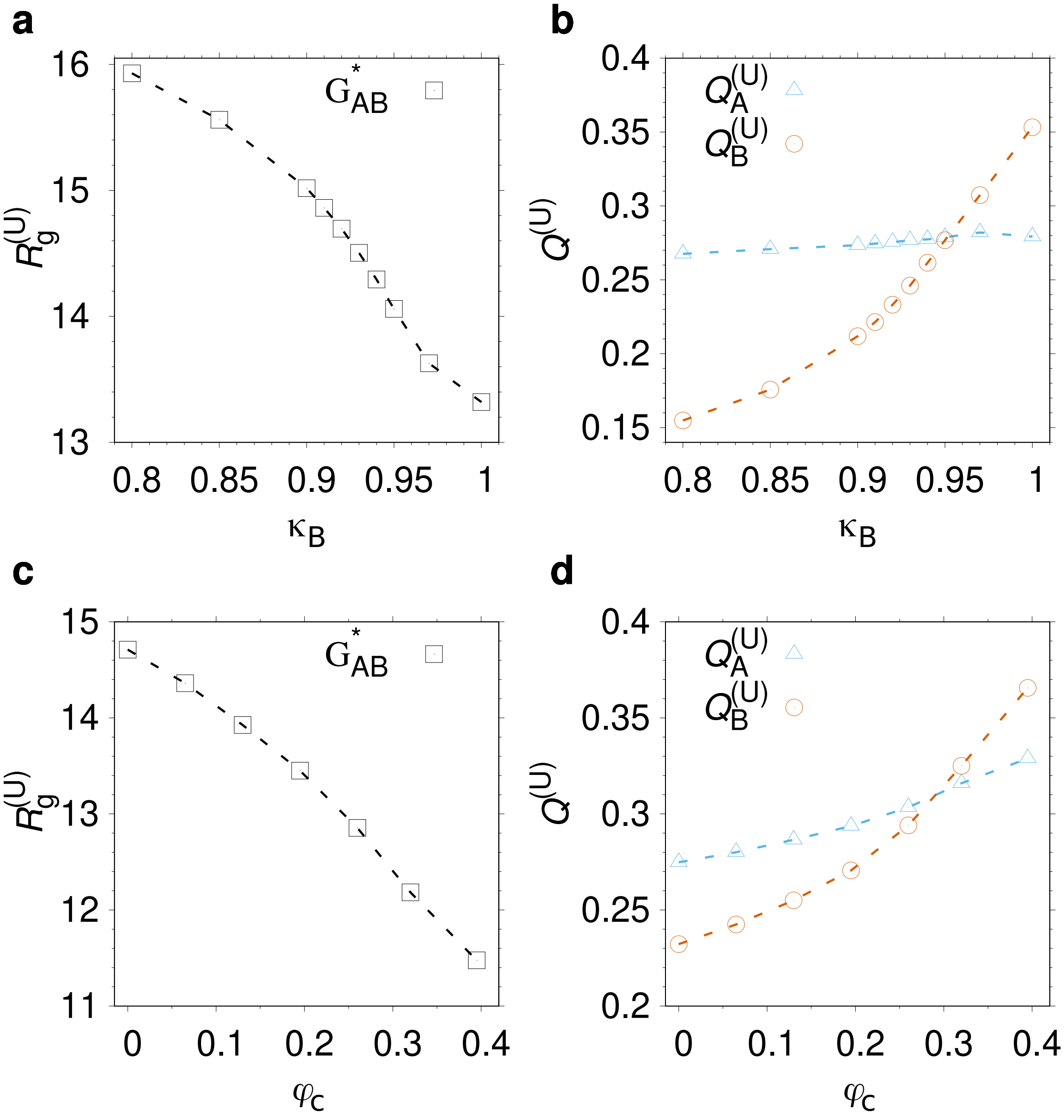

Figure 6.5: The unfolded state changes character across the fold switch. (a) $R_{\mathrm{g}}^{(\mathrm{U})}$, (b) $Q_{\mathrm{A}}^{(\mathrm{U})}$ (triangles) and $Q_{\mathrm{B}}^{(\mathrm{U})}$ (circles) as functions of the contact strength κ_{B} at $\phi_{\mathrm{c}} = 0$, where $R_{\mathrm{g}}^{(\mathrm{U})}$, $Q_{\mathrm{A}}^{(\mathrm{U})}$, and $Q_{\mathrm{B}}^{(\mathrm{U})}$ are the radius of gyration, fraction of GA contacts, and fraction of GB contacts, respectively, determined for the unfolded state, U. (c) $R_{\mathrm{g}}^{(\mathrm{U})}$, (d) $Q_{\mathrm{A}}^{(\mathrm{U})}$ (triangles) and $Q_{\mathrm{B}}^{(\mathrm{U})}$ (circles) as functions of ϕ_{c}, obtained for GAB* ($\kappa_{\mathrm{B}} = 0.92$).

bridge sequences at the border between folds may hamper evolutionary transitions [16, 41, 42]. A recent study suggests fold switching within the context of multidomain proteins, in which non-switching domains can act as stabilizing templates, may help stabilize such bridge sequences and facilitate fold transitions [13]. Our results suggest that additional stabilization may be provided by crowding effects.

Our study opens up for additional experimental and theoretical investigations into the effects of crowding on fold switching. Advances in the field are improving our understanding of fold switching within functional [27, 43] and evolutionary [3, 12, 13] contexts. These efforts will include also a characterization of the impact of macromolecular crowding on equilibrium as well as kinetic properties of fold switching proteins.

6.4 Model and Methods

6.4.1 Native structures and contact maps

The experimentally determined structures of GA95 (PDB id 2KDL) and GB95 (2KDM) [15] were downloaded from the Protein Data Bank (PDB). Both structures were submitted to the SMOG webserver (https://smog-server.org/) to obtained contact maps as prescribed by the shadow map method [44]. The two contact maps contain 106 and 145 contacts, respectively.

6.4.2 Observables

The fractions of native contacts formed, Q_A or Q_B, are determined using the following contact criterion: Two amino acids i and j is considered in contact if $r_{ij} < 1.2r^0_{ij}$, where r_{ij} is the distance between the C_α atoms and r^0_{ij} is the distance in the native structure. The root-mean-square deviations, RMSD, is calculated over all C_α positions of the chain.

6.4.3 Coarse-grained model for protein fold switching

Simulations are carried out using a dual-basin structure-based model in which each amino acid is represented by a single bead located on the C_α position. The starting point for developing this model is a modified version of the single-basin structure-based model in Ref. [18] with a potential energy function with 5 terms, $E = E_\mathrm{bond} + E_\mathrm{bend} + E_\mathrm{torsion} + E_\mathrm{rep} + E_\mathrm{cont}$, representing bond stretching, bond flexing, torsional rotations, repulsions between bead pairs, and attractive native contact interactions. We apply this model separately to the native structures of GA95 and GB95 resulting in two structure-based energy functions, $E^{(\mathrm{A})}$ and $E^{(\mathrm{B})}$, with single basins of attraction (either the GA fold or the GB folds). Using the exponentially-weighted mixing approach of Best et al. [45], we then merge $E^{(\mathrm{A})}$ and $E^{(\mathrm{B})}$ into a single (dual basin) energy function, $E^{(\mathrm{db})}$. The strength of GA and GB contacts, κ_A and κ_B, are left as free parameters in $E^{(\mathrm{db})}$, allowing the relative depth of the GA and GB basins of attraction to be controlled. Further details of the model are given in Appendix A.4.

