UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

ENHANCED SAMPLING METHODS FOR MOLECULAR DYNAMICS SIMULATIONS OF PROTEINS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

NATHAN BERNHARDT
Norman, Oklahoma
2018

ENHANCED SAMPLING METHODS FOR MOLECULAR DYNAMICS SIMULATIONS OF PROTEINS

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF CHEMISTRY AND BIOCHEMISTRY

BY

Dr. U.H.E. Hansmann, Chair

Dr. Liangliang Huang

Dr. Charles Rice

Dr. Rakhi Rajan

Dr. Indrajeet Sharma

I dedicate this text to my loving wife and family whom without their support I would not be where I am today.

Table of Contents

List of Tables

List of Figures

Abstract

Due to the many folds existing in nature, proteins are able to perform a multitude of functions within cells. This relationship between structure and function is one of great importance and has added to the understanding of human disease. With the development of the energy landscape theory, a theoretical description of protein folding has emerged and it is now understood that proteins exhibit a funnel-shaped energy landscape. This model has been widely used for interpreting experimental data but is not able to explain folding for all proteins. For example, there now exists evidence in support of a multi-funnel theory such as that from studies of the proteins GA98, GB98 and RfaH. Probing these landscapes is challenging and many details of their topology, such as transition and intermediate states, remain unknown. Aiding in this task have been computer simulations and special algorithms designed to enhance exploration of protein landscapes like replica exchange molecular dynamics (REMD). While REMD enables calculation of thermodynamic quantities, the method is restricted to rather small protein systems. For this reason, the replica-exchange-with-tunneling (RET) method is introduced in this text and enables fast and efficient exploration of protein landscapes for protein systems of increased size. Using RET and a variant of the multiscale essential sampling (MSES) method combined with a Go-model, simulations of the fold switching proteins GA98, GB98 and RfaH-CTD are performed thus revealing the biophysical interactions driving their behavior. This method is then modified to include the use of a coarse-grained model thereby extending the use of RET to a broader class of protein. Continuing in this direction, a single step resolution exchange method, referred to as resolution-exchange-with-tunneling (ResET), is introduced and makes possible the highly efficient exploration of any protein landscape.

Chapter 1: The Protein Folding Problem

1.1 Introduction

Proteins play an important role in the cell, participating in a variety of processes ranging from neuronal transmission at the synapse to myosin mediated movement of vesicles along actin filaments.[1,2] Indeed, proteins perform numerous roles, even within a single cell. However, despite having many functions, all proteins are made of the same building blocks (the 20 amino acids).[1] These polymers of amino acids are built within cells from instructions stored in DNA. What's more, the rules by which cellular machinery convert genetic instructions into protein (transcription/translation) are known in remarkable detail.[2] Following this process, proteins undergo structural rearrangements, a process called folding, until a stable configuration is reached. While much is known about the folding process, scientists are not yet able to predict protein structure from sequence data alone. Instead, structures must be determined by experiment. All the more troubling, the factors selecting these folds (the driving forces) are often poorly understood.

In the following sections, several factors known to influence protein folding are discussed. Specifically, in section 1.2, the relationship between primary structure[3] and folding is examined. Then, in section 1.3, an introduction to secondary and tertiary structure is given. In this section many of the physical interactions known to drive folding are determined. This is followed by a discussion in section 1.4, on the importance of interactions between proteins and their environment. Because, both the peptide sequence and its environment influence protein structure, a folding theory considering these variables is required. Thus, in section 1.5 the current energy landscape theory[4–6] of protein folding is examined and a brief description of chal-

lenges currently faced by protein scientists is given. Some of these challenges may be overcome by computational models[7] which offer a molecular level view of proteins. However, current simulations fail to describe, for many proteins, a complete picture of the energy landscape. It is thus a major focus of this text to develop and test new computer algorithms better suited to handle the protein folding problem.[8–10]

1.2 Primary Structure, Mutations and the Sequence Space

While non-covalent interactions between proteins and their environment are important (this topic is discussed in section 1.4), aside from these factors, the information necessary for proteins to fold is encoded in the polypeptide sequence. Such a dependence on sequence may be understood by an analysis of the 20 amino acids. Because each differ only by their side-chain group, it is these atoms and the complex interplay between them that determines how a protein folds. If substitution of one side-chain is made for another, also called a mutation, important interactions once favoring a specific fold may be lost and those leading to a new fold acquired. Through the process of natural selection, the genetic code has been sculpted to produce a variety of protein folds that are biologically active and beneficial to the host. These native structures are at least marginally stable[9] and quick to fold.[11] In contrast, some mutations result in a protein that no longer behaves correctly or has deleterious function. In these cases, a disease state (possibly fatal) may emerge.

Take for example the cystic fibrosis disease.[12,13] Subjects with cystic fibrosis carry a mutation in the CFTR gene.[14,15] In healthy tissue, CFTR encodes a protein that functions as an ion channel to chlorine and other small ions.[12] However, certain mutations of the CFTR gene lead to a protein that fails to fold correctly and its subsequent degradation.[16] As a consequence, chlorine ions become concentrated

in affected cells, a condition that draws in water molecules from the surrounding tissues, leaving them dehydrated. Under these conditions, patients with cystic fibrosis become prone to infection of the lungs and have a shortened life expectancy.[12] While medically relevant, membrane proteins like CFTR and other large proteins are notoriously difficult to study, either experimentally or computationally. For this reason, CFTR is not investigated in this text but is mentioned here to illustrate the harmful effects of mutations. Instead, model proteins,[17–22] in which sufficient data may be obtained using the current generation of computer resources, are employed. However, it is often the case that the underlying principles identified in studies of model proteins can be directly applied to more complex protein systems.

While mutations within the genetic code are the main force driving evolution by natural selection, there is no requirement that such changes be random. In fact, there is a growing interest in the design of proteins.[23–25] This area of research has great potential and could lead to the discovery of new protein folds with properties currently unbeknown to nature. However, progress in this field will likely require an efficient means of exploring the sequence space (a space describing how a proteins fold and function changes by mutations) for proteins. One possibility is to search this space using computer simulations and in chapters 4 and 6 this principle is demonstrated by performing mutations in silica. In this manner, the degree in which select residues affect protein structure may be determined.

In this section, the connection between primary structure and protein folding was established. Continuing in this direction, secondary and tertiary structure are introduced next. While proteins are known to exhibit higher levels of structure (quaternary structure), this topic is not discussed in this text.

1.3 Secondary/Tertiary Structure and Driving Forces

With the use of x-ray crystallography[26] and NMR techniques,[21,22] scientists have resolved the structure of more than 130,000 proteins. From these structures, it is known that proteins contain both local and non-local order, also called secondary and tertiary structure.[1] In contrast to tertiary structure, which merely refers to the global arrangement of the polypeptide chain, secondary structure is strictly defined. The two most common types of secondary structure are the α-helix and the β-sheet (see figure 1.1). For α-helices, the backbone atoms of a protein spiral around an imaginary axis, running parallel to the helix. These structures are stabilized by reoccurring hydrogen bonds between backbone atoms. β-sheets on the other hand, require the backbone atoms be extended in a zig-zag pattern. As with α-helices, β-sheets are stabilized through hydrogen bonding. However, with β-sheets, the bonds occur between adjacent β-strands which may be located far away in the peptide sequence. Depending on the direction of adjacent β-strands, a β-sheet may be classified as parallel or antiparallel (see figure 1.1 for an example of each).

Given the occurrence of α-helices and β-sheets in most protein folds, it might seem tempting to suppose their formation the driving force controlling tertiary structure.[27] However, predicting secondary structure from sequence alone remains challenging and it seems that tertiary structure may play as much a role in determining secondary structure as the other way around.[28,29] Indeed, it is unlikely that protein folding is ever controlled by a single factor (like the formation of secondary structure) alone but instead results from many driving forces working together such as the van der Waals interaction, hydrogen bonding, formation of salt bridges and the hydrophobic effect.[9] Due to the importance of these factors, they are often quan-

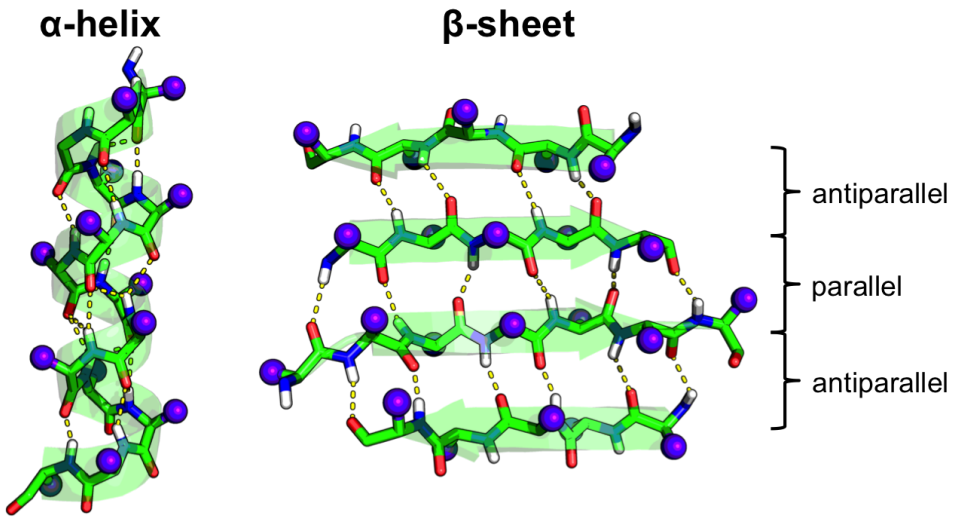

Figure 1.1: Common types of secondary structure. Shown on the left is an α-helix and on the right a β-sheet containing both parallel and antiparallel sheets. Dotted lines highlight stabilizing hydrogen bonds. Side-chain positions are shown as purple spheres.

tified in protein simulations and used to explain structural stability. What's more, many of these factors are through space interactions and occur between the protein and other molecules. For this reason, the chemical environment plays a major role in determining protein structure. This topic is discussed in the next section.

1.4 The Protein Environment and Free Energy

While many factors[9] contribute to the stability of a given protein fold, including hydrogen bonding, the van der Waals interaction and electrostatics, the hydrophobic effect is perhaps most interesting. It is known from experiments of small model proteins that the movement of hydrophobic side-chains from an aqueous medium to the organic phase results in a 1-2 kcal/mol drop in free energy.[9] Because any given protein may contain numerous hydrophobic side-chains, their isolation from

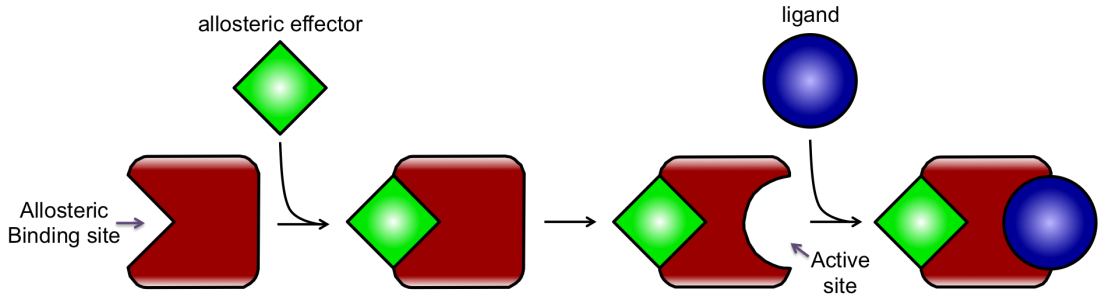

Figure 1.2: Activation of an allosteric protein by binding of an allosteric effector.

water results in a significant reduction of free energy. As a consequence, proteins burry most of these residues in their interior core.[10] What's more, the placement of pre-folded protein in an organic phase, where the hydrophobic interior now interacts favorably with the solvent, leads to an unfolding of the protein.[30] These results demonstrate how factors other than peptide sequence can influence protein structure and highlight the importance of interactions between proteins and their environment.

On the other hand, not all environmental effects are as harmful to protein function as their placement in an organic solvent. In fact, interactions between proteins and other biomolecules are instrumental to the regulation of cellular processes.[2] Take for example allosteric proteins.[1,31,32] An allosteric protein, the word allostery being Greek for "other shape", refers to any protein that undergoes a structural change that alters its activity or function subject to binding of one or more ligands. Such changes in activity are induced by binding of an allosteric effector at the allosteric site, resulting in structural changes to binding pockets distal this site (see figure 1.2). These changes can have either positive or negative effects on binding to other molecules. For this reason, allostery is frequently used by nature to regulate protein activity. In this way, proteins may be readily available but turned on or

off as needed by the cell. Due to the many cellular pathways[2] regulated allosterically, the determination of molecules which mimic these effects may be important for the development of new pharmaceuticals. However, predicting changes in protein structure as a result of binding by other molecules remains a challenge for both computationalists and experimentalists. Computationally, the problem is to collect sufficient data as to allow accurate computation of free energy changes. Toward this end, the importance of inter-domain contacts in selecting the fold of the model protein RfaH-CTD is investigated in chapter 7. These simulations take advantage of the enhanced sampling techniques developed in chapters 5 and 6.

Yet another example demonstrating the power environmental conditions exert on proteins comes from the study of amyloids,[33–35] a protein state associated with many diseases including Alzheimer's disease and type II diabetes.[34,35] Under healthy conditions, proteins assume a so-called native state in which the protein is correctly folded and maintains proper function. However, under certain environmental conditions, proteins may be driven from their native fold thereby exposing their hydrophobic core and leaving them open to aggregate with other protein, the end result being the formation of a stable complex known as an amyloid fiber[33–35] (see figure 1.3). These fibers are characterized not only by their toxicity to cellular tissues but also structural features such as a repeating array of β-strands in perpendicular alignment to the fibril axis.[36,37] Amyloid fibers are incredibly stable, owing to a network of hydrogen bonds and salt bridges, although the important interactions could vary depending on the fiber. Studies show fiber formation generally takes years in-vivo[38] and even hours to days in the lab under irregular conditions.[39] This lag phase preceding fiber formation can be explained as the time required for formation of a stable seed[40] (see figure 1.3). Lag times can be shortened; however, if

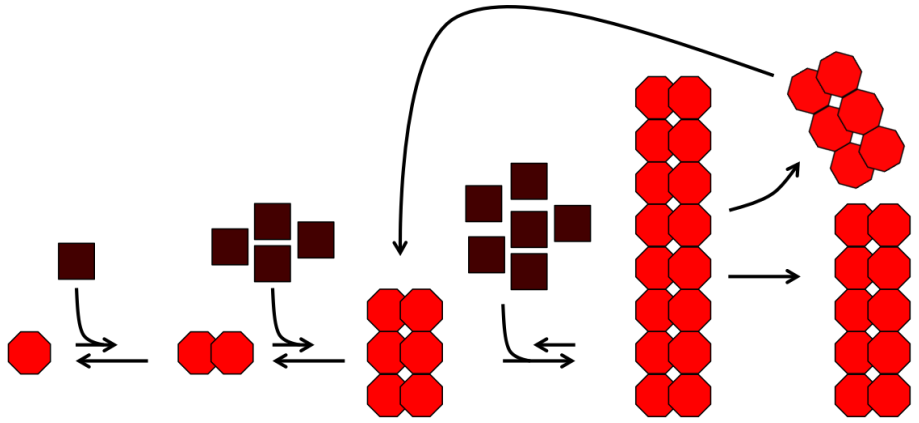

Figure 1.3: Formation of amyloid fibers from monomers. Arrows indicate the favored direction. Also shown, is the breaking of a fiber, a process that generates additional seeds. This figure follows a model by Harper and Lansbury.[40]

pre-formed oligomers are added to monomers,[38] demonstrating the ability of amyloid seeds to denature folded protein. What's more, amyloid seeds of a particular type can seed fibrils from monomers of a different type, a phenomenon called cross-seeding.[41] In chapter 4, the molecular mechanism of amyloid cross-seeding observed for the amylin[36,41–43] protein is investigated. This study uses molecular dynamics simulations to determine the factors promoting efficient cross-seeding. Then, in chapter 6, the important interactions promoting fiber formation are determined for the small 13-residue fragment of the amyloidogenic protein serum amyloid A.[20,44–46] These later simulations use the enhanced sampling methods introduced in chapter 5 of this text.

Together, the examples listed in this section demonstrate the ability of molecules in the cellular environment to alter protein structure. Taking into consideration the driving forces discussed previously, a picture now emerges where many factors are seen to influence protein structure. This picture may be simplified; however, by computation of a quantity introduced at the beginning of this section, the free

energy. With this variable, the problem of folding is reduced to finding the structure with the lowest free energy. In the following section, this topic is discussed further by an introduction of the energy landscape theory of protein folding.

1.5 The Energy Landscape Theory and Current Challenges

Despite numerous examples demonstrating the direct ties between protein structure and function,[3,13,47] a theoretical description of the folding phenomena remains incomplete. In early experiments by Anfinsen,[48] it was shown that many denatured proteins can refold to their native states on an experimentally observable timescale. These observations influenced Levinthal in his work on the folding problem. Levinthal concluded that, given the vast number of conformational states accessible to a typical protein, a random walk through configuration space (even if spending a very short time in each state) would take an astronomically large amount of time to reach the native state.[11] Levinthal's time estimates for folding by a random walk require each configuration of the protein to be equally probable or rather the free energy landscape surrounding the native fold be flat. Because proteins are not so slow to fold it is now understood that most exhibit instead a funnel-shaped energy landscape.[4–6] In the funnel theory, proteins are thought to follow a folding pathway[49] or pathways which guide the protein down the funnel toward the native state housed at the bottom. This model has been very powerful in explaining data[4] from folding experiments as well as computer simulations. However, the single-funnel theory is not able to explain folding for all proteins and there is now mounting evidence[21,22,50,51] in support of a multi-funnel landscape theory.

