Computational Study of the Biophysics of Protein Conformational Switching

by

© AINA, KAYODE ADEKUNLE

A Thesis submitted to the School of Graduate Studies in
partial fulfillment of the requirements for the degree of

Master of Science in Physics

Department of Physics and Physical Oceanography

Memorial University of Newfoundland

St. John's, Newfoundland

August 2018

Abstract

One of the basic tenets of biophysics is that a globular protein, under physiological conditions, folds spontaneously into a unique three-dimensional structure called the native state and that it dictates the biological function of the protein. However, recent experimental observations show that some proteins can undergo drastic structural rearrangements that lead to a complete change of their native folds to alternative functional folds. In order to access the underlying biophysical principles of this conformational switch, we develop and test a generalized-ensemble algorithm for biomolecular simulations that is able to calculate the thermodynamic behavior of many sequences in a single run. By applying this method to a coarse-grained model for protein folding, we explore the folding of thousands of (model) protein sequences and find that successive point mutations can lead to abrupt fold switching. Our method helps to unravel some of the biophysical properties of mutational pathways between elementary (distinct) folds and thus provide a physical explanation of the effects of mutations in conformational switching. In addition, we employ an atomistic model to characterize the fold-switching tendency in the naturally occurring protein RfaH. Our results suggest that the all-α to all-β fold switch of its carboxyl-terminal domain, in agreement with *in vitro* experiments, is thermodynamically favored. Providing a physical basis for protein fold switching, and ultimately the ability to design them, may have an extensive impact in biology and biotechnology.

Co-authorship Statement

The following people contributed to the publication of work undertaken as part of this thesis:

Adekunle Aina, *Memorial University*: **Candidate**

Stefan Wallin, *Memorial University*: **Supervisor**

Author details and their roles:

Paper: A. Aina and S. Wallin, "Multisequence algorithm for coarse-grained biomolecular simulations: exploring the sequence-structure relationship of proteins", *J. Chem. Phys.* **147 (9)**, 095102 (2017).

Location: Chapter 2

Contribution: Candidate was the primary author and the supervisor contributed to the conception and design of the research project and drafted significant parts of the paper. Candidate performed all computer experiments and simulations. Candidate wrote the computer code for data analysis. Candidate and supervisor contributed to the analysis and interpretation of the research data. In particular, candidate analyzed the computational efficiency of the multisequence algorithm and produced some of the figures. Candidate and supervisor wrote and critically revised the manuscript. Candidate specifically wrote the "Results section" and proof read the whole manuscript. Candidate contributed approximately 60% to the planning, execution and preparation of the paper for publication.

Acknowledgements

I would like to thank my supervisor, Dr. Stefan Wallin, for his guidance and tutelage, the Natural Sciences and Engineering Research Council of Canada and Memorial University for funding this research work, and the Department of Physics and Physical Oceanography for availing me this opportunity. For these, I am most grateful.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

One of the most established facts in biology is that certain biomolecules known as proteins are largely responsible for the many observed characteristics in life. Living organisms including plants and animals have genetic information encoded in the form of deoxyribonucleic acid (DNA) that is passed down almost perfectly from one generation to another. There are two main biochemical processes which occur in living cells: one involving the transcription of DNA to messenger ribonucleic acid (mRNA) and the other translates mRNA to proteins, such that the main information flow in a cell can be summarized as:

$$\text{DNA} \xrightarrow{\text{transcription}} \text{mRNA} \xrightarrow{\text{translation}} \text{PROTEIN}$$

Thus, transcription and translation are essential for the continuous existence of each cell and ultimately that of the host organism.

Proteins are molecular machines that perform wide spectrum of functions in living organisms. They are the building blocks and arms of all living cells [1]. They catalyze biochemical reactions, control gene expression, mediate intracellular signals, transport and store other molecules [2]. Moreover, proteins also provide and maintain the needed

structural support in cells and tissues. The vast biological functions performed by proteins make them essential for life and important entities to physically explore.

A fundamental tenet of biophysics is that a protein assumes (or folds into) a unique, stable, and functional three-dimensional structure under physiological conditions, involving a structural transition from a disordered state to an ordered state. Structural rearrangement from a disordered state to an ordered state is ubiquitous in the study of proteins [3], which is not surprising since the folding process is central to their functionality. This assumption of a unique and stable, so-called native structure led (and reasonably so) to the search for "folding pathways" for many years. Folding pathway refers to a linear sequence of events between the unfolded and folded protein. A native structure implies that a protein is expected to perform a specific function. Truly, the overwhelming majority of globular proteins meet this expectation [4].

However, recent experimental observations [5, 6] suggest that some proteins have the interesting ability to rearrange their native conformation into an alternative functional fold. This can be as a result of mutational changes, or interaction with a different biological environment, or even for non-obvious reasons. Such large-scale rearrangement can involve major changes in secondary structures, repacking of the protein core, and exposure of new surfaces [7].

One of the most dramatic examples of a protein that undergoes structural rearrangement is *Lymphotactin*, which exists in two forms (Ltn10 and Ltn40) in almost equal amounts under native conditions [8]. While Ltn10 adopts a monomeric chemokine fold, Ltn40 has a dimeric β-sandwich fold. Another example of a fold-switching protein is RfaH, which is the focus of one of the projects in this thesis. RfaH is a compact two-domain protein from *Escherichia coli*. Its C-terminal domain has been shown experimentally to be able to undergo a complete conformational change from an α-helix bundle to a β-barrel structure [6]. A number of other studies [5, 9–12]

have indicated that alternative native conformations can co-exist in equilibrium. This cumulative data make stronger the argument that a class of proteins demonstrate the ability to switch their fold, in contrast to classical belief of maintaining a specific fold, thereby expanding their functional capability.

While it is true that the observations of protein fold switching are rare, this phenomenon can be a window through which we expand our understanding of how new folds and functions may arise in evolution. After considering some examples of conformational switching in both naturally occurring and designed proteins, Bryan and Orban [3] highlighted three common features; (i) the structural transitions require states with marginal stability, (ii) disordered regions can facilitate these structural transitions, and (iii) a new binding surface is exposed in the alternative folds.

A question one may genuinely ask is why is protein fold switching important to study? Or to put it into perspective, why is it important to understand the physics of conformational switching in proteins? While there is no strict answer to this question, what we can say is that if an observation as radical as fold switching, which challenges a basic tenet of biophysics and long-time dogma of biochemistry is made, that potentially changes everything from our understanding of protein folding itself to protein function, and protein evolution. Its implications [7] may extensively impact areas such as computational and structural biology, human disease, protein design, and biotechnological applications.

The purpose of this thesis is to understand the phenomenon of protein fold switching, to test a new algorithm for biomolecular simulation, to study in particular fold switching tendency in a naturally occurring protein, and to discuss how fold switching may lead to the evolution of new folds and functions. The biophysical properties of proteins are key to understanding their biological and physical implications. Consequently, in this thesis we seek to explore them. Specifically, this thesis focuses on

the biophysics of protein fold switching, as well as the computational tools needed to explore it.

We first consider a generalized-ensemble algorithm [13] for coarse-grained simulations of biomolecules which allows the thermodynamic behavior of two or more protein sequences to be determined in a single "multisequence" run. To explore the biophysical mechanism underlying fold switching through point mutation, we test the method on an intermediate-resolution coarse-grained model for protein folding with three amino acid types and apply the method to sets of more than a thousand sequences. The resulting thermodynamic data allow us to carry out a more systematic analysis of the biophysical properties of sequences along mutational pathways connecting two pairs of folds than has been previously possible. We then utilize an atomistic model to computationally study the fold-switching tendency in the RfaH protein. The relatively low stability of the α-helical bundle form of its C-terminal domain suggests that it may be primed to switch into the β-barrel structural form.

In the following sections, a short background on proteins is given. First, we highlight amino acids as the building blocks of proteins followed by how a linear chain of these amino acids uniquely specifies a protein. Then, the physical interactions in proteins and how they result in an important biological process of folding is presented. What remains in this introductory chapter includes the fundamental physics, molecular models and the simulation methods deployed to gain insight into the phenomenon of protein conformational switching.

1.1 Amino acids: The building blocks of proteins

A class of chemical molecules having both the carboxilic acid (-COOH) and amine ($-NH_2$) groups are referred to as amino acids. Examples include Glycine, Leucine,

Proline, Cysteine, Tyrosine, Aparagine etc. They have the general structure as shown in Fig.1.1.

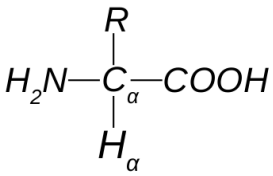

Figure 1.1: General structure of amino acids

The -R (side chain) group differentiates one amino acid from the others. There are 20 naturally occurring amino acids [2]. All of them, except proline, have the general structure shown in Fig.1.1. Proline is an exception in the sense that its -R group bonds the N-atom stripping it of an H-atom. Glycine is also unique because its C_α (α-carbon) is not chiral like the other nineteen. That is, it is symmetric with respect to reflection. A chiral carbon is covalently bonded to four different groups whereas glycine (with R=H) has two identical H-atoms.

The chemical composition of the side chain determines the properties of each amino acid [2]. Amino acids can be polar, non-polar, charged or uncharged.

1.2 Protein: Linear chain of amino acids

Proteins are heterogeneous chain polymers (Fig.1.2). Each consists of tens or hundreds and sometimes thousands of amino acid residues[†].

All proteins are synthesized from the 20 naturally occurring amino acids. The residues are linked together by rigid covalent peptide bonds between the C and N atoms, which makes (free) rotation about them practically very difficult with the exception of proline having both cis and trans conformations.

[†]A "residue" is the portion of a free amino acid that remains after polymerization.

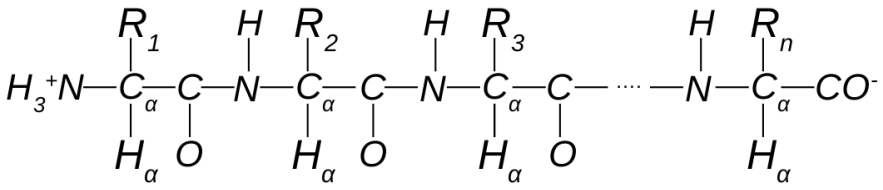

Figure 1.2: Polymer chain of proteins are formed after polymerization of n amino acids. One amino acid residue is covalently bonded to the other by a peptide bond. These are single bonds but have a double bond character which makes them rigid [1].

As can be deduced from Fig.1.2, the linear sequence of amino acids specifies a protein and this is unique in the sense that no two proteins have the same amino acid sequence. Because proteins typically have long chains and there are 20 different naturally occurring amino acids, the possible number of sequences is enormous. Presently, the amino acid sequences of more than 1.5 million proteins [14] are known and even more are being determined. Once the linear chain of a protein has been formed, physical interactions take place in and around the chain, which leads to an important biophysical process known as protein folding.

1.3 Physical interactions in proteins

The main interactions between atoms of the same amino acid or between adjacent amino acid residues are strong covalent (peptide) bonds. That is, covalent bonds are the dominant forces holding the atoms of a protein molecule together. However, other individual non-covalent interactions that are weaker than covalent bonds act collectively to provide strong attractive forces also. These non-covalent interactions include hydrogen bonds, ionic bonds and van der Waals interactions.

There is also a disulfide bond which sometimes exists between pairs of cysteine amino acids. The most important by far among these interactions is the hydrogen

bond owning to its effective ability to stabilize structural elements [15,16]. Hydrogen bonds between water molecules are also responsible for the hydrophobic effect that is essential for protein folding. Hydrophobic effect results from the non-interaction of protein molecules with surrounding water molecules. Because of their importance in protein folding and stability of regular structures, a brief look at hydrogen bonds is discussed next.

1.3.1 Hydrogen bond

A hydrogen (H) atom that is covalently bonded to highly electronegative atoms of oxygen (O) or nitrogen (N) acquires a partial positive charge while the O or N atom acquires a partial negative charge. In the presence of another electronegative atom there exists an electrostatic attraction to the H atom. This attractive force is the so-called hydrogen bond (Fig.1.3) .

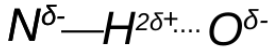

Figure 1.3: Hydrogen bond.

The protein backbone includes -NH and -CO groups that are potentially capable of participating in H-bonds, whereby the N atom is the donor and the O atom is the acceptor (of H atom). Some side chains can also participate in H-bonds. Hence, there are backbone-backbone, backbone-side chain or side chain-side chain H-bonds. However, backbone-backbone H-bonds are the most common [1]. They are significantly important in the sense that they are responsible for the stability of structural motifs like α -helices and β -sheets frequently found in protein structures. H-bonds are highly directional because of their polar nature. Hence, they are strongest when the N, H and O atoms are aligned. A major consequence of H-bonds in an aqueous

environment is the hydrophobic effect [17].

1.3.2 Hydrophobic effect

The natural surrounding of globular proteins is mainly water whose molecules are networked together by H-bonds with O atoms acting as both donors and acceptors respectively. This loose network of H-bonds of water is disrupted by non-polar molecules that can not participate in it. As a result, group of non-polar molecules assemble together in an aqueous environment in order to minimize the surface area exposed to water. This is the hydrophobic effect [17]. It is entropically driven since it allows maximum entropy (freedom) possible for the solvent. That is, there is a minimum number of "trapped" water molecules close to a non-polar surface.

Protein molecules typically have both polar and hydrophobic (non-polar) amino acids. In an aqueous environment, the non-polar amino acids are usually buried inside the protein structure to form a hydrophobic core. Consequently, polar amino acids are mostly found on the surface.

1.4 Protein folding: From linear chain to 3D structure

Many proteins *in vivo* do not remain in their primary linear chain conformation but undergo an important biophysical process of folding into essentially unique three dimensional structures on a biological time scale. That is, a protein which is usually a long chain of many amino acids, falls over itself to form a compact so-called native structure within milliseconds [15]. This observation alone is incredible but there is more. In the early 1960's, Christian Anfinsen *et al.* in a series of experiments [18, 19] showed that proteins can fold reversibly suggesting that the native structures

are thermodynamically favorable states. Out of the indefinite number of possible conformations available, how does the protein chain identify its native conformation? And how does this happen within a very short time merely by random searching? This reasoning, termed "Levinthal paradox", led to the search for specific folding pathways [20, 21]. Because the number of possible states grows exponentially with chain length [22], it therefore seems that achieving the global free energy minimum and to do so in biological time scales are mutually exclusive. By contrast, the "funnel view" of protein folding via energy landscape folding funnels (rather than specific folding pathways) was discussed extensively by Dill and Chan [22]. Contrary to the folding pathway hypothesis, which requires a linear sequence of events, the folding process is parallel in the funnel view. This implies a many-pathway process where ensemble of conformations determines the folding behavior.

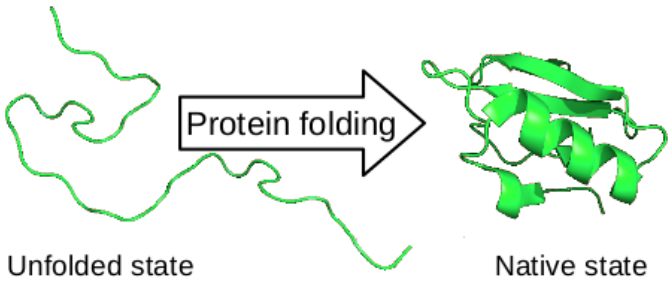

Figure 1.4: Protein folding: A linear sequence of amino acids folds reversibly into a 3D structure under native conditions, suggesting sequence encodes structure. Figure adapted from www.commons.wikimedia.org

How a protein chain is guided[†] towards the native structure lies in the physical interactions within and around the chain. H-bonds, electrostatic interactions, van der Waals interactions and the hydrophobic effect, all play part in this folding process.

[†]The estimated time [20] required for a random search of the native structure in the indefinite conformational space suggests that the folding process is guided.

However, it is widely believed that the hydrophobic effect is the overwhelming force responsible for the collapse of protein chains [17].

1.5 Structural motifs in proteins

As mentioned in the previous section, many proteins fold into 3D native structures which corresponds to the global minimum of their accessible free energy. Although these structures vary widely from one protein chain to another, there are commonly repeating structural elements that constitute what are called secondary structures.

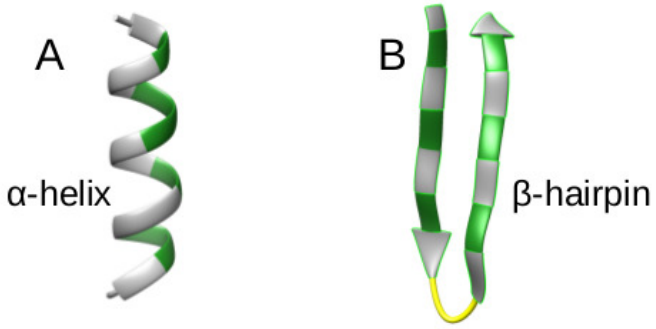

Figure 1.5: Secondary structures commonly found in proteins, (A) α-helix and (B) β-sheets. They are stabilized mainly by a regular H-bond interaction between the back-bone donors and acceptor atoms.

A regular H-bond between -NH and -CO groups of different amino acid residues respectively in a protein back-bone gives rise to regular structures such as α-helices and β-sheets shown in Fig.1.5. The α-helix is mainly (partly) stabilized[†] by H-bonds between the -NH group of one amino acid residue i and the -CO group of the amino acid residue $i-4$. On the other hand β-structures which could be parallel or anti-parallel are stabilized by H-bonds between adjacent strands and usually have turns.

[†]There are also helices without any hydrogen bonds. Their tight energetically favourable arrangement is stabilized by van der Waals interactions only [1].

A β-structure with a single turn as shown in Fig.1.5B is so-called β-hairpin. Other regular structures like π-helices, 3_{10}-helices and 2_7-helices are sometimes observed but are much less frequent due to conformational strain or steric constraints [1].

1.6 Some fundamental physics

This section presents the underlying physics of the simulation methods used in this thesis.