6.4.4 Excluded volume crowders

Crowder-crowder and crowder-bead interactions are modeled using the potential function [29]

$$V(r) = \epsilon \left(\frac{\sigma}{r - \rho + \sigma} \right)^{12} \tag{6.3}$$

for distances $r > \rho - \sigma$, and $V(r) = \infty$ otherwise. Hence, our crowders have a soft repulsive shell over a hard core. The parameters ρ and σ control the range of the interaction and the width of the soft repulsive shell, respectively. For crowder-crowder interactions, we set $\rho = 2R_{\mathrm{cr}}$ and $\sigma = 2\sigma_{\mathrm{cr}}$, where $\sigma_{\mathrm{cr}} = 3$ Å controls the width of the soft shell of the crowders. For crowder-bead interactions, we set $\rho = R_{\mathrm{cr}} + \sigma_{\mathrm{b}}$ and $\sigma = \sigma_{\mathrm{cr}} + \sigma_{\mathrm{b}}$, where $\sigma_{\mathrm{b}} = 4$ Å is the bead radius. Crowder concentration is quantified as the fraction ϕ_{c} of the total simulation volume V occupied by the crowders, i.e., $\phi_{\mathrm{c}} = 4\pi {R_{\mathrm{cr}}}^3 N_{\mathrm{cr}}/3V$. The number of crowding particles N_{cr} in our simulations range from 9 for $\phi_{\mathrm{c}} = 0.06$ to 54 for $\phi_{\mathrm{c}} = 0.39$.

6.4.5 Langevin dynamics

Conformational sampling is carried out using Langevin dynamics, following the approach of Ref. [18]. The time evolution of the system is governed by the equation, $m\dot{v}(t) = F_{\mathrm{conf}} - m\gamma v(t) + \eta(t)$, where m, v, $\dot{v}$, γ, F_{conf} and $\eta(t)$ are the mass, velocity, acceleration, friction coefficient, conformational force and random force, respectively. For computational reasons, simulations are carried out in the low-friction (underdamped) limit, where $-m\gamma v(t)$ is small relative to the inertial term $m\dot{v}(t)$. In this limit, a natural unit of time for the dynamics is $\tau = \sqrt{ml^2/\epsilon}$ [46], where ϵ is the magnitude of typical interactions and l is a length scale, which we set to 4 Å. The friction coefficient for beads is taken to be $\gamma_{\rm b} = 0.05\tau^{-1}$. Units are set so that the mass of a bead is $m_{\rm b} = 1.0$. The random force $\eta(t)$ is drawn from a Gaussian distribution, the variance of which sets the temperature of the system. Numerical integration of the equation of motion is carried out using the velocity form of the Verlet algorithm [47] with an integration time step $\delta t = 0.005\tau$. For crowders, the mass and friction coefficient are set to $m_{\rm c} = 9.0$ and $\gamma_{\rm c} = 0.017\tau^{-1}$.

6.4.6 Simulation and analysis details

Simulations were carried out by placing the protein and crowders in a cubic box with side 100 Å. Periodic boundary conditions were applied. Langevin dynamics simulations were used to determine the equilibrium behavior of various systems characterized by different GB contact strengths $\kappa_{\rm B}$ and crowder concentrations $\phi_{\rm c}$. Simulations were performed at either fixed temperature or using simulated tempering [48], in which temperature changes dynamically between a predetermined set of temperatures. In the simulated tempering runs, temperatures were updated every 100 time steps. For each system, 5-10 independent runs of $(4-5)\times 10^9$ time steps each were carried out and used to estimate averages and statistical uncertainties. All simulations were initiated from a random protein conformation (random torsional angles ϕ_i) and random crowder positions, followed by a Monte Carlo-based relaxation step in which all hard core steric clashes were removed.

6.4.7 Scaled particle theory

Some theories have been used to predict the effect of excluded volume crowders on folding free energy. One of these theories is the scaled particle theory (SPT). According to the SPT, the free energy cost of inserting a hard sphere of radius R in a hard sphere fluid of particles with radius R_{cr} is [31]

$$\beta F = (3x + 3x^2 + x^3)\psi + (\frac{9x^2}{2} + 3x^3)\psi^2 + 3x^3\psi^3 - \ln(1 - \phi_{\mathrm{c}})\,, \qquad (6.4)$$

where $\beta = 1/k_{\mathrm{B}}T$, T is the temperature, k_{B} is the Boltzmann constant, $x = \frac{R}{R_{\mathrm{cr}}}$, $\psi = \frac{\phi_{\mathrm{c}}}{1-\phi_{\mathrm{c}}}$, and ϕ_{c} is fluid volume fraction. Minton showed that SPT predicts a strong stabilizing effect on the stability of native state of proteins if the unfolded state is modeled as a random Gaussian chain [33]. Here we apply SPT to model the free energy cost of switching between two folds of different radii.

Bibliography

[1] P. N. Bryan, and J. Orban. Proteins that switch folds. *Curr Opin Struct Biol*, 20 :482–488, 2010.

[2] A. F. Dishman, and B. F. Volkman. Unfolding the mysteries of protein metamorphosis. *ACS Chem Biol*, 13:1438–1446, 2018.

[3] A. A. Kim, and L. L. Porter. Functional and regulatory roles of fold-switching proteins. *Structure*, 29:6–14, 2021.

[4] I. Artsimovitch, and C. A. Ramírez-Sarmiento. Metamorphic proteins under a computational microscope: Lessons from a fold-switching RfaH protein. *Comput Struct Biotechnol J*, 20:5824–5837, 2022.

[5] A. G. Murzin. Biochemistry. Metamorphic proteins. *Science*, 320:1725–1726, 2008.

[6] S. H. Knauer, I. Artsimovitch, and P. Rösch. Transformer proteins. *Cell Cycle*, 11:4289–4290, 2012.

[7] B. M. Burmann, S. H. Knauer, A. Sevostyanova, K. Schweimer, R. A. Mooney, R. Landick, I. Artsimovitch, and P. Rösch. An α helix to β barrel domain

switch transforms the transcription factor RfaH into a translation factor. *Cell*, 150:291–303, 2012.

[8] L. L. Porter, and L. L. Looger. Extant fold-switching proteins are widespread . *Proc Natl Acad Sci USA*, 115:5968–5973, 2018 .

[9] R. L. Tuinstra, F. C. Peterson, S. Kutlesa, E. S. Elgin, M. A. Kron, and B. F. Volkman. Interconversion between two unrelated protein folds in the lymphotactin native state. *Proc Natl Acad Sci USA*, 105:5057–5062, 2008.