In the multi-funnel theory, different states compete with one another leading to a funnel with multiple basins of similar depth. Evidence of such funnels comes

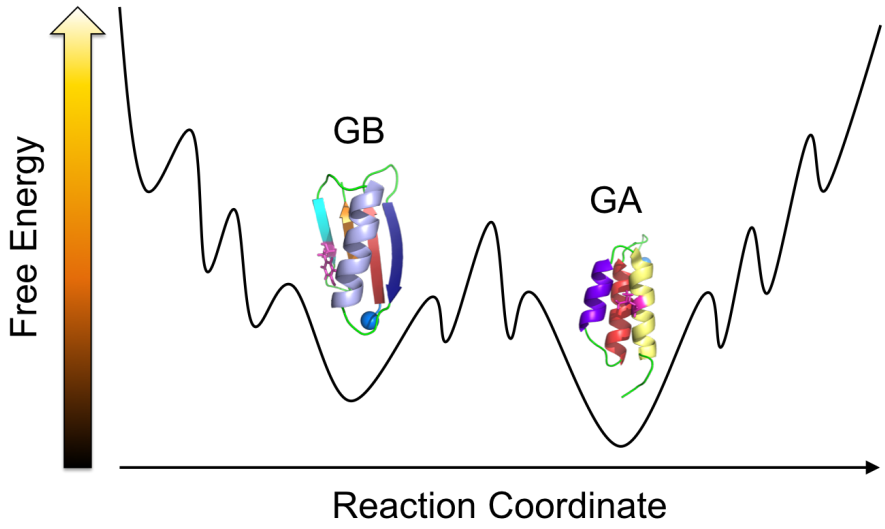

Figure 1.4: Hypothesized multi-funnel energy landscape of GA98. Shown in the dominant basins are the GA and GB folds.

from, at least, two sets of studies. The first of these is a series of mutational experiments performed by Orban et al.[21,50,51] In Orban's work, the sequence space separating two subdomains of protein G,[52] termed GA and GB, was studied. Protein G contains two domains that bind to serum proteins in the blood. The first of these (the GA domain) is a 45 residue polypeptide that binds to human serum albumin (HAS)[53] whereas the second (the GB domain) is a 56 residue polypeptide that binds to a region of Immunoglobulin G (IgG).[54] The natural versions of GA and GB share no significant sequence homology and have different folds, GA forming a bundle of three α-helices[55] and GB four β-sheets and an α-helix[26] (see the structures displayed in figure 1.4). Orban proceeded to increase sequence homology by making mutations in both sequences and testing the mutants for their ability to bind the respective ligands. Secondary structure of mutants was also assessed by circular dichroism[21] and high-resolution NMR[56] structures were determined for key mutants. The resulting mutants of this study,[21] named GA98 and GB98, share 98% sequence

identity and differ by a single amino acid at position 45, GA98 having a leucine and GB98 having a tyrosine. These mutants both retain the fold of their parent proteins GA and GB.[56] However, GA98 displays a small affinity to bind with IgG, indicating that a small portion of GA98 occupies the GB fold.[21] These results suggest the energy landscape of GA98 contains two funnels, one for the GA fold and the other the GB, which are both significantly populated (see figure 1.4). In chapter 6, the multi-funnel energy landscape of GA98 is probed further using computational methods developed in chapter 5.

The second example supporting the multi-funnel theory comes from experiments of the transcription factor RfaH.[22,57–62] This regulatory protein contains both a C- and N-terminal domain, which assume very different globular structures and together play an important role in the regulation of transcription and translation within *Escherichia coli*. More specifically, in the absence of RfaH, transcription of DNA sequences containing an ops element (operon polarity suppressor)[62] is cut short by Rho-dependent termination[63,64] (figure 1.5). However, binding of the RfaH N-terminal domain to RNA-polymerase (RNAP) blocks the binding site of the Rho recruiter NusG and transcription is rescued.[22,64] Subsequent recruitment of the ribosomal subunit 30S by the RfaH C-terminal domain then initiates translation at the growing RNA chain and protein is produced as normal.[22] In its inactive form, the RfaH C-terminal domain is locked in a helix hairpin fold by interactions with the N-terminal domain.[22,59] However, upon loss of these interactions or as an isolated protein,[22] RfaH C-terminal domain undergoes a spontaneous structural rearrangement ending in a β-barrel conformation (also the active form that binds the 30S ribosomal subunit). Initially, it was hypothesized that the β-barrel form of RfaH C-terminal domain is the lowest energy fold but, due to constraints imposed

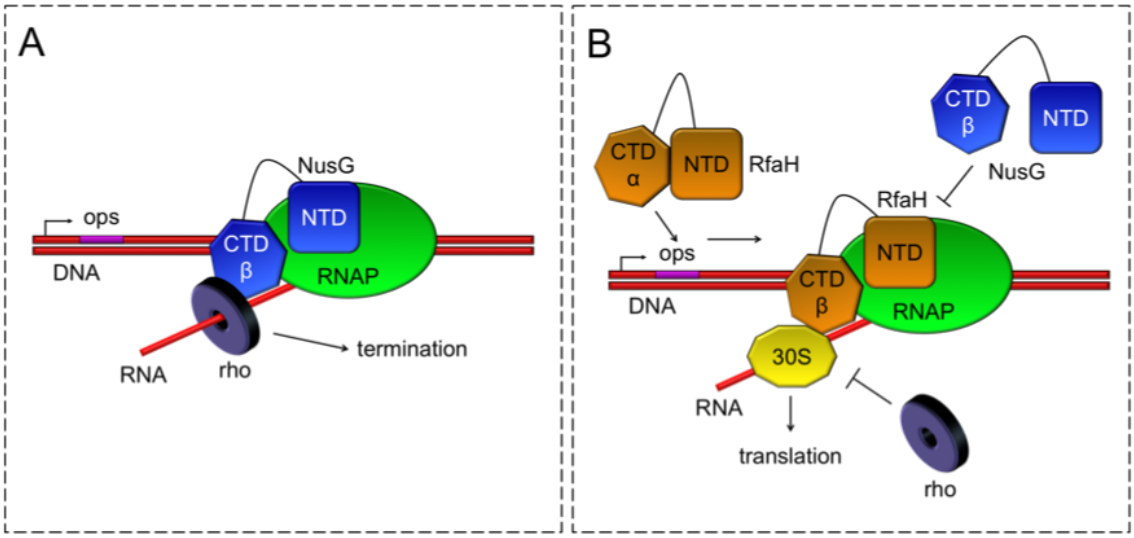

Figure 1.5: Rho mediated termination of transcription in the absence of RfaH. (A). RfaH binding to RNA polymerase rescues transcription and initiates translation. (B).

by the N-terminal domain, is not realized on a timescale observable by experiment. However, experiments[57] that reversed the peptide sequence, and thus the order in which the RfaH domains are transcribed thereby giving the C-terminal domain time to fold by itself, show that the helix bundle is still preferred. These results suggest the helix bundle of RfaH C-terminal domain is not kinetically trapped but is the thermodynamically favored fold in the presence of inter-domain contacts. Such a description would imply that both folds are separated by a marginal free energy difference and that contacts with the N-terminal domain lower the basin containing the helix bundle, thus selecting that fold. The free energy landscape of the isolated RfaH C-terminal domain is examined further in chapter 7 using the methods developed in chapters 5 and 6.

Both experimental data from GA98 and RfaH experiments suggest the existence of dual funnel energy landscapes, for at least some proteins. However, further prob-

ing of protein landscapes by experiment remains a challenge leaving many details of their topology unknown. Specifically, the following questions should be addressed.

1. What are the important structures along the funnel and what are their relative free energies?

2. What barriers separate relevant folds and what do the transition state structures look like?

Because both intermediate folds and transition states are difficult to observe by experiment, the push for computer models capable of producing accurate protein landscapes has become ever more enticing.

1.6 Summary

As a consequence of the many protein folds existing in nature, proteins are able to perform a multitude of functions within cells.[1,2] Because protein activity depends heavily on structure, the identification of properties affecting protein plasticity is an important step toward understanding disease and developing protein-based materials. Progress has been made toward this end and, with the aid of the energy landscape theory[4–6] and advances in experimental methodologies,[21,22,26] a theoretical description of folding is now understood. However, a complete picture of protein funnels remains inaccessible for most protein systems leaving many details of their topology blank. Important information regarding intermediate structures and transition states is especially difficult to obtain by experiment leading to the need for accurate computer models.

In the following chapter, the computer methodologies commonly used in protein studies are introduced. While their application has seen great success, these capa-

bilities are demonstrated in chapter 4 with the investigation of amylin oligomers by molecular dynamics simulation, there are limits to their use. These shortcomings are also examined in chapter 2 and, in chapters 5 through 8, solutions to some of these challenges are discussed. Specifically, in chapter 5 the replica-exchange-with-tunneling (RET) method is introduced. This method allows for quick exploration and production of the energy landscape for many protein systems of significant size. With RET, the dual funnel energy landscapes of GA98 and GB98 are examined in chapter 6. The interactions promoting fibrilization of the 13-residue fragment of serum amyloid A are also determined in this chapter and validation of the method is performed. Similarly, in chapter 7 the energy landscape of the isolated RfaH C-terminal domain is investigated and a transition pathway, connecting the helix hairpin and β-barrel folds, is described. And finally, in chapter 8 additional methods combining RET with a coarse-grained model are introduced, thus enabling investigation of a broader class of proteins. The resolution-exchange-with-tunneling (ResET) method is also introduced in chapter 8 and demonstrates how multi-scale simulations of protein systems may be performed using only modest computational resources. Coarse-grained and multi-scale models are discussed in section 2.3. For further discussion of the research strategy taken in this text the reader may refer to chapter 3.

Chapter 2: Computer Simulations to Study Biomolecules

2.1 Quantum Mechanics and Molecular Dynamics

Because proteins are made of atoms, they are governed by the laws of quantum mechanics.[65–69] Therefore, any theoretical description of protein folding must be rooted in quantum theory. For the purpose of this text, Schrodinger's wave theory[70] will suffice. By introduction of the Schrodinger equation, scientists have the potential to model any chemical system of interest, although the mathematics involved may be non-trivial. Aiding in this task, have been computers and computational methods making possible quantum calculations in-silica. Nevertheless, quantum calculations remain computationally expensive and for some systems, like a protein in an aqueous environment, are not practical. Thus, for protein simulations, one instead turns to molecular mechanics also called molecular dynamics (MD).

With MD, atomic systems are handled classically by Newton's equations. Replacement of the time-dependent Schrodinger equation by Newton's $F = MA$ is made possible because the motion of the nuclei within a molecule is slow compared to that of the electrons. Thus, for any nuclear arrangement, the electronic effects may be averaged out. For this reason, it is possible to compute, for any system, a ground state potential energy as a function of the nuclear positions only.[7] Such a function is referred to as a force field. The development of atomic force fields is an active area of research and proposal of new force fields continues today.[71,72] Still, there is no single force field currently accepted by the scientific community as the correct one for describing all aspects of any molecule. With that said, there are many force fields in use that accurately predict, for a broad class of molecules including proteins, some molecular properties like folding dynamics.

When building a force field, there are no strict rules requiring it take a particular form. However, the following terms are common

$$V\left(r^N\right) = \sum_{bonds} \frac{k_b}{2}\left(I - I_o\right)^2 + \sum_{angles} \frac{k_\theta}{2}\left(\theta - \theta_o\right)^2 + \sum_{torsions} \frac{k_\phi}{2}\left(1 + cos\left(n\phi - \gamma\right)\right) + \sum_{i=1}\sum_{j=i+1}\left(4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}\right). \tag{2.1}$$

In 2.1, k_b, k_θ, k_ϕ, n, γ, ϵ_{ij} and σ_{ij} are constants that must be determined for each pair-type or group interaction and q_i and q_j are partial charges assigned to each atom. The process by which these constants are determined is called force field parameterization. What's more, there are two philosophies when it comes to parameterizing a force field. With the first approach, parameters are fit to existing experimental data[72] as opposed to the second, where parameters are determined by quantum calculations.[71] Many force fields have been parameterized by a mix of these strategies.

The inclusion of 2.1 in modern force fields may be attributed to the simplicity of these terms as well as the efficiency in which they are computed and the accuracy of the resulting dynamics. All the more appealing, each term of 2.1 may be understood as an energy contribution stemming from a particular type of motion (see figure 2.1) or, for some of the terms, the physical origin is known. Take, for example, the first three terms of equation 2.1. These terms describe the energetics of short-range interactions and are each characterized by a different type of molecular motion. Specifically, the first term $\left(\frac{k_b}{2}\left(I_i - I_o\right)^2\right)$ adds an energy penalty for any bond length that deviates from a reference value (I_o). While an anharmonic potential is known to more accurately describe the energy profile of a stretching bond, a simple harmonic

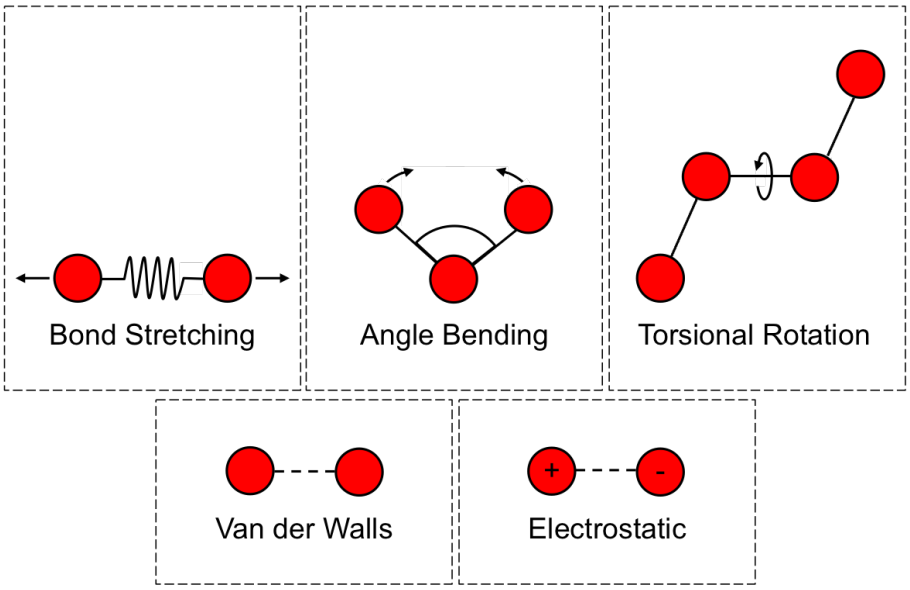

Figure 2.1: Molecular motions and interactions described by the potential energy functions of equation 2.1. Motions due to bonded interactions are shown in the top panels whereas non-bonded interactions are shown in the bottom pannels

term, as used here, aptly characterizes bond lengths near the equilibrium value and is cheaper to compute. The second term $\left(\frac{k_\theta}{2}\left(\theta_i - \theta_o\right)^2\right)$ penalizes for valance angle deviations from a reference value (θ_o), where a valance angle is the angle between three bonded atoms. The third term $\left(\frac{k_\phi}{2}\left(1 + cos\left(n\phi - \gamma\right)\right)\right)$ models energy due to steric interactions associated with rotation about a chemical bond.

The last set of terms in 2.1 characterize long-range or "through space" interactions and have origins rooted in well understood physical phenomena. For example, the so-called Leonard Jones 6-12 terms $\left(4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]\right)$ are used to model the van der Walls interaction and may be broken into 2 parts, one attractive in nature and the other repulsive. The attractive terms $\left(4\epsilon_{ij}\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right)$ act over a long-range and model dispersion forces whereas the repulsive terms $\left(4\epsilon_{ij}\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}\right)$ act over a short-range and arise from the Pauli principle. Finally, the last terms in

the double sum $\left(\frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}\right)$ model the coulombic interaction. Because different atoms possess different electronegativities, most molecules have a non-uniform charge distribution. A common way in which the electrostatic properties of a molecule may be captured by MD is to assign partial charges to the nucleus of each atom. However, the methods for determining these partial charges are complex and will not be discussed here. For more information regarding force fields and the methods by which they are parameterized see references 7, 73 and 74.

As a side note, it is worth mentioning that the force fields discussed so far were assumed to represent each atom of the system explicitly and are thus referred to as fine-grained (FG) models. There are; however, force fields that do not represent every atom of a molecule but present instead a reduced representation. These force fields are called coarse-grained[75] (CG) models and are discussed in section 2.3. It is assumed for the remainder of this section that a fine-grained model is used. Now, with the energy of the system well defined, forces are computed in the usual manner

$$F_i = -\frac{dV\left(r^N\right)}{dr_i}. \tag{2.2}$$

Using these forces and the current state of the system, one can solve Newton's equations by means of numerical integration. A numerical integrator commonly used in MD simulations is the leapfrog update. With leapfrog, the coordinates and momentum are updated in separate steps (see figure 2.2) by

$$V\left(t+\frac{1}{2}\Delta t\right) = V\left(t-\frac{1}{2}\Delta t\right) + A\left(t\right)\Delta t \tag{2.3}$$

and

$$X\left(t+\Delta t\right) = X\left(t\right) + V\left(t+\frac{1}{2}\Delta t\right)\Delta t. \tag{2.4}$$

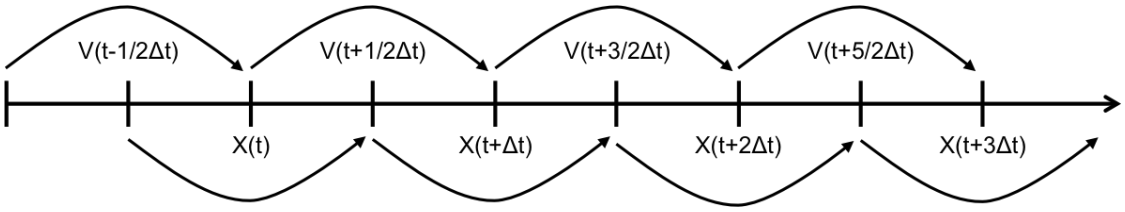

Figure 2.2: Leapfrog updating schematic.