1.6.1 Free energy

All physical processes including biophysical reactions are constrained by the laws of thermodynamics. While many of these processes require the conservation of energy within the system of interest, others do not but rather interact with the surrounding and proceed under constant temperature. It is notable that a system at constant temperature is capable of extracting heat for free from its surrounding and if annihilated, the entropy[†] of the system is disposed into the surroundings as heat. Therefore, the energy required to create such a system from nothing (or the energy that can be extracted from such a system as useful work) is given by the Helmholtz free energy

$$F = E - TS \tag{1.1}$$

where E, T and S represent the total internal energy, temperature and entropy of the system, respectively. If in addition the system is under constant pressure, the free

[†]Entropy S is a measure of the possible number of conformations in a system. $S = k_B \ln\Omega$. Where k_B is the Boltzmann constant and Ω is the number of conformations.

energy in this case becomes

$$G = E - TS + PV \tag{1.2}$$

where P is the pressure and V is the volume. The thermodynamic quantity G is known as the Gibbs free energy. Physical processes occurring at constant temperature are favorable when either F or G decreases. Hence, free energy is a force towards equilibrium [23].

1.6.2 Boltzmann distribution

For a system in thermal equilibrium with a surrounding heat bath at constant temperature, a useful question is what is the probability of finding it in one of its possible conformations? The answer is given by the Boltzmann distribution

$$P_{\mathrm{B}}(r) = \frac{\exp\left[-E(r)/k_B T\right]}{Z} \tag{1.3}$$

where $\exp\left[-E(r)/k_{\mathrm{B}}T\right]$ is the Boltzmann factor and $Z = \sum_r \exp\left[-E(r)/k_{\mathrm{B}}T\right]$ is the partition function. $E(r)$ is the energy of the system in conformation r, and the sum is taken over all conformations r. The thermodynamic average of an observable O can be computed using

$$\langle O \rangle = \sum_r O(r) P_{\mathrm{B}}(r) = \frac{1}{Z} \sum_r O(r) \exp\left[-E(r)/k_{\mathrm{B}}T\right] \tag{1.4}$$

It can be shown that the Helmholtz free energy F is related to the partition function Z. Indeed, this formula turns out to be

$$F = -k_{\mathrm{B}} T \ln Z \tag{1.5}$$

or equivalently,

$$Z = \exp\left[-F/k_{\mathrm{B}}T\right]. \tag{1.6}$$

1.7 Molecular simulation methods

Computer simulations have, for many years now, become an important tool for solving problems that are otherwise difficult and sometimes practically impossible to solve analytically or providing information that is difficult to obtain experimentally. Molecular dynamics (MD) and Monte Carlo (MC) techniques are two of the most common conformational sampling methods. The general idea of any molecular simulation is to numerically calculate, based on relevant model and theory, the properties of a physical system.

1.7.1 Molecular dynamics

A very common method to simulate the time evolution of a macromolecular system is molecular dynamics (MD)[†]. MD generates states by numerically solving Newton's equations of motion for many-body system interacting via a particular force field. Thus the time evolution of the system is determined by solving

$$\mathbf{F}_i = m_i \frac{\partial^2 \mathbf{r}_i}{\partial t^2} \tag{1.7}$$

for each particle i in the system.

The force $\mathbf{F}_i$, on particle i, is calculated from $\mathbf{F}_i = -\nabla_i U(\mathbf{r}^N)$, where $U(\mathbf{r}^N)$ is the interaction potential which may include bonding potentials and long-range potentials such as electrostatics and van der Waals interactions.

[†]The first molecular dynamics simulations [24–26] were performed shortly after Metropolis and co-workers introduced Markov chain Monte Carlo method (discussed in the next section).

The time average of a physical observable O is calculated by averaging over time along a trajectory, i.e.,

$$\bar{O} = \lim_{t\to\infty} \frac{1}{t} \int_0^t d\tau \; O(r(\tau),\; \dot{r}(\tau)). \tag{1.8}$$

Ergodicity principle states that the average over periods of time along a given trajectory of an observable is, at the limit, identical to the ensemble average of the same observable. That is, $\bar{O} = \langle O \rangle$. Ergodicity is important when performing MD simulations [27]. MD simulation must be sufficiently long such that it is longer than the relevant relaxation time in the system.

1.7.2 Markov chain Monte Carlo method

The basic idea of a Markov chain Monte Carlo (MC) method is to generate a sequence of conformations $r_1, r_2, \ldots, r_i, \ldots$ that are biased according to a desired probability distribution. Each new conformation r_i is generated based only on the previous conformation r_{i-1}, i.e., without using $r_1, r_2, \ldots, r_{i-2}$. In 1953, Metropolis *et al.* [28] put forward a Markov chain method that generates states weighted according to the Boltzmann distribution in the limit of large N, i.e.,

$$\lim_{\mathrm{N}\to\infty} P(r = r_{\mathrm{N}}) \to P_{\mathrm{B}}(r). \tag{1.9}$$

In that case, the thermodynamic average of an observable O then becomes

$$\langle O \rangle = \frac{1}{N} \sum_{i=1}^{N} O(r_i) \quad \text{for large N.} \tag{1.10}$$

This approximation requires a proper sampling of the conformational space such that the conformations are practically uncorrelated.

The "Metropolis method" chooses a configuration $r^{'}$ given that a system currently has a conformation r with an acceptance probability $A(r \longrightarrow r^{'})$. The sufficient but not necessary requirement [27, 29] for this method to generate the representative configurations $\{r_1, r_2 \ldots r_N\}$ is to satisfy the condition of detailed balance[†]

$$P_{\mathrm{B}}(r)T(r \longrightarrow r^{'})A(r \longrightarrow r^{'}) = P_{\mathrm{B}}(r^{'})T(r^{'} \longrightarrow r)A(r^{'} \longrightarrow r) \tag{1.11}$$

where $T(r \longrightarrow r^{'})$ is the probability of proposing the configuration $r^{'}$ given that the current configuration is r. If the proposal probability is symmetrical, that is, $T(r \longrightarrow r^{'}) = T(r^{'} \longrightarrow r)$ then equation 1.11 becomes

$$\frac{A(r \longrightarrow r^{'})}{A(r^{'} \longrightarrow r)} = \frac{P(r^{'})}{P(r)} = \exp\left[-(E(r^{'}) - E(r))/k_B T\right] \tag{1.12}$$

or

$$A(r \longrightarrow r^{'}) = A(r^{'} \longrightarrow r)\exp\left[-\Delta E/k_B T\right] \tag{1.13}$$

where $\Delta E = E(r^{'}) - E(r)$. The Metropolis condition for this to happen is given by $A(r \longrightarrow r^{'}) = \min\left[1, \exp\left(-\Delta E/k_B T\right)\right]$ [28]. It implies that a higher energy conformation is accepted with probability $\exp\left[-\Delta E/k_B T\right]$ while a lower energy conformation is always accepted. Practically, this can be achieved by drawing a random number $\rho \in [0, 1]$ from a uniform distribution and accepting the new state if $\rho < \exp\left[-\Delta E/k_B T\right]$.

A general problem with this Metropolis method, especially at low temperatures, is that successive conformations $r_i^{'}s$ are typically correlated. This is particularly the case for proteins, due to the rough energy landscapes and several local energy minima. Thus, the system can easily get trapped in small region of conformational space for a long time. To solve this inherent problem, many generalized ensemble methods like

[†]The term "detailed balance" was coined by Hansmann and Okamato [30]

multi-canonical ensemble [31,32], umbrella sampling [33], expanded ensembles [34,35] and replica exchange [36] have been developed[†].

1.7.3 The method of generalized ensembles

The fundamental idea of the generalized ensemble method is to include additional variables, or what are sometimes called dynamic parameters, beyond the physical variables (e.g. position and momentum), thereby increasing the size of the configuration space. This approach helps to accelerate simulations of systems with rough energy landscape [35] by allowing deep local minima to be more easily escaped. Two generalized ensembles are used in this thesis: simulated tempering [35] and the multisequence algorithm [13].

Simulated tempering

In simulated tempering, the temperature is made a dynamic parameter. Thus, the probability distribution becomes

$$P(r,m) = \frac{1}{\check{\mathrm{Z}}} \exp[-E(r)/k_B T_m + g_m] \tag{1.14}$$

where the normalization factor $\check{\mathrm{Z}} = \sum_r \sum_m \exp[-E(r)/k_B T_m + g_m]$. T_m is a set of temperatures that the system is allowed to visit and g_m is a set of simulation parameters that determine the marginal probability distribution

$$P(m) = \frac{1}{\check{\mathrm{Z}}} \sum_r \exp[-E(r)/k_B T_m + g_m] \tag{1.15}$$

[†]See ref. [37] for a recent but brief history of the introduction of generalized ensemble to Markov chain Monte Carlo simulations.

which can also be written as $P(m) = \frac{1}{\tilde{Z}} \exp[-F_m/k_BT_m + g_m]$. Here, F_m is the free energy at temperature T_m. In order to ensure all temperatures are frequently visited, the g_m parameters must be chosen very carefully. This is done by one or more "tuning" runs. The g_m parameters are updated with values from a previous run until an approximately uniform distribution is achieved or equivalently, when the input and the output values of these g_m are roughly the same. At this point, $g_m \approx F_m/k_BT_m$ and $P(m) \approx \frac{1}{\tilde{Z}}$ for all temperatures. The joint probability $P(r, m)$ is simulated by using separate ordinary Monte Carlo updates of conformation r and temperature m.

Multisequence algorithm

In this thesis, we test a generalized ensemble method where the biological sequence of a macromolecule (instead of the temperature as in simulated tempering) is made a dynamic parameter. Thus, the method simulates the joint probability distribution

$$P(r, s) = \frac{1}{Z} \exp[-E(r, s)/k_BT + h(s)] \tag{1.16}$$

where the normalization factor $Z = \sum_r \sum_{s \epsilon S} \exp[-E(r, s)/k_BT + h(s)]$ and S is a set of pre-selected sequences. The simulation parameters $h(s)$, similar to g_m in simulated tempering, determine the marginal distribution

$$P(s) = \frac{1}{Z} \sum_r \exp[-E(r, s)/k_BT + h(s)] \tag{1.17}$$

which is equivalent to $P(s) = \frac{1}{Z} \exp[-F_s/k_BT + h(s)]$, where F_s is the free energy for sequence s at temperature T. Again, for a roughly flat distribution of $P(s)$, $h(s)$ is conveniently chosen to be approximately equal to F_s/k_BT. In practice, this is achieved by updating $h(s)$ until the input and the output values of these $h(s)$ are roughly the same. In this thesis, we show that though simulated tempering and multisequence

algorithms give similar results, the latter is computationally more efficient.

1.8 Coarse-grained molecular models

Ever since the advent of computers and subsequent increase in computing power, the development of computational models has made possible a new avenue of scientific enquiry, namely computational simulations of physical processes. Computational models have allowed the study of complex systems with high accuracy, thereby extending their applicability as well as increasing their predictive power. As explained by van Gunsteren *et al.* [38] "Any model involves a choice of the essential degrees of freedom, of the interactions governing the motion along these degrees of freedom, of a method to generate configuration of the degrees of freedom, and of the way in which the interaction with the outside world is represented . Models with a range of resolutions can be used to describe the same physical system, with each model's resolution providing the context to interpret its representation" [39]. The different levels of resolution and representation may include nucleons and electrons, nuclei and electrons, atoms, molecules and supra-molecules.

Coarse-grained models are a common type of model that enable the simulation of huge system sizes. They provide access to long time scales in simulations of biomolecular processes. The basic idea of coarse-grained models is to extend the properties of a system that can be studied by simplifying the physical system but retaining the essential physics. Such models are coarse-grained with respect to atomistic (or fine-grained) models. They use fewer particles than their corresponding atomistic models to represent the same physical system. A proper coarse-grained representation preserves the features that are necessary to describe the phenomena of interest while simultaneously eliminating atomic details that are considered unimportant [40].

Usually, coarse-grained models for biomolecular simulations are those in which the particles which constitute the degrees of freedom represent more than one non-hydrogen atom [41]. By focusing on essential features, while averaging over less important details, coarse-grained models provide significant computational advantage. That is, they provide greater efficiency than atomically detailed models [42], as a result of fewer degrees of freedom combined with larger time steps and faster conformational sampling due to smoother energy landscape. The main motivation for the use of coarse-grained models is the relatively fast sampling they provide [43]. This perhaps is the chief reason why in spite of the tremendous recent advances in atomically detailed models and computational resources [44], coarse-grained models continue to gain popularity.

Despite the fact that coarse-grained models provide computational efficiency, they introduce new challenges [40]. Notably, the most significant challenge is how to ensure the model reflects the correct underlying physical principles. A coarse-grained model should not only give the right results, but also provide them for the correct reasons [45, 46]. There are also limitations associated with most coarse-grained models [43]; they may be too biased and as a result are not transferable to different situations, only parameterized for specific class of molecules, and too coarse to capture certain properties.

Computational modeling of biological systems is particularly challenging because of the many spacial and temporal scales involved. Using coarse-grained models has aided large-scale biomolecular simulations on time scales that are otherwise inaccessible. Additionally, coarse-graining for biomolecules is a challenge because of their heterogeneity [41], i.e., the scale invariance observed in largely homogeneous polymers is absent in biomolecules, which consist of different complex structural units that interact in different ways. In particular, coarse-graining of proteins is specially

challenging because they also exhibit transitions into specific structures.

Many coarse-grained protein models have been developed since the pioneering models [21, 47, 48] for protein folding were introduced a few decades ago. It is not uncommon amongst these models that each amino acid is represented with one or a few sites which have properties that are completely specified by the amino acid type. Often times, a whole site is associated with the α-carbon of the amino acid because this helps with detailed representation of the protein backbone [49, 50]. Some of the models do not distinguish between all types of amino acids, but instead classify them based upon, for instance, hydrophobicity [51]. A typical example of this is the coarse-grained model [52] we employed in this thesis, which has only three amino-acid types; hydrophobic, non-hydrophobic (or polar) and a turn type which essentially represent glycine. Inspite of their simplicity, these models have helped in understanding some general principles that govern, e.g., protein folding and interactions [53, 54].

Bibliography

[1] Alexei Finkelstein and Oleg Ptitsyn. *Protein Physics*. Academic press, Elsevier science, USA, 2002.

[2] Alberts Bruce, Johnson Alexander, Lewis Julian, Raff Martin, Roberts Keith, and Walter Peter. *Molecular biology of the cell*. Garland science publishing, 5th edition, 2008.

[3] Philip N. Bryan and John Orban. Proteins that switch folds. *Current Opinion in Structural Biology*, 20(4):482–488, 2010.

[4] C Holzgräfe and S Wallin. Smooth functional transition along a mutational pathway with an abrupt protein fold switch. *Biophysical Journal*, 107(5):1217–

1225, 2014.

[5] Stephen C Harrison. Viral membrane fusion. *Nature Structural & Molecular Biology*, 15(7):690–698, 2008.

[6] 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(2):291–303, 2012.

[7] P N Bryan and J Orban. Implications of protein fold switching. 23(2):314–316, 2013.

[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. *Proceedings of the National Academy of Sciences*, 105(13):5057–5062, 2008.

[9] James Mottonen, Arne Strand, Jindrich Symersky, Robert M. Sweet, Dennis E. Danley, Kieran F. Geoghegan, Robert D. Gerard, and Elizabeth J. Goldsmith. Structural basis of latency in plasminogen activator inhibitor-1. *Nature*, 355(6357):270–273, 1992.

[10] Xuelian Luo, Zhanyun Tang, Guohong Xia, Katja Wassmann, Tomohiro Matsumoto, Josep Rizo, and Hongtao Yu. The Mad2 spindle checkpoint protein has two distinct natively folded states. *Nature Structural & Molecular Biology*, 11(4):338–345, 2004.

[11] Dene R. Littler, Stephen J. Harrop, W. Douglas Fairlie, Louise J. Brown, Greg J. Pankhurst, Susan Pankhurst, Matthew Z. DeMaere, Terence J. Campbell, Asne R. Bauskin, Raffaella Tonini, Michele Mazzanti, Samuel N. Breit, and

Paul M. G. Curmi. The Intracellular Chloride Ion Channel Protein CLIC1 Undergoes a Redox-controlled Structural Transition. *Journal of Biological Chemistry*, 279(10):9298–9305, 2004.

[12] Per A. Bullough, Frederick M. Hughson, John J. Skehel, and Don C. Wiley. Structure of influenza haemagglutinin at the pH of membrane fusion. *Nature*, 371(6492):37–43, 1994.

[13] A. Aina and S. Wallin. Multisequence algorithm for coarse-grained biomolecular simulations: Exploring the sequence-structure relationship of proteins. *The Journal of Chemical Physics*, 147(9):095102, 2017.

[14] Joji M Otaki, Shunsuke Ienaka, Tomonori Gotoh, and Haruhiko Yamamoto. Availability of short amino acid sequences in proteins. *Protein science : a publication of the Protein Society*, 14(3):617–25, 2005.

[15] Thomas Creighton. *Proteins: Structure and Molecular Properties*. Freeman, New York, 1993.

[16] L. Pauling, R. B. Corey, and H. R. Branson. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences*, 37(4):205–211, 1951.

[17] W Kauzmann. Some factors in the interpretation of protein denaturation. *Advances in Protein Chemistry*, 14:1–63, 1959.

[18] C B Anfinsen, E Haber, M Sela, and F H White. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America*, 47:1309–14, 1961.

[19] C. B. Anfinsen. Principles that Govern the Folding of Protein Chains. *Science*, 181(4096):223–230, 1973.

[20] Cyrus Levinthal. Are there pathways for protein folding? *Journal de Chimie Physique*, 65:44–45, 1968.

[21] Ken A. Dill. Theory for the folding and stability of globular proteins. *Biochemistry*, 24(6):1501–1509, 1985.

[22] Ken A. Dill and Hue Sun Chan. From Levinthal to pathways to funnels. *Nature Structural & Molecular Biology*, 4(1):10–19, 1997.

[23] Daniel V. Schroeder. *An introduction to thermal physics*. Addison Wesley, 2000.

[24] B. J. Alder and T. E. Wainwright. Phase Transition for a Hard Sphere System. *The Journal of Chemical Physics*, 27(5):1208–1209, 1957.

[25] B. J. Alder and T. E. Wainwright. Studies in Molecular Dynamics. I. General Method. *The Journal of Chemical Physics*, 31(2):459–466, 1959.