[10] D. R. Littler, S. J. Harrop, W. D. Fairlie, L. J. Brown, G. J. Pankhurst, S. Pankhurst, M. Z. DeMaere, T. J. Campbell, A. R. Bauskin, R. Tonini, M. Mazzanti, S. N. Breit, and P. M. Curmi. The intracellular chloride ion channel protein CLIC1 undergoes a redox-controlled structural transition. *J Biol Chem*, 279:9298–9305, 2004.

[11] Y. G. Chang, S. E. Cohen, C. Phong, W. K. Myers, Y. T. Kim, R. Tseng, J. Lin, L. Zhang, J. S. Boyd, Y. Lee, S. Kang, D. Lee, S. Li, R. D. Britt, M. J. Rust, S. S. Golden, and A. LiWang. A protein fold switch joins the circadian oscillator to clock output in cyanobacteria. *Science*, 349:324–328, 2015.

[12] A. F. Dishman, R. C. Tyler, J. C. Fox, A. B. Kleist, K. E. Prehoda, M. M. Babu, F. C. Peterson, and B. F. Volkman. Evolution of fold switching in a metamorphic protein. *Science*, 371:86–90, 2021.

[13] B. Ruan, Y. He, Y. Chen, E. J. Choi, Y. Chen, D. Motabar, T. Solomon, R. Simmerman, T. Kauffman, D. T. Gallagher, and others. Design and characterization of a protein fold switching network. *Nat Commun*, 14:431, 2023.

[14] I. Yadid, N. Kirshenbaum, M. Sharon, O. Dym, D. S. Tawfik. Metamorphic proteins mediate evolutionary transitions of structure. *Proc Natl Acad Sci USA*, 107:7287–7292, 2010.

[15] P. A. Alexander, Y. He, Y. Chen, J. Orban, P. N. Bryan. A minimal sequence code for switching protein structure and function. *Proc Natl Acad Sci USA*, 106:21149–21154, 2009.

[16] Y. He, Y. Chen, P. A. Alexander, P. N. Bryan, J. Orban. Mutational tipping points for switching protein folds and functions. *Structure*, 20:283–291, 2012.

[17] L. L. Chavez, J. N. Onuchic, and C. Clementi. Quantifying the roughness on the free energy landscape: entropic bottlenecks and protein folding rates. *J Am Chem Soc*, 126:8426–8432, 2004.

[18] S. Wallin, H. S. Chan. Conformational entropic barriers in topology-dependent protein folding: perspectives from a simple native-centric polymer model. *J Phys Condens Matter*, 18:S307, 2006.

[19] S. G. Estáciocio, C. S. Fernandes, H. Krobath, P. F. Faísca, and E. I. Shakhnovich. Robustness of atomistic Gō models in predicting native-like folding intermediates. *J Chem Phys*, 137:085102, 2012.

[20] A. Kluber, T. A. Burt, and C. Clementi. Size and topology modulate the effects of frustration in protein folding. *Proc Natl Acad Sci USA*, 115:9234–9239, 2018.

[21] M. Kouza, and U. H. Hansmann. Folding simulations of the A and B domains of protein G. *J Phys Chem B*, 116:6645–6653, 2012.

[22] L. Sutto, and C. Camilloni. From A to B: a ride in the free energy surfaces of protein G domains suggests how new folds arise. *J Chem Phys*, 136:185101, 2012.

[23] C. A. Ramirez-Sarmiento, J. K. Noel, S. L. Valenzuela, and I. Artsimovitch. Interdomain contacts control native state switching of RfaH on a dual-funneled landscape. *PLoS Comput Biol*, 11:e1004379, 2015.

[24] L. Xiong, and Z. Liu. Molecular dynamics study on folding and allostery in RfaH. *Proteins: Struct Funct Genet*, 83:1582–1592, 2015.

[25] B. Seifi, and S. Wallin. The C-terminal domain of transcription factor RfaH: Folding, fold switching and energy landscape, *Biopolymers*, 112:e23420, 2021.

[26] P. Galaz-Davison, E. A. Román, and C. A. rez-Sarmiento. The N-terminal domain of RfaH plays an active role in protein fold-switching. *PLoS Comput Biol*, 17:e1008882, 2021.

[27] P. K. Zuber, T. Daviter, R. Heißmann, U. Persau, K. Schweimer, and S. H. Knauer. Structural and thermodynamic analyses of the β-to-α transformation in RfaH reveal principles of fold-switching proteins. *Elife*, 11:e76630, 2022.

[28] M. H. Cordes, R. E. Burton, N. P. Walsh, C. J. McKnight, and R. T. Sauer. An evolutionary bridge to a new protein fold, *Nat Struct Mol Biol*, 7:1129–1132,2000.

[29] J. Mettle, R. B. Best. Dependence of protein folding stability and dynamics on the density and composition of macromolecular crowders. *Biophys J*, 98:315–320, 2010.

[30] T. Sikosek, E. Bornberg-Bauer, and H. S. Chan. Evolutionary dynamics on protein bi-stability landscapes can potentially resolve adaptive conflicts. *PLoS Comput Biol*, 8:e1002659, 2012.

[31] J. L. Lebowitz, and J. S. Rowlinson. Thermodynamic properties of mixtures of hard spheres. *J Chem Phys*, 41:133–138, 1964.

[32] H. X. Zhou. Protein folding in confined and crowded environments. *Arch Biochem Biophys*, 469:76–82, 2008.

[33] A. P. Minton. Models for excluded volume interaction between an unfolded protein and rigid macromolecular cosolutes: macromolecular crowding and protein stability revisited. *Biophys J*, 88:971–985, 2005.

[34] A. P. Minton. The effect of volume occupancy upon the thermodynamic activity of proteins: some biochemical consequences. *Mol Cell Biochem*, 55:119–140, 1983.

[35] H. S. Samanta, P. I. Zhuravlev, M. Hinczewski, N. Hori, S. Chakrabarti, D. Thirumalai. Protein collapse is encoded in the folded state architecture. *Soft Matt*, 13:3622–3638, 2017.

[36] J. Hong, and L. M. Gierasch. Macromolecular crowding remodels the energy landscape of a protein by favoring a more compact unfolded state. *J Am Chem Soc*, 132:10445–10452, 2010.

[37] T. Mikaelsson, J. Adén, L. B. Johansson, and P. Wittung-Stafshede. Direct observation of protein unfolded state compaction in the presence of macromolecular crowding. *Biophys J*, 104:694–704, 2013.

[38] A. G. Guseman, G. M. P. Goncalves, S. L. Speer, G. B. Young, and G. J. Pielak. Protein shape modulates crowding effects. *Proc Natl Acad Sci USA*, 115:10965–10970, 2018.

[39] M. Sarkar, C. Li, and G. J. Pielak. Soft interactions and crowding. *Biophys Rev*, 5:187–194, 2013.

[40] H. X. Zhou. Polymer crowders and protein crowders act similarly on protein folding stability. *FEBS lett*, 587:394–397, 2013.

[41] C. Holzgräfe, and S. Wallin. Smooth functional transition along a mutational pathway with an abrupt protein fold switch. *Biophys J*, 107:1217–1225, 2014.

[42] C. Holzgräfe, and S. Wallin. Local versus global fold switching in protein evolution: insight from a three-letter continuous model. *Phys Biol*, 12:026002, 2015.