The leapfrog integrator has unique qualities making it well suited for MD simulations. Specifically, the method is time reversible and preserves phase space volume. As a consequence, simulations employing the leapfrog integrator are able to conserve energy for long periods of time. This enables acquisition of large trajectories and accurate calculation of time average quantities using

$$\langle A \rangle = \lim_{\tau \to \infty} \frac{1}{\tau} \int_{t=0}^{\tau} A(P^N(t), r^N(t)), \tag{2.5}$$

where $r^N(t)$ and $P^N(t)$ are the coordinates and momentum of the system respectively and depend on time. Taking the limit in 2.5, the time and ensemble averages become equal. Thus, for sufficiently long trajectories, MD simulations can be used to sample states in accordance with a desired statistical ensemble. For the microcanonical ensemble, the method described thus far is sufficient. However, if one wishes to sample from other ensembles, such as the canonical or isobaric ensemble, a thermostat or barostat must be introduced. A deeper discussion of thermostats and barostats is beyond the scope of this text and will not be considered here.

With adequate sampling, thermodynamic properties may be determined. For example, the heat capacity of the system is computed by

$$C_v = \left(\frac{\partial U}{\partial T} \right). \tag{2.6}$$

Using 2.6, the heat capacity may be found by performing many simulations of the system of interest, each differing in temperature, and analyzing how the average energy changes. In other ways, the remaining thermodynamic quantities may be calculated. Of particular interest to protein scientists is the change in free energy of the system along some internal coordinates. Such a projection of the free energy is called a potential of mean force (PMF).[7,76,77] With the selection of appropriate reaction coordinates, a PMF may be used to identify folding pathways as well as the transition states separating important folds. For this reason, PMF plots are considered frequently throughout this text.

While MD simulations make possible the computation of thermodynamic quantities, these calculations are often inaccurate (for protein systems). Most of the time, this is because the trajectory collected is too short or the sampling insufficient. To prevent this problem, MD trajectories should cover the same time scale as the phenomena studied. For protein folding, this ranges from a microsecond to a millisecond. However, acquisition of such trajectories on general purpose computers is usually not feasible. This is because the time step used in equations 2.3 and 2.4 is limited by the fastest degrees of freedom of the system, i.e. the vibration of hydrogen atoms. While there has been work done to remove these motions,[78,79] the time step used in protein simulations rarely exceeds two femtoseconds, making difficult simulations on the microsecond timescale. This problem is exacerbated by the inclusion of an explicit solvent,[80] as the number of calculations for each update becomes very large. For this reason, alternative approaches have been introduced for enhancing the exploration of protein landscapes. These methods are discussed in the following sections.

With the first two methods discussed, markov chain monte carlo and replica

exchange, smooth trajectories of MD simulations are abandoned for a discontinuous sampling of states in accordance with a desired distribution. These methods are discussed in detail in section 2.2. Then, methods that reduce the degrees of freedom of the system, thus lowering computational cost required to produce long trajectories, are introduced. These are the coarse-grained models. Because removing degrees of freedom results in a loss of accuracy, multi-scale models have been developed to offset this problem and are also discussed here. With the final approach considered, structure-based models are introduced which, while unphysical, have added to the understanding of protein folding. Coarse-grained, multi-scale and structure-based models are discussed in section 2.3. While each of these approaches enhances the sampling of states, these enhancements come with a cost. Therefore, the methods introduced in chapters 5 through 8 aim to reduce these effects while retaining the enhanced sampling properties.

2.2 Markov Chain Monte Carlo and Replica Exchange

An alternative method to MD capable of producing Boltzmann weights for ensemble members is the markov chain monte carlo (MCMC) simulation. While the main focus of this text is not toward MCMC, the principles learned here are directly applicable to the general ensemble methods discussed at the end of this section. Thus, a brief introduction to MCMC is warranted.

With MCMC, states are no longer found deterministically but are sampled in a probabilistic fashion. This is accomplished by constructing a markov process so that the future state of the system depends only on the current state and not the past (see figure 2.3). With MCMC, samples are either thrown away or stored to make a markov chain. These samples are selected such that, for sufficiently large chains,

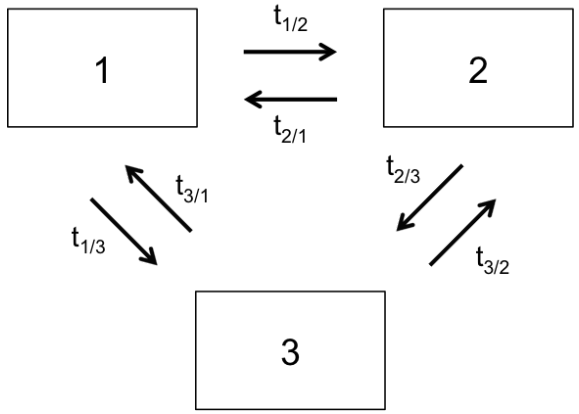

Figure 2.3: Transition between states as governed by the transition matrix T. The members of $t_{i/j}$ contain the probability that the system moves from state i to j. For simplicity a system containing only 3 states is shown.

a desired distribution is asymptotically approached.[81] When building this chain, a molecular system is evolved over a discrete set of states (k in total) as driven by the transition matrix T and probability density vector X[7]

$$T = \begin{bmatrix} t_{1/1} & t_{2/1} & \dots & t_{k/1} \\ t_{1/2} & t_{2/2} & \dots & t_{k/2} \\ t_{1/3} & t_{2/3} & \dots & t_{k/3} \\ \vdots & \vdots & \ddots & \vdots \\ t_{1/k} & t_{2/k} & \dots & t_{k/k} \end{bmatrix} \qquad X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_k \end{bmatrix}, \tag{2.7}$$

where $t_{i/j}$ is the probability of moving to state j if at i and x_i is the probability density for state i. If one imagines n non-interacting copies of the system, then (for large n) nTX governs how the copies are redistributed between the states. Ignoring n momentarily, T acts on X to redistribute probability density

$$TX_m = X_{m+1}. \tag{2.8}$$

In the limit that T acts on X many times, it is possible that the densities of X converge to some distribution. When further acts of T on X no longer change its value, X is said to be stationary

$$TX_m = X_m = X^*. \tag{2.9}$$

If a stationary distribution exists, represented in equation 2.9 as X^*, T must be selected so that it is unique. This is accomplished by requiring the markov process to be ergodic.

One way to test that a distribution is stationary is to check, for all i and j, the condition of detailed balance

$$x_i t_{i/j} = x_j t_{j/i} \qquad (for\ \ all\ \ i\ \ and\ \ j)\,. \tag{2.10}$$

When detailed balance is satisfied, it can be shown that equation 2.9 is also satisfied. However, detailed balance is most easily understood as an equilibrium between states. To see this, let n be the total number of systems distributed over the k states. Multiplying equation 2.10 by n gives $nx_i t_{i/j} = nx_j t_{j/i}$. This equation may be simplified by recognizing that nx_i is the number of systems in state i (n_i). Making this substitution gives $n_i t_{i/j}$, which is the number of systems moving from i to j ($n_{i/j}$). Thus, equation 2.10 reduces to a statement of equilibrium

$$n_{i/j} = n_{j/i}. \tag{2.11}$$

Together, equations 2.9 through 2.11 state that if n non-interacting copies of the system are allowed to evolve, as driven by the transition matrix T, then after some

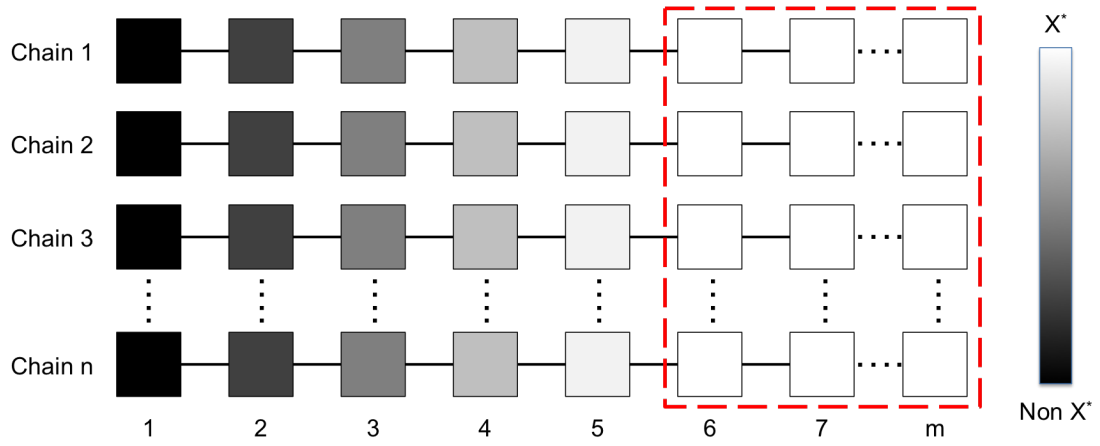

Figure 2.4: Cartoon representation of n replicate systems evolving (independent of each other) as driven by the transition matrix T. The color indicates how closely the distribution of the n replicas (over the states) matches X^*. The dotted box encloses the segment of the simulation occurring after this distribution has converged.

time their distribution over the states will closely match X^* (see figure 2.4). These results suggest a strategy for determining the weight of each state. That is, one can determine these weights by simulating many copies of the system in parallel and then counting (once stationary) how the copies are distributed over the states. While a possibility, this is not how weights are determined. Instead, it is sufficient to simulate a single system for an extended period of time and count how the states are distributed within the resulting markov chain. To show that this approach will sample states with the right frequency, it is necessary to count how often a state i appears in a single markov chain. To accomplish this, consider a simulation containing n independent copies of the system each driven by T (figure 2.4). Because each copy is driven by the same transition matrix, the individual chains (if sufficiently long) will be identically distributed. What's more, the number of times state i is visited by all the copies may be counted (beginning only after the systems have converged to the stationary distribution i.e. using parts of the markov chains contained in the

dotted box in figure 2.4) by

$$n_i = (m-5)\, n x_i^*. \tag{2.12}$$

Because all the chains are identically distributed, it is found that the number of times state i appears in a single chain is

$$n_i = \frac{1}{n}(m-5)\, n x_i^* = (m-5)\, x_i^*. \tag{2.13}$$

And, so a single chain itself will be distributed according to X^*. For this reason, one typically simulates for MCMC a single chain for an extended period of time while discarding for analysis the initial non-equilibrium part of the chain.

When performing MCMC simulations, one has to determine the transition probabilities $\left(t_{i/j}\right)$ of T. As shown by Hastings,[82] these terms may be split into two independent probabilities $p_{i/j}$ and $a_{i/j}$, where $p_{i/j}$ is the probability of proposing to move from i to j and $a_{i/j}$ is the probability of accepting such a move. In most cases, it is desirable to sample states of the system according to the Boltzmann distribution $\pi_i = \frac{1}{Z} exp\left(\frac{-E_i}{k_B T}\right)$, where Z is the partition function, k_B is the Boltzmann constant and T is the temperature. Substitution of π for x into the detailed balance equation allows one to ensure the resulting stationary distribution is Boltzmann

$$\pi_i p_{i/j} a_{i/j} = \pi_j p_{j/i} a_{j/i}. \tag{2.14}$$

It follows that, for MCMC, samples are proposed according to $p_{i/j}$ and accepted with probability

$$a_{i/j} = min\left(1, \frac{\pi_j p_{j/i}}{\pi_i p_{i/j}}\right). \tag{2.15}$$

It should be noted that equation 2.15 does not actually require probabilities be

known but instead the exponential factors since the partition function cancels. This makes MCMC very powerful, as the partition function is generally difficult to calculate.

In the earliest MCMC simulations,[83] samples were randomly generated so the change in position of a selected particle from its previous position would be small. While such a strategy makes the proposal of highly energetic states unlikely, it also means the updates will be correlated. This makes, for systems with a rugged energy landscape (like that of proteins), difficult the crossing of barriers, as these states are exponentially suppressed. Fortunately, more elaborate schemes may be used and one can mix and match how samples are generated so long as each move leaves the distribution invariant[84] (when this is true one can show that the resulting method is correct by following the same counting procedure that was discussed before).

A powerful sampling method taking this strategy is replica exchange[85] also called Parallel Tempering[84] (PT). With replica exchange, one manufactures an artificial system composed of N copies of the chemical construct under investigation. As the latter name suggests, these copies vary in temperature and are typically ordered from lowest to highest. Because the systems do not interact with each other, the probability of finding the multi-replica system in a given state, which is defined by the set of coordinates from each replica, is given as the product of their Boltzmann probabilities

$$\pi\left(MRS\right) = \prod_{i}^{N} \frac{1}{Z_i} exp\left(-\beta_i E_i\right), \tag{2.16}$$

where the subscript i is used to denote beta i and the current energy for that replica. Evolution of the N systems is driven by either MCMC moves that are accepted in accordance with 2.15 or by MD, the latter method being called replica exchange molecular dynamics[86,87] (REMD). For both methods, an additional update

step is performed periodically, where neighboring replicas attempt the exchange of configurations. Being that this move is symmetric, meaning $p_{i/j}$ and $p_{j/i}$ are equal, the move may be accepted with probability

$$a_{i/j} = min\left(1, exp\left(-\Delta\beta\Delta E\right)\right), \tag{2.17}$$

with $\Delta\beta = \beta_j - \beta_i$ and $\Delta E = E_j - E_i$. Because, for REMD, updates are made by MD, one must also account for the momentum of the system. Thus, velocities are uniformly rescaled following exchange moves such that

$$V_i^{'} = \sqrt{\frac{T_i}{T_j}} V_j \qquad and \qquad V_j^{'} = \sqrt{\frac{T_j}{T_i}} V_i. \tag{2.18}$$

Owing to each replica having a unique β value, the multi-replica system does not itself sample from the canonical ensemble but instead from a "generalized ensemble" in accordance to 2.16. However, the individual chains will be canonical for their respective temperature. What's more, the coupling of chains results in only weakly correlated states and faster convergence at low-temperature compared to MCMC. This is because simulations at high-temperature, both MCMC and MD, are able to more easily cross barriers than their low-temperature counterparts. Thus, sampling is improved for high-temperature replicas allowing for rapid conversion between local energy minima.

While replica exchange improves sampling in protein simulations compared to MD and MCMC, the observed gains in efficiency are typically below those predicted by theory. Specifically, the computational costs of REMD simulations are found to grow by a power law with exponents of order 4 and often a large pre-factor. For this reason, REMD simulations of large systems, like a protein in explicit solvent,

require the use of many replicas (on the order of 50 or more). Failure to provide enough replicas will result in an insufficient exchange rate, the so-called bottleneck problem,[88] or exploration of too narrow a temperature range. For this reason, REMD simulations remain practical only for protein systems of modest size. These problems are addressed further in chapter 5 by the introduction of replica-exchange-with-tunneling (RET).

As a last note, it is worth mentioning that the replica exchange method is not restricted to the exploration of temperature space but instead any property of the system under investigation may be altered. For example, the system Hamiltonian is commonly varied. As such, this type of replica exchange is called hamiltonian replica exchange[89] (HRE). HRE is a powerful method and is used extensively in the investigation of the dual funnel switching proteins GA98 and GB98 as well as RfaH-CTD. These simulations are discussed in detail in chapters 6 and 7. In the next section 2.3, a third type of replica exchange simulation is discussed called resolution exchange,[90,91] where replicas differ now in the degree of coarseness, ranging from a fine-grained resolution on one end of the spectrum to a coarse-grained representation on the other.