[26] B. J. Alder and T. E. Wainwright. Studies in Molecular Dynamics. II. Behavior of a Small Number of Elastic Spheres. *The Journal of Chemical Physics*, 33(5):1439–1451, 1960.

[27] Eric Paquet and Herna L. Viktor. Molecular Dynamics, Monte Carlo Simulations, and Langevin Dynamics: A Computational Review. *BioMed Research International*, 2015:1–18, 2015.

[28] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[29] Vasilios I. Manousiouthakis and Michael W. Deem. Strict detailed balance is unnecessary in Monte Carlo simulation. *The Journal of Chemical Physics*, 110:2753, 1999.

[30] Ulrich H. E. Hansmann and Yuko Okamoto. The generalized-ensenble approach for protein folding simulations. In *Annual Reviews of Computational Physics VI*, pages 129–157. World scientific, 1999.

[31] Bernd A. Berg and Thomas Neuhaus. Multicanonical algorithms for first order phase transitions. *Physics Letters B*, 267(2):249–253, 1991.

[32] Bernd A. Berg and Thomas Neuhaus. Multicanonical ensemble: A new approach to simulate first-order phase transitions. *Physical Review Letters*, 68(1):9–12, 1992.

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

[34] A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov, and P. N. Vorontsov. New approach to Monte Carlo calculation of the free energy: Method of expanded ensembles. *The Journal of Chemical Physics*, 96(3):1776–1783, 1992.

[35] Enzo Marinari and Giorgio Parisi. Simulated Tempering: A New Monte Carlo Scheme. *Europhysics Letters (EPL)*, 19(6):451–458, 1992.

[36] Koji Hukushima and Koji Nemoto. Exchange Monte Carlo Method and Application to Spin Glass Simulations. *Journal of the Physical Society of Japan*, 65(6):1604–1608, 1996.

[37] Bernd A. Berg. A brief history of the introduction of generalized ensembles to Markov chain Monte Carlo simulations. *The European Physical Journal Special Topics*, 226(4):551–565, 2017.

[38] Wilfred F. van Gunsteren, Dirk Bakowies, Riccardo Baron, Indira Chandrasekhar, Markus Christen, Xavier Daura, Peter Gee, Daan P. Geerke, Alice Glättli, Philippe H. Hünenberger, Mika A. Kastenholz, Chris Oostenbrink, Merijn Schenk, Daniel Trzesniak, Nico F. A. van der Vegt, and Haibo B. Yu. Biomolecular Modeling: Goals, Problems, Perspectives. *Angewandte Chemie International Edition*, 45(25):4064–4092, 2006.

[39] Jacob W. Wagner, James F. Dama, Aleksander E. P. Durumeric, and Gregory A. Voth. On the representability problem and the physical meaning of coarse-grained models. *The Journal of Chemical Physics*, 145(4):044108, 2016.

[40] W. G. Noid. Perspective: Coarse-grained models for biomolecular systems. *The Journal of Chemical Physics*, 139(9):090901, 2013.

[41] Sereina Riniker, Jane R. Allison, and Wilfred F. van Gunsteren. On developing coarse-grained models for biomolecular simulation: a review. *Physical Chemistry Chemical Physics*, 14(36):12423, 2012.

[42] Dominik Fritz, Claudia R. Herbers, Kurt Kremer, and Nico F. A. van der Vegt. Hierarchical modeling of polymer permeation. *Soft Matter*, 5(22):4556, 2009.

[43] Helgi I. Ingólfsson, Cesar A. Lopez, Jaakko J. Uusitalo, Djurre H. de Jong, Srinivasa M. Gopal, Xavier Periole, and Siewert J. Marrink. The power of coarse graining in biomolecular simulations. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 4(3):225–248, 2014.

[44] Tamar Schlick, Rosana Collepardo-Guevara, Leif Arthur Halvorsen, Segun Jung, and Xia Xiao. Biomolecular modeling and simulation: a field coming of age. *Quarterly Reviews of Biophysics*, 44(02):191–228, 2011.

[45] Steve O Nielsen, Carlos F Lopez, Goundla Srinivas, and Michael L Klein. Coarse grain models and the computer simulation of soft materials. *J. Phys.: Condens. Matter*, 16(04):481–512, 2004.

[46] Larry Scott, H. Modeling the lipid component of membranes. *Current Opinion in Structural Biology*, 12:495–502, 2002.

[47] Michael Levitt and Arieh Warshel. Computer simulation of protein folding. *Nature*, 253(5494):694–698, 1975.

[48] M Levitt. A simplified representation of protein conformations for rapid simulation of protein folding. *Journal of Molecular Biology*, 104(1):59–107, 1976.

[49] J. Maupetit, R. Gautier, and P. Tuffery. SABBAC: online Structural Alphabet-based protein BackBone reconstruction from Alpha-Carbon trace. *Nucleic Acids Research*, 34(Web Server):W147–W151, 2006.

[50] Piotr Rotkiewicz and Jeffrey Skolnick. Fast procedure for reconstruction of full-atom protein models from reduced representations. *Journal of Computational Chemistry*, 29(9):1460–1465, 2008.

[51] Scott Brown, Nicolas J Fawzi, and Teresa Head-Gordon. Coarse-grained sequences for protein folding and design. *Proceedings of the National Academy of Sciences of the United States of America*, 100(19):10712–7, 2003.

[52] Arnab Bhattacherjee and Stefan Wallin. Coupled folding-binding in a hydrophobic/polar protein model: impact of synergistic folding and disordered flanks. *Biophysical Journal*, 102(3):569–578, 2012.

[53] Jon M Sorenson and Teresa Head-Gordon. Toward minimalist models of larger proteins: a ubiquitin-like protein. *Proteins*, 46(4):368–79, 2002.

[54] Miriam Friedel, Daniel J. Sheeler, and Joan-Emma Shea. Effects of confinement and crowding on the thermodynamics and kinetics of folding of a minimalist β-barrel protein. *The Journal of Chemical Physics*, 118(17):8106–8113, 2003.

Chapter 2

Multisequence algorithm for coarse-grained biomolecular simulations[†]

Abstract

We consider a generalized-ensemble algorithm for coarse-grained simulations of biomolecules which allows the thermodynamic behavior of two or more sequences to be determined in a single multisequence run. By carrying out a random walk in sequence space, the method also enhances conformational sampling. Escape from local energy minima is accelerated by visiting sequences for which the minima are more shallow or absent. We test the method on an intermediate-resolution coarse-grained model for protein folding with 3 amino acid types and explore the potential for large-scale coverage of sequence space by applying the method to sets of more than 1,000 sequences. The resulting thermodynamic data is used to analyze the structures and stability properties of sequences covering the space between folds with different secondary structures.

[†]This chapter is a modified version of the publication; A. Aina and S. Wallin, "Multisequence algorithm for coarse-grained biomolecular simulations: exploring the sequence-structure relationship of proteins", *J. Chem. Phys.* **147 (9)**, 095102 (2017).

2.1 Introduction

Recent years have seen important advances in biomolecular simulation methods, including improvements to standard molecular dynamics force fields [1], the advent of several alternative atomistic simulation approaches [2–5], and new techniques for conformational sampling [6]. Together with the ever-increasing availability of computational resources, these advances have triggered a few major efforts [7–11] to characterize the dynamics of biomolecular systems of various sizes, e.g., a small native protein on the millisecond scale [10] and a comprehensive model cytoplasm on the nanosecond scale [11]. While encouraging and insightful, these large-scale simulations have also highlighted that severe tradeoffs in size and time scales will likely persist for the foreseeable future.

One way to expand the range of biomolecular simulations is to turn to coarse-grained (CG) models, where the basic aim is to simplify the description of physical interactions while retaining the essential physics of the system under study [12]. Ingolfsson *et al.* list four main factors that make CG models computationally fast: reduction in the number of degrees of freedom, faster simulation dynamics, emphasis on short-range interactions and the ability of using larger integration time steps [13]. To this list can be added that a CG representation of either the interaction potential or the molecular geometry often opens up for alternative sampling schemes beyond traditional molecular dynamics approaches, which can further speed up conformational sampling. Examples of such sampling schemes include activation-relaxation kinetics [14], discrete molecular dynamics [15] and various Monte Carlo (MC)-based techniques such as cluster moves [16].

The challenges of achieving representative conformational sampling of individual biomolecular systems notwithstanding, many biologically motivated problems naturally call for the investigation and comparison of molecular variants, e.g., deter-

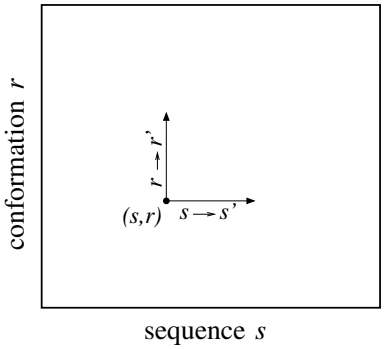

Figure 2.1: The two types of Monte Carlo updates in the multisequence Monte Carlo algorithm.

mining the molecular mechanisms of specificity in protein-protein [17, 18] or protein-nucleotide [19] interactions, or the role of mutations in molecular disease processes [20]. Another example is protein folding, where unique insight has been achieved by comparing sequences within and between protein families [21, 22]. In a situation with extremely rapid growth of sequence information [23], it is of interest to explore ways to efficiently sample multiple sequences in biomolecular simulations.

To this end, we consider in this work an MC-based algorithm that can calculate the thermodynamics of multiple sequences in a single run and apply it to a coarse-grained model for protein folding [24]. This multisequence (MS) method was originally developed in the context of homo- and heteropolymer simulations [25] and was later adapted for the characterization of peptide-protein binding specificity [26, 27]. To our knowledge, it has not been previously tested in realistic protein folding simulations. The MS algorithm carries out a simulation in a generalized ensemble that performs a random walk in sequence space. Hence, there are two main types of updates: conformational updates $r \to r'$ and sequence updates $s \to s'$. This strategy is straightforward when r and s are "perpendicular" coordinates, as illustrated in Fig.2.1, such that the potential energy of the model can be written in terms of two independent variables, $E(s, r)$.

As a test application of the MS algorithm, we selected the phenomenon of protein fold switching which recently was demonstrated in a handful of natural and engineered proteins. These special proteins exhibit a unique ability to reversibly switch between entirely different folds, with accompanying changes in secondary structure, hydrophobic core packing and overall shape [28]. The fold switching transitions found in natural proteins typically play a functional role. For example, rare transitions to an alternative fold in the protein KaiB provide a crucial time delay mechanism in the circadian clock cycle of cyanobacteria [29]. Fold switching can occur either spontaneously [30] or be triggered by various signals including changes to solution conditions [31], subdomain detachment [32], ligand binding [33] and point mutations [34]. Computational studies, using CG [35–38] or atomistic [39–41] models, have attempted to explain how proteins can exhibit multiple folding funnels and how they are altered in response to binding events or changes in sequence.

The discovery that proteins can be driven to switch folds through an accumulation of point mutations, in particular, holds implications for protein evolution as it suggests a simple mechanism of fold evolution [42]. Alexander *et al.* demonstrated that the similarly sized but structurally distinct A (3α) and B ($4\beta + \alpha$) domains of protein G could, after extensive mutations leaving their respective folds undisturbed, be triggered to switch folds by applying one additional mutation, Y45L, located at the edge of the hydrophobic core in the B domain [34]. This remarkable discovery suggests the possibility that the two domains might be evolutionary related despite a lack of detectable similarity in either sequence or structure in wild-type protein G [43], although this has yet to be proven. Moreover, it is unclear how common such fold-to-fold transitions are and how they might occur in evolutionary processes [44]. In previous work [35, 36] Holzgräfe and Wallin showed that mutational paths with abrupt fold switching exist between two other pairs of smaller protein folds within

the framework of our CG model [24].

In demonstrating mutation-induced fold switching in our model we characterized the folding of a set of 144 different model sequences with 16 amino acids. This set (denoted here $S16_{144}$) was constructed to sparsely span the sequence space between two ideally designed sequences, A1 and N1, folding into an α-helix and a β-hairpin, respectively. Here we use the set $S16_{144}$ to validate the MS method and compare its computational efficiency to a standard generalized-ensemble method [45, 46]. We thereafter greatly enlarge $S16_{144}$ to a set with 1,024 sequences as well as another set of the same size spanning two 35-amino acid sequences, A2 and TN, that fold into two-helical bundle and mixed α-β structures, respectively. Besides demonstrating that the MS method can be applied to large numbers of sequences, the results allow us to carry out a more systematic analysis of the biophysical properties of sequences along mutational pathways connecting these two pairs of basic folds than has been previously possible.

2.2 Theory

2.2.1 Generalized-ensemble algorithms and simulated tempering

Conventional Metropolis Monte Carlo simulations of the canonical distribution is problematic at low temperatures for many physical systems because simulations tend to become trapped in local energy minima and hamper representative sampling of configurational space. The basic idea of generalized-ensemble algorithms is to alleviate this trapping problem by sampling states using a non-Boltzmann weight factor and/or expand the state space with additional dynamical parameters [47] (for a recent his-

torical account see Ref. 48). Generalized-ensemble methods that have been frequently used for biomolecular simulations include simulated tempering (ST) [45, 46], replica exchange [49], or parallel tempering [50], and metadynamics [51].

ST is a direct extension to the Metropolis algorithm in which the temperature T becomes a dynamic parameter. In this way, frequent visits to high T allow simulations to readily escape from local energy minima. The algorithm thus simulates the joint probability distribution

$$P(m, r) = \frac{1}{\hat{Z}} e^{-\beta_m E(r) + g_m} , \tag{2.1}$$

where $\beta_m = 1/k_{\mathrm{B}} T_m$, $\{T_m\}_{m=1}^{\mathrm{M}}$ a set of temperatures and k_{B} is Boltzmann's constant. The normalization constant in equation 2.1 is

$$\hat{Z} = \sum_r \sum_{m=1}^{\mathrm{M}} e^{-\beta_m E(r) + g_m} , \tag{2.2}$$

where the first sum is over all conformations r. The simulation parameters g_m control the marginal probability distribution

$$P(m) = \frac{1}{\hat{Z}} \sum_r e^{-\beta_m E(r) + g_m} , \tag{2.3}$$

and must therefore be carefully chosen. A common and convenient choice is $g_m \approx \beta_m F_m$, where F_m is the free energy at temperature T_m. With this choice, $P(m)$ becomes approximately flat ensuring all temperatures are frequently visited.

2.2.2 Multisequence algorithm

The basic idea of the MS algorithm for biomolecular simulation is to let the sequence s become a dynamic parameter rather than the temperature as in ST. A dynamic s is technically feasible when the potential energy can be written as $E(s, r)$, where s

and r are independent variables. This is the case in our coarse-grained protein model which has only backbone degrees of freedom. It can also be achieved in some more detailed models [26, 27].

Similarly to ST, the MS algorithm simulates the joint probability distribution

$$P(s,r) = \frac{1}{Z} e^{-\beta E(s,r)+h(s)}\,, \tag{2.4}$$

where

$$Z = \sum_{s\in \mathrm{S}} \sum_{r} e^{-\beta E(s,r)+h(s)} \tag{2.5}$$

and S is a set of pre-selected sequences, i.e., the sequences to which visits are allowed during a simulation. The simulation parameters $h(s)$, similar to the parameters g_m in ST, control the marginal distribution $P(s) = Z^{-1}\sum_r e^{-\beta E(s,r)+h(s)} = Z^{-1}e^{-\beta F(s)+h(s)}$ and a roughly flat $P(s)$ can be achieved by choosing $h(s) \approx \beta F(s)$, where $F(s)$ is the free energy of sequence s at temperature T.

Two types of MC updates are required to sample from the distrubution in equation 2.4, ordinary conformational updates $r \to r'$ and sequence updates $s \to s'$. The acceptance probability for the latter becomes

$$P_{\mathrm{acc}}(s \to s') = \min[1, \exp\{-\beta\Delta E + \Delta h\}]\,, \tag{2.6}$$

where $\Delta E = E(s',r) - E(s,r)$ and $\Delta h = h(s') - h(s)$.

Picking a new sequence s' in a sequence update $s \to s'$ can be done in several ways. One possibility is to draw s' randomly from the set S, such than $s' \neq s$. Alternatively, a type of "mutational" move can be used where an amino acid position is first picked and then assigned a new amino acid type. The selection of position and type would have to be chosen such that s' does not end up outside S. In this work, we use the

former update which is general and guarantees that ergodicity is fulfilled for any S. Importantly, both updates fulfill detailed balance and therefore lead to the same estimates of equilibrium quantities, such as native state stabilities, for the different sequences in S.

2.3 Model and Methods

2.3.1 Coarse-grained 3-letter model for protein folding

All calculations were carried out using the coarse-grained model for protein folding developed in Ref. 24. In this model, there are 3 different amino acid types: hydrophobic (h), polar (p) and turn-type (t). The backbone chain is represented atomistically by the N, H, C_α, $H_{\alpha 1}$, C' and O atoms. By contrast, the sidechain represention is simplified to a single enlarged C_β atom, which is geometrically identical for h and p types. The sidechain is absent for the t type which instead has an $H_{\alpha 2}$ atom. The t type is therefore closely related to glycine. All bond lengths, bond angles, and peptide plane angles (180°) are held fixed. Hence, an N-amino acid chain conformation r can, for any sequence s, therefore be described by the set of $2N$ backbone torsional angles $\{\phi_i, \psi_i\}_{i=1}^{N}$.

This geometrical description is paired with a simplified but finely tuned energy function with 4 terms: $E = E_{\rm ev} + E_{\rm loc} + E_{\rm hb} + E_{\rm hp}$. The first two, $E_{\rm ev}$ and $E_{\rm loc}$, represent excluded-volume effects and local electrostatic effects, respectively. The hydrogen-bond energy, $E_{\rm hb}$, represents directionally dependent interactions between NH and CO groups and is necessary for secondary structure formation. Finally, the "hydrophobicity" term, $E_{\rm hp}$, implements pairwise Lennard-Jones-like interactions between the C_β atoms of h amino acids which are necessary for driving chain collapse during folding. Various model parameters, e.g., the strengths of hydrophobic attrac-

tions and hydrogen bonding, were determined based on the ability of the model to spontaneously fold a set of model sequences with 18-54 amino acids into structurally diverse and thermodynamically stable native states with both β and α-structure. As it turned out, this strategy made the model robust enough to fold sequences designed to have mixed α and β structures.