[43] Y. G. Chang, S. E. Cohen, C. Phong, W. K. Myers, and others. Circadian rhythms. A protein fold switch joins the circadian oscillator to clock output in cyanobacteria. *Science*, 349:324–328, 2015.

[44] J. K. Noel, M. Levi, M. Raghunathan, H. Lammert, R. L. Hayes, J. N. Onuchic, and P. C. Whitford. SMOG 2: a versatile software package for generating structure-based models. *PLoS Comput Biol*, 12:e1004794, 2016.

[45] R. B. Best, Y. G. Chen, and G. Hummer. Slow protein conformational dynamics from multiple experimental structures: the helix/sheet transition of arc repressor. *Structure*, 13:1755–1763, 2005.

[46] T. Veitshans, D. Klimov, D. Thirumalai. Protein folding kinetics: timescales,

pathways and energy landscapes in terms of sequence-dependent properties. *Fold Des*, 2:1–22, 1997.

[47] W. S. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J Chem Phys*, 76:637–649, 1982.

[48] E. Marinari, and G. Parisi. Simulated Tempering: a new Monte Carlo scheme *Europhys Lett*, 19:451–458, 1992.

Chapter 7

Summary and outlook

In this study, our aim has been to characterize the biophysical properties of metamorphic proteins and the large-scale structural transitions they undergo. We focused on the metamorphic protein RfaH and the GA/GB fold switch system.

The energy landscape of metamorphic proteins has been the subject of many computational studies. Most studies depict energy landscapes with dual funnels, which must be the case when both folds have significant populations. To study the energy landscape of the CTD of RfaH protein, we developed a hybrid all-atom model that combines a physics-based model with a dual-basin structure-based potential. Applying this model to the isolated CTD, we found that the β-barrel fold is more favorable than the α-helical hairpin, and the energy landscape has a single funnel toward the β-barrel fold. Hence, we have found that the RfaH CTD on its own does not exhibit a dual-funnel energy landscape, contrary to the expectation for metamorphic proteins.

Using the same model, we observed a relatively high α-helix structure content in the unfolded state of isolated CTD. Moreover, this domain exhibited transient formation of α-helical structure during folding to its stable β-barrel state.

In addition, we have investigated the effect of the N-terminal domain of RfaH on domain dissociation and fold switching of RfaH. We tested a hypothesis that a change in the orientation of the $\beta3$-$\beta4$ extended hairpin plays a key role in the dissociation of CTD from NTD which is the trigger for fold switching in this protein.

The effect of macromolecular crowders on protein stability has been studied by many groups [1–5]. Under crowded conditions, the folded state of proteins is usually entropically favored. We developed a coarse-grained C_α model with a dual basin SBM and applied the model to GA/GB fold switch system to study the effect of macromolecular crowders on fold switching. We found that increasing the concentration of crowders increases the total stability of the folded states and shifts the folded population towards the GB state. Our analysis showed that it is the presence of intrinsically disordered tails, which only appear in the GA structure, that drives the population shift. It would be interesting to compare our results with an experimental study of GA/GB sequences in the presence of various types of crowders.

To further our research, we continue to explore how RNAP triggers the fold switching in RfaH. Given the complexity of the RfaH-RNAP interaction, to make further progress into this question will likely require experimental collaboration. We also plan to further investigate macromolecular effects, for example, the impact on fold switching rates and the effect of including attractive interactions between protein and crowders. Additionally, we are intrigued by the potential impact of inter-domain attractive contacts on domain dissociation in RfaH, which we can explore using our hybrid model.

Our simulation methods, which we have developed at both coarse-grained and all-atom levels, are capable of being applied to other metamorphic proteins with multiple native states. Furthermore, we can conduct fold switching tests in a more realistic crowded environment using different crowder particles, such as crowders made up of other proteins, which would be particularly interesting.

Bibliography

[1] A. J. Guseman, G. M. P. Gerardo, S. L. Speer, G. B. Young, and G. J. Pielak. Protein shape modulates crowding effects. *Proc Natl Acad Sci USA*, 115:10965–10970, 2018.

[2] H. X. Zhou. Effect of mixed macromolecular crowding agents on protein folding. *Proteins Struct Funct Bioinf*, 72:1109–1113, 2008.

[3] G. Ping, J. M. Yuan, Z. Sun, and Y. Wei. Studies of effects of macromolecular crowding and confinement on protein folding and protein stability. *J Mol Recognit*, 17:433–440, 2004.

[4] D. Gomez,K. Huber, and S. Klumpp. On protein folding in crowded conditions. *J Phys Chem Lett*, 10:7650–7656, 2019.

[5] A. Christiansen, Q. Wang, M. S. Cheung, and P. Wittung-Stafshede. Effects of macromolecular crowding agents on protein folding in vitro and in silico. *Biophys Rev*, 5:137–145, 2013.

Appendix A

Appendix

A.1 Supporting Information for chapter 3

Structural fluctuations and mechanical stabilities of the metamorphic transcription factor RfaH