2.3 Coarse-Grained, Multi-Scale and Go-Type Models

Another way in which protein landscapes may be explored (that is faster and more efficient than running a high-resolution MD simulation) is to perform a coarse-grained (CG) simulation. It is noted that, while coarse-grained simulations may use either MD or MCMC for updating the system, this text will assume MD is used. With coarse-grained models, select degrees of freedom are removed from the system thus lowering the computational costs required for each update. Because the fastest

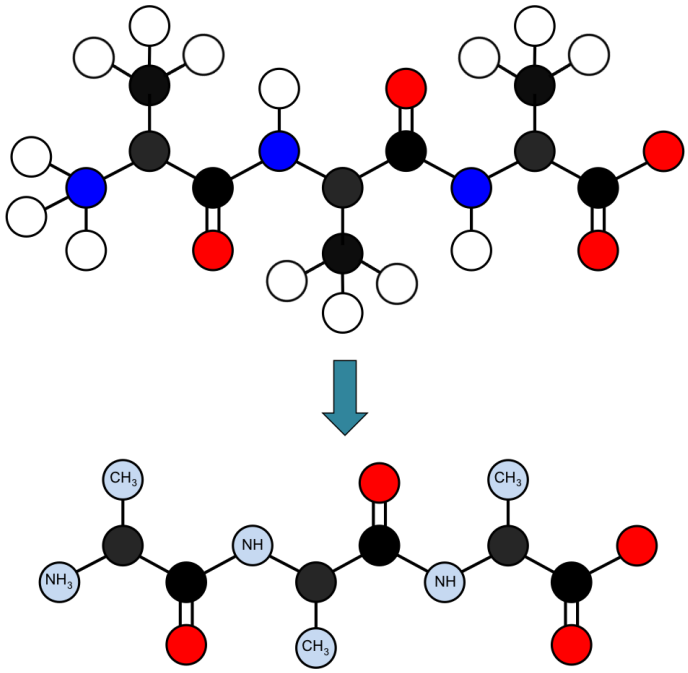

Figure 2.5: A typical example demonstrating how a coarse-grain model (bottom) could be constructed from a fine-grained (top) representation of poly-alanine. Atoms are color coded with red representing oxygen, black carbon, blue nitrogen and white for hydrogen. Pseudo-atoms are labeled according to the atoms grouped.

degrees of freedom are sometimes removed, a large integration step (compared to that used to simulate a fine-grained model) is not uncommon. For these reasons, coarse-grained simulations are able to reach beyond the nanosecond time scale while using fewer resources than their fine-grained counterparts.[92,93]

While details vary, coarse-grained models often take a form similar to that of 2.1.[94–96] Such a similarity between coarse-grained and high-resolution force fields is not surprising considering how these models are built. By starting with a fine-grained model, one can coarsen the system by lumping groups of atoms together to make a pseudo-atom (figure 2.5). The size of this group defines the resolution. Over the years, models have been developed that span this spectrum ranging from extremely coarse models on one end of the spectrum to those that only remove select

hydrogen atoms on the other.

Because removing degrees of freedom from the system lowers its entropy, these effects must be offset by a reduction in enthalpy.[75] Coarse-grained models that do this well will give a description of the energy landscape comparable to that of real proteins. However, building a coarse-grained model that matches exactly the accuracy of a fine-grained one is no easy task. As a result, coarse-grained simulations produce (at best) a qualitative description of protein landscapes.[75] For this reason, experts now work on multi-scale methods[90,91,97–100] with the hope of exploiting the computational efficiency and broad sampling observed in coarse-grained simulations[92,93] while retaining the accuracy of a fine-grained model. In these simulations, a coarse-grained model usually acts to propose states for the fine-grained model thereby improving sampling at high-resolution.

One method taking this approach is the resolution exchange[90,91] simulation. With resolution exchange, a multiple-replica system (similar to that of REMD[86,87]) is simulated. However, instead of varying the temperature of each replica, the resolution now differs. Similar to REMD, Resolution Exchange aims to reduce the convergence time of high-resolution simulations by introducing exchange moves between neighboring replicas. However, with Resolution Exchange, either a complete set of coordinates (compromising the low-resolution model) or only a subset of coordinates (within the high-resolution model) is exchanged. Furthermore, because low-resolution models do not contain explicitly the information needed to construct a fine-grained representation, algorithms that reintroduce these missing degrees of freedom must be used.[90,91,101–103] Unfortunately, the methods for doing this generate biased samples[102,103] or lead to the proposal of high-energy states likely to be rejected.[90,91,101] For this reason, resolution exchange simulations are susceptible to

exchange bottleneck problems[88] (similar to that observed in REMD simulations), a problem that can be alleviated by introducing the RET method developed in chapter 5.

In addition to resolution exchange, other multi-scale methods have been developed for protein simulations.[97–100] For example, the recently proposed multiscale essential sampling[104,105] (MSES) method attempts to bypass the problem of reintroducing lost degrees of freedom holding back resolution exchange. This is accomplished by simulating a compound system (coarse- and fine-grained model together) and using a restraining potential[105] to send information between the two models. The potential energy of such a system takes the form

$$E_{pot} = V_{FG}(q_{FG}) + V_{CG}(q_{CG}) + \lambda E_{\lambda}(q_{FG}, q_{CG}), \tag{2.19}$$

where $V_{FG}(q_{FG})$ and $V_{CG}(q_{CG})$ are the potential energy of the fine- and coarse-grained models respectively and $E_{\lambda}(q_{FG}, q_{CG})$ is the restraining potential, typically a function capable of enforcing structural similarity between the two models. The λ term, as introduced here, is a control parameter that may be varied and dictates the degree in which the two models are coupled. Introduction of HRE[89] now results in a random walk through λ space where the fine- and coarse-grained models track each other through configurational space. In this way, the fine-grained models are able to escape local traps resulting in a broader sampling of states. What's more, replicas where λ equals zero are not biased by the coarse-grained model and sample from the canonical ensemble. Finally, it should be mentioned that regardless the functional form used for E_{λ} (details on this matter are given in chapters 6, 7 and 8), large λ values are likely required to enforce a strict agreement between fine- and coarse-grained models. As a consequence, MSES simulations require the use of

many replicas, a problem that may be circumvented by the inclusion of the RET method introduced in chapter 5. This is demonstrated repeatedly in chapters 6 through 8.

As a last note, It should be known that the MSES[104,105] method does not require the use of a coarse-grained model. Instead, a fine-grained model in combination with an arbitrary type is possible. One possibility, well suited for studying transitions between competing attractors, is to use a structure-based model (SBM), also called a Go-model.[75,106] Construction of a Go-model requires knowledge of a protein configuration, typically the native fold, for determining force field parameters.[107] With details of the native fold at hand, a force field can be constructed that contains long-range attractive interactions only for pairs of atoms involved in native contacts. By construction, Go-models are able to fold into a predetermined configuration very quickly. However, the energetics of non-native folds are unphysical. What's more, the bias toward a predetermined fold may be removed by introduction of the MSES method. With MSES and RET, the quick folding properties of Go-models are exploited in chapters 6 and 7 to determine the transition pathways connecting the GA and GB folds of GA98 and GB98 as well as the helix hairpin and β-barrel folds of RfaH-CTD. The method is also used to determine the importance of select residues in promoting fiber formation in the 13-residue fragment of serum amyloid A.

2.4 Summary

When studying large bimolecular systems like a protein in explicit solvent, one typically approximates the dynamics of the molecules by switching from a quantum treatment to a classical one. Making this switch greatly reduces the resources needed to model the system but allows for the collection of a finite trajectory, often

shorter in time than that in which the phenomena of interest occurs. To enable investigation of processes that occur on a large time scale, such as protein folding and aggregation, more elaborate models and simulation protocols have been developed. The markov chain monte carlo (MCMC) method enables sampling of energy landscapes while maintaining strict Boltzmann weights for samples. However, samples generated by MCMC are highly correlated. As a result, the method is unable to sample sufficiently for molecular systems with a rough energy landscape like proteins. To overcome this problem, general ensemble methods like replica exchange molecular dynamics have been developed and allow for a much broader exploration of protein landscapes. Because exchange rates decrease with the system size, REMD simulations are practical only for moderately sized protein systems. Still, this problem may be alleviated by the introduction of the replica-exchange-with-tunneling protocol developed in chapter 5 of this text.

In a different approach, exploration of protein landscapes is improved by reducing the system degrees of freedom. As a consequence, larger trajectories compared to fine-grained simulations may be acquired but the results are less accurate. To offset this problem, multi-scale methods have been developed like resolution exchange and multiscale essential sampling. With MSES, the problem of rebuilding missing degrees of freedom common to resolution exchange is circumvented by the introduction of a restraining potential E_λ which acts to pass information between coarse- and fine-grained models in real time. Because the MSES method relies on hamiltonian replica exchange, the method is susceptible to exchange bottleneck problems similar to that in REMD simulations making its use impractical for most protein systems. It is the major focus of chapters 5 through 8 to develop protocols such as RET and the resolution-exchange-with-tunneling (ResET) method that enable sufficient

exploration of protein landscapes so that real protein systems may be investigated using modern computing clusters.

Chapter 3: Research Overview

3.1 Summary of the First Two Chapters

In chapter 1, the protein folding problem was introduced and the challenges faced by protein scientists discussed. Of importance but also difficult to determine, is the free energy landscape of a given protein system. This landscape holds information regarding which states of the protein are populated at equilibrium but also kinetic information about folding. Because many topological details of protein landscapes, such as transition states and intermediate folds, are difficult to determine by experiment, computational methods have been developed to give their theoretical description.

In chapter 2 several such methods were described. The simplest of these was molecular dynamics (MD) simulations. While incredibly powerful, MD simulations are not currently able to predict the native fold for most proteins. This is because folding and other important processes occur on a timescale inaccessible to high-resolution MD simulations using present-day computer resources. For this reason, enhanced sampling methods like replica exchange molecular dynamics (REMD) have been developed. While REMD can sample from the canonical ensemble, the method is currently restricted to moderate sized chemical environments leaving many important protein systems inaccessible.

Also discussed in chapter 2, were coarse-grained models. Despite the broad sampling observed in coarse-grained simulations, the resulting landscapes are often inaccurate resulting in a smooth topology and low barriers compared to real proteins. Improving upon these simulations, are multi-scale methods which exploit a coarse-grained model for sampling but use a more detailed description for determining

weights. With multiscale essential sampling, the problem of reintroducing missing degrees of freedom associated with other methods like resolution exchange is avoided by introducing a restraining potential E_λ. Acting through this potential, information is exchanged between a coarse- and fine-grained model thus enhancing sampling for the high-resolution model. Because MSES relies on hamiltonian replica exchange for removing bias introduced by addition of E_λ, the method is susceptible to exchange bottlenecks similar to REMD simulations.

3.2 Targeting Specific Problems

While each of the computational methods discussed in chapter 2 have contributed to the understanding of proteins and their dynamics, use of these methods with large protein systems remains a challenge. It is thus the purpose of the research presented in this text to build upon the ideas introduced in chapter 2. Specifically, the exchange bottleneck problem of REMD is addressed in chapter 5 by the introduction of the replica-exchange-with-tunneling (RET) method. With RET many of the problems identified in chapter 2 are overcome. For example, the RET method is not only applicable to REMD but may also be used in hamiltonian replica exchange simulations, such as the MSES method, as well as resolution exchange. Thus, RET has the potential to alleviate many of the problems holding back these methods. Combined with RET, a variant of the MSES method we call "Go-model feeding" is introduced in chapter 6 and the free energy landscapes of several proteins, including GA98, GB98 and the 13-residue fragment of serum amyloid A, are constructed. The validity of the "Go-model feeding" method is also asserted in this chapter. Continuing in this direction, the "Go-model feeding" method is used again in chapter 7 to determine the transition pathway connecting the helix-hairpin and β-barrel folds

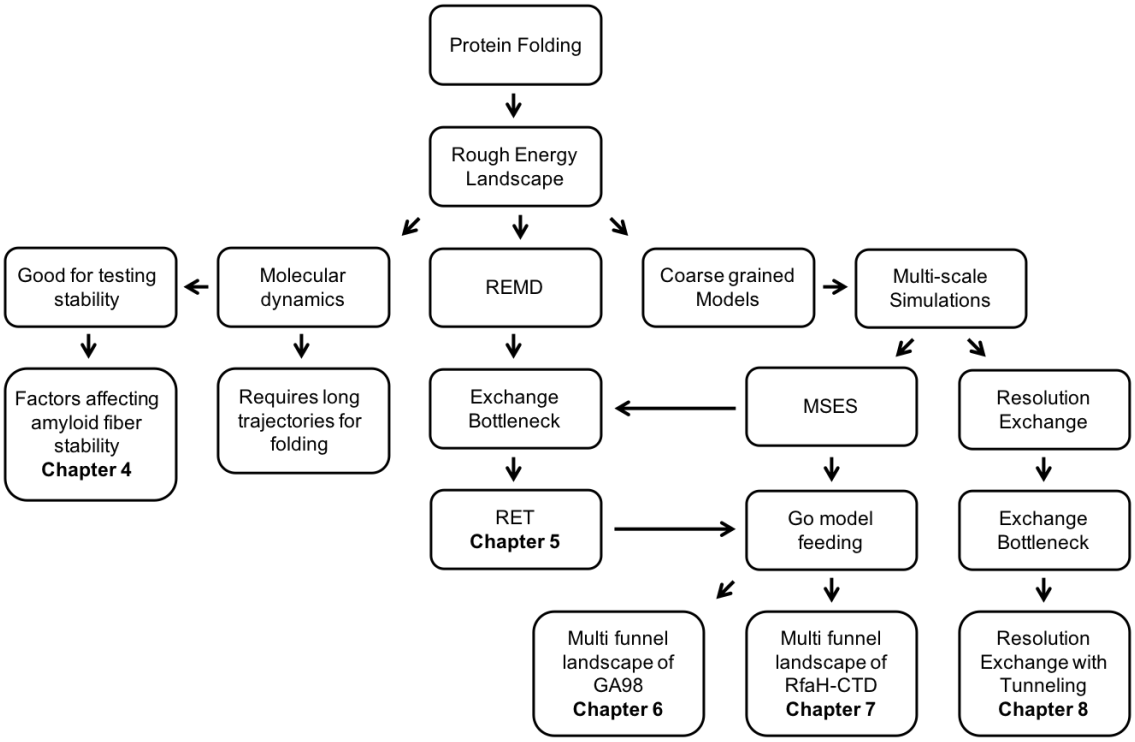

Figure 3.1: Flow chart showing current methods used in protein simulations and the challenges faced when using these methods. Connected to these are the methods introduced in later chapters and the protein systems studied with them.

of the isolated RfaH C-terminal domain. Because use of Go-models in MSES simulations requires knowledge of a protein structure the "Go-model feeding" method is limited to proteins with known structure. Thus, in chapter 8 similar methods are described that use in place of Go-models a coarse-grained representation. These simulations enable fast and efficient exploration of protein landscapes for any protein regardless of prior knowledge of its native fold. Also introduced in chapter 8 is a single step resolution exchange simulation called resolution-exchange-with-tunneling (ResET). With ResET, a multi-scale simulation can be performed using only two replicas making these simulations less expensive than other methods. This research strategy is summarized in figure 3.1.

3.3 About the Next Chapters

Before continuing it should be noted that the following chapters (4 through 7) are taken from published work. The work presented in chapter 8 is currently unpublished but a manuscript is in preparation. The details regarding chapters taken from published work are as follows:

- Chapter 4 was published in The Journal of Physical Chemistry B as the article: Nathan A Bernhardt, Workalemahu M. Berhanu, and Ulrich H. E. Hansmann. Mutations and seeding of amylin fibril-like oligomers. The Journal of Physical Chemistry B, 117(50):16076–16085, 2013.

- Chapter 5 was published in The Journal of Chemical Physics as article: Fatih Yasar, Nathan A. Bernhardt, and Ulrich H. E. Hansmann. Replica-exchange-with-tunneling for fast exploration of protein landscapes. J. Chem. Phys., 143(22):224102, Dec 2015.

- Chapter 6 was published in The Journal of Chemical Theory and Computation as the article: Nathan A. Bernhardt, Wenhui Xi, Wei Wang, and Ulrich H. E. Hansmann. Simulating protein fold switching by replica exchange with tunneling. J. Chem. Theory. Comput., 12(11):5656–66, 2016.

- Chapter 7 was published in The Journal of Physical Chemistry B as the article: N. A. Bernhardt and U. H. E. Hansmann. Multifunnel landscape of the fold-switching protein rfah-ctd. J Phys Chem B, 122:1600–1607, 2018.

In the next chapter (4), the power of MD simulations is demonstrated by simulation of the amyloid cross-seeding phenomena observed in amylin. In this chapter

the factors determining compatibility between amyloid seeds and the addition of further monomeric protein are determined. Then in the remaining chapters, the enhanced sampling techniques discussed previously are introduced.

Chapter 4: Mutations and Seeding of Amylin Fibril-Like Oligomers

The following chapter was published in The Journal of Physical Chemistry B by the author of this dissertation as the following article: Nathan A Bernhardt, Workalemahu M. Berhanu, and Ulrich H. E. Hansmann. Mutations and seeding of amylin fibril-like oligomers. The Journal of Physical Chemistry B, 117(50):16076–16085, 2013. All text and figures are taken with the permission of the publisher.

Author Contributions: Dr. Workalemahu Berhanu is credited for his contribution of double layer simulations and free energy calculations to this chapter. All single layer simulations were contributed by the author of this dissertation.

4.1 Introduction

Amyloid aggregates are implicated in at least 30 distinct diseases.[34,35] These aggregates result from failure of a specific peptide or protein to maintain its native (functional) conformation.[33,35] Instead, they form amyloid filaments characterized by β-strands that are oriented perpendicularly to the fibril axis. The strands are connected through a dense hydrogen-bonding network and side chain interactions between strands that drive their lateral association[37] to supramolecular β-sheets. The rate-limiting step in the growth process of these aggregates is the formation of an initial nucleus,[40] and in vitro, the seeding of the protein solution with preformed fibrils leads to dramatically faster fibril growth.[38] Human amyloids can be biochemically mixed, implying the possibility that one amyloidic peptide can cross-seed another one in vivo.[41] Such cross-seeding may account for the observed

correlations between amyloid diseases. Hence, detailed knowledge of the molecular mechanism of fibril seeding would provide a platform for the rational design and therapeutic intervention of disease states associated with mature amyloid fibrils and their precursors.