2.3.2 Model sequences

Six of the model sequences studied in this work, A1, N1, R1, R2, A2, and TN, are given in Table 2.1. In addition, we study two sequence sets $S16_{1024}$ and $S35_{1024}$ with 1,024 sequences each derived from the A1-N1 and A2-TN pairs, respectively, through mutational combinations, as well as the set $S16_{144}$ taken from Ref. 35.

Table 2.1: List of 6 model sequences of different lengths N studied in this work.

Name	N	Sequence
A1	16	pphpphhpphpphhpp
N1	16	phphphpttphphphp
R1	16	pphhphptthpphhpp
R2	16	ppphphhtthhphppp
A2	35	(A1)ttt(A1)
TN	35	(A1)ttt(N1)

2.3.3 Monte Carlo simulation parameters and updates

Both ST and MS simulations are carried out with two types of conformational updates $r \to r'$: (1) a global pivot move (20%) which randomly picks a ϕ_i angle or ψ_i angle and assign a new value between $-\pi$ and π; and (2) a semi-local move (80%) which turns the ϕ_i and ψ_i-angles of 4 consecutive amino acids in a coordinated manner [52].

In our MS simulations, sequence updates $s \to s'$ are carried out in the following way. First, a new sequence s' is picked randomly from the set of pre-selected (allowed) sequences S, such that $s' \neq s$. This new sequence s' therefore differs from s in one or more amino acid positions. Thereafter, the sidechains of the protein, which remains in an unchanged (backbone) conformation r, is re-built according to the new sequence s'. Practically this means that, at the position(s) where the amino acid type has changed, the sidechain is altered according to the type change. For example, if p $\to$ t, the C_β atom is removed and replaced with an $H_{\alpha 2}$ atom or, if p $\to$ h, the C_β remains in place but its character is changed to hydrophobic. Finally, the change in total energy ΔE is calculated and the accept-reject criterion in equation 2.6 is applied. If rejected, the old state (s, r) is restored.

Table 2.2: List of simulations carried out in this work.

Runs	**Algorithm**	$\mathbf{k_B T}$	**MC steps/run**[a]	**Sequences**
32	ST	0.43–0.65	1×10^7	A1
32	ST	0.43–0.65	1×10^7	N1
32	ST	0.43–0.65	1×10^7	R1
32	ST	0.43–0.65	1×10^7	R2
32×8[b]	MS	0.43–0.65	18×10^7	$S16_{144}$
16	MS	0.43	5×10^9	$S16_{1024}$
16	MS	0.46	4×10^9	$S35_{1024}$

[a] Excludes a thermalization step with 10^7 MC steps/run.
[b] 32 runs per temperature at 8 different temperatures.

A sequence update is attempted every $1,000$ MC steps while temperature updates $m \to m'$ are attempted every 100 steps. The computational cost for sequence updates is somewhat higher than for temperature updates. The latter update does not require any energy calculation and is thus extremely rapid. For the purpose of comparing computational efficiencies of ST and MS, we therefore chose sequence updates to be

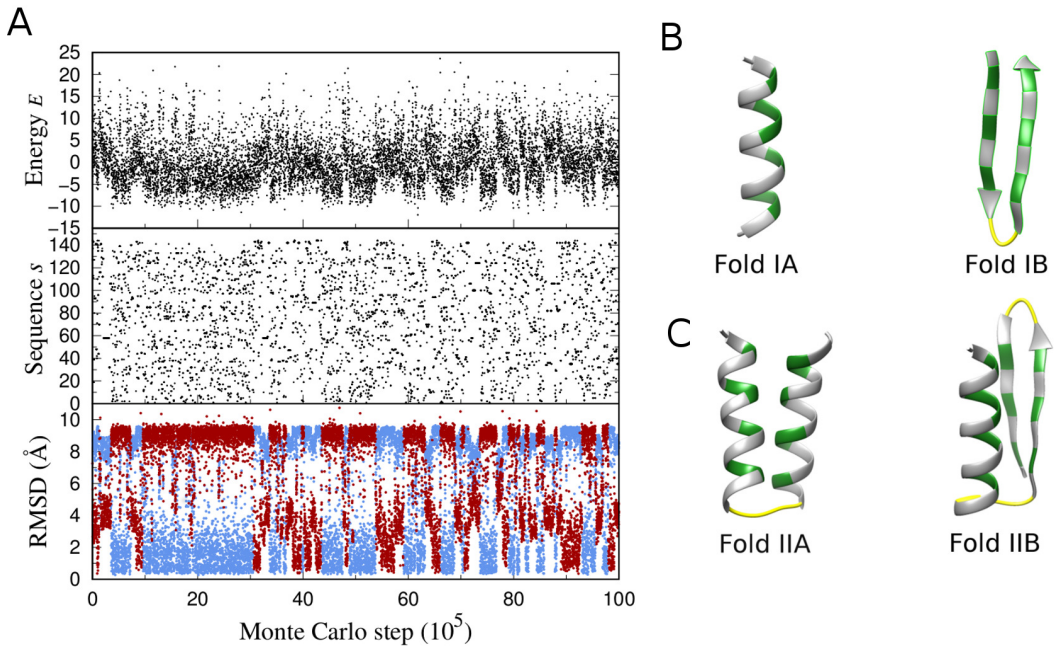

Figure 2.2: (A) Example of an MS simulation of the sequence set $S16_{144}$ carried out at $k_BT = 0.43$. The plot shows the MC evolution of the total potential energy E, the sequence s (numbered 1–144), and the root-mean-square deviation (RMSD) calculated against the representative fold IA (light blue) and fold IB (dark red) structures in (B). Representative structures of folds (B) IA, IB, (C) IIA and IIB, chosen to be the minimum-energy conformations found for the sequences A1, N1, A2 and TN, respectively.

slightly less frequent than temperature updates while both are fairly frequent. All simulations carried out in this work are summarized in Table 2.2.

2.3.4 Observables

Fold stabilities are calculated as in Ref. 36 and described briefly below. First we define two structural similarity measures Q_{IA} and Q_{IB} for folds IA and IB (Fig.2.2B), respectively, indicating the fraction of the fold-specific backbone-backbone hydrogen bonds that have been formed. The fold IA-hydrogen bonds are (2,6), (3,7), (4,8), (5,9), (6,10), (7,11), (8,12), (9,13), (10,14), (11,15) and the fold IB-bonds are (3,14), (5,12), (7,10), (10,7), (12,5), (14,3), where (i,j) indicates a hydrogen bond between

the CO group of amino acid i and the NH group of amino acid j. The stabilities of folds IA and IB are defined as the probabilities $P_{\rm IA} = P(Q_{\rm IA} \geq 0.8)$ and $P_{\rm IB} = P(Q_{\rm IB} \geq 0.8)$, respectively, i.e., the probability that at least 80% of the fold's hydrogen bonds are formed. $P_{\rm IA}$ and $P_{\rm IB}$ thus depend on both sequence s and temperature T. For example, $P_{\rm IA} = 0.875 \pm 0.003$ for A1 and $P_{\rm IB} = 0.785 \pm 0.008$ for N1 at $k_{\rm B}T = 0.43$. Structural similarity measures for 35-amino acid folds IIA and IIB are defined as $Q_{\rm IIA} = (Q_{\rm IA}^{1-16} + Q_{\rm IA}^{20-35} + Q_{\rm tert})/3$ and $Q_{\rm IIB} = (Q_{\rm IA}^{1-16} + Q_{\rm IB}^{20-35} + Q_{\rm tert})/3$, respectively, where superscripts on $Q_{\rm IA}$ and $Q_{\rm IB}$ indicate over which amino acid positions those measures are applied to within the longer 35 amino acid sequences and $Q_{\rm tert}$ is a measure that counts the number of C_β-C_β contacts between the two secondary structure elements of these folds. In analogy with $P_{\rm IA}$ and $P_{\rm IB}$, we define the stabilities of folds IIA and IIB as $P_{\rm IIA} = P(Q_{\rm IIA} \geq 0.8)$ and $P_{\rm IIB} = P(Q_{\rm IIB} \geq 0.8)$, respectively. The root-mean-square-deviation, RMSD, is calculated over all C_α atoms.

2.4 Results

2.4.1 Computational efficiency

We start by applying the MS algorithm to the set $S16_{144}$ across a range of temperatures T (Table 2.2). Two of the sequences in $S16_{144}$ are A1 and N1 (Table 2.1) which fold into stable α-helix and β-hairpin structures, respectively, as shown in Fig.2.2B. Because A1 and N1 differ at 10 positions, 10 consecutive point mutations can transform A1 into N1, and vice versa. The binary sequence space between A1 and N1 in which any combination of these mutations have been carried out, therefore contains $2^{10} = 1,024$ sequences. The 144 sequences in $S16_{144}$ were selected from this binary space with the constraints that the total number of hydrophobic amino acids are not too high and that they are not too unevenly distributed along the sequence.

Figure 2.2 illustrates a typical MS simulation trajectory carried out at the lowest studied temperature which is below the folding temperature of both A1 and N1 [35,36]. From the MC evolution of the total energy E, sequence index s, and RMSD values from the representative structures in Fig.2.2B, it is evident that visits to various sequences drive transitions into a range of structural states. In particular, there are frequent visits to α-helix and β-hairpin structures and transitions between them are accompanied by a shift in which sequences are preferably visited. For example, visits to high s-indices, including N1 with index 144, tend to coincide with formation of β-hairpin structures as required to generate correct equilibrium conformational ensembles.

One might have suspected that the MS algorithm would be hampered by poor acceptance rates for sequence updates. However, this is not the case in our model. We carry out updates $s \to s'$ by picking a new random sequence $s' \neq s$ from the set of allowed sequences. The (average) acceptance rate depends on both T and the step in sequence space Δh, i.e., the number of amino acid positions changed, as shown in Fig.2.3. At the lowest T and highest Δh, acceptance rates are only around 0.1-0.2. However, for most other T and Δh the overall acceptance rates are substantially higher and often above the oft-quoted rule-of-thumb value 0.25 [53] (Fig.2.3B). An increased acceptance rate can easily be achieved by restricting proposed updates such that $\Delta h \leq \Delta h_{\max}$, where $\Delta h_{\max}$ is a maximum step size, which might be necessary for longer chains. For example, $\Delta h_{\max} = 1$ would be equivalent to applying only a "mutational" update, i.e., picking a random (allowed) position and changing the amino acid type at that position.

We now compare the results from our MS calculations with simulated tempering (ST) simulations carried out on 4 of the 144 sequences, namely A1 and N1 and two random sequences, R1 and R2, chosen at distances $h = 4$ and $h = 6$ from A1,

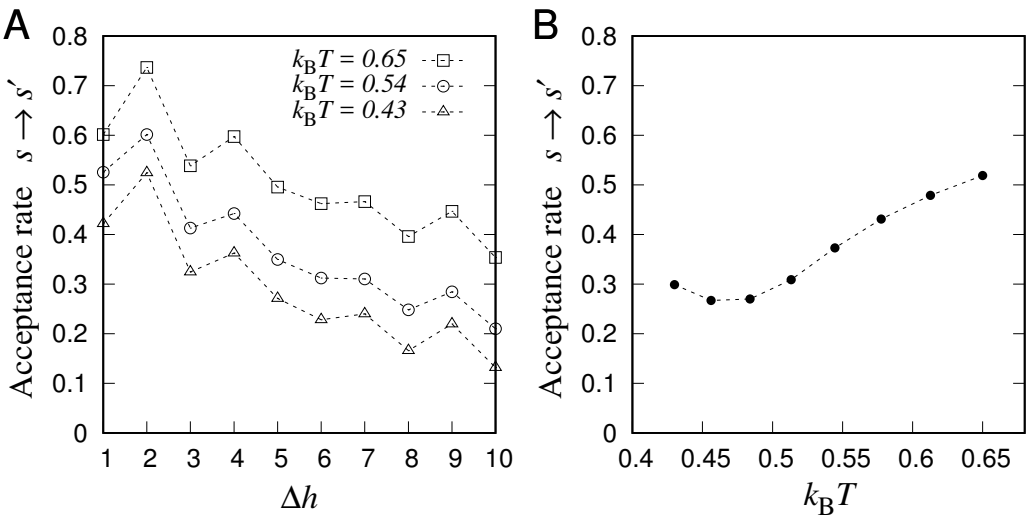

Figure 2.3: Acceptance rates for $s \rightarrow s'$ updates in MS simulations of the $S16_{144}$ sequence set as a function of (A) the number of changed amino acid positions Δh and (B) temperature T. Acceptance rates for 3 different T's are shown in (A).

respectively (Table 2.1). While ST provides the thermodynamics of a given sequence across a range of T in a single run, an MS simulation provides the thermodynamics of all 144 sequences at one T. We adjust the simulation lengths for ST and MS runs such that roughly the same number of sampled conformations are obtained for each s and T combination, thus ensuring that similar computational resources are used for the two algorithms (Table 2.2). We first validate the MS algorithm by comparing the average total energy, $\langle E \rangle$, calculated for these 4 sequences with the two different methods (Fig.2.4). The two sets of results are entirely consistent showing that, for a given s and T, the MS and ST algorithms indeed sample the same (Boltzmann) distribution.

As a way to assess conformational sampling efficiency, we compare in Fig.2.5 the statistical error, $\sigma_{\langle E \rangle}$, of the average energy $\langle E \rangle$ for the 4 sequences obtained using ST and MS, respectively. Because approximately the same number of sampled conformations were obtained for each combination of s and T, we compare the statistical errors directly. At the highest studied T, which is well above the folding temperature of both

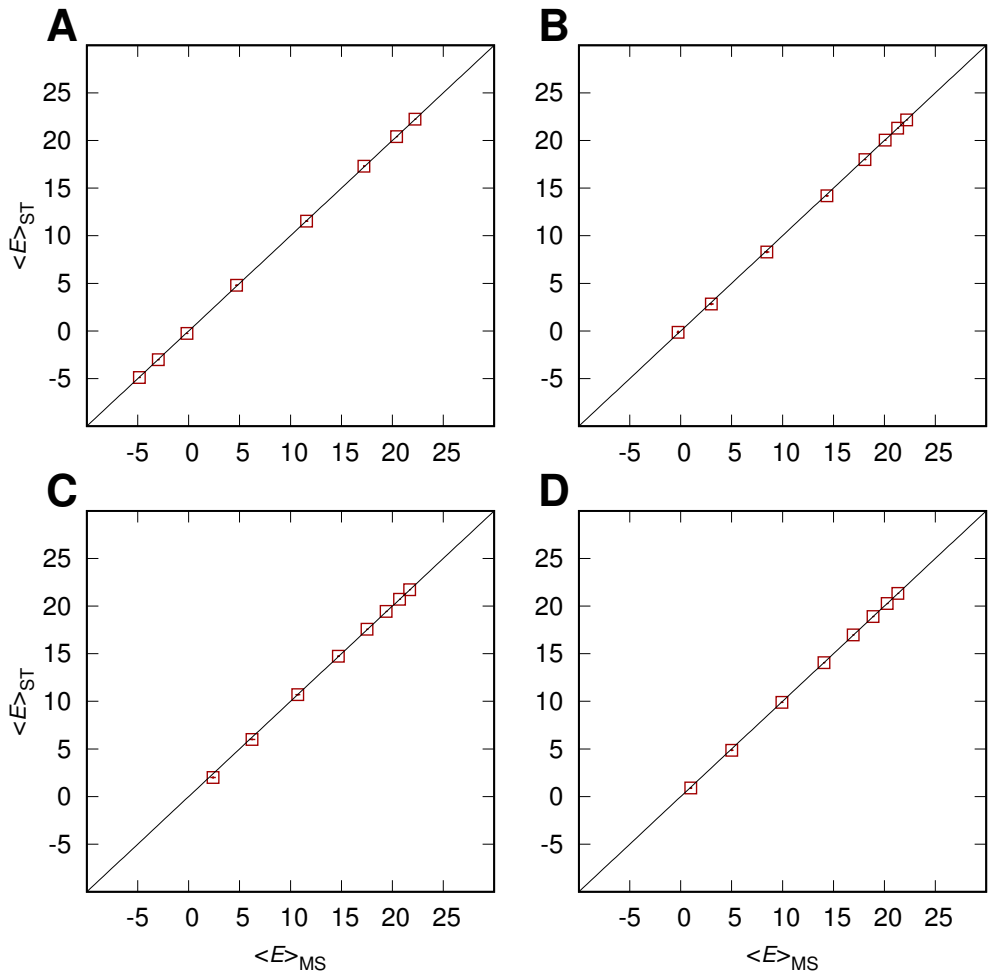

Figure 2.4: Comparison of the average energy, $\langle E \rangle$, as calculated at 8 different temperatures by the ST and MS algorithms for the 4 model sequences (A) A1, (B) N1, (C) R1, and (D) R2. Statistical 1σ errors, estimated from 32 independent runs for each method (Table 2.2), are shown but are smaller than the plot symbols for all points.