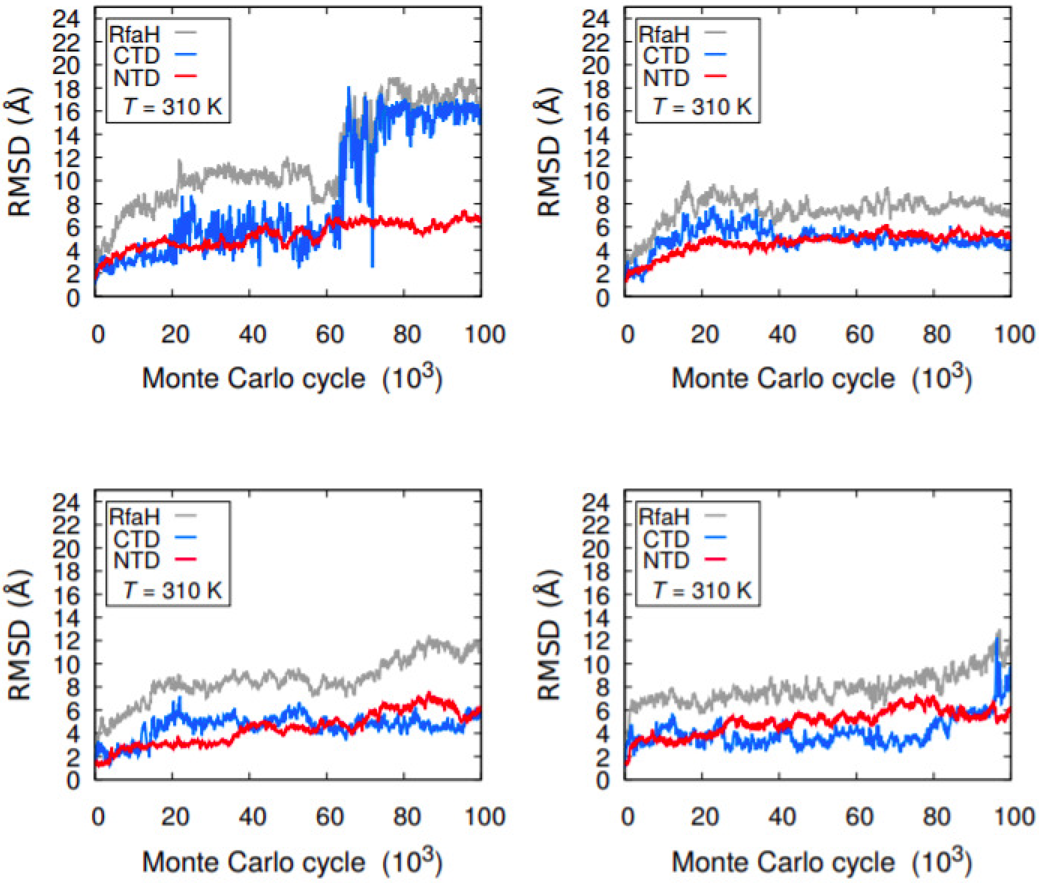

Figure S1: Shown is the root-mean-square deviation, RMSD, as a function of MC time, for 4 out of the 30 simulations carried out of domain-closed RfaH at $T = 310K$. The RMSD is determined with respect to the experimental structure of free RfaH (PDB id 2oug; see Fig. 1(left)), taken over residues 1-162 (RfaH), 1-100 (NTD) or 115-156 (CTD). All runs were initialized from a regularized version of 2oug (see Methods).

A.2 Supporting Information for chapter 4

The C-terminal domain of transcription factor RfaH: Folding, fold switching and energy landscape

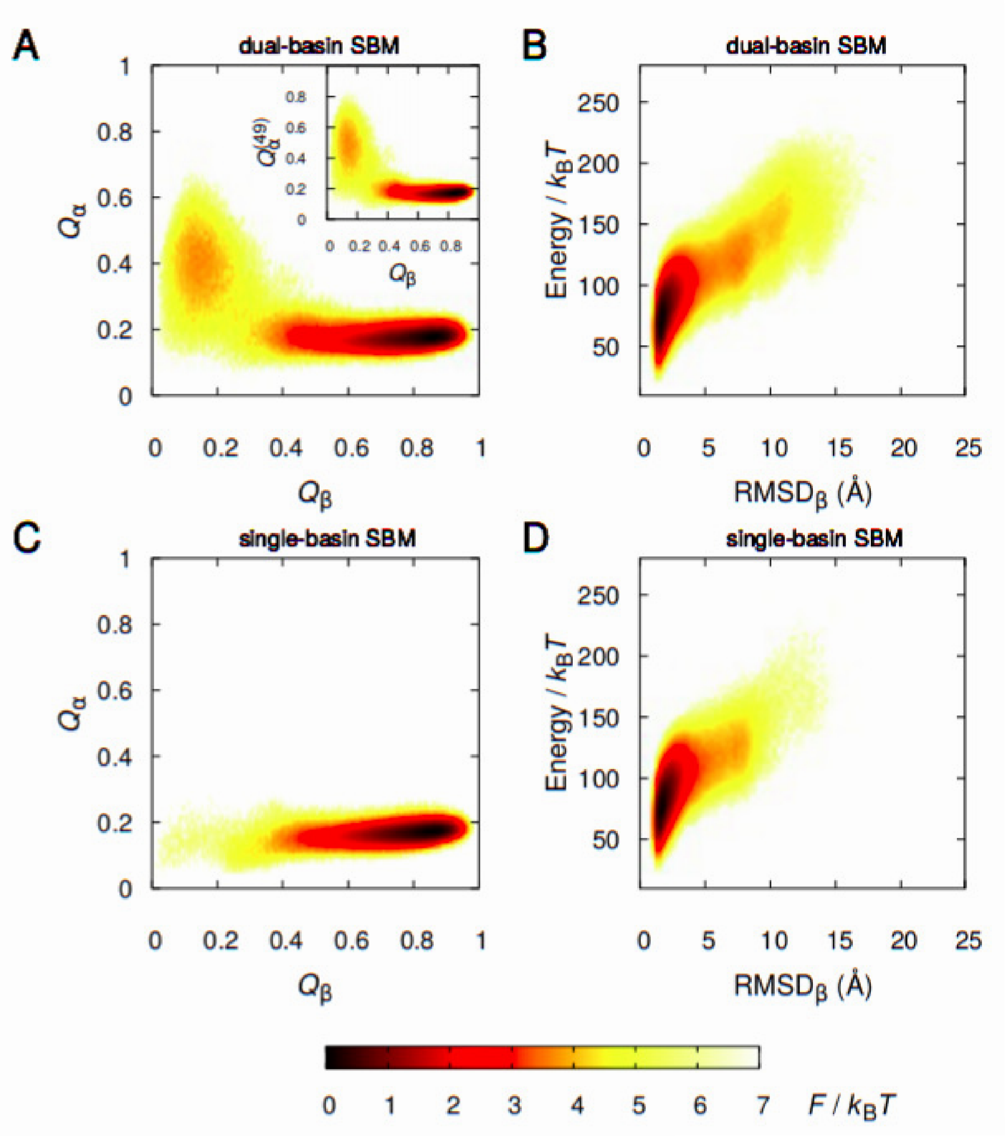

Figure S2: Free energy surfaces at low T: single-basin SBM vs dual-basin SBM. Shown are free energy surfaces $F(X_1, X_2) = -k_BT \ln P(X_1, X_2)$, with (A, B) $X_1 = Q_\alpha$ and $X_2 = Q_\beta$ or (B, D) X_1 equal to the total energy and $X_2 = \text{RMSD}_\beta$. The probability distributions $P(X_1, X_2)$ are taken at the lowest simulated temperature, i.e., $T = 365$ K for the dual-basin SBM case and $T = 370$ K for the single-basin SBM case. Insert shows the same surface $F(X_1, X_2)$ as in (A) except that $X_1 = Q_\alpha^{(49)}$ (see main text).

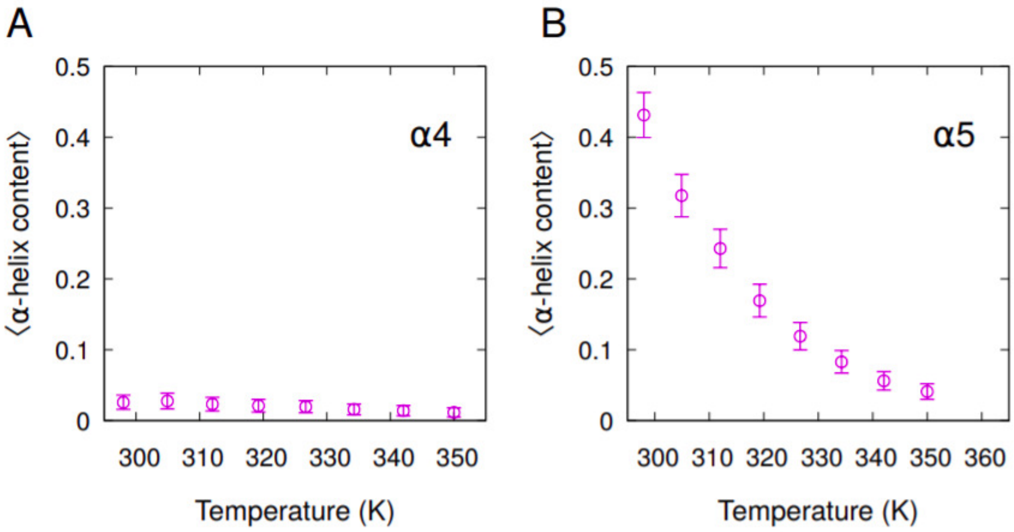

Figure S3: Temperature dependence of the average α-helix content of (A) the 15-amino acid sequence VIITEGAFEGFQAIF and the 21-amino acid sequence GEARSMLLLNLINKEIKHSVK, as obtained by our physics-based model (no SBM term included). The two sequences correspond to the α4 and α5 regions of the all-α fold of RfaH, respectively. The α-helix contents are determined using STRIDE. Averages and statistical errors are estimated using 10 independent simulated tempering runs of each 10^9 Monte Carlo step cycles.