However, an accurate description of the aggregation process and seeding of fibrils is still missing, as these processes are difficult to explore in experiments.[108] For this reason, we rely in the present study on an alternative approach. Using molecular dynamics simulations, we study the molecular mechanisms of cross-seeding by probing the effects of mutations on the seeding of fibrils. Our underlying assumption is that crossseeding depends on the similarity of the fibrils resulting from the different components. Hence, studying "cross-seeding" between wild type and mutants having similar fibrillar structures allows comparison of the various interactions that enable cross seeding. Key candidates are the stacking of aromatic residues along the outside of the fiber, hydrophobicity of key residues, and structural similarity between adjacent polypeptide strands.[109–111]

Our test system is amylin, a 37 residue hormone produced in the pancreas that is highly amyloidogenic and associated with type-2 diabetes mellitus through membrane permeabilization induced β-cell loss.[39,112] A fibril model of the full-length human amylin has been extrapolated from X-ray diffraction data of cross-β spine structures of two segments of human amylin (NNFGAIL and SSTNVG).[36] The topology of this model is similar to that reported by Luca et al.[36] and Bedrood et al.,[113] and exhibits a β-strand-loop-β-strand motif, consisting of an N terminal β-stand residue 8-19, with the loop region located at residues 20-23 and C terminal β-stand comprising residues 24-36.[36] Experimental and computational studies have shown that the human amylin proto-fibril pair has molecular polymorphism[36,43,113–118] in which

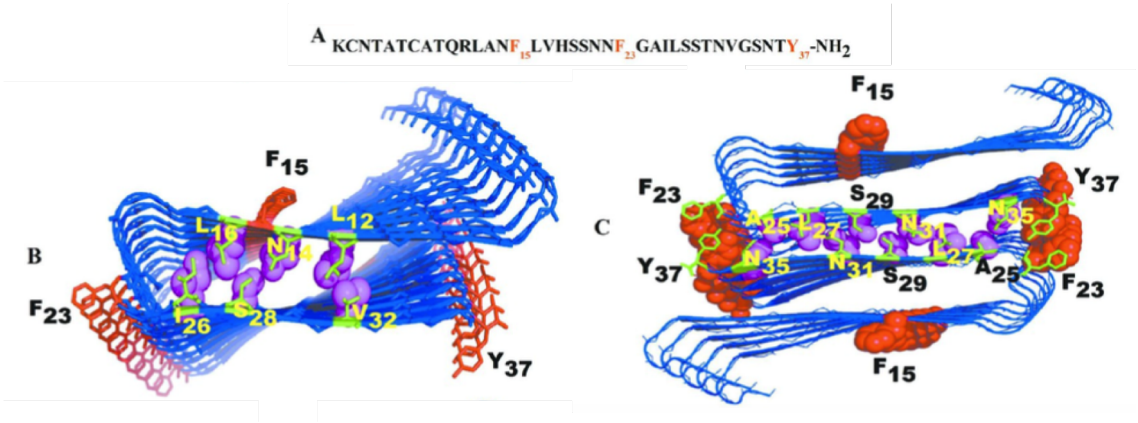

Figure 4.1: Structure of human amylin fibril model.[110] (A) Sequence of human amylin, the aromatic amino acids (F15, F23, and Y37) are colored in red. (B) Single layer amylin decamer. Intramolecular face-to-face side chain interaction that determines the stability of the U shape in wild type, mutants, and seeded oligomers are emphasized by representing the corresponding side chains as balls and sticks. (C) Double layered amylin fibril model.

the two layer models could be packed in either antiparallel or parallel fashion. Depending on sample preparation, Middleton et al.[115] found parallel and antiparallel mixed polymorphism, while Nielson et al.[114] found only antiparallel fibrils. Other simulations have shown the stability of the packing of proto-fibril pairs of amylin where one layer is shifted against the other by two residues.[43,118] This suggests that two-layer models such as in Figure 4.1C, packed in an antiparallel fashion, are good structural descriptions[43,118] of such oligomers. Molecular dynamics simulations also indicate that X-ray models[42,117] with more closely interdigitated (interlocked β-strands that tighten the binding of two β-sheets) side chains are more compact and stable than the NMR models.[117] Amylin contains three aromatic residues: F15 is the only one that resides in the β-sheet core, while F23 is located in a bend, and Y37 is exposed at the C-terminus. The fibril model of amylin has an in-register alignment of matching residues, generating a tight packing that maximizes favorable

hydrophobic and van der Waals side chain contacts along the long axis of the fibril, and may be stabilized in addition through π-π stacking of the F15, F23, and Y37 aromatic rings in the β-strand, loop, and C terminal regions (Figure 4.1).

We will use molecular dynamics simulations to study the role of interactions involving the three aromatic residues F15, F23, and Y37 by comparing the wild type with mutants where a single aromatic residue is replaced by leucine (F15L, F23L, and Y37L). Tu et al.[119] have already shown that fibril seeds from such mutants can still seed amyloid growth of wild-type amylin, suggesting that similarity in fiber structures is a key requirement for efficient mixed growth. Amyloid formation of the F15L mutant is almost twice as rapid than for the wild-type, while it is almost three times slower for the Y37L mutant, and two times slower for F23L mutant.[119] Molecular dynamics simulations of stability and conformational changes of such amyloid assemblies will enable us to measure directly changes in structural stability induced by specific alterations of the amino acid sequence of amyloid polypeptides. In vivo conditions are mimicked and information that is currently unobtainable through experiment, such as real time structural data, will be obtained, making visual inspection possible. Especially, we aim to answer the following questions:

1. Is similarity in fibril structure the key requirement for effective cross seeding of single mutant seeds with wild type amylin?

2. Can differences in lag times observed in the self-assembly of single mutants amylin fibers be explained by the relative stability of the mutant structures during the simulation?

3. Does the loss of π-π interactions affect the stability of amylin mutant fibers? What is the effect of such mutation on the face-to-face hydrophobic contacts

in the interior core?

Answering these questions may point the way to in silico development of molecules that, by binding to the amylin fibril, shift the equilibrium of amylin toxic oligomer to the nontoxic fibers.[120] Finally, we propose new mutation experiments that investigate the role of pairs of amino acids involved in face-to-face interactions between the β sheet regions.

4.2 Materials and Methods

For the wild type, we base the start configuration on the fibril models of amylin by the Eisenberg group,[36] and build out of these a single layer decamer (Figure 4.1B) and double layer decamer models of amylin consist of two pentamer layers (Figure 4.1C). This choice is motivated by recent work by Kahler et al.[121] which showed that decamers of Aβ retained the fibrillar states and gain in stability with oligomer growth[122] and are more stable than pentamers.[121] Since amylin has similar size and the same β-strand-loop-β-strand motif as Aβ, we assume that the same observations apply for amylin. We denote the so designed wild-type amylin decamer as WT. Mutant decamers are designed by replacing the aromatic residues, F15, F23, and Y37 with leucine, and denote them as F15L, F23L, and Y37L, respectively (Table 4.1). These mutants are obtained from wild type coordinates by replacing the side chains of the targeted residues while retaining the original backbone conformations of the wild-type.[123] The structure of the designed mutants is minimized for 5000 steps using the steepest decent algorithm with the backbone of the protein restrained. The hetero-assemblies of wild type and mutants are a 1:1 mixture of wild-type (the first five strands) and mutants (the last five strands, consisting of F15L, F23L, or

Y37L in the above notation), and are denoted as WT-F15L, WT-F23L, and WT-Y37L, respectively.

Our molecular dynamics simulations rely on a combination of the AMBER ff99SB force field[124] with explicit water (TIP3P),[125,126] a common choice for exploring amyloid peptide aggregation,[127,128] as implemented in the GROMACS program version 4.5.5.[129] Hydrogen atoms are added with the pdb2gmx module of the GROMACS suite. The start configurations for all proteins are put in the center of a cubic box, with at least 12 Å between the solute and the edge of the box. Periodic boundary conditions are employed, and electrostatic interactions are calculated with the PME algorithm.[130,131] We use a time step of 2 fs. Hydrogen atoms are constrained with the LINCS[78] algorithm while for water the Settle algorithm is used.[79] The temperature of 310 K is kept constant by the Parrinello-Donadio-Bussi algorithm[132] (τ = 0.1 fs) which is similar to Berendsen coupling but adds a stochastic term that ensures a proper canonical ensemble.[132,133] In a similar way, the pressure is kept constant at 1 bar by the Parrinello-Rahman algorithm[134] (τ = 1 fs). After energy-minimizing first the solvated start configuration using the steepest descent method, followed by conjugate gradient, the system is equilibrated in two steps of 500 ps, first in an NVT ensemble and second in an NPT ensemble at 1 bar. After equilibrization, 200 ns of trajectories are analyzed for each system to monitor how the oligomer structures evolve with time. Data are saved at 4.0 ps intervals for further analysis. For each system (Table 4.1), we run three distinct simulations of 200 ns with different initial velocity distributions. This allows us to test whether we reached equilibrium and guarantees three independent sets of measurements.

The molecular dynamics trajectories are analyzed with the tool set of the GROMACS package. Especially, we monitor conformational changes and stability of

System	# atoms of peptide/ # atoms water/Cl^-	simulation box dimensions (x, y, z [Å])	total simulation time, ns
WT (SL)	5350/36665/20	106.8, 106.8, 106.8	$600ns(200x3)$
F15L (SL)	340/36675/20	106.8, 106.8, 106.8	$600ns(200x3)$
F23L (SL)	5340/36396/20	106.6, 106.6, 106.6	$600ns(200x3)$
Y37L (SL)	5330/36695/20	106.8, 106.8, 106.8	$600ns(200x3)$
WT-F15L* (SL)	5345/36667/20	107.0, 107.0, 107.0	$600ns(200x3)$
WT-F23L* (SL)	5345/36677/20	106.9, 106.9, 106.9	$600ns(200x3)$
WT-Y37L* (SL)	5340/36681/20	106.9, 106.9, 106.9	$600ns(200x3)$
WT (DL)	5350/39644/20	108.1, 108.1, 108.1	$600ns(200x3)$
F23L (DL)	5340/39666/20	108.0, 108.0, 108.0	$600ns(200x3)$
Y37L (DL)	5330/39660/20	108.0, 108.0, 108.0	$600ns(200x3)$

Table 4.1: The simulations are performed at 310 K. The symbol * marks the cross-seeded aggregates where the first five strands are from the wild type amylin and the last five strands are from one of the mutants (i.e., F15L, F23L, and Y37L, respectively). SL marks single layer decamers and DL double layer decamers, where each layer consists of five strands.

the oligomer models through the time evolution of root mean square deviations of the C_α atoms (RMSD), root-mean-square fluctuation (RMSF), hydrophobic contacts distances, and hydrogen bonds, measured with the g_{hbond} and g_{dist} modules in GROMACS. Hydrogen bonds are defined by a distance cut off between donor and acceptor of 0.36 nm and an angle cut off of 30^o. Configurations are visualized using PyMOL.[135]

4.3 Results and Discussion

The purpose of our simulations is to examine the stability of oligomers of amylin wild-type and single aromatic mutants in comparison with mixed decamers. The initial conformations and final structures obtained from the molecular dynamics simulations of the single layer aggregates[111] are shown in Figure 4.2. We find by visual inspection that the β-strand-loop-β-strand topology of the decamers, the

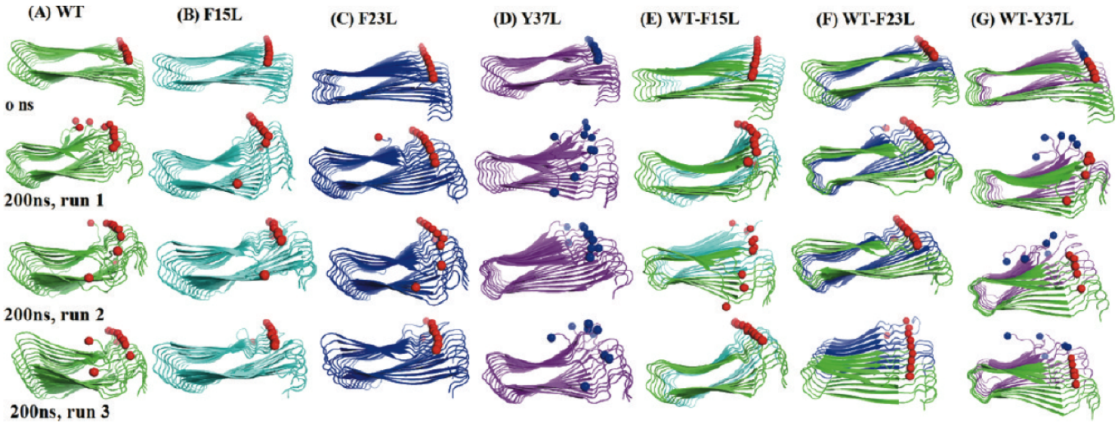

Figure 4.2: Snapshots of the amylin decamers of wild type, mutants, and their seeded assembly before and after 200 ns (water molecules are omitted for clarity; the snapshots are from three independent simulations). (A) Wild type (WT, red ball: Y37); (B) F15L (red ball: L37); (C) F23L (red ball: Y37); (D) Y37L (red ball: L37); (E) WT-F15L (red ball: Y37); (E) WT-F23L (red ball: L37); and (G) WT-Y37L (red ball: Y37 and blue ball: L37). The balls mark the residues 37 in the C-termini of each strand of the peptides..

main structural feature of amyloidogenic fibrillar state, is maintained throughout the simulation (Figure 4.2). The structures from wild-type oligomer mixed with the mutants (the heteropolymers WT-F15L, WT-F23L, and WT-Y37L) exhibit the same shape and topology as the wild-type oligomer (WT) and mutants oligomer, pointing to structural similarity as major factor for the seeding observed in previous experiments.[119]

In order to obtain a more detailed picture of the stability of the various amylin decamers, we show in Figure 4.3A the C_α-RMSD as calculated from the molecular dynamic trajectory. The reference structure is the equilibrated start structure (i.e., the configuration at 0 ns). The faster this quantity grows throughout the simulation, the less stable and the more dynamic a decamer is. The average RMSD values, calculated over three independent trajectories of 200 ns for each system, are within

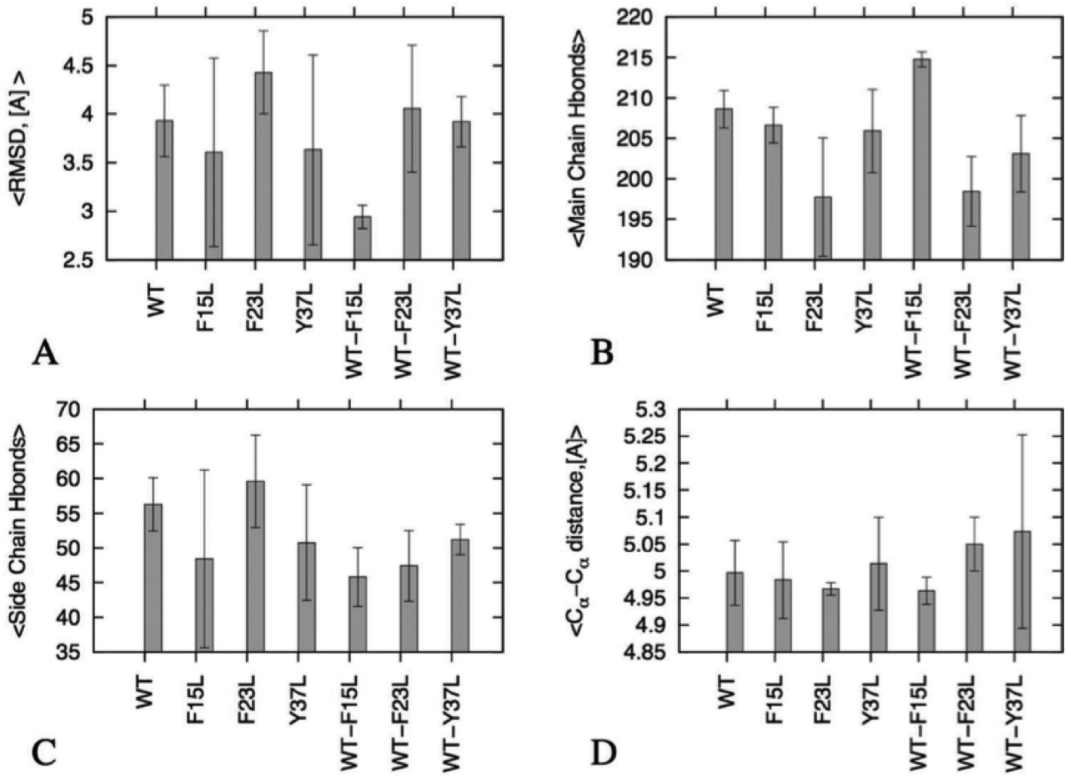

Figure 4.3: RMSD, $C_\alpha - C_\alpha$ distances along strand 5 and 6 and hydrogen bonds of the studied oligomer systems. Average $\langle RMSD \rangle$ in Å (A); Average number of main chain hydrogen bonds (B); side chain hydrogen bonds (C); and Average $\langle C_\alpha - C_\alpha distances \rangle$, in Å, between strands 5 and 6 (D).

the range of 3.0 to 4.5 Å. These values are smaller than the RMSD value of more than 5.5 Å measured for the single layer amylin pentamer,[136] and close to the 4.5 Å observed for double layer amylin decamer.[117] Thus, the simulated oligomers appear to be stable, with no discernible difference in stability of the examined structures (Figure 4.3A).

We then examined the average interstrand distance between strand 5 and strand 6 across the U-shaped β-strand-loop-β-stand motif for the wild type, mutants, and mixed oligomers (Figure 4.3D). This distance has been selected because in the mixed assemblies the two strands are located along the interface between wild type and

mutant chain. This distance quantifies how close the two strands and therefore how strong and favorable the wild type and mutants are. On the other hand, an increase in the distance over that in the initial structures indicates detachment of the heteroassembly system and unfavorable contacts between the peptides. The measured distances for all three cases, wild type, mutants, and mixed oligomers, are around $\sim$5 Å range, indicating that the interstrand distance is close to the experimental interstrand distance of 4.8 Å for amyloid fibrils and oligomers.[36,137] Hence, in all three forms of oligomers the peptides interact strongly at this interface.