A1 and N1, the two algorithms give almost identical statistical errors. This can be understood by noting that at high-T the free-energy landscape is relatively smooth and conformational space requires little difficulty to sample. The benefit of adding a dynamic parameter, whether s or T, is apparently minimal under these conditions. However, at lower T, the $\sigma_{\langle E \rangle}$ values from MS is often smaller than those from ST and

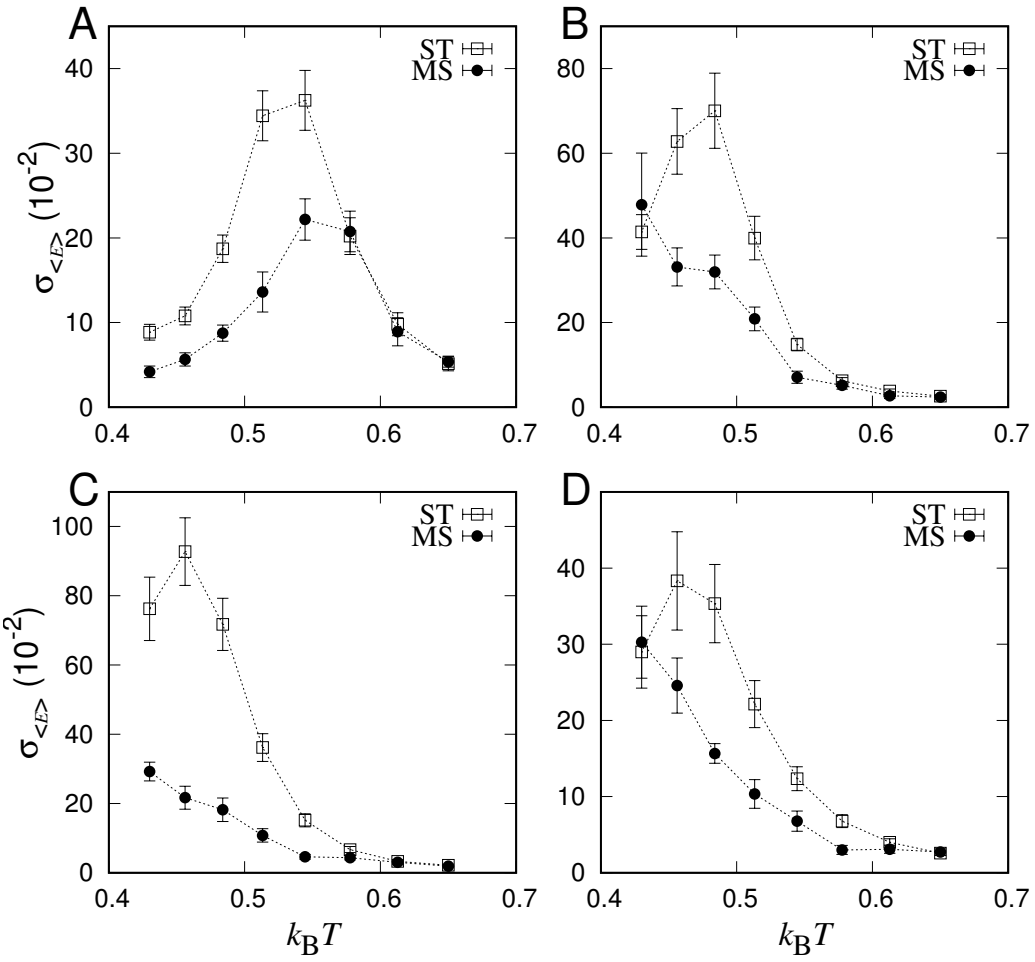

Figure 2.5: Comparing sampling efficiencies of the MS and ST algorithms. Statistical errors $\sigma_{\langle E \rangle}$ of the average total energy $\langle E \rangle$ obtained for the sequences (A) A1, (B) N1, (C) R1 and (D) R2 (Table 2.1) at different temperatures T. Simulation lengths in the two methods are adjusted such that the number of conformations sampled per sequence and temperature is roughly the same.

never significantly higher. For example, at the lowest T, the precision in the estimate of $\langle E \rangle$ is roughly twice as high in MS than ST for A1 and R1, and roughly the same for N1 and R2.

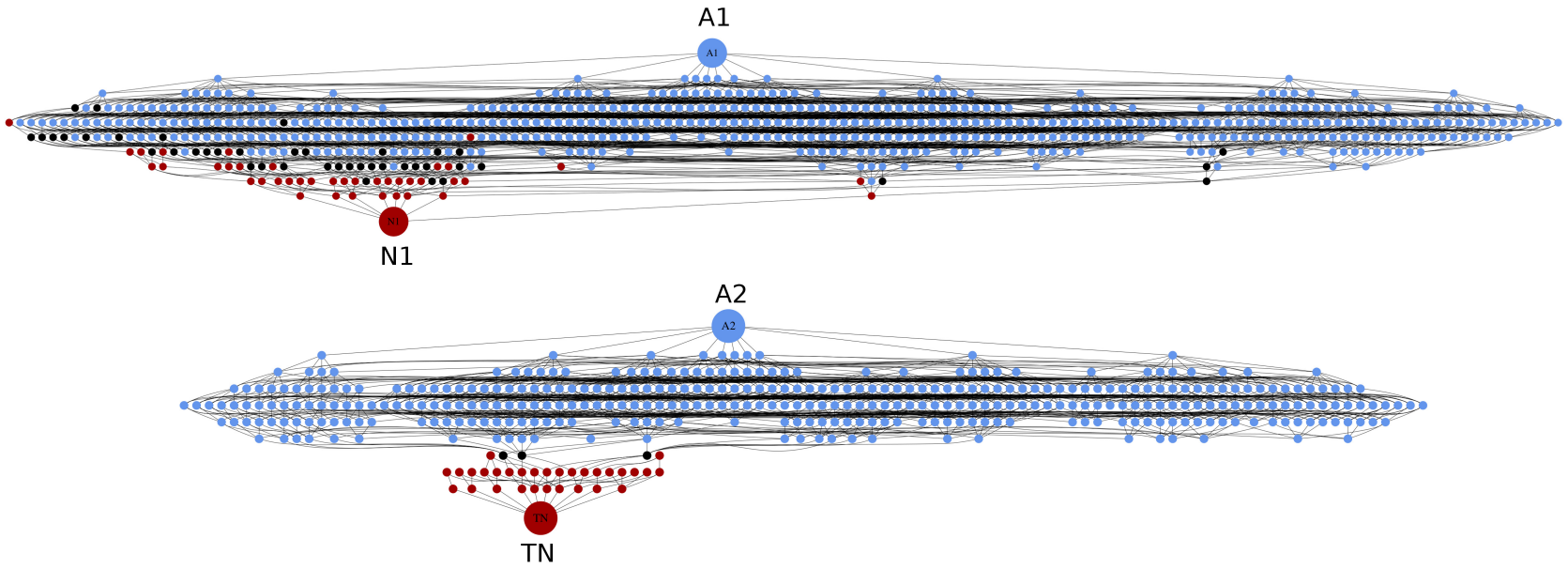

Figure 2.6: Networks of sequences connecting folds IA and IB (top) and folds IIA and IIB (bottom). Each node represents a stable sequence ($P_{\text{tot}} \geq P_{\text{cut}}$ where $P_{\text{cut}} = 0.50$) that folds into either IA or IIA (light blue), IB or IIB (dark red), or is classified as bistable ($B > 0.5$, black). A line between two nodes indicates that the sequences differ at only one position. Graph created using the tool Graphviz [54] obtained from www.graphviz.org.

2.4.2 Exploring sequence space: IA/IB and IIA/IIB fold connectivities

We now turn to the full binary sequence sets $S16_{1024}$ and $S35_{1024}$ with 1,024 sequences each. By applying the MS method to these two sets (Table 2.2), we determine the low-T thermodynamic behavior of each included sequence. In particular, we calculate the stabilities of folds IA and IB, $P_{\rm IA}$ and $P_{\rm IB}$, for all sequences in $S16_{1024}$ and the stabilities of folds IIA and IIB, $P_{\rm IIA}$ and $P_{\rm IIB}$, for all sequences in $S35_{1024}$ (see Methods). The relative statistical errors on these quantities vary but are only a few percent at the most, despite the large number of sequences included.

Having calculated these fold stabilities, we are in a position to determine if there are pathways in sequence space that lead to abrupt IA-IB or IIA-IIB fold changes, i.e., paths that do not pass through any unstable intermediate sequence. To this end, we construct graphs in which each stable sequence is represented by a node and any two nodes are connected if their sequences differ at a single amino acid position. To determine if a sequence is stable we use the criterion $P_{\rm tot} > P_{\rm cut}$, where $P_{\rm tot} = P_{\rm IA} + P_{\rm IB}$ and $P_{\rm IIA} + P_{\rm IIB}$ for the IA-IB and IIA-IIB fold pairs, respectively; $P_{\rm tot}$ thus indicates the total stability of a sequence across the two competing folds. The precise network depends, of course, on the cut-off value $P_{\rm cut}$ and a higher $P_{\rm cut}$ generally means a selection of more stable pathways.

Fig.2.6 illustrates the networks obtained with $P_{\rm cut} = 0.50$ showing that both the IA-IB and IIA-IIB fold pairs are connected in sequence space at this stability threshold. A precise analysis shows that there are 516,972 viable IA-IB paths and 57,912 viable IIA-IIB paths. These paths represent 14.2 % and 1.6 % of all possible paths, respectively, because there are a total of $10! = 3,628,800$ possible paths between start and end points in both cases. Hence, folds IA and IB are rather highly connected in

our model for $P_{\text{cut}} = 0.50$. For $P_{\text{cut}} = 0.60$, the numbers are 104,640 paths (2.9%) for IA-IB and 22,512 (0.6%) paths for IIA-IIB. We find that there are no possible IA-to-IB or IIA-to-IIB paths when $P_{\text{cut}} \geq 0.74$ and ≥ 0.66, respectively.

2.4.3 Biophysical properties of fold-to-fold mutational pathways

An apparently general characteristic of designed and natural proteins that exhibit mutation-induced fold switching is a reduced stability near the switch point [34, 37, 38, 40, 43]. Our model proteins exhibit a similar trend. Fig.2.7A and B show the average total stability P_{tot} for sequences found at different Hamming distances h from the starting point. Intermediate sequences are less stable than sequences at distances $h = 0$ (i.e. A1 or A2) and $h = 10$ (i.e. N1 or TN), although there are large variations between sequences as indicated by the upper and lower bounds. There is nonetheless a clear statistical trend that sequences become gradually less stable as successive mutations are applied to any of the 4 start and end points until a minimum is reached.

However, the smooth stability trends in Fig.2.7A and B belie the real character of the individual mutational pathways which tend to exhibit an abrupt switch between the two folds. To see this and to further examine the character of the fold transitions in our model, we make a distinction between two types of stable sequences: those that fold into a single unique fold, thus behaving as classical proteins, and those that display substantial stabilities of both folds. Such "bistable" sequences are interesting from a biophysical perspective in that they are able to fold into two alternative folds. Indeed, bistable sequences have been proposed to play a role in the evolution of new protein folds [55]. We consider a sequence to be bistable if $B > 0.5$, where B is a

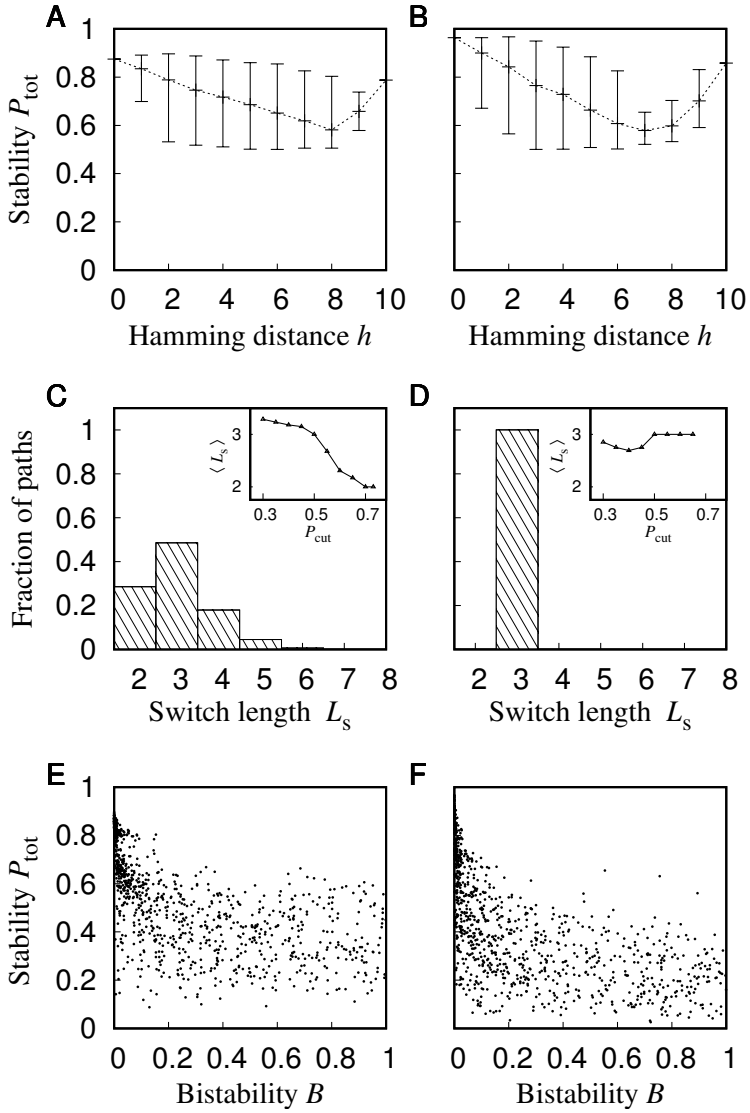

Figure 2.7: Stability properties of mutational pathways. The total stability $P_{\rm tot}$ as a function of the distance h from A1 averaged over all (A) IA-IB and (B) IIA-IIB mutational paths obtained with $P_{\rm cut} = 0.50$. Error bars indicate maximum and minimum $P_{\rm tot}$ values. The distribution of switch lengths $L_{\rm s}$ for the (C) IA-IB and (D) IIA-IIB mutational paths ($P_{\rm cut} = 0.50$). C and D insets: Average switch length $\langle L_{\rm s} \rangle$ across all paths as a function of $P_{\rm cut}$. Scatter plots of $P_{\rm tot}$ versus bistability B for all sequences in (E) $S16_{1024}$ and (F) $S35_{1024}$, where $B = 1 - \Delta P/P_{\rm tot}$ and $\Delta P = |P_{\rm IA} - P_{\rm IB}|$ or $|P_{\rm IIA} - P_{\rm IIB}|$.

bistability measure (Fig.2.7 legend). In principle, a fold transition can then occur directly between two classical proteins with unique native folds, or it can proceed via

one or more intermediate bistable sequences which populate both folds. We define the switch length of a mutational pathway $L_s = 2 + n_B$, where n_B is the number of bistable sequences in between the two classical sequences that define the switch point. Hence, a path with $L_s = 2$ accomplishes a fold switch in a single mutational step without going through a bistable point. From the distributions of L_s in Fig.2.7C and D, taken over all pathways with $P_{cut} = 0.50$, it can be seen that fold switching along individual pathways are typically completed in only 1-2 mutations and a single step is often sufficient to switch between the IA and IB folds. Hence, fold switching is typically abrupt and, for $P_{cut} = 0.5$, it is fairly common that viable pathways pass through one or two bistable sequences.

Interestingly, switching between folds IA and IB through one or more bistable sequences become less and less frequent as selections for more stable pathways are made. This can be seen from the decrease in $\langle L_s \rangle$ as a function of P_{cut} (Fig.2.7C(inset)). For $P_{cut} \geq 0.70$, there is no longer any remaining path between the α-helix and β-hairpin that passes through a bistable sequence because $\langle L_s \rangle = 2$. An underlying reason for the occurrence of sharper fold switches for more stable mutational pathways is apparent from a comparison between P_{tot} and B across all sequences in $S16_{1024}$. As shown in Fig.2.7E, sequences with the highest P_{tot} tend to exhibit very little bistability. Hence, highly stable paths are therefore forced to go through abrupt switch points where they transition directly between folds in a single step. The situation for the IIA-IIB fold pair is more complicated. We find that, just as for $S16_{1024}$, sequences in $S35_{1024}$ follow the trend that the highest P_{tot} values occur for only classical, low-B proteins. One might therefore expect that selection of more stable IIA-IIB paths would decrease $\langle L_s \rangle$, however, this is not the case as such abrupt switch points between the IIA and IIB are not available for $P_{cut} \geq 0.50$ (cf. Fig.2.6 bottom). As a result, bistable sequences do play a crucial role in bridging the IIA and IIB folds, although passing

though these sequences lead to additional reduction in stability at the switch point.

2.5 Discussion

We have evaluated a biomolecular simulation algorithm that works by making the biological sequence a dynamic parameter. As a test, we applied it on a CG model for protein folding. The results indicate that there are two main benefits of this approach. Firstly, it provides a convenient way to sample the canonical distributions of large numbers of sequences in a single run and, secondly, it enhances the sampling of conformational space meaning it can be applied directly to low temperatures. The conformational sampling efficiency can be assessed from the comparison with ST. Although there is no single "fair" way to compare the two methods, we chose as a measure of efficiency the statistical error of the total energy, $\sigma_{\langle E\rangle}$, obtained with roughly the same computational cost per temperature and sequence. At the highest studied temperatures, we find that the statistical errors $\sigma_{\langle E\rangle}$ are basically the same. This finding is not unexpected because conformational sampling of short polymers at high T does not involve crossing any major energy barriers. As a result, successively sampled conformations for a given combination of sequence s and temperature T, are likely uncorrelated in both methods which leads to similar $\sigma_{\langle E\rangle}$ values.

At lower temperatures, we find that the MS simulations often yields significantly smaller $\sigma_{\langle E\rangle}$ than ST simulations. It is notable that this acceleration in conformational sampling vis-á-vis ST is achieved despite the simulations being carried out at constant T. Therefore, rather than promoting escape from local minima by visits to higher T, as in ST [45, 46] or temperature replica-exchange [49], MS simulations must escape the minima occurring for one sequence through visits to other sequences. How this process works can be envisioned by considering an MS simulation that is visiting a

sequence s and is trapped in a local minimum, requiring that a high free-energy barrier is overcome for escape. The trapped state could be, e.g., a compact β-sheet rich state with a particular register. Eventually, sequence updates will carry the simulation to other sequences $\hat{s}$ while it conformationally still remains in the trapped state. However, the free energy barrier of escape might be substantially lower for $\hat{s}$ than for s, or the barrier may even be altogether absent if the trapped state is unstable for $\hat{s}$, leading to rapid escape from the minimum. The above reasoning also implies that the performance of the MS algorithm likely depends on the size of the sequence set S as well as the conformational properties of the sequences. Specifically, the performance of MS simulations of proteins at low T may benefit from the inclusion of at least a few sequences with poor stability properties, such that partial unfolding of the chain is regularly triggered and thus promoting transitions to new conformational states.