A.3 Supporting Information for chapter 5

Examining the effect of the N-terminal domain of RfaH on domain dissociation and fold switching

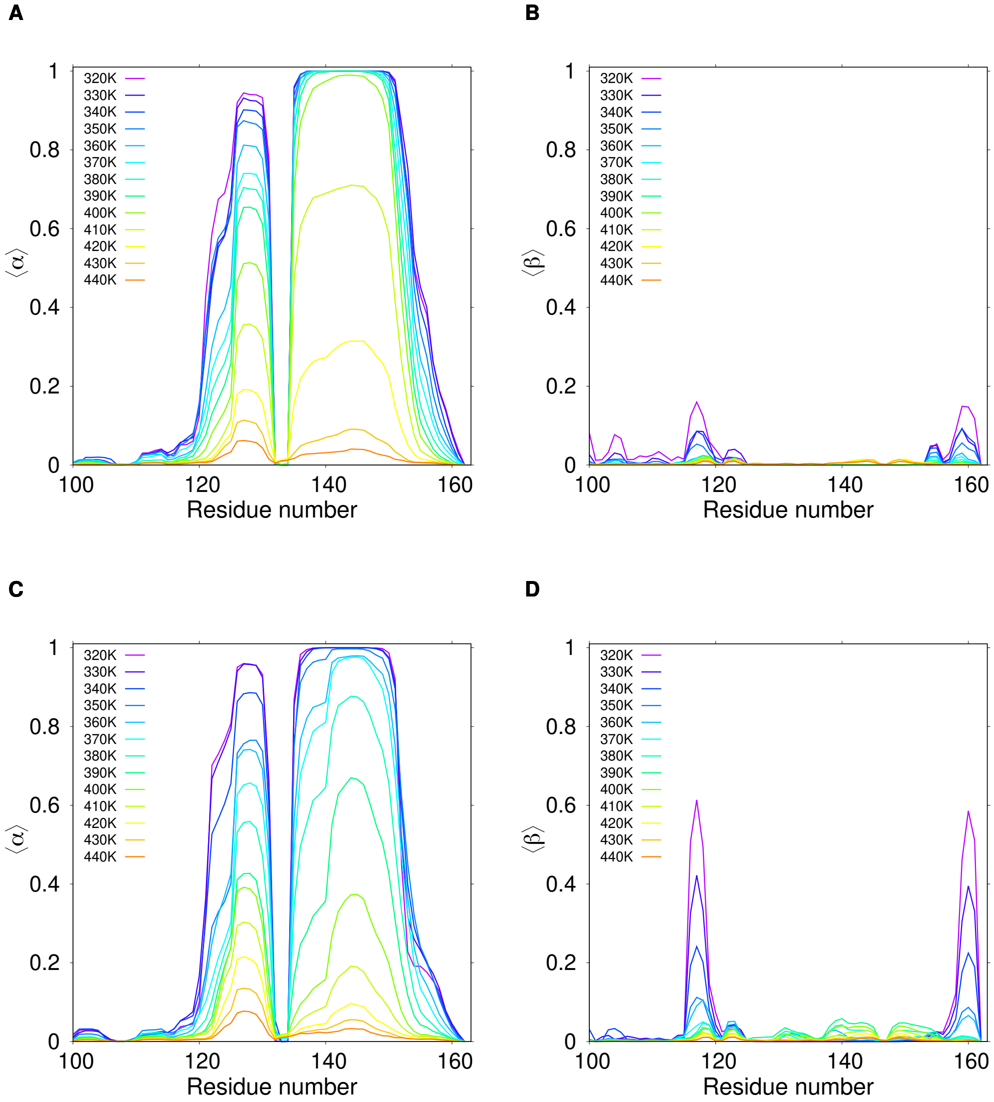

Figure S4: Stabilities of secondary structure elements in α-CTD and β-CTD for free RfaH and H_1 structures. The graph illustrates the average content of α-helix and β-sheet as a function of sequence position for 10 independent simulations started in free RfaH structure (A and B), and H_1 structure (C and D). Brackets ⟨⟩ represent an average over the 10 simulations, and data correspond to MC cycles more 2×10^6 of the simulations. Results are given for simulations carried out at 13 different temperatures from 320 to 440K. The residue numbers correspond to those of the full-length RfaH.

A.4 Supporting Information for chapter 6

Effect of crowding on a fold-switching protein is controlled by its disordered tails

A.4.1 Single-basin structure-based model for protein folding

As a starting point for the development our dual-basin structure-based model (see next section), which we apply in this work to the GA/GB switch, we take a previous single-basin model for protein folding developed in Ref. [1]. We start by describing the model in [1], along with a modification introduced to enhance the conformational specificity of native contact interactions. We find that the enhanced contact specificity is necessary to make the two folds structurally well defined in the dual-basin model. Geometrically, the protein is represented by single beads located at the C_α atom positions. The conformation of an N-amino-acid chain can therefore be described by the bead positions $\mathbf{r}_i$, where $i = 1, ..., N$. Alternatively, a conformation can also be described by the bond lengths, b_i, bond angles θ_i, and dihedral angles, ϕ_i, defined by the $N-1$ (pseudo) C_α-C_α bonds of the chain. We denote by b_i^0, θ_i^0, and ϕ_i^0 the values of b_i, θ_i and ϕ_i in the native conformation. The potential energy E can be written as a sum of five terms:

$$\begin{aligned} E &= \sum_i^{\text{bonds}} K_\mathrm{b}(b_i - b_i^0)^2 + \sum_i^{\text{angles}} K_\theta(\theta_i - \theta_i^0)^2 \\ &+ \sum_i^{\text{dihedrals}} K_\phi^{(1)}[1 - \cos(\phi_i - \phi_i^0)] + K_\phi^{(3)}[1 - \cos 3(\phi_i - \phi_i^0)]) \\ &+ \sum_{i<j-3}^{\text{nonnative}} \epsilon \left(\frac{\sigma}{r_{ij}}\right)^{12} + \sum_{i<j-3}^{\text{native}} \epsilon(h_{ij} - f_{ij}) \,, \end{aligned} \tag{A.1}$$

where ϵ sets the energy scale of the model and $r_{ij} = |\mathbf{r}_j - \mathbf{r}_i|$. The first three terms represent bonded interactions with strengths set to $K_\mathrm{b} = 100\epsilon$, $K_\theta = 20\epsilon$, $K_\phi^{(1)} = \epsilon$ and $K_\phi^{(3)} = 0.5\epsilon$. The fourth term represents steric repulsions between bead pairs that do not form a contact in the native structure. The repulsion range is set to $\sigma = 4$ Å. These first four terms in Eq. A.1 are identical to Ref. [1].