Protein aggregates are stabilized through a network of side chain hydrophobic and hydrogen bonding interactions.[37,138] Main chain (interbackbone) hydrogen bonds link the β-strands within an amyloid fibril, while the side chain hydrogen bonds modulate the intermolecular packing arrangement within and between the β-sheets of the fibril core.[139] For this reason, we have counted main chain and side chain hydrogen bonds during the simulations, and averaged the values over the three trajectories (Figure 4.3B and 4.3C). Again, we find little differences in the average side chain and main chain hydrogen bond networks between wild type, mutants, and mixed oligomers. On average, all decamer structures are stabilized by about 200 main chain and 50 side chain hydrogen bonds (Figure 4.3B and C).

An analysis of the root-mean-square fluctuation (RMSF), computed for the wild type, mutants, and mixed decamers, shows that in all cases the residues in the turn region are more flexible than those in the β-strand regions. The exceptions are residues near the N-/C-termini that are exposed to the solvent (Figure 4.4). This is similar to previous amylin simulations.[117] The RMSF describes where in the sequence of a protein structural stability is gained or lost, and therefore allows one to relate this change in stability to specific mutations. The smallest per residue fluc-

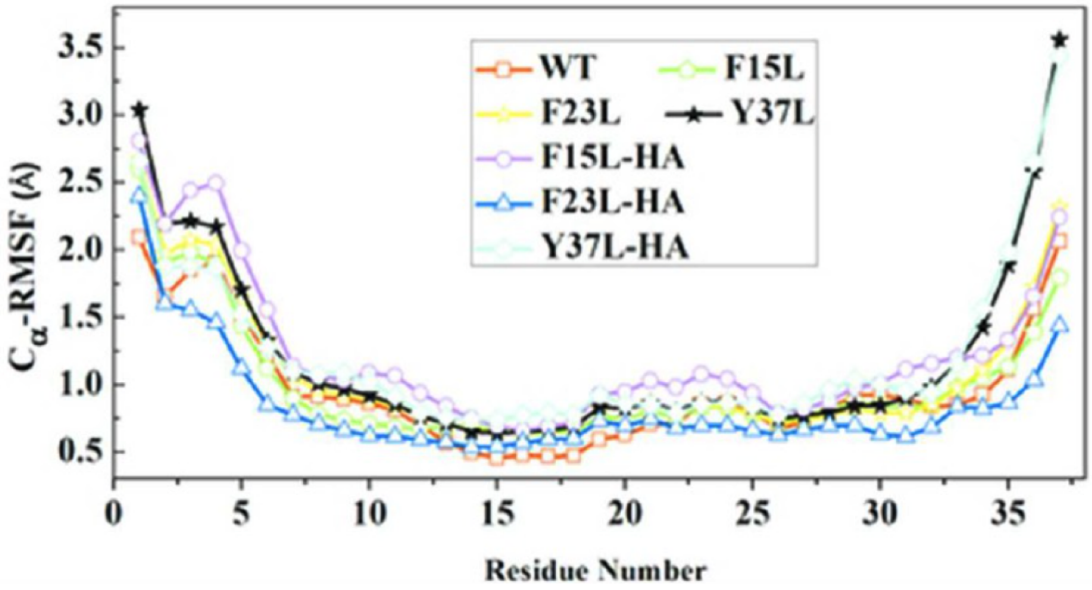

Figure 4.4: Average RMSF values for the 10 β-strands in single layer amylin, its mutants and cross-seeded aggregates. In the mixed aggregates the first five strands are from the wild type, while the last five strands are from mutants. The results are calculated from three independent trajectories of 200 ns.

tuation is observed in β_1 region that includes residues $L_{12}ANFLV_{17}$ and residues $F_{23}GAIL_{27}$ of β_2 region for all the systems studied indicating these regions are crucial for the stability of the oligomers. Tyrosine to Leucine substitution at position 37 causes decreased stability at the C terminal (Figure 4.4), which may explain the experimentally observed longer lag time in the growth of this mutant. This can be seen by both visual inspection and comparison of the RMSF values of each structure (Figures 4.4 and 4.2). This increased flexibility of the C terminal region may hinder the formation of contact between adjacent Y37 and between Y37 and F23 at the interface of the double layer oligomer during fibril elongation, therefore contributing to the increased lag phase observed in experiments. The F15 is not involved in the sheet-to-sheet contact while F23 and Y37 are involved. The distance between

	WT oligomer			F15L oligomer			F23L oligomer		
Y_{37}/Y_{37} distance (Å)	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3
Chains-2_3	11.14(4.0)	7.14(2.8)	5.16(0.3)	5.18(0.3)	6.22(1.5)	5.28(0.5)	5.25(0.4)	5.44(0.7)	10.68(4.9)
Chains-3_4	6.02(2.5)	5.13(0.3)	5.06(0.2)	5.05(0.2)	5.14(0.3)	5.03(0.3)	5.14(0.3)	5.14(0.3)	5.14(0.3)
Chains-4_5	6.92(1.8)	5.04(0.2)	5.02(0.2)	5.00(0.2)	4.99(0.2)	5.00(0.2)	5.00(0.2)	5.02(0.2)	5.02(0.3)
Chains-5_6	5.14(0.4)	5.02(0.3)	5.00(0.3)	4.95(0.2)	4.95(0.2)	4.99(0.2)	4.95(0.2)	4.97(0.2)	4.99(0.2)
Chains-6_7	5.06(0.2)	5.05(0.3)	5.40(0.8)	4.99(0.2)	4.97(0.2)	4.92(0.2)	4.98(0.2)	4.99(0.2)	5.02(0.2)
Chains-7_8	5.01(0.1)	5.19(0.4)	8.70(4.0)	5.06(0.2)	5.03(0.2)	4.93(0.2)	4.98(0.2)	5.55(1.4)	5.05(0.2)
Chains-8_9	5.05(0.3)	6.45(2.7)	8.74(6.4)	5.76(0.8)	5.17(0.4)	5.00(0.3)	4.99(0.2)	9.93(4.2)	5.06(0.2
Mean ± SD	**6.33 ± 2.2**	**5.57 ± 0.8**	**6.15 ± 1.8**	**5.14 ± 0.3**	**5.21 ± 0.4**	**5.02 ± 0.1**	**5.04 ± 0.1**	**5.86 ± 1.8**	**5.85 ± 2**

	Y37L oligomer				WT-F15L hetero assembly oligomer		
L_{37}/L_{37} distance (Å)	Run 1	Run 2	Run 3	Res_{37}/Res_{37}, distance (Å)	Run 1	Run 2	Run 3
Chains-2_3	10.35(1.1)	13.01(2.4)	10.67(3.2)	Chains-2_3 (WT)	5.22(0.5)	16.21(3.4)	5.10(0.2)
Chains-3_4	6.08(0.5)	7.73(0.8)	6.54(1.1)	Chains-3_4 (WT)	5.09(0.3)	5.45(1.1)	4.99(0.2)
Chains-4_5	7.25(0.7)	6.58(1.0)	7.79(1.6)	Chains-4_5 (WT)	5.06(0.3)	5.32(0.5)	4.92(0.2)
Chains-5_6	6.51(0.9)	5.58(0.3)	8.98(3.2)	Chains-5_6 (interface)	5.47(0.3)	5.58(0.5)	4.88(0.2)
Chains-6_7	13.20(6.5)	7.50(0.8)	10.77(3.7)	Chains-6_7 (F15L)	6.19(1.2)	7.64(1.5)	4.88(0.2)
Chains-7_8	17.89(6.5)	5.24(0.3)	12.17(6.2)	Chains-7_8 (F15L)	6.30(1.7)	11.62(3.0)	4.98(0.2)
Chains-8_9	9.86(2.9)	5.61(0.4)	14.09(5.4)	Chains-8_9 (F15L)	5.35(0.4)	10.66(2.0)	5.09(0.2)
Mean ± SD	**10.16 ± 4.2**	**7.3 ± 2.7**	**10.1 ± 2.6**	**Mean ± SD**	**5.52 ± 0.5**	**8.92 ± 4.1**	**4.98 ± 0.1**

	WT-F23L hetero assembly oligomer				WT-Y37L hetero assembly oligomer		
Res_{37}/Res_{37} distance (Å)	Run 1	Run 2	Run 3	Res_{37}/Res_{37}, distance (Å)	Run 1	Run 2	Run 3
Chains-2_3	5.16(0.3)	5.36(0.6)	5.11(0.3)	Chains-2_3 (WT)	7.37(1.5)	16.48(0.2)	16.89(4.4)
Chains-3_4	5.02(0.2)	5.04(0.2)	5.00(0.2)	Chains-3_4 (WT)	9.44(1.5)	9.15(0.2)	11.62(4.0)
Chains-4_5	4.93(0.2)	4.94(0.2)	4.97(0.2)	Chains-4_5 (WT)	6.74(1.4)	9.92(0.4)	6.39(0.7)
Chains-5_6	4.94(0.2)	4.90(0.2)	4.91(0.2)	Chains-5_6 (interface)	7.92(0.2)	9.51(0.2)	7.57(0.9)
Chains-6_7	5.00(0.2)	4.98(0.2)	4.97(0.2)	Chains-6_7 (Y37L)	5.40(0.5)	5.34(0.4)	5.19(0.4)
Chains-7_8	.5.05(0.3)	5.03(0.2)	5.06(0.2)	Chains-7_8 (Y37L)	5.55(0.5)	5.19(0.3)	5.02(0.2)
Chains-8_9	5.42(1.0)	4.97(0.3)	5.04(0.3)	Chains-8_9 (Y37L)	7.76(1.6)	5.17(0.4)	5.03(0.2)
Mean ± SD	**5.07 ± 0.2**	**5.03 ± 0.2**	**65.01 ± 0.1**	**Mean ± SD**	**7.17 ± 1.4**	**8.68 ± 4.1**	**8.24 ± 4.5**

Table 4.2: The distances are measured between residue 37 of each strand and the adjacent successive one, excluding the less stable terminal chains 1 and 10.

the centers of mass of residues 37 of the C-terminal of the three different kinds of decamers change little during the 200 ns simulation from the initial distance (see Table 4.2). However, the distance between centers of mass of residues 37 of the C-terminal Y37L mutant of amylin become within the range of 7-10 Å, therefore much larger than the value of about 4.8 Å of the initial conformation (i.e., at 0 ns of the simulation). Hence, mutation of Y37 to L results in the largest fluctuation in the C terminal of each monomer. This higher flexibility could slow down the elongation of the Y37L mutant and therefore contribute to the larger lag phase in its aggregation observed in a previous experimental study.[119] The simulation of decameric double layer wild-type amylin and its F23L and Y37L mutants (F23 and Y37 participate in sheet-to-sheet contacts between the β-sheets at the interface of the two U-shaped oligomers, Figure 4.1C) shows no discernible difference in stability (Figure 4.5). The lack of susceptibility to destabilization of the initially preformed structure of

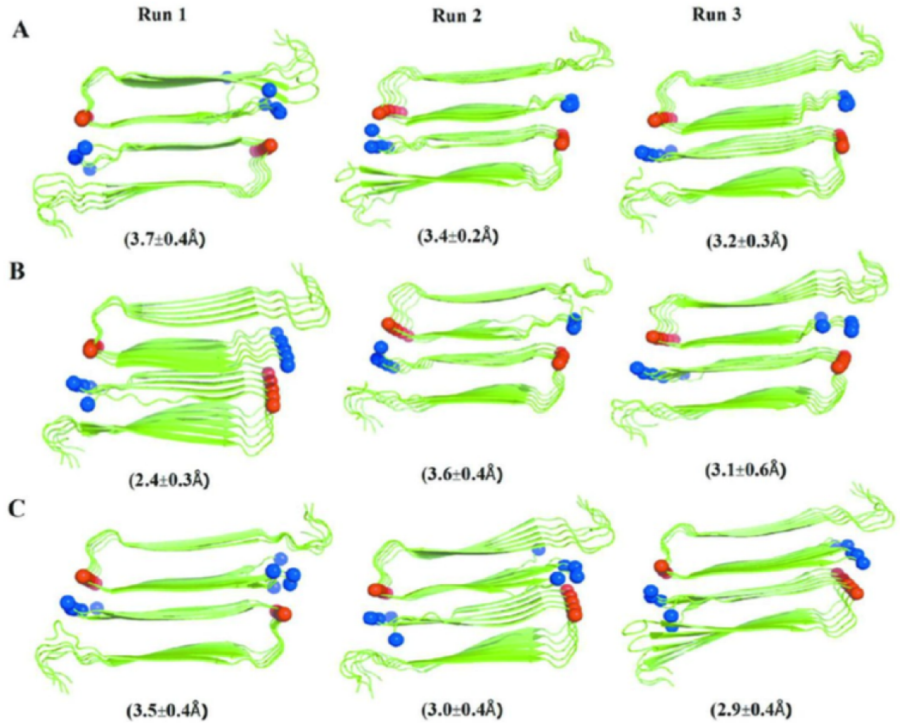

Figure 4.5: C_α root-mean-square deviations (RMSDs) with respect to the corresponding minimized start configurations and the average structures of (A) WT, (B) F23L, and (C) Y37L. The average structures are calculated from the position of all heavy atoms of the protein over the 200 ns of each trajectory using the program g_covar of the Gromacs 4.5.5 package.

the decamer double layer oligomers of both wild type and mutants suggests that the aromatic π-stacking interactions are not critical for aggregation stability as Leu is similar in size and hydrophobicity to F and Y but not capable of π-stacking. Note that our result does not exclude a role of the aromatic side chain of F23 and Y37 for favoring aggregation nucleation, an effect which could contribute also to the experimentally observed differences in aggregation kinetics.[119] However, a combination of much longer simulation times and enhanced conformational techniques would be required to determine the equilibrium structures[140] of the wild type and mutant monomers and dimers, and to get an insight into the influence of π-stacking during

		models of amylin, mutants, and heteroassembly decameric oligomers						
average $C_\alpha - C_\alpha$ distances	runs	WT	F15L	F23L	Y37L	WT-F15L	WT-F23L	WT-Y37L
$\langle L_{16} - I_{26} \rangle$	1	9.21 ± 0.6	9.14 ± 0.3	9.03 ± 0.3	9.51 ± 1.0	9.12 ± 0.4	9.09 ± 0.3	9.09 ± 0.4
	2	9.43 ± 0.8	9.34 ± 0.8	9.08 ± 0.4	9.10 ± 0.5	9.49 ± 1.2	9.38 ± 0.9	9.56 ± 1.1
	3	9.10 ± 0.4	9.33 ± 0.6	9.92 ± 1.7	9.05 ± 0.4	8.98 ± 0.2	9.22 ± 0.7	9.45 ± 1.0
$Mean \pm SD^a$		9.25 ± 0.2	9.27 ± 0.1	9.34 ± 0.5	9.22 ± 0.2	9.20 ± 0.3	9.23 ± 0.1	9.37 ± 0.2
$\langle N_{14} - S_{28} \rangle$	1	8.42 ± 1.5	8.43 ± 0.3	7.97 ± 0.4	9.36 ± 1.4	8.42 ± 0.4	8.98 ± 1.0	8.48 ± 0.4
	2	8.74 ± 1.1	8.74 ± 1.0	8.12 ± 1.2	8.68 ± 0.7	9.26 ± 1.7	9.21 ± 1.2	8.76 ± 1.7
	3	8.44 ± 1.2	8.89 ± 0.9	9.44 ± 1.7	8.33 ± 0.4	7.67 ± 0.4	9.19 ± 0.9	9.18 ± 1.7
$Mean \pm SD^a$		8.53 ± 0.2	8.69 ± 0.2	8.51 ± 0.8	8.79 ± 0.5	8.45 ± 0.8	9.13 ± 0.1	8.81 ± 0.4
$\langle L_{12} - V_{32} \rangle$	1	9.12 ± 0.4	9.53 ± 0.2	8.87 ± 0.4	8.89 ± 0.3	8.75 ± 0.3	8.80 ± 0.4	9.23 ± 0.1
	2	8.98 ± 0.3	8.76 ± 0.4	9.27 ± 0.4	8.90 ± 0.5	9.11 ± 0.3	8.75 ± 0.3	9.18 ± 0.5
	3	8.98 ± 0.3	8.72 ± 0.4	8.91 ± 0.5	9.44 ± 0.2	9.25 ± 0.3	8.87 ± 0.2	9.01 ± 0.2
$Mean \pm SD^a$		9.02 ± 0.1	9.00 ± 0.5	9.02 ± 0.2	9.08 ± 0.3	9.04 ± 0.3	8.81 ± 0.1	9.14 ± 0.1

Table 4.3: Mean values and standard deviation (SD) are calculated from the three values obtained by averaging over the 200 ns of each of the three independent runs of each model. Results are listed for the single layer models of amylin wild type, mutants, and cross-seeded decamers. The listed values are averages over three independent trajectories, and are calculated using all ten strands.

the slow nucleation phase.