Individual protein sequences that exhibit spontaneous transitions between widely different competing conformational states, such as fold switching proteins, are especially challenging to molecular simulation methods. Representative sampling in such cases requires multiple transitions between highly different states, which can be a slow process. This problem has been addressed by using Hamiltonian replica-exchange techniques to couple Gō-like terms, i.e., energy terms with an artificial bias towards a given structure, with a physical force field [56–58]. This way exchange moves "feed" diverse conformations into the replica corresponding to the physical force field and enhance sampling [58]. We did not specifically study sampling efficiency for sequences exhibiting competing states, such as bistable sequences. It appears likely, however, that a very similar type of "feeding" of conformations would take place in MS simulations although the coupling occurs instead with other sequences rather than with Gō-type terms.

We emphasize that the MS method should not be seen as a general technique

to speed up conformational sampling. However, our results indicate that for CG models that permit sequence updates to be carried out as a simple Metropolis step and when visits to higher T is unwanted (or unnecessary), our method can be a highly efficient way to sample the equilibrium behavior of many sequences. This opens up the posibility of using MS for various applications, such as exploring sequence effects on the conformational properties of disordered [59] or denatured [60] proteins, or as a tool in efforts to combine population genetics and molecular simulations [61]. While applied to proteins in this work, we note that the theoretical framework of the MS algorithm is equally valid for other bio-macromolecules, including DNA and RNA. The ability to promote conformational sampling without resorting to an increase in T may make the method useful in simulations of biomolecules in ordered phases, such as lipid bilayers or double-stranded DNA, where escape from local minima can be especially challenging [62, 63] and elevated T is typically avoided in simulations because it may lead to unwanted perturbations or unraveling of the basic underlying structure.

2.6 Conclusion

We have evaluated an algorithm for biomolecular simulations that allows the thermodynamics of multiple sequences to be calculated in a single run. We applied the algorithm to protein folding and showed that the thermodynamic behavior of more than 1,000 amino acid sequences with up to 35 amino acids could be determined in an intermediate-resolution CG model. The method performs a random walk in sequence space which is especially useful at low temperature as it promotes escape from local minima present in the free energy landscapes of individual sequences. The method might be suitable for CG simulations of various other biomolecular systems, such as

peptides in phospholipid bilayers, where sampling at elevated temperatures is not desirable.

Bibliography

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

[2] F Ding, D Tsao, H Nie, and N V Dokholyan. Ab initio folding of proteins with all-atom discrete molecular dynamics. *Structure*, 16(7):1010–1018, 2008.

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

[4] A Verma and W Wenzel. A free-energy approach for all-atom protein simulation. *Biophysical Journal*, 96(9):3483–3494, 2009.

[5] J S Yang, W W Chen, J Skolnick, and E I Shakhnovich. All-atom ab initio folding of a diverse set of proteins. *Structure*, 15(1):53–63, 2007.

[6] R C Bernardi, M C Melo, and K Schulten. {E}nhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochimica et Biophysica Acta*, 1850(5):872–877, 2015.

[7] S R McGuffee and A H Elcock. Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm. *PLoS Comput Biol*, 6(3):e1000694, 2010.

[8] Y Miao, J E Johnson, and P J Ortoleva. All-atom multiscale simulation of cowpea chlorotic mottle virus capsid swelling. *J. Phys. Chem, B.*, 114(34):11181–11195, 2010.

[9] J R Perilla, J A Hadden, B C Goh, C G Mayne, and K Schulten. All-atom molecular dynamics of virus capsids as drug targets. *Journal of Physical Chemistry Letters*, 7(10):1836–1844, 2016.

[10] K Lindorff-Larsen, S Piana, R O Dror, and D E Shaw. How fast-folding proteins fold. *Science (New York, N.Y.)*, 334(6055):517–520, 2011.

[11] I Yu, T Mori, T Ando, R Harada, J Jung, Y Sugita, and M Feig. Biomolecular interactions modulate macromolecular structure and dynamics in atomistic model of a bacterial cytoplasm. *eLife*, 5, 2016.

[12] Sereina Riniker, Jane R. Allison, and Wilfred F. van Gunsteren. On developing coarse-grained models for biomolecular simulation: a review. *Physical Chemistry Chemical Physics*, 14(36):12423, 2012.

[13] Helgi I. Ingólfsson, Cesar A. Lopez, Jaakko J. Uusitalo, Djurre H. de Jong, Srinivasa M. Gopal, Xavier Periole, and Siewert J. Marrink. The power of coarse graining in biomolecular simulations. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 4(3):225–248, 2014.

[14] L K Beland, P Brommer, F El-Mellouhi, J F Joly, and N Mousseau. Kinetic activation-relaxation technique. 84(4):46704, 2011.

[15] E A Proctor, F Ding, and N V Dokholyan. Dicrete molecular dynamics. 1:80–92, 2011.

[16] A Vitalis and R V Pappu. Methods for Monte Carlo simulations of biomacromolecules. 5:49–76, 2009.

[17] A Zarrinpar, S H Park, and W A Lim. Optimization of specificity in a cellular protein interaction network by negative selection. *Nature*, 426(6967):676–680, 2003.

[18] L. Hakes, S. C. Lovell, S. G. Oliver, and D. L. Robertson. Specificity in protein interactions and its relationship with sequence diversity and coevolution. *Proceedings of the National Academy of Sciences of the United States of America*, 104(19):7999–8004, 2007.

[19] R Rohs, X Jin, S M West, R Joshi, B Honig, and R S Mann. Origins of specificity in protein-DNA recognition. *Annual Review of Biochemistry*, 79:233–269, 2010.

[20] C A Ross and M A Poirier. Protein aggregation and neurodegenerative disease. *Nature Medicine*, 10 Suppl:S10—-17, 2004.

[21] F O Tzul, D Vasilchuk, and G I Makhatadze. Evidence for the principle of minimal frustration in the evolution of protein folding landscapes. *Proceedings of the National Academy of Sciences*, 114(9):E1627–E1632, 2017.

[22] B G Wensley, S Batey, F A Bone, Z M Chan, N R Tumelty, A Steward, L G Kwa, A Borgia, and J Clarke. Experimental evidence for a frustrated energy landscape in a three-helix-bundle protein family. *Nature*, 463(7281):685–688, 2010.

[23] O G Vukmirovic and S M Tilghman. Exploring genome space. *Nature*, 405(6788):820–822, 2000.

[24] Arnab Bhattacherjee and Stefan Wallin. Coupled folding-binding in a hydrophobic/polar protein model: impact of synergistic folding and disordered flanks. *Biophysical Journal*, 102(3):569–578, 2012.

[25] A Irbäck and F Potthast. Studies of an off-lattice model for protein folding: Sequence dependence and improved sampling at finite temperature. 103:10298, 1995.

[26] A Bhattacherjee and S Wallin. Exploring protein-peptide binding specificity through computational peptide screening. *PLoS Computational Biology*, 9(10):e1003277, 2013.

[27] S Wallin. Binding specificity profiles from computational peptide screening. 1561:201–211, 2017.

[28] Philip N. Bryan and John Orban. Proteins that switch folds. *Current Opinion in Structural Biology*, 20(4):482–488, 2010.

[29] 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(6245):324–328, 2015.

[30] S Meier, P R Jensen, C N David, J Chapman, T W Holstein, S Grzesiek, and S Ozbek. Continuous molecular evolution of protein-domain structures by single amino acid changes. 17(2):173–178, 2007.

[31] E S Kuloglu, D R McCaslin, J L Markley, and B F Volkman. Structural rearrangement of human lymphotactin, a C chemokine, under physiological solution conditions. *Journal of Biological Chemistry*, 277(20):17863–17870, 2002.

[32] 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. 150(2):291–303, 2012.

[33] Xuelian Luo, Zhanyun Tang, Guohong Xia, Katja Wassmann, Tomohiro Matsumoto, Josep Rizo, and Hongtao Yu. The Mad2 spindle checkpoint protein has two distinct natively folded states. *Nature Structural & Molecular Biology*, 11(4):338–345, 2004.

[34] P A Alexander, Y He, Y Chen, J Orban, and P N Bryan. A minimal sequence code for switching protein structure and function. *Proceedings of the National Academy of Sciences*, 106(50):21149–21154, 2009.

[35] C Holzgräfe and S Wallin. Smooth functional transition along a mutational pathway with an abrupt protein fold switch. *Biophysical Journal*, 107(5):1217–1225, 2014.

[36] C Holzgräfe and S Wallin. Local versus global fold switching in protein evolution: insight from a three-letter continuous model. *Physical Biology*, 12(2):26002, 2015.

[37] M Kouza and U H Hansmann. Folding simulations of the A and B domains of protein G. *Journal of Physical Chemistry B*, 116(23):6645–6653, 2012.

[38] 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. 136(18):185101, 2012.

[39] César A. Ramírez-Sarmiento, Jeffrey K. Noel, Sandro L. Valenzuela, and Irina Artsimovitch. Interdomain Contacts Control Native State Switching of RfaH on a Dual-Funneled Landscape. *PLOS Computational Biology*, 11(7):e1004379, 2015.

[40] Tobias Sikosek, Heinrich Krobath, and Hue Sun Chan. Theoretical Insights into the Biophysics of Protein Bi-stability and Evolutionary Switches. *PLOS Computational Biology*, 12(6):e1004960, 2016.

[41] N Hansen, J R Allison, F H Hodel, and W F van Gunsteren. Relative free enthalpies for point mutations in two proteins with highly similar sequences but different folds. 52(29):4962–4970, 2013.

[42] T Sikosek and H S Chan. Biophysics of protein evolution and evolutionary protein biophysics. *Journal of The Royal Society Interface*, 11(100):20140419, 2014.

[43] Y He, Y Chen, P A Alexander, P N Bryan, and J Orban. Mutational tipping points for switching protein folds and functions. *Structure*, 20(2):283–291, 2012.

[44] 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. *Biophysical Journal*, 108(1):154–162, 2015.

[45] Enzo Marinari and Giorgio Parisi. Simulated Tempering: A New Monte Carlo Scheme. *Europhysics Letters (EPL)*, 19(6):451–458, 1992.

[46] A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov, and P. N. Vorontsov. New approach to Monte Carlo calculation of the free energy: Method of expanded ensembles. *The Journal of Chemical Physics*, 96(3):1776–1783, 1992.

[47] A Mitsutake, Y Sugita, and Y Okamoto. Generalized-ensemble algorithms for molecular simulations of biopolymers. *Biopolymers*, 60(2):96–123, 2001.

[48] Bernd A. Berg. A brief history of the introduction of generalized ensembles to Markov chain Monte Carlo simulations. *The European Physical Journal Special Topics*, 226(4):551–565, 2017.

[49] Robert H. Swendsen and J-S. Jian-Sheng Wang. Replica Monte Carlo simulation of spin-glasses. *Physical Review Letters*, 57(21):2607–2609, 1986.

[50] U H E Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chemical Physics Letters*, 281:140–150, 1997.

[51] A Laio and M Parrinello. Escaping free-energy minima. 99(20):12562–12566, 2002.

[52] G Favrin, A Irbäck, and F Sjunnesson. Monte Carlo update for chain molecules: biased Gaussian steps in torsional space. *Journal of Chemical Physics*, 114(18):8154–8158, 2001.

[53] A Gelman, G O Roberts, and W R Gilks. Efficient Metropolis jumping rules. 5:599–608, 1996.

[54] Emden R Gansner and Stephen C North. An open graph visualization system and its applications to software engineering. *Software – practice and experience*, 30(11):1203–1233, 2000.

[55] T Sikosek, E Bornberg-Bauer, and H S Chan. Evolutionary dynamics on protein bi-stability landscapes can potentially resolve adaptive conflicts. *PLoS Computational Biology*, 8(9):e1002659, 2012.

[56] J H Meinke and U H E Hansmann. Protein simulations combining an all-atom force field with a Go term. *Journal of Physics: Condensed Matter*, 19:285215, 2007.

[57] W Zhang and J Chen. Accelerate sampling in atomistic energy landscapes using topology-based coarse-grained models. *Journal of Chemical Theory and Computation*, 10(3):918–923, 2014.

[58] Nathan A. Bernhardt, Wenhui Xi, Wei Wang, and Ulrich H. E. Hansmann. Simulating Protein Fold Switching by Replica Exchange with Tunneling. *Journal of Chemical Theory and Computation*, 12(11):5656–5666, 2016.

[59] R K Das and R V Pappu. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proceedings of the National Academy of Sciences*, 110(33):13392–13397, 2013.

[60] W Meng, B Luan, N Lyle, R V Pappu, and D P Raleigh. The denatured state ensemble contains significant local and long-range structure under native conditions: analysis of the N-terminal domain of ribosomal protein L9. 52(15):2662–2671, 2013.

[61] A W Serohijos and E I Shakhnovich. Merging molecular mechanism and evolution: theory and computation at the interface of biophysics and evolutionary population genetics. *Current Opinion in Structural Biology*, 26:84–91, 2014.

[62] T Bereau and M Deserno. Enhanced sampling of coarse-grained transmembrane-peptide structure formation from hydrogen-bond replica exchange. *Journal of Membrane Biology*, 248(3):395–405, 2015.

[63] J Curuksu and M Zacharias. Enhanced conformational sampling of nucleic acids by a new Hamiltonian replica exchange molecular dynamics approach. *The Journal of Chemical Physics*, 130(10):104110, 2009.

Chapter 3

Escherichia coli's RfaH studied by all-atom Monte Carlo simulation

Abstract

RfaH is a compact two-domain protein of the bacteria *Escherichia coli* . Its C-terminal domain (CTD) has been shown experimentally to be able to undergo a complete conformational change from an α-helix bundle to a β-barrel structure. The α-helix bundle to β-barrel fold switch may account for the observed dual role of RfaH, whereby it regulates transcription as well as enhances translation. We employ all-atom Monte Carlo simulations to investigate the stabilities of the two structural forms of RfaH and the character of transition between them. Our simulations reveal that the stand-alone α-helix CTD is relatively unstable. However, it is stabilized by interactions with the N-terminal domain (NTD). Moreover, we observe the stability of the stand-alone β-barrel conformation to be always higher than the α-helix bundle structure. Thus, we conclude that the α-helix bundle to β-barrel fold switch of the CTD in RfaH is thermodynamically favored in our model.

3.1 Introduction

Under physiological conditions, most proteins fold into a unique and stable three-dimensional conformation, the so-called native structure [1]. The native conformation is widely believed to determine a single specific biological function. However, some moonlighting proteins [2–5] like NusE/S10 [6,7] are able to perform multiple functions, while remaining in the same fold, by utilizing different interfaces, through domain separation or oligomerization [4,8]. Presumably, however, remaining in the same fold puts limits on the range of functions that can be carried out.

Recently, a new class of proteins has been discovered with remarkable ability to switch reversibly between two or more folds, giving the capability to further extend functional abilities. For example *lymphotactin* (Ltn) [9] exists in two forms almost equally populated under physiological conditions. One form is a monomeric chemokine fold (Ltn10) and the other a dimeric β-sandwich fold (Ltn40) [9,10]. While the Ltn10 fold acts as an *in vivo* agonist of the G-protein coupled XCR1 receptor, the Ltn40 fold binds glycosaminoglycans [11]. Other examples of proteins that undergo dramatic structural rearrangements include chloride intracellular channel1 (CLIC1) [12], Mad2 spindle checkpoint protein [13,14], Aquifex aeolicus ribosomal protein (L20) [15] and RfaH [16].

RfaH (PDB-ID: 2OUG, Fig.3.1A) belongs to the NusG protein family [17]. This two-domain transcription antiterminator forms a compact structure with closely interfacing N-terminal (NTD) and C-terminal (CTD) domains, which are connected by a flexible linker. This is contrary to the general transcription factor NusG, where the CTD does not interface with the NTD at all. Expectedly, being of the same family, both *Escherichia coli (E. coli)* paralogs RfaH and NusG have a similar NTD conformation, which interact identically with ribonucleic acid polymerase (RNAP) [16]; the enzyme responsible for transcription of deoxyribonucleic acid (DNA) to RNA.

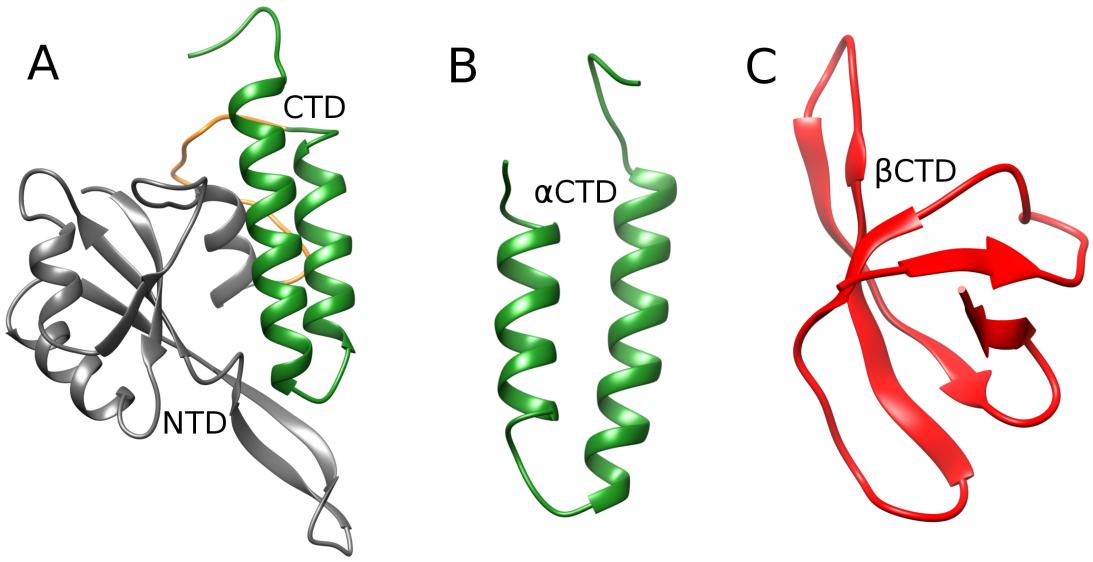

Figure 3.1: Crystal structures. Crystal structures of (A) full-length RfaH (NTD in gray, CTD in green, linker in dark orange), and the CTD of RfaH in (B) α-helix bundle (helix 1 [short], helix 2 [long]) and (C) β-barrel conformations. Missing residues including those of the linker were built with Modeller [18] and the structures were rendered using UCSF Chimera [19].