The final term in Eq. A.1 represents native contact interactions, which in the previous model [1] were described by the Lennard-Jones potential $f_{\mathrm{LJ}}(r_{ij}) = (r^0_{ij}/r_{ij})^{12} - 2(r^0_{ij}/r_{ij})^6$. Here we separate the interaction into a repulsive part (h_{ij}) and an attractive part (f_{ij}), such that they can be independently controlled. The repulsive part is described by a Weeks-Chandler-Anderson type function,

$$h_{ij} = \begin{cases} \left(\frac{r^0_{ij}}{r_{ij}}\right)^{12} - 2\left(\frac{r^0_{ij}}{r_{ij}}\right)^{6} + 1, & \text{if } r_{ij} < r^0_{ij}\,, \\ 0, & \text{if } r_{ij} \geq r^0_{ij}\,, \end{cases} \tag{A.2}$$

where r^0_{ij} is the distance between beads i and j in the native structure. The attractive part takes the form

$$f_{ij} = g_{\xi_1}(r_{ij}) g_{\xi_2}(r'_{ij}) g_{\xi_2}(r''_{ij})\,, \tag{A.3}$$

where $g_\xi(r) = \exp[-(r - r^0)^2/2\xi^2]$. With the construct in Eq. A.3, the distance r_{ij} as well as the two nearest neighbor distances, r'_{ij} and r''_{ij}, (see Figure S6) must assume their respective native values r^0_{ij}, r'^0_{ij} and r''^0_{ij} for ij to become a fully formed native contact, which then contributes $-\epsilon$ towards the total potential energy E. The parameter ξ_1 sets the width of the attractive well $-\epsilon g_{\xi_1}(r_{ij})$. The combination of this attractive well and the repulsive part of the interaction results in a function, $h_{ij} - g_{\xi_1}$, with gross features similar to a Lennard-Jones potential (see Fig. S5).

The factor $g_{\xi_2}(r'_{ij}) g_{\xi_2}(r''_{ij})$ is included in f_{ij} in order to increase the conformational specificity of native interactions. It promotes the local chain segments ($i-1$, i, $i+1$) and ($j-1$, j, $j+1$) to adopt a relative orientation close to that found in the native structure. The strength of this effect is controlled by the parameter ξ_2. It is weak when $\xi_2 \gg \xi_1$ and becomes strong when $\xi_2 \approx \xi_1$. Test simulations on a few small single domain proteins show that decreasing ξ_2 leads to increased folding cooperativity. We

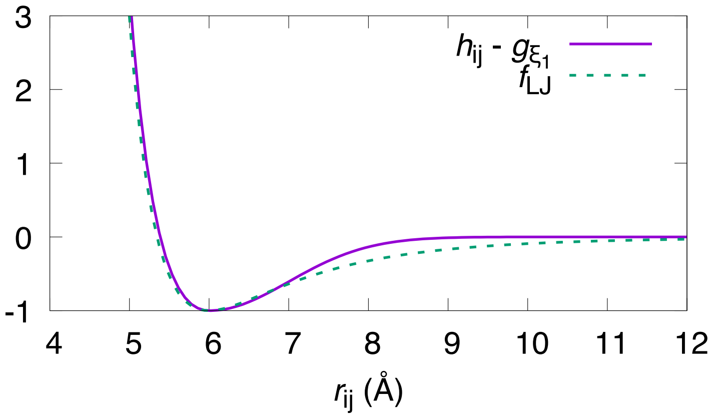

Figure S5: The potentials $h_{ij} - g_{\xi_1}$ and f_{LJ} (see text) as functions of r_{ij} using $r^0_{ij} = 6$ Å.

picked $\xi_1 = 1.0$ Å and $\xi_2 = 5.0$ Å. We note also that there are terms in Eq. A.3 for which r'_{ij} or r''_{ij} is undefined because i or j is a terminal bead. In those cases, we set the corresponding factor $g = 1$.

The effect from the factor $g_{\xi_2}(r'_{ij})g_{\xi_2}(r''_{ij})$ in Eq. A.3 is similar to so-called local-nonlocal coupling, which also leads to increased folding cooperativity [2]. Our effect is not exactly the same, however, because it does not provide a direct constraint on the local internal conformation around beads i and j, which exists in local-nonlocal coupling.

A.4.2 Dual-basin structure-based model for fold switching

Next we extend the model of the previous section to a dual-basin (db) model, which provides bias towards two different reference structures "(a)" and "(b)". Such a bias can be achieved by first obtaining the two single-basin energy potentials $E^{(\mathrm{a})}$ and $E^{(\mathrm{b})}$ using Eq. A.1, and thereafter merging them into a single energy surface, $E^{(\mathrm{db})}$.

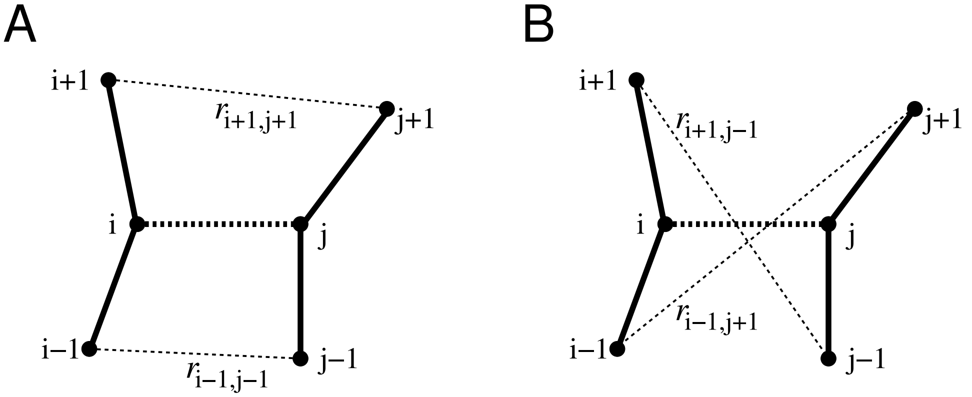

Figure S6: A contact between two non-terminal positions, i and j, (thick dashed line) has four different nearest neighbor-nearest neighbor distances (thin dashed lines): (A) $r_{i-1,j-1}$ and $r_{i+1,j+1}$ and (B) $r_{i-1,j+1}$ and $r_{i+1,j-1}$. In evaluating the factor $g_{\xi_2}(r'_{ij})g_{\xi_2}(r''_{ij})$ in Equation A.1, r'_{ij} and r''_{ij} are the distances shown in (A), if $\Sigma_{\mathrm{A}} < \Sigma_{\mathrm{B}}$, or in (B), if $\Sigma_{\mathrm{B}} < \Sigma_{\mathrm{A}}$, where $\Sigma_{\mathrm{A}} = r^0_{i-1,j-1} + r^0_{\mathrm{i+1,j+1}}$ and $\Sigma_{\mathrm{B}} = r^0_{i-1,j+1} + r^0_{i-1,j+1}$.