Face-to-face interactions between β-sheets are common in proteins and amyloids.[37] They involve hydrophobic surfaces with good shape complementarity held together through van der Waals and hydrophobic interactions.[37] For this reason, we have also monitored the face-to-face contacts between β-sheets (Figure 4.1B) for amylin wild type and its mutants, selecting residue pairs known to disrupt amylin amyloid formation.[141] These contacts are calculated also in the mixed aggregates. Results for all three cases are shown in Table 4.3. The side chains of L12, N14, and L16 (projecting from the lower face of the β_1-region) in β_1-region interact with the side chains of S28, I26, and V32 (projecting on the upper face of β_2-region) in β_2-region located on the opposite side of β_1-region (Figure 4.1). The average face-to-face distances are measured for all ten strands of each system and are within 8.5 to 9.5 Å (Table 4.3) in agreement with experimental results 8-11 Å.[142] Hence, in all oligomers, the β-strand-turn-β-strand motifs are stabilized by such face-to-face hydrophobic interactions. The face-to-face interactions between amino acid side chains

(hydrogen bonds between aligned N14 with S28, or T30 residues; hydrophobic interactions between aligned L16, I26, L12, V32, or polar interaction between N14 and S28 residues) in β_1 and β_2 regions is important in keeping the U-shaped structure. The cross-β-strand topology structure of the fibril, which is the main structural feature of amyloid fibril, is maintained throughout the simulation for both the hetero and homoaggregates indicating stabilization by a network of hydrogen bond and hydrophobic interactions. The stability of the aggregates is a strong indicator of the role of structural similarity in explaining the efficient cross seeding.[119] Pinpointing the amino acids (the hydrophobic cores; $L_{12}ANFLV_{17}$ and $F_{23}GAIL_{27}$ and Y37 in our simulation) that can alter the stability of amylin could contribute to a rational design of aggregation inhibitors that trap the toxic species in the fibril form.[143,144] Our simulation indicates that the selected face-to-face contacts within the two β-sheet regions are preserved, indicating that such side chain contacts are important for retaining an overall U-shaped structure. Interior contact distances are not significantly changed by the three substitutions. Future experiments and computer simulation involving single and double mutants of those amino acids mutations at residues 16, 26, 14, 28, 12, or 32 will elucidate the role of these residues in the stability of fibril structures.

Using the DSSP[145] software, we have compared the changes in secondary structure encountered by wild types, mutants, and their complex. Our analysis compares in each case the first and last 50 ns of the molecular dynamics trajectory. Our data listed in Table 4.4 indicate that the β-sheet content of the aggregates is maintained for wild type, mutants, and their complex. This is another indication for the stability of the β-hairpin topology and provides additional evidence for the importance of fibril structural similarity in efficient seeding.[119]

	secondary structurea, first 50 ns			secondary structure, last 50 ns		
system	β-sheet	helix	turn	β-sheet	helix	turn
WT (SL)	60.24(1.21)	0.04(0.02)	39.73(1.20)	56.11(0.50)	0(0)	43.89(0.50)
F15L (SL)	62.09(2.27)	0(0)	37.91(2.28)	59.44(1.80)	0.04(0.07)	40.53(1.82)
F23L (SL)	61.80(1.94)	0(0)	38.20(1.94)	59.97(2.99)	0.02(0.05)	42.34(1.17)
Y37L (SL)	62.34(3.22)	0(0)	40.76(3.53)	55.71(2.11)	0.01(0.02)	44.28(2.13)
WT-F15La (SL)	60.68(0.38)	0(0)	39.32(1.07)	57.63(1.20)	0.03(0.05)	42.34(1.17)
WT-F23La (SL)	62.25(1.73)	0(0)	37.85(1.73)	0.03(0.06)	60.67(1.99)	39.30(2.02)
WT-Y37La (SL)	59.03(1.04)	0.09(0.15)	40.88(0.96)	56.11(0.50)	0(0)	43.89(0.50)
WT (DL)	60.04(0.38)	0(0)	39.96(0.38)	59.00(1.23)	0(0)	41.00(1.23)
F23L (DL)	57.81(1.60)	0.01(0.01)	42.17(1.58)	56.47(2.90)	0.03(0.03)	43.50(2.90)
Y37L (DL)	61.80(2.54)	0(0)	38.20(2.54)	57.87(4.43)	42.12(4.42)	0.01(0.01)

Table 4.4: β-sheet = β-strand + β-bridge, helix = α-helix +3^{10}-helix + π-helix, turn = turns + bend + coil. The results are the averages of three independent simulations and the standard deviation is given in parentheses.

Using single trajectory MM-PBSA[146] we have also estimated the binding free energy for symmetrically segmented proto-filaments (single layer, between the first pentameric and the second pentameric units) and proto-filament pairs (double layer, between the upper pentameric and the lower pentameric β-hairpin units). This allows us to characterize favorable association between wild type and different mutants and related it to the experimentally observed efficient seeding between mutant and wild type amylin.[119] We have taken for this analysis an average of over 2000 equally spaced (at an interval of 20 ps) snapshots from the 40 ns production trajectory. Assuming equal entropic factors for the oligomers, their binding will be stronger the larger their negative free energy, whereas a positive number or small negative number signifies a weaker binding.[138,147,148] The calculated binding free energies and energy components are shown in Table 4.5. The results show that the binding affinity of the heteroassembly single layer complexes is comparable to the homogeneous fibrils. However, the binding energy of the double layer assembly of the single aromatic amino acid mutants is less favorable than that of the wild type (Table 4.5), suggesting contribution of aromatic interactions to the thermodynamic stabil-

energy components	WT (SL)	F15L (SL)	F23L (SL)	Y37L (SL)	WT- F15L (SL)
ΔE_{elec}	1038.7 ± 182.0	1055.1 ± 94.9	1006.1 ± 38.2	$1047.9 \pm 146.2)$	$1148.5 \pm 118.5.4$
ΔE_{vdw}	-188.9 ± 6.1	-187.3 ± 2.6	-187.5 ± 11.1	-187.0 ± 4.5	-190.8 ± 5.0
ΔE_{PB}	-1004.7 ± 179.0	-1008.3 ± 95.0	-969.1 ± 26.2	-1007.7 ± 145.0	-1104.1 ± 117.0
ΔE_{SA}	108.3 ± 0.8	108.9 ± 0.1	108.3 ± 6.4	107.1 ± 2.1	109.1 ± 3.2
ΔE_{polar}	34.4 ± 2.7	46.9 ± 0.1	37.1 ± 12.0	40.2 ± 1.3	44.4 ± 1.5
$\Delta E_{nonpolar}$	-118.3 ± 49.9	-109.4 ± 41.3	-79.2 ± 4.6	-80.1 ± 2.4	-146.4 ± 6.6
$\Delta G_{binding}$	-46.3 ± 2.6	-31.6 ± 2.6	-42.1 ± 7.4	-39.9 ± 1.1	-37.3 ± 3.3
energy components	WT- F23L (SL)	WT- Y37L (SL)	WT (DL)	F23L (DL)	Y37L (DL)
ΔE_{elec}	960.9 ± 9.9	1057.5 ± 162.8	453.7 ± 20.0	570.2 ± 12.0	482.8 ± 25.4
ΔE_{vdw}	-186.7 ± 6.2	-190.9 ± 0.1	-206.5 ± 17.7	-98.8 ± 12.0	-211.8 ± 3.7
ΔE_{PB}	-921.5 ± 5.4	-1020.7 ± 159.1	-395.5 ± 24.9	-531.4 ± 18.6	-397.4 ± 14.8
ΔE_{SA}	-78.5 ± 3.4	108.3 ± 0.1	120.9 ± 0.5	58.5 ± 7.3	126.7 ± 2.6
ΔE_{polar}	39.4 ± 4.4	36.7 ± 3.7	58.2 ± 4.7	38.8 ± 6.6	85.4 ± 10.6
$\Delta E_{nonpolar}$	-78.5 ± 3.4	-82.5 ± 0.3	-85.6 ± 17.2	-38.4 ± 6.7	-85.2 ± 6.4
$\Delta G_{binding}$	-39.1 ± 7.8	-45.8 ± 3.4	-27.5 ± 12.5	0.4 ± 0.2	0.2 ± 4.2

Table 4.5: The data are averages of two independent 40 ns simulation with the corresponding standard deviations. All values are in kcal/mol. The polar term is the sum of Coulomb interaction energy (E_{elec}) and polar contribution to the solvation free energy (E_{PB}). The nonpolar term consists of the van der Waals interaction energies (E_{vdW}) and the nonpolar contribution to the solvation free energy (E_{SA}).

ity. The polar energy term is unfavorable for all the aggregates while the nonpolar energy term contributes favorably to their association and stability. Note that the binding energy addresses only the stability of the preformed aggregates; it does not describe adequately the initial interaction between the chains (i.e., the kinetic process). Hence, it is difficult to compare the values calculated by MMPBSA with experimental[149] rates of aggregate formation and lag times.

The commonly accepted mechanism for their toxicity is that amyloid oligomers interact with cell membranes compromising their structural integrity and lowering their permeability barrier.[150] This ability of amyloid oligomers to form membrane pores or channels is then responsible for their neurotoxicity.[151] Recent structural studies of the toxic oligomer models suggest that the hydrophobic surface facilitates insertion into the membrane while water pores formed from the hydrophilic groups interfere with cellular homeostatic by enabling water and ion transport.[152] The

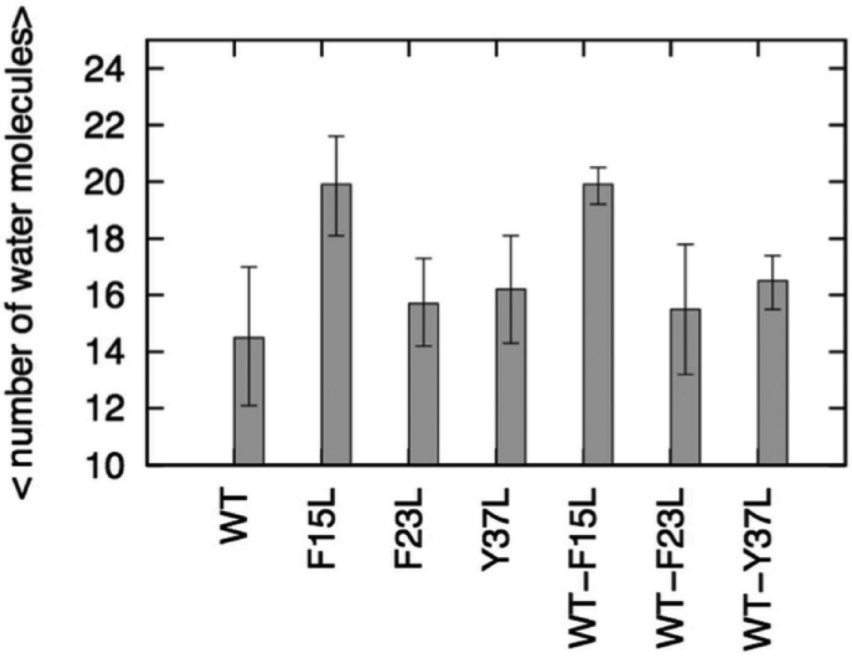

Figure 4.6: Average number of water molecules inside the hydrophilic cavities. Mean values and standard deviation (SD) are calculated for each model from the three values obtained by averaging each independent run over the full 200 ns.

hydrophobic surfaces ($L_{12}ANFLV_{17}$ and $F_{23}GAIL_{27}$) of the fibril-like oligomer assembly of amylin may provide a means by which these structures can insert into lipid bilayer membranes, while the hydrophilic water channel observed in a recent simulation could be involved with toxicity of amylin oligomer. Studies have shown that amyloid aggregates in lipid bilayers adopt structures similar to that in an aqueous medium.[153–155] The experimentally determined structures of amyloid fibril[116] (including amylin) do not contain water molecules, while various molecular dynamic simulations[147,156] revealed embedded water molecules are integral part of the fibril models. Thus, we have examined the presence of the water trapped in amylin wild-type, mutants, and cross-assembly pores. The results are shown in Figure 4.6 and Figure 4.7. The water molecules in amylin, its mutants, and their hetero-assembly

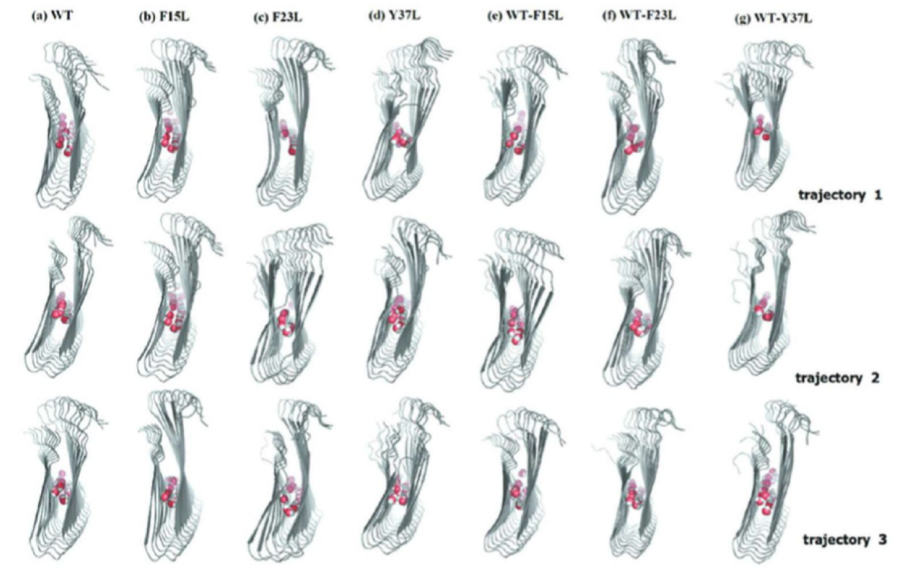

Figure 4.7: Water residing in the hydrophilic cavities. Snapshots taken after 10 ns for (a) wild type (WT); (b) Y37L; (c) WT-Y37L; (d) F15L; (e) F23L; (f) WT-F15L; (g) WT-F23L. The red and white spheres represent water oxygen and hydrogen atoms, respectively.

are found in the middle of the two β-strands near a group of polar amino acids whose side chains point toward the interior of the oligomer cavity (N14, S28, and T30). The location of the hydration channel in our simulation is similar to that found in a previous study of amylin alone and in heteroassembly with Aβ.[147]

4.4 Conclusions

We investigate stability and conformational changes of amyloid heteroassemblies through molecular dynamics simulations as the structure and stability of heteropolymeric fibrils are difficult to probe in experiments. Our all atom explicit solvent simulations give molecular level insight into the cross seeding between amyloids. We find no significant differences in the structure of the wild type, mutants, and

their heteroassembly, as all of them retain the original U-shaped fibril conformation over the 200 ns time trajectories. Hence, amyloids with similar side chain packing at the β-sheet interface are structurally compatible, acting as a good template for the congruent incorporation of homologue peptides, which underlie efficient mixed growth. This points to structure similarity as a key determinant for an efficient cross seeding between wild type and mutants. Replacement of aromatic amino acids with nonaromatic residues of similar size and hydrophobicity is not critical at position 15, which is not involved in any intersheet steric zipper interaction. On the other hand, the replacement of tyrosine with leucine at position 37 leads to a fibril structure with a greater flexibility resulting in a loss of steric zipper interaction. This loss of a steric zipper explains the slow growth of the aggregates of Y37L mutants compared to other mutants and wild type. Our results indicate that the residues $L_{12}ANFLV_{17}$ and $F_{23}GAIL_{27}$ which are located in the β-strand domain are more rigid and could be a pharmacophore for ligand binding,[120,157] targeted to stabilize amylin fibril and thereby reducing its toxicity. We therefore propose to use these two segments of amylin for combined computational screening and experimental tests to find small molecules that affect the aggregation and toxicity of amylin. Jiang et al.[120] recently found this approach to be successful in searching for compounds that reduced Aβ cytotoxicity.

4.5 Acknowledgments

This work is supported by the National Institutes of Health under Grant No. GM62838 and by the National Science Foundation under Grant No. EPS-0814361. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of En-

ergy under contract no. DE-AC02- 05CH11231. Other parts of the simulations were done on the BOOMER cluster of the University of Oklahoma. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, the National Institutes of Health, or the Department of Energy.

Chapter 5: Replica-Exchange-with-Tunneling for Fast Exploration of Protein Landscapes

Reproduced from Fatih Yasar, Nathan A. Bernhardt, and Ulrich H. E. Hansmann. Replica-exchange-with-tunneling for fast exploration of protein landscapes. J. Chem. Phys., 143(22):224102, Dec 2015., with the permission of AIP Publishing.

Author Contributions: The majority of the work presented in this chapter is credited to Dr. Fatih Yasar. The author of this dissertation played a supporting role in this project but did help with the implementation of the replica-exchange-with-tunneling method.

5.1 Introduction

Cellular processes are often controlled by transient interactions between proteins that are difficult to trace in experiments. Simulations can probe such interactions, but the accuracy of macroscopic observables predicted by Monte Carlo or molecular dynamics is still limited. Despite increased computing power, the problem remains that biomolecular motions often cover time scales that are not achievable in atomistic simulations. This is because the computational requirements increase exponentially with system size. It has become popular to try alleviating this problem through the use of replica-exchange[85] and other generalized-ensemble techniques.[158,159] While these approaches have become ubiquitous, there has grown a disenchantment with them as the theoretically expected gain in efficiency is realized rarely in practical applications: the exponential growth in computational costs is replaced by a power law, but the exponents are typically of order four, and pre-factors are often large,

restricting the application of such approaches to rather small systems. In the present paper, we focus on replica-exchange molecular dynamics (REMD).[86,87,160] In order to enable its efficient use in multiscale and explicit solvent simulations, we introduce a Replica-Exchange-with-Tunneling (RET) method by integrating ideas of hybrid MC/MD[161] into the replica-exchange protocol. We test this approach for the Trp-cage protein,[17,162] which has been previously investigated with REMD[85–87,163] by various groups[164–166] allowing a comparison with our data.