RfaH-CTD has an α-helix bundle crystal structure when interacting with the NTD resulting in the blockage of the RNAP interaction surface.

Despite having approximately 17% identical amino acid sequences [20], the CTDs of RfaH and NusG have radically different structures. While RfaH-CTD is a two-helix bundle, NusG-CTD is a β-barrel. However, it has been shown experimentally [16] that upon being released from the NTD, thereby allowing the interaction of RNAP, RfaH-CTD is able to transmogrify from an α-helix bundle (Fig.3.1B) to a β-barrel structure (PDB-ID: 2LCL, Fig.3.1C). Also, it is suggested that the binding of RfaH to its DNA target operon polarity suppressor (ops) DNA, triggers the RfaH-CTD release *in vivo* [21,22]. In their experiments using solution nuclear magnetic resonance (NMR) spectroscopy, Burmann *et al.* [16] observed that this β-barrel structure is, essentially, identical to NusG-CTD.

The transition of RfaH-CTD from all-α to all-β structure is essential to the functions of RfaH [23]. When RfaH-CTD adopts the β-barrel conformation, which requires it to be unbundled from the NTD, it enhances translation by recruiting a ribosomal protein S10 (RPS10) to a messenger RNA (mRNA) that lacks a ribosome binding site [24]. Therefore, the fold switch from an α-helix bundle to a β-barrel structure transforms the transcription factor RfaH into a translation factor [16]. That is to say, not only is RfaH able to regulate transcription but it is also capable of enhancing translation at the same time. This complete conformational change of the CTD in RfaH, equipping it to perform different functions, makes it a transformer protein [20].

Several computational studies [23–29] have been done previously on RfaH using a variety of molecular dynamics (MD) simulations including targeted MD, steered MD, replica exchange MD, *et cetera*. A common observation with all these methods suggests that the β-barrel conformation of RfaH-CTD is more stable than its α-helical bundle structure.

Here, we take a different simulation approach by employing Monte Carlo (MC) method armed with an atomistic protein model [30]. To our knowledge, this is the first MC simulation of the RfaH protein. We do this curiously, first, to see if the result from our MC simulations corroborate previous MD simulation results, in which case we provide additional supporting data from a different approach. Secondly, we potentially seek new insights into (I) the thermal stability of (a) the whole RfaH protein with particular attention to its CTD; (b) the isolated α-helix bundle (αCTD) and (c) the stand-alone β-barrel (βCTD) conformations, (II) the folding and unfolding dynamics of αCTD and βCTD, and (III) the αCTD $\rightarrow$ βCTD fold switching tendency in RfaH.

3.2 Methods

3.2.1 Experimental structures

The Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB) (www.rcsb.org) [31] was used to obtain the crystal structures of RfaH (PDB-ID: 2OUG [32], Fig.3.1A) and its isolated β-barrel C-terminal domain (βCTD) (PDB-ID: 2LCL [16], Fig.3.1C). The missing residues Pro101 - Pro112 and Thr157 - Leu162, due to the intrinsic disorder of these segments in 2OUG, were modeled using protein structure modeling program Modeller [18] integrated into the molecular graphics program UCSF Chimera [19].

3.2.2 Computational method

All simulations were done using the software package PROFASI (Protein Folding and Agrregation Simulator [30]) implementing the all-atom protein model described in Ref.33. A single model was first selected from each PDB file and then regularized to satisfy the constraints (such as the bond length and bond angle values) imposed by PROFASI. Moreover, regularization helps to identify the best approximation of the crystal structures which correspond to the minimum of PROFASI's force fields. We used two different types conformational updates: rotamer (ROT) and Biased Gaussian Step (BGS). ROT acts on side chain degrees of freedom (DOFs) alone while BGS change backbone DOFs by taking up to 8 consecutive torsion angles and making a coordinated rotation to execute a semi-local deformation of the protein chain.

3.2.3 Stability properties

To study stability properties, we performed 80 Basic Monte Carlo (MC) simulations each for the full-length RfaH, stand-alone α-helix bundle (αCTD) and β-barrel

MQSWYLLYCK[10]RGQLQRAQEH[20]LERQAVNCLA[30]PMITLEKIVR[40]
GKRTAVSEPL[50]FPNYLFVEFD[60]PEVIHTTTIN[70]ATRGVSHFVR[80]FG
ASPAIVPS[90]AVIHQLSVYK[100]PKDIVDPATP[110]YPGDKVIITE[120]GAF
EGFQAIF[130]TEPDGEARSM[140]LLLNLINKEI[150]KHSVKNTEFR[160]KL

Figure 3.2: The amino acid sequence of full-length RfaH (PDB-ID: 2OUG; color code: α-helix (green), β-sheet (red), unstructured regions (dark orange), numbers (black) to guide in locating the positions of residues on the chain).

(βCTD) C-terminal domain conformations started from the respective (regularized) native structures. Each 80 MC simulations consists of 10 simulations each for eight different temperatures. Every simulation was 10^6 MC cycles with 100 MC steps per cycle. We included an observable to measure the C_α root mean square deviation (RMSD) of the current conformation from the starting structure. In order to avoid unrealistic global conformational changes, we turned off the pivot updates and allowed only ROT and BGS MC moves for all simulations. We set the probability of the conformational updates to 70% and 30% for ROT and BGS, respectively.

3.3 Results

3.3.1 Full-length RfaH

RfaH exists in its native state as a two-domain protein with NTD (residue Met1 - Lys100) and CTD (residue Gly113 - Leu162) linked together by an unstructured loop (residue Pro100 - Pro112). The amino acid sequence of full-length RfaH is shown in Fig.3.2. In order to study stability, we performed a set of unfolding simulations started from the X-ray structure of RfaH (Fig.3.1A) at a range of temperatures between 273-340K.

We observe substantial variations between the different trajectories even at the

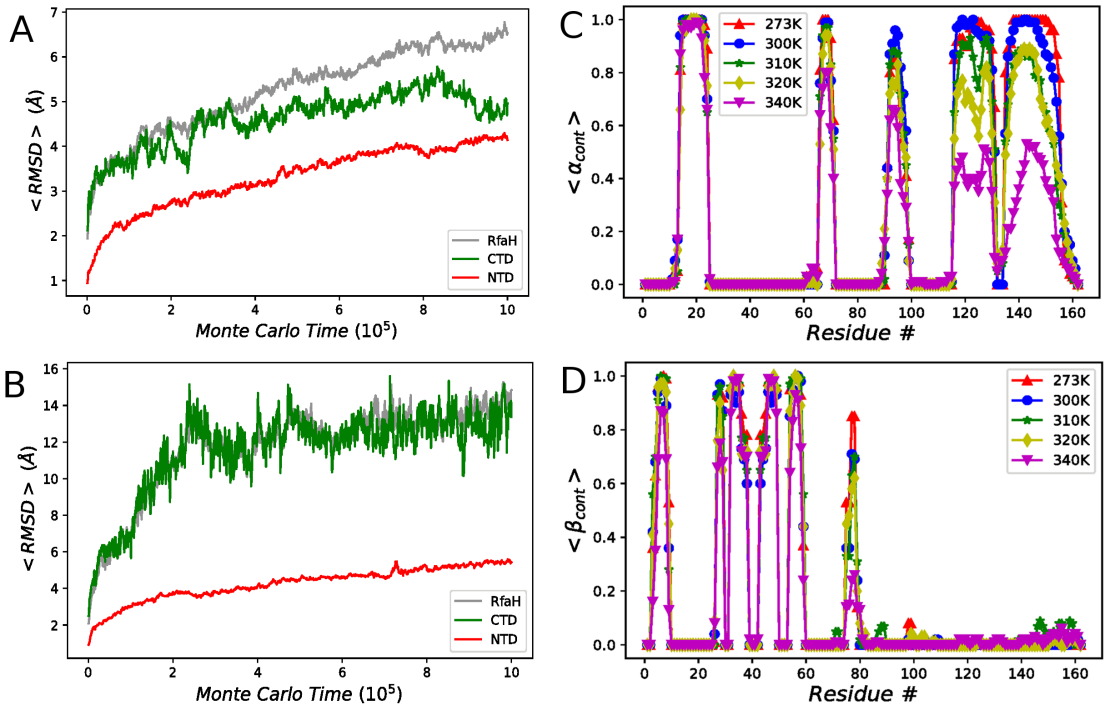

Figure 3.3: Stability of full-length RfaH. Temporal dependence of the average RMSD of conformations assumed during 10 different simulations started from X-ray structure of RfaH at (A) 273K and (B) 300K. The ensemble average of (C) α-helix content profile, and (D) β-sheet content profile of amino acids in full-length RfaH over the range of temperatures; $273K$ (red), $300K$ (blue), $310K$ (green), $320K$ (yellow), $340K$ (magenta).

same temperature. For this reason, to assess stability of the native structure of RfaH, we consider the average RMSD taken over the 10 runs at each temperature. Figure 3.3A and B show $\langle$RMSD$\rangle$ versus Monte Carlo step for full-length RfaH at 273 and 300K respectively. At 273K, both the NTD and CTD of RfaH remain close to the native structure with $\langle$RMSD$\rangle$ below 4Å and 6Å, respectively and RfaH is thus relative stable for the entire duration of simulation. However, Fig.3.3B shows that there is a slow but gradual unfolding of RfaH at 300K. Most of this loss of structure occur in the CTD of RfaH, while its NTD remains close to the native structure, as shown by average RMSD calculated over the NTD and CTD, respectively.

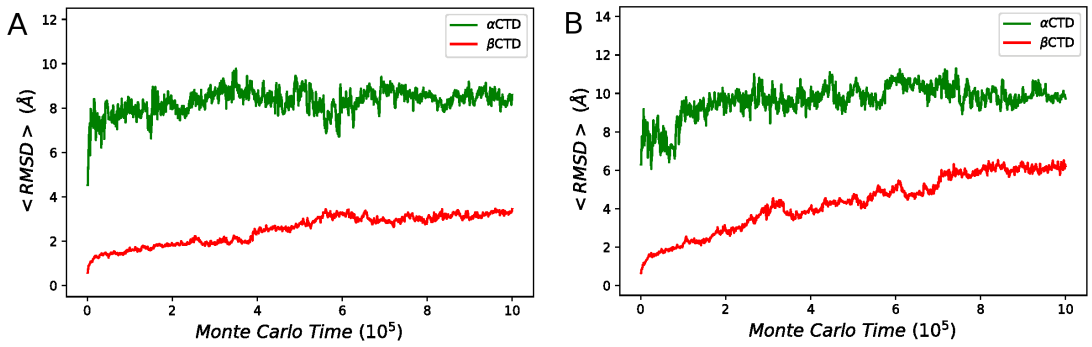

Figure 3.4: Stability of the α-helical bundle and β-barrel structural forms of isolated RfaH-CTD. Time evolution of average RMSD for simulations started with α-helical bundle (green), and β-barrel (red) populations at (A) $273K$ and (B) $300K$. The ensemble average is over 10 different all-atom Monte Carlo simulations, showing that αCTD is less stable than βCTD.

Figure 3.3C and D show the result of the secondary structure content measurement. The average α-helix content (α_{cont}) and β-sheet content (β_{cont}) for each amino acid residue were determined using the secondary structure assignment program DSSP [34, 35]. α-helices have high α_{cont} and low β_{cont} profiles while β-sheets have low α_{cont} and high β_{cont} profiles. Over the range of temperatures, the secondary structure of NTD is well conserved even at higher temperatures, while the α-helices of the CTD "melt" at lower temperatures. Both helices [helix 1 (Ile118 - Thr131) and helix 2 (Gly135 - Lys155)], however, appear to have similar stability.

3.3.2 Isolated C-terminal domain

We now turn our attention to the isolated C-terminal domain of RfaH. As for the full-length protein, we assess stability by performing a set of unfolding simulations started from both the α-helical bundle (Fig.3.1B) and β-barrel (Fig.3.1C) X-ray structures at different temperatures. The average RMSD (Fig.3.4) shows, clearly, that βCTD is more stable than αCTD with the former having a lower $\langle$RMSD$\rangle$ values than the

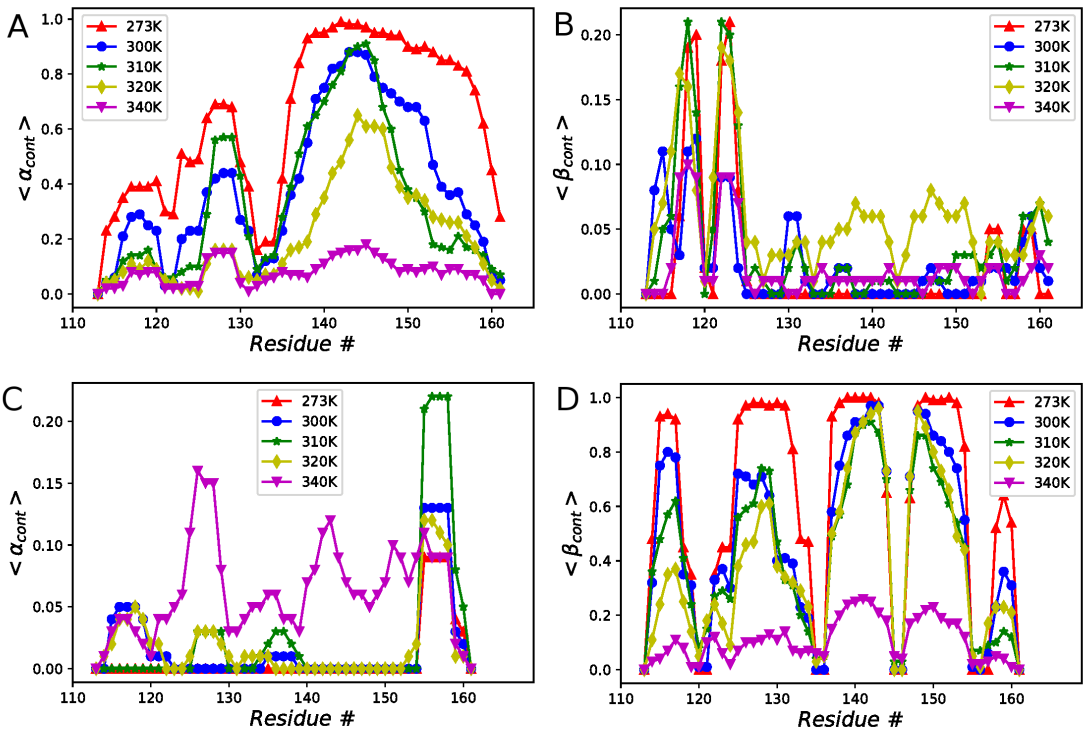

Figure 3.5: Secondary structure content of the isolated CTD. Shown are the α-helix (α_{cont}) and β-sheet (β_{cont}) contents as a function of chain position from simulations started in the α-helix bundle (A,B) and β-barrel (C,D) forms. Results are over the range of temperatures; $273K$ (red), $300K$ (blue), $310K$ (green), $320K$ (yellow), $340K$ (magenta). Residue numbers correspond to those on the full-length RfaH.

latter. For instance, at 273K, the βCTD structural form is relatively stable with $\langle \text{RMSD} \rangle < 4$Å for the entire duration of simulation (Fig.3.4A). In contrast, the αCTD structural form is highly unstable with $\langle \text{RMSD} \rangle > 5$Å, after only a few Monte Carlo cycles. Result at 300K give a similar picture (Fig.3.4B).

To determine the relative stabilities between different structural elements in αCTD and βCTD, we plot the average α-helix and β-sheet contents at different positions along the chain, as shown in Fig.3.5.

The α-helix profile for αCTD (Fig.3.5A)shows that, without the inter-domain interaction with the NTD, the α-helices lose much of their stabilities. In particular,

helix 1 becomes unstable in the isolated fragment. For example, at 320K, the α-helical content of helix 1 is only around ≈ 0.1 while it is > 0.5 in the full-length RfaH at the same temperature.

Both helices have a stronger tendency to lose stability as temperature increases in the isolated CTD than in the full-length protein. However, this tendency is much stronger for helix 1, which not only become less stable but loses most of its helicity even at 273K (Fig.3.5A). This observation of higher stability for helix 2 compared to helix 1 is in agreement with experimental studies by Burmann *et al*. [16] where they used solution nuclear magnetic resonance (NMR). In contrast, helix 1 and indeed the entire CTD is still very much in its folded state at low temperatures in the presence of NTD interactions (Fig.3.3C). Our results thus suggest that the NTD-CTD inter-domain tertiary interactions are necessary to stabilize the CTD in RfaH especially its helix 1. The β-sheet content remains overall quite low, however, some tendencies for β-sheet formation can be seen, especially at 320K (Fig.3.5B).

If instead we begin our simulations with βCTD population, Fig.3.5D when compared with Fig.3.5A shows that the CTD, by surviving higher temperature in this all-β conformation, is thermally more stable than when it assumes αCTD configuration. A closer look at Fig.3.5D reveals an especially high stability of the β-hairpin formed by residue Arg138 - Lys155) and lesser stability of strand 5 (Phe159 - Lys161) in βCTD.

3.4 Discussion

We have used all-atom Monte Carlo (MC) simulations to investigate stability properties of *Escherichia coli*'s RfaH, including the full-length protein and its C-terminal domain (CTD) both in the α-helix bundle (αCTD) as well as in the β-barrel (βCTD)

configurations. The use of MC sampling and an effective energy function [33], has allowed us to carry out multiple simulations under identical conditions, such that average behavior can be assessed. To monitor stability, we computed two measures; (I) the secondary structural content profiles, α_{cont} and β_{cont}, and (II) the root mean square deviation (RMSD) from the starting X-ray structures in our simulations. The results for full-length RfaH revealed that its double α-helical CTD is less stable than the N-terminal domain (NTD) in agreement with experimental observation [16]. Nonetheless, we found that the NTD-CTD inter-domain tertiary interactions give some stability to the αCTD especially at low temperature. In particular, we observed that helix 1 gains stability from the interactions between CTD and NTD. In the absence of the NTD, helix 1 of αCTD is especially unstable. This is contrary to the observation made by Xiong *et al.* [23] using coarse-grained model, in which the melting temperatures for both helices in the isolated αCTD segment is similar, implying similar stabilities. However, Jeevan *et al.* [24] and Li *et al.* [25], independently, reached similar conclusion to ours using biased all-atom molecular dynamics simulations. Interestingly, the relatively higher stabilization of helix 1 by NTD occurs despite both helices having comparable interactions with the NTD [28].