Naively, one may attempt to put $E^{(\mathrm{db})} = E^{(\mathrm{a})} + E^{(\mathrm{b})}$. However, this strategy is problematic for some types of interactions, as pointed out by Ramirez-Sarmiento et al. [22]. For example, the sum of two quadratic bond terms $K_{\mathrm{b}}[(b_i - b_i^{(\mathrm{a})})^2 + (b_i - b_i^{(\mathrm{b})})^2]$ is another quadratic function with minimum at $(b_i^{(\mathrm{a})} + b_i^{(\mathrm{b})})/2$. Hence, this would abolish both minima. We combine the two single-basin potentials $E^{(\mathrm{a})}$ and $E^{(\mathrm{b})}$ using the procedure described below, which avoids these problems. This procedure is then applied to the GA and GB folds to produce the dual-basin potential used in this work.

Bonded terms. The bonded interactions are represented by the first three terms in Eq. A.1. Consider two individual energy terms, $e^{(\mathrm{a})}(x)$ and $e^{(\mathrm{b})}(x)$, with global minimum at $x = x^{\mathrm{a}}$ and $x = x^{\mathrm{b}}$, respectively. The functions $e^{(\mathrm{a})}(x)$ and $e^{(\mathrm{b})}(x)$ could be, e.g., the bond angle terms corresponding to a particular bond, in which case

$x = \theta_i$. To "mix" $e^{(\mathrm{a})}(x)$ and $e^{(\mathrm{b})}(x)$ into a single function $e(x)$, we use [3]

$$e(x) = \beta_{\mathrm{mix}}^{-1} \ln \left[e^{-\beta_{\mathrm{mix}} e^{(\mathrm{a})}(x)} + e^{-\beta_{\mathrm{mix}} e^{(\mathrm{b})}(x)} \right] , \tag{A.4}$$

where β_{mix} is a parameter controlling the smoothness of the mixing. We pick $\beta_{\mathrm{mix}} = 10$ for the bond term, and $\beta_{\mathrm{mix}} = 5$ for the angle and torsion terms. Examples of three different terms for the GA and GB folds are given in Fig. S7.

Non-bonded terms. For the native contact term, we include all contact interactions present in either $E^{(\mathrm{a})}$ or $E^{(\mathrm{b})}$. Although this is straightforward in principle, care must be taken to avoid double counting interactions for common contacts, i.e., contacts that occur in both structure (a) and structure (b). Moreover, we want to insert parameters κ_{A} and κ_{B} such that depths of the attractive wells $-\epsilon f_{ij}^{(\mathrm{a})}$ and $-\epsilon f_{ij}^{(\mathrm{a})}$ can be controlled. Hence, our dual-basin contact term becomes

$$\sum_{ij}^{(\mathrm{a})} \epsilon(h_{ij}^{(\mathrm{a})} - \kappa_{\mathrm{A}} f_{ij}^{(\mathrm{a})}) + \sum_{ij}^{(\mathrm{b})} \epsilon(h_{ij}^{(\mathrm{b})} - \kappa_{\mathrm{B}} f_{ij}^{(\mathrm{b})}) + \sum_{ij}^{\mathrm{common}} \epsilon \left\{ \tilde{h}_{ij} - \max \left[\kappa_{\mathrm{A}} f_{ij}^{(\mathrm{a})}, \kappa_{\mathrm{B}} f_{ij}^{(\mathrm{b})} \right] \right\} .$$

In the above equation, the first two sums are taken over native contacts in (a) and native contacts in (b), respectively, but exclude all common contacts. The final sum, which is taken over these common contacts, retains only the energetically most favorable attraction for each contact. The repulsive part, $\tilde{h}_{ij}$, is evaluated as h_{ij} using the reference (native) distance $r_{ij}^0 = \min \left[r_{ij}^{(\mathrm{a})}, r_{ij}^{(\mathrm{b})} \right]$. This smaller range of the repulsion is necessary to guarantee that these contacts are able to be formed in both conformations (a) and (b), without being sterically excluded by the repulsive part of the interaction. Note that the reference distance r_{ij}^0 for common contacts can be calculated before a simulation and that $\tilde{h}_{ij}$ does not change form during the simulation. The nonnative repulsive energy term, i.e., the fourth term in Eq. A.1, is evaluated over all pairs ij that are not present in either (a) or (b).

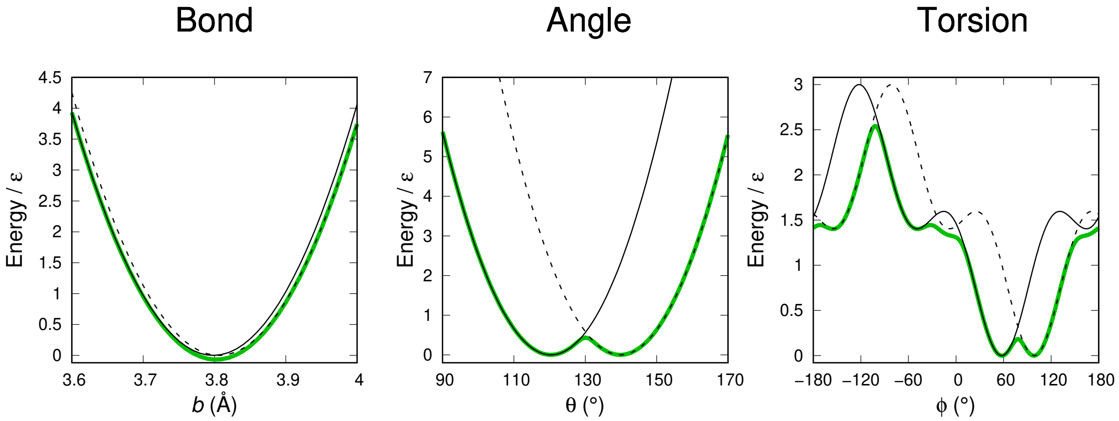

Figure S7: Examples of the merging of different bonded potentials for GA and GB (thin black solid/dashed curves) into a single potential (thick solid green curves) using the “mixing” equation A.4.

Bibliography

[1] S. Wallin, and H. S. Chan. Conformational entropic barriers in topology-dependent protein folding: perspectives from a simple native-centric polymer model. *JPCM*, 18:S307, 2006.

[2] H. S. Chan, Z. Zhang, S. Wallin, and Z. Liu. Cooperativity, local-nonlocal coupling, and nonnative interactions: principles of protein folding from coarse-grained models. *Annu Rev Phys Chem*, 62:301–326, 2011.

[3] R. B. Best, Y. G. Chen, and G. Hummer. Slow protein conformational dynamics from multiple experimental structures: the helix/sheet transition of arc repressor. *Structure*, 13:1755–1763, 2005.