5.2 Methods

In replica-exchange sampling, replicas of the protein system evolve in parallel at different temperatures. At certain times, replicas are exchanged between neighboring temperatures T_i and $T_j = i + 1$ with a probability,

$$\begin{aligned} w(\ \mathbf{C}^{old} \rightarrow\ \mathbf{C}^{new}) &= \min(1, \exp(-\beta_i E(C_j) - \beta_j E(C_i) + \beta_i E(C_i) + \beta_j E(C_j))) \\ &= \min(1, \exp(\Delta\beta\Delta E)) \end{aligned} \tag{5.1}$$

with $\beta = 1/k_B T$. The resulting random walk through temperature yields an enhanced exploration of configurations at low temperatures. However, despite many successful applications, replica-exchange sampling is often restricted by severe limitations. Take as an example simulations of proteins in explicit solvent where the number of required replicas increases rapidly with protein size. As the time to sample independent configurations increases quadratically with the number of replicas, it follows that both many replicas and long trajectories are required to generate sufficient statistics at temperatures of interest. While exchange schemes have been developed that target specifically this problem,[167,168] all-atom folding simulations in

explicit solvent are still restricted to rather short proteins (of $\approx 50 - 80$ residues). Note also that the rapid growth of replica number is not restricted to explicit solvent simulation but will appear for all sufficiently large systems.

Here, and in many other practical applications, the use of replica-exchange sampling is held back because the exchange move leads to a proposed state C^{new} of the multiple replica system that is exponentially suppressed. However, once such an exchange move is accepted, the two replicas will quickly evolve and the compound system will assume a new state that has a weight comparable to that before the exchange move. Hence, the problem is to "tunnel" through the unfavorable "transition state" generated by the exchange move.

We propose to tackle this problem by combining replica-exchange molecular dynamics with ideas from hybrid MC/MD.[161] In the latter technique, one starts with a configuration q_i and velocities v_i corresponding to the selected temperature. A short molecular dynamics run leads to a configuration q_{io}, v_{io} that is accepted or rejected by a Metropolis step. As the time reversibility of the Verlet algorithm guarantees detailed balance, the Monte Carlo step ensures that the sampled configurations are distributed according to the chosen temperature. Utilizing in a similar way the time reversibility of the Verlet algorithm our RET replaces a configuration A by $\hat{B}$ at temperature T_1, and B by $\hat{A}$ at temperature T_2,

$$\begin{aligned} T_1: \quad & A \longrightarrow A' \quad \searrow\nearrow \quad B'' \longrightarrow \hat{B}, \\ T_2: \quad & B \longrightarrow B' \quad \nearrow\searrow \quad A'' \longrightarrow \hat{A}, \end{aligned}$$

where $A = (q_A, v_A)$ denotes a state characterized by coordinates q_A of all its atoms and the associated velocities v_A. The crossing arrows mark the exchange step.

Hence, the RET move consists of four parts:

1. In the first part, the configuration A (B) evolves by a short microcanonical molecular dynamics run to a configuration $A' = (q'_A, v'_A)$ (B') with total energy $E_{tot} = E_{Pot} + E_{kin} = E_1$ (E_2).

2. In the second step, the two replicas are provisionally exchanged, and at the same time their velocities rescaled such that the total energies at the two temperatures stay the same:

$$E_{tot}(B'') = E_1 \quad \text{and} \quad E_{tot}(A'') = E_2\,. \tag{5.2}$$

   Here, $A'' = (q'_A, v''_A)$ and $B'' = (q'_B, v''_B)$. This is achieved by rescaling the velocities according to[169,170]

$$v''_A = v'_A \sqrt{\frac{E_2 - E_{pot}(q'_A)}{E_{kin}(v'_A)}} \quad \text{and} \quad v''_B = v'_B \sqrt{\frac{E_1 - E_{pot}(q'_B)}{E_{kin}(v'_B)}}\,. \tag{5.3}$$

3. The above exchange move generates a "transition state" in the multiple replica system where the unfavorable potential energies at the two temperatures are compensated by the rescaled velocities. In the third step, each of the configurations $A''(B'')$ evolve again by a short microcanonical molecular dynamics run to a configuration $\hat{A} = (\hat{q}_A, \hat{v}_A)(\hat{B})$ where the velocity distribution corresponds now again to the target temperatures, and the potential energies are comparable to the ones found at the respective temperatures before the exchange move.

4. Finally, in the last step, this set of configurations $\hat{A}, \hat{B}$ is compared with the

set A, B and is accepted with probability

$$\begin{aligned} w(\ \mathbf{C}^{old} \rightarrow\ \mathbf{C}^{new}) \ &= \ \min(1, \exp(-\beta_1 \left(E_{pot}(\hat{q}_B) - E_{pot}(q_A)\right) \\ &\quad -\beta_2 \left(E_{pot}(\hat{q}_A) - E_{pot}(q_B)\right))). \end{aligned} \tag{5.4}$$

If rejected, the molecular dynamics simulations will continue at $T_1(T_2)$ with configuration $A(B)$. In both cases, new velocities are drawn from a distribution corresponding to the respective temperatures. Again, the time reversibility of the trajectories $A \longrightarrow A'$ $(B \longrightarrow B')$ and $A'' \longrightarrow \hat{A}$ $(B'' \longrightarrow \hat{B})$ ensures detailed balance and convergence to the correct distribution.

In order to see that RET leads to an increased exchange rate, we write the acceptance probability of Eq. (5.4) as

$$\begin{aligned} w(\ \mathbf{C}^{old} \rightarrow\ \mathbf{C}^{new}) \ &= \ \min(1, \exp\left(\Delta\beta\Delta E\right) \times \exp(-\beta_1 \left(E_{pot}(\hat{q}_A) - E_{pot}(q_A)\right) \\ &\quad -\beta_2 \left(E_{pot}(\hat{q}_B) - E_{pot}(q_B)\right))), \end{aligned} \tag{5.5}$$

where $\Delta E = E_{pot}(\hat{q}_B) - E_{pot}(\hat{q}_A)$. The first factor is the acceptance rate for regular replica-exchange sampling. Hence, the acceptance probability is enhanced by

$$\frac{w^{RET}}{w^{REMD}} = \exp\left(-\beta_1 \left(E_{pot}(\hat{q}_A) - E_{pot}(q_A)\right) - \beta_2 \left(E_{pot}(\hat{q}_B) - E_{pot}(q_B)\right)\right) \tag{5.6}$$

where $E_{pot}(\hat{q}_A) - E_{pot}(q_A)$ and $E_{pot}(\hat{q}_B) - E_{pot}(q_B)$ have opposite sign. Assuming both terms to be similar in magnitude, the enhancement factor can be approximated as

$$\frac{w^{RET}}{w^{REMD}} \approx \exp\left(-\Delta\beta\delta E\right) \tag{5.7}$$

where we have defined $\delta E = |E_{pot}(\hat{q}_A) - E_{pot}(q_A)| \sim |E_{pot}(\hat{q}_B) - E_{pot}(q_B)|$ and

$\Delta\beta = \beta_2 - \beta_1 < 0$. As this enhancement factor is always larger or equal one, it follows that the acceptance rate will be always better than in traditional replica-exchange moves, but will depend on the length of microcanonical step which controls how much the potential energies $E_{pot}(\hat{q}_A)$ ($E_{pot}(\hat{q}_B)$) differ from $E_{pot}(q_A)$ ($E_{pot}(q_B)$).

Our test case is the designed 20-residue Trp-cage miniprotein[17,162] (Protein Data Bank Identifier 1L2Y). As one of the smallest proteins with a defined tertiary structure it is often used to evaluate new sampling schemes.[171,172] Using the same force field, implicit solvent, and temperature distribution, we compare the results from our RET simulations with previous simulations[173] that rely on regular REMD.[85–87,163] Out of the many Trp-cage replica-exchange studies,[164–166] these are chosen by us because their setup leads to a melting temperature of 400 K, closer to the experimental values of 315 K[162] than found in other implicit solvent simulations. We use the molecular dynamics program package GROMACS,[174] version 4.6.5, either in its original version or modified to implement our RET approach. The modified version is available as supplementary material[175] and from the authors. Interactions between the atoms in the protein are described by the Amber force field 94,[176] and the interaction of the protein with the surrounding solvent is approximated by a generalized Born surface area implicit solvent.[177] The N- and C-termini are capped with methyl groups. We use the LINCS algorithm[78] to constrain hydrogen atoms to their bonded heavy atoms. van der Waals and Coulomb energies are calculated using twin range cutoffs. The equations of motion are integrated with a time step of 1 fs for RET and 2 fs for regular REMD. We use either 22, or 12 replicas distributed over a range of temperatures from 250 to 605 K. For 22 replicas, the selected temperatures are the same as in Ref.,[173] i.e., 250, 255, 260, 265, 273, 284, 298, 315, 333, 353, 373, 393, 413, 433, 454, 473, 493, 513, 533, 555, 580, and 605 K. The thermo-

stat temperature is controlled by the v-rescale method.[132] Exchanges are attempted every 200 ps, which for RET includes two segments of 1 ps where the system evolves at constant energy. All simulations start from unfolded configurations and physical quantities are calculated after discarding the first 50 ns.

5.3 Results and Discussions

In RET, the exchange move generates a "transition state" where unfavorable potential energies at the two temperatures are compensated by the rescaled velocities. Obviously, the rescaled velocities will not be distributed according to the corresponding temperatures. However, RET assumes that the subsequent short microcanonical segment leads to a redistribution of energies, i.e., to configurations with more favorable potential energies and velocities whose distribution corresponds to the target temperatures. We demonstrate that this assumption is indeed valid by showing in Figure 5.1 the distribution of velocities as measured after these microcanonical segments and contrast this distribution with the one of the velocities measured before the RET move, i.e., the one generated by the thermostat. Our data are for a temperature of 250 K. Both distribution overlap with each other within the error bars, and the inset shows that the differences in frequencies are random.

Having shown the correctness of the RET approach we compare now its efficiency with that of regular REMD. For this purpose, we show first in Figure 5.2 for both regular REMD and RET a typical run through temperature space. Note that we use for both simulations the same temperature distribution of 22 replicas, force field and implicit solvent as in previous work.[173] This figure suggests a faster walk through temperature space in RET than seen in regular replica-exchange molecular dynamics. We quantify this observation by measuring the number of tunneling

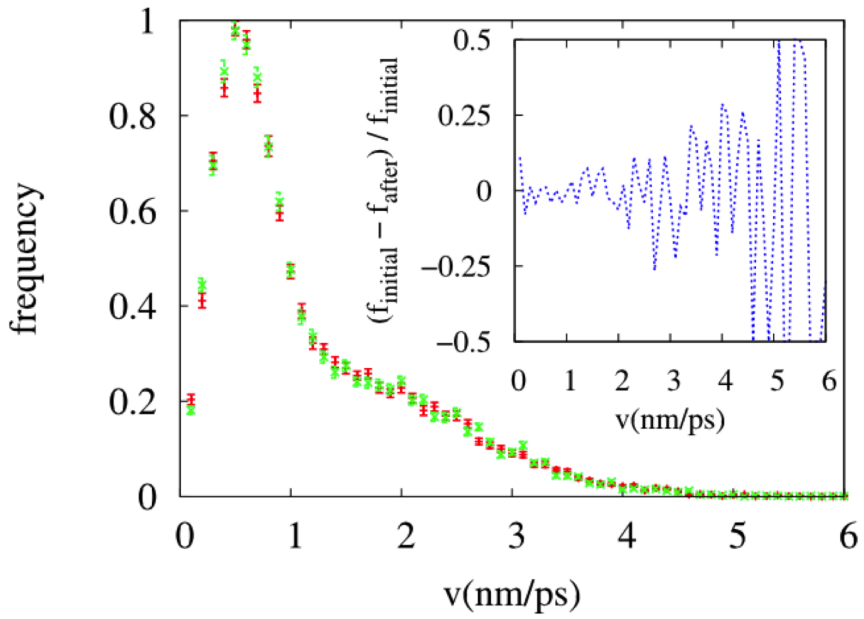

Figure 5.1: Distribution of velocities as measured before and after RET exchange protocol. The inset shows the differences between the two curves.

events in both cases. Here, a tunneling event is defined as a walk of a replica from the lowest temperature of 250 K to the highest temperature (605 K), and back. Note that we begin with measuring this quantity only after discarding the first 50 ns for equilibrization. While we observe only six such events in regular replica-exchange molecular dynamics, we find 17 tunneling events in our RET simulations of the same length. These tunneling events are listed in Table 5.1.

On average it takes in regular REMD least 103.7(21.5) ns to cross the whole ladder of temperatures while only 62.9 (9.1) ns in RET. Note that the number of tunneling events is more easily to define than the average time it take to go from lowest to highest temperature (or in opposite direction) as some replicas did not walk at all along the whole temperature space, and other did not have time to finish an ongoing walk. The listed times are therefore lower boundaries. Note also that

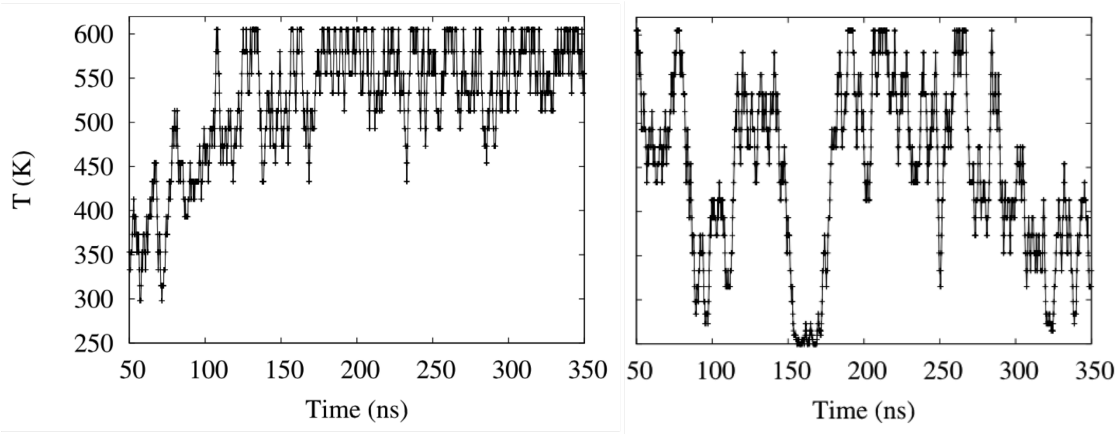

Figure 5.2: Time series for a single replica for REMD (left) and RET (right).

the times are much smaller if one considers only completed tunneling events. These take on average only 31 (13.2) ns in regular REMD and 30.6 (8.6) ns in RET. As the times for completed tunneling events differ little between the two approaches, we conclude that RET enhances sampling by helping to overcome bottlenecks that otherwise inhibit the temperature walk of a replica. If correct, our conjecture would imply a better mixing of replicas in RET, a point we discuss below.

We remark also that the absolute number of tunneling events depends to a certain degree on the time between exchange moves. We have therefore also tested that changing this time does not qualitatively alters the picture by continuing both the regular REMD trajectory and the RET run for an additional 25 ns, allowing now for only 10 ps (instead of 200 ps) between exchange moves. With this reduced time between exchange moves we find within the 25 ns four tunneling events for regular REMD and 49 such tunneling events for RET. Hence, while a shorter time between exchange moves leads for both approaches to faster movement in temperature space, it does not diminish the advantage of RET over regular REMD.

The consequence of this faster walk through temperature is a better mixing of

	REMD			RET		
Tunneling event	Start time (ps)	End time (ps)	Duration (ps)	Start time (ps)	End time (ps)	Duration (ps)
1	170 400	217 400	47 000	88 200	124 800	36 600
2	192 000	221 800	29 800	114 600	133 600	19 000
3	228 400	258 600	30 200	177 400	191 200	13 800
4	239 200	262 600	23 400	79 400	213 600	34 200
5	292 400	310 200	17 800	182 200	205 200	23 000
6	294 000	331 800	37 800	200 600	238 000	37 400
7				221 400	245 600	24 200
8				229 400	258 400	29 000
9				235 800	285 200	49 400
10				244 200	329 200	85 000
11				252 200	274 800	22 600
12				255 400	272 800	17 400
13				259 800	314 000	54 200
14				281 800	303 200	21 400
15				313 000	338 000	25 000
16				317 000	336 000	19 000
17				328 200	338 400	10 200

Table 5.1: Tunneling events for regular REMD and the newly developed RET.

the replicas. This can be seen in Figure 5.3 where we show the frequency with that a certain replica resides at a given temperature. Note the more even distribution in RET while in classical REMD some replicas stay confined in certain temperature regions.

The faster walk through temperature space and better mixing of replicas leads to an improved sampling of protein configurations at low temperatures that should allow calculations of thermal averages with at least the same accuracy and efficiency as regular REMD. This is demonstrated in Figure 5.4 where we show the average frequency of configurations that are within 2.7Å to the first entry of the NMR ensemble of the native Trp-cage structures as deposited in the Protein Data Bank under identifier 1L2Y. Data are calculated using the last 300 ns of our trajectories obtained with either regular replica-exchange molecular dynamics or RET. Comparing both kind of simulations, which sample the same set of 22 temperatures, we see that within the error bars the averages agree with each other. We note that our thermal averages also agree with the earlier work in Ref.[173] This demonstrates