In our αCTD simulations, it is not surprising that little β-sheet structure is formed at low T, as the protein remains locked into α-helical conformations (Fig.3.5B), and at high T, when the protein is expected to globally unfold. Interestingly, however, there is a striking similarity of the β-sheet profiles at intermediate temperature $T = 320K$ between the αCTD and βCTD. This similarity indicates that some level of convergence has been achieved in our simulations, even though β_{cont} remains higher for the βCTD simulations (cf. Fig.3.5B and D).

Although Jeevan *et al.* [24], Li *et al.* [25] and Xiong *et al.* [23] have all previously concluded that βCTD is more stable than αCTD using different measures and

analyses including free energy surface, hydrogen bonding, melting temperature and cooperativity respectively. We show this by simple observation of the time evolution of the RMSD as well as the secondary structural content profiles from separate unfolding simulations of each conformation. As shown by Balasco *et al*. [28], the leucine-rich segments (residues Leu141 - Leu145) in helix 2 may account for its higher stability. Indeed, we find this segment to be especially stable in our simulations (Fig.3.5A).

The stability of full-length RfaH at low temperature in the face of highly unstable stand-alone αCTD suggests that the NTD-CTD inter-domain interactions stabilize the alpha-helix conformation of RfaH-CTD. At say, 300K, it is observed that the RMSD is still < 5Å upto 6×10^5 MC cycles for βCTD (Fig.3.4B) unlike αCTD which is immediately destabilized. Interestingly, the full-length RfaH protein is almost immediately destabilized as well at 300K (Fig.3.3B). Because we know from our previous analysis of the α_{cont} and β_{cont} that the RfaH-NTD remains close to its native conformation, the sudden increase in RMSD for RfaH from ≈ 2Å to ≈ 5Å almost immediately starting the simulation indicates that RfaH-CTD is only marginally stable even while in close proximity to the NTD and interacting with it.

3.5 Conclusion

We have employed all-atom Monte Carlo simulations to investigate the stability of the naturally occurring protein RfaH. Our results revealed that not only is RfaH-CTD less stable than the N-terminal domain (NTD) but it is (marginally) stabilized by it via the NTD-CTD tertiary inter-domain interactions. We also found both helices in the CTD to have similar stability which is disrupted in the absence of the NTD interactions with helix 1 becoming considerably less stable than helix 2. The unfolding times observed from the average RMSD time evolution indicated that βCTD conformation

is always more stable than αCTD structure even with the stabilizing force of the NTD. The relatively low stability of αCTD indicates that it may be primed to switch into the βCTD structural form upon disruption of the stabilizing interface with the NTD.

Bibliography

[1] C. B. Anfinsen and C. B. Principles that Govern the Folding of Protein Chains. *Science*, 181(4096):223–230, jul 1973.

[2] C J Jeffery. Moonlighting proteins. *Trends in biochemical sciences*, 24(1):8–11, jan 1999.

[3] Constance J. Jeffery. Moonlighting proteins—an update. *Molecular BioSystems*, 5(4):345, apr 2009.

[4] Constance J Jeffery. Molecular mechanisms for multitasking: recent crystal structures of moonlighting proteins. *Current Opinion in Structural Biology*, 14(6):663–668, dec 2004.

[5] Shelley D. Copley. Moonlighting is mainstream: Paradigm adjustment required. *BioEssays*, 34(7):578–588, jul 2012.

[6] Catherine L. Squires and Dmitry Zaporojets. Proteins Shared by the Transcription and Translation Machines. *Annual Review of Microbiology*, 54(1):775–798, oct 2000.

[7] B. M. Burmann, K. Schweimer, X. Luo, M. C. Wahl, B. L. Stitt, M. E. Gottesman, and P. Rosch. A NusE:NusG Complex Links Transcription and Translation. *Science*, 328(5977):501–504, apr 2010.

[8] N. Tokuriki and D. S. Tawfik. Protein Dynamism and Evolvability. *Science*, 324(5924):203–207, apr 2009.

[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. *Proceedings of the National Academy of Sciences*, 105(13):5057–5062, apr 2008.

[10] Brian F. Volkman, Tina Y. Liu, and Francis C. Peterson. Lymphotactin Structural Dynamics. In *Methods in enzymology*, volume 461, pages 51–70. 2009.

[11] Philip N. Bryan and John Orban. Proteins that switch folds. *Current Opinion in Structural Biology*, 20(4):482–488, aug 2010.

[12] Dene R. Littler, Stephen J. Harrop, W. Douglas Fairlie, Louise J. Brown, Greg J. Pankhurst, Susan Pankhurst, Matthew Z. DeMaere, Terence J. Campbell, Asne R. Bauskin, Raffaella Tonini, Michele Mazzanti, Samuel N. Breit, and Paul M. G. Curmi. The Intracellular Chloride Ion Channel Protein CLIC1 Undergoes a Redox-controlled Structural Transition. *Journal of Biological Chemistry*, 279(10):9298–9305, mar 2004.

[13] Xuelian Luo, Zhanyun Tang, Guohong Xia, Katja Wassmann, Tomohiro Matsumoto, Josep Rizo, and Hongtao Yu. The Mad2 spindle checkpoint protein has two distinct natively folded states. *Nature Structural & Molecular Biology*, 11(4):338–345, apr 2004.

[14] Marina Mapelli, Lucia Massimiliano, Stefano Santaguida, and Andrea Musacchio. The Mad2 conformational dimer: structure and implications for the spindle assembly checkpoint. *Cell*, 131(4):730–43, nov 2007.

[15] Youri Timsit, Fréderic Allemand, Claude Chiaruttini, and Mathias Springer. Coexistence of two protein folding states in the crystal structure of ribosomal protein L20. *EMBO reports*, 7(10):1013–8, oct 2006.

[16] 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. 150(2):291–303, 2012.

[17] Rachel Anne Mooney, Kristian Schweimer, Paul Rösch, Max Gottesman, and Robert Landick. Two Structurally Independent Domains of E. coli NusG Create Regulatory Plasticity via Distinct Interactions with RNA Polymerase and Regulators. *Journal of Molecular Biology*, 391(2):341–358, aug 2009.

[18] Andrej Šali and Tom L. Blundell. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology*, 234(3):779–815, dec 1993.

[19] Eric F. Pettersen, Thomas D. Goddard, Conrad C. Huang, Gregory S. Couch, Daniel M. Greenblatt, Elaine C. Meng, and Thomas E. Ferrin. UCSF Chimera?A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612, oct 2004.

[20] Stefan H Knauer, Irina Artsimovitch, and Paul Rösch. Transformer proteins. *Cell cycle (Georgetown, Tex.)*, 11(23):4289–90, dec 2012.

[21] Georgiy A Belogurov, Rachel A Mooney, Vladimir Svetlov, Robert Landick, and Irina Artsimovitch. Functional specialization of transcription elongation factors. *The EMBO Journal*, 28(2):112–122, jan 2009.

[22] Georgiy A. Belogurov, Anastasia Sevostyanova, Vladimir Svetlov, and Irina Artsimovitch. Functional regions of the N-terminal domain of the antiterminator RfaH. *Molecular Microbiology*, 76(2):286–301, apr 2010.

[23] Liqin Xiong and Zhenxing Liu. Molecular dynamics study on folding and allostery in RfaH. *Proteins: Structure, Function, and Bioinformatics*, 83(9):1582–1592, sep 2015.

[24] GC B. Jeevan, Yuba R. Bhandari, Bernard S. Gerstman, and Prem P. Chapagain. Molecular Dynamics Investigations of the α-Helix to β-Barrel Conformational Transformation in the RfaH Transcription Factor. *The Journal of Physical Chemistry B*, 118(19):5101–5108, may 2014.

[25] Shanshan Li, Bing Xiong, Yuan Xu, Tao Lu, Xiaomin Luo, Cheng Luo, Jingkang Shen, Kaixian Chen, Mingyue Zheng, and Hualiang Jiang. Mechanism of the All-α to All-β Conformational Transition of RfaH-CTD: Molecular Dynamics Simulation and Markov State Model. *Journal of Chemical Theory and Computation*, 10(6):2255–2264, jun 2014.

[26] GC B. Jeevan, Bernard S. Gerstman, and Prem P. Chapagain. The Role of the Interdomain Interactions on RfaH Dynamics and Conformational Transformation. *The Journal of Physical Chemistry B*, 119(40):12750–12759, oct 2015.

[27] César A. Ramírez-Sarmiento, Jeffrey K. Noel, Sandro L. Valenzuela, and Irina Artsimovitch. Interdomain Contacts Control Native State Switching of RfaH on a Dual-Funneled Landscape. *PLOS Computational Biology*, 11(7):e1004379, jul 2015.

[28] Nicole Balasco, Daniela Barone, and Luigi Vitagliano. Structural conversion of the transformer protein RfaH: new insights derived from protein structure pre-

diction and molecular dynamics simulations. *Journal of Biomolecular Structure and Dynamics*, 33(10):2173–2179, oct 2015.

[29] Sangni Xun, Fan Jiang, and Yun-Dong Wu. Intrinsically disordered regions stabilize the helical form of the C-terminal domain of RfaH: A molecular dynamics study. *Bioorganic & Medicinal Chemistry*, 24(20):4970–4977, oct 2016.

[30] 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.

[31] H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. The Protein Data Bank. *Nucleic acids research*, 28(1):235–42, jan 2000.

[32] Georgiy A. Belogurov, Marina N. Vassylyeva, Vladimir Svetlov, Sergiy Klyuyev, Nick V. Grishin, Dmitry G. Vassylyev, and Irina Artsimovitch. Structural Basis for Converting a General Transcription Factor into an Operon-Specific Virulence Regulator. *Molecular Cell*, 26(1):117–129, apr 2007.

[33] Anders Irbäck, Simon Mitternacht, and Sandipan Mohanty. An effective all-atom potential for proteins. *PMC biophysics*, 2(1):2, apr 2009.

[34] Wouter G Touw, Coos Baakman, Jon Black, Tim A H Te Beek, E Krieger, Robbie P Joosten, and Gert Vriend. A series of PDB-related databanks for everyday needs. *Nucleic Acids Research*, 43, 2015.

[35] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, dec 1983.

Chapter 4

Summary and outlook

4.1 Putting it all together

While the amino acid sequence of a globular protein encodes its native structure, the native structure and associated structural fluctuations of such globular protein are directly responsible for its biological function. Therefore, understanding the effects of mutations in protein sequence is of fundamental importance to the study of protein structure evolution. By extension, the evolution of novel structures is an important means by which new functions may emerge. Although many mutations leave a protein fold unchanged, studies [1, 2] show that a single amino acid substitution can lead to large conformational changes in the native state. The underlying physics of this switching behavior is, however, not yet understood.

The foregoing necessitated us to consider a coarse-grained model for protein folding which provides a framework within which the relationship between sequence and native structure can be explored. Moreover, we developed a generalized-ensemble algorithm for coarse-grained biomolecular simulations, which allowed a systematic study, in large-sequence space, of how novel protein fold may arise from preexisting

folds via series of point mutations. To this end, we characterized the folding of a set of 1024 model sequences ($S16_{1024}$) with 3 amino acid types and another set of the same size ($S35_{1024}$) with each sequence having 16 amino acids in the former and 35 amino acids in the latter. While $S16_{1024}$ was constructed to sparsely span the sequence space between two ideally designed sequences, A1 and N1, folding into an α-helix (IA) and a β-hairpin (IB), respectively, $S35_{1024}$ span the sequence space between A2 and TN, that fold into two-helical bundle (IIA) and mixed α-β (IIB) structures, respectively.

Our results [3] show that intermediate sequences along mutational pathways between two distinct folds are less stable. Particularly, there is, on the average, a minimum stability near the switch points in agreement with experimental and theoretical observations of fold switching in natural and designed proteins. By considering bistable sequences that are able to populate two different folds simultaneously with varying probabilities, we found that, fold switching along individual mutational pathways can either be abrupt, whereby switching occurs directly from one fold to another, or requires one or more bistable sequences. Notably, it was observed that the selection of more stable mutational pathways is accompanied by less frequent bistable sequences in the IA-IB fold pair, which generally is not the case for the IIA-IIB fold pair. In fact, highly stable pathways indicated that fold switching between dissimilar folds are abrupt. In contrary, bistable sequences may always play an important role in connecting similar folds. Our method helps to provide a physical explanation of the effects of mutations and conformational switching in proteins.

Although mutations can lead to conformational switching as we have shown in the study of two fold pairs, there are other causes of this phenomenon. Fold switching can occur either spontaneously [4] or be triggered by changes in biological environment [5], subdomain separation [6], and ligand binding [7]. Thus, we extended our study to a naturally occurring protein in *E. coli* called RfaH whose C-terminal domain (CTD)

has been shown experimentally to be able to undergo an all-α to all-β fold switch by binding to its DNA target operon polarity suppressor (ops) DNA. By employing atomistic Monte Carlo simulations, we observed that RfaH-CTD is less stable than its N-terminal domain. Moreover, the all-α structural form of RfaH-CTD was found to be much less stable than the all-β structural form, suggesting that the all-α to all-β fold switch is thermodynamically favored in our model.

The ability of some proteins to switch their fold and the capacity to understand the underlying physical principles may have important implications in areas including structural biology, human disease, protein design, protein evolution and biotechnological applications [8]. For instance, a direct application of binding-induced fold switching may be the development of a new and more specific drug that is able to hide a function until the target is reached. Furthermore, the ability to predict potential fold switches may lead to novel ways for interpreting genetic polymorphisms and other disease related events [8]. In summary, understanding the physics of protein fold switching may have extensive impact in biology and indeed biophysics.

4.2 Future study

Biased potential for the all-α to all-β fold switch in RfaH

Due to the limitations in our knowledge of protein energetics, folding simulations with our all-atom model PROFASI [9] or other molecular dynamics simulations [10–13] have not yet reproduced the all-α to all-β fold switching behavior in RfaH. Such simulations could provide insight into the mechanism of RfaH switching.

In order to address this challenge, we turn to Gō-like approach to model proteins [14], which has been heavily used in protein folding studies [15]. In a Gō-like model, contacts between amino acids present in the native structure are made artificially

favorable, while all other possible contacts are either neutral or even repulsive. The approach that we are working on is to create a hybrid of the transferable physics-based potential in PROFASI and Gō-like potential that favor the two structural forms of RfaH-CTD. If successful, this may help to more deeply unravel the character of transition between these two structural forms.

Bibliography

[1] P A Alexander, Y He, Y Chen, J Orban, and P N Bryan. A minimal sequence code for switching protein structure and function. *Proceedings of the National Academy of Sciences*, 106(50):21149–21154, 2009.

[2] Patrick A Alexander, Yanan He, Yihong Chen, John Orban, and Philip N Bryan. The design and characterization of two proteins with 88% sequence identity but different structure and function. *Proceedings of the National Academy of Sciences of the United States of America*, 104(29):11963–8, 2007.

[3] A. Aina and S. Wallin. Multisequence algorithm for coarse-grained biomolecular simulations: Exploring the sequence-structure relationship of proteins. *The Journal of Chemical Physics*, 147(9):095102, 2017.

[4] S Meier, P R Jensen, C N David, J Chapman, T W Holstein, S Grzesiek, and S Ozbek. Continuous molecular evolution of protein-domain structures by single amino acid changes. 17(2):173–178, 2007.

[5] E S Kuloglu, D R McCaslin, J L Markley, and B F Volkman. Structural rearrangement of human lymphotactin, a C chemokine, under physiological solution conditions. *Journal of Biological Chemistry*, 277(20):17863–17870, 2002.

[6] 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. 150(2):291–303, 2012.

[7] Xuelian Luo, Zhanyun Tang, Guohong Xia, Katja Wassmann, Tomohiro Matsumoto, Josep Rizo, and Hongtao Yu. The Mad2 spindle checkpoint protein has two distinct natively folded states. *Nature Structural & Molecular Biology*, 11(4):338–345, 2004.

[8] P N Bryan and J Orban. Implications of protein fold switching. 23(2):314–316, 2013.

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

[10] GC B. Jeevan, Yuba R. Bhandari, Bernard S. Gerstman, and Prem P. Chapagain. Molecular Dynamics Investigations of the α-Helix to β-Barrel Conformational Transformation in the RfaH Transcription Factor. *The Journal of Physical Chemistry B*, 118(19):5101–5108, 2014.

[11] Shanshan Li, Bing Xiong, Yuan Xu, Tao Lu, Xiaomin Luo, Cheng Luo, Jingkang Shen, Kaixian Chen, Mingyue Zheng, and Hualiang Jiang. Mechanism of the All-α to All-β Conformational Transition of RfaH-CTD: Molecular Dynamics Simulation and Markov State Model. *Journal of Chemical Theory and Computation*, 10(6):2255–2264, 2014.

[12] Nicole Balasco, Daniela Barone, and Luigi Vitagliano. Structural conversion of the transformer protein RfaH: new insights derived from protein structure pre-

diction and molecular dynamics simulations. *Journal of Biomolecular Structure and Dynamics*, 33(10):2173–2179, 2015.

[13] Liqin Xiong and Zhenxing Liu. Molecular dynamics study on folding and allostery in RfaH. *Proteins: Structure, Function, and Bioinformatics*, 83(9):1582–1592, 2015.

[14] N Gō and H Taketomi. Respective roles of short- and long-range interactions in protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, 75(2):559–63, 1978.

[15] Ronald D Hills, Charles L Brooks, Charles L. Brooks, and III. Insights from coarse-grained Gō models for protein folding and dynamics. *International journal of molecular sciences*, 10(3):889–905, 2009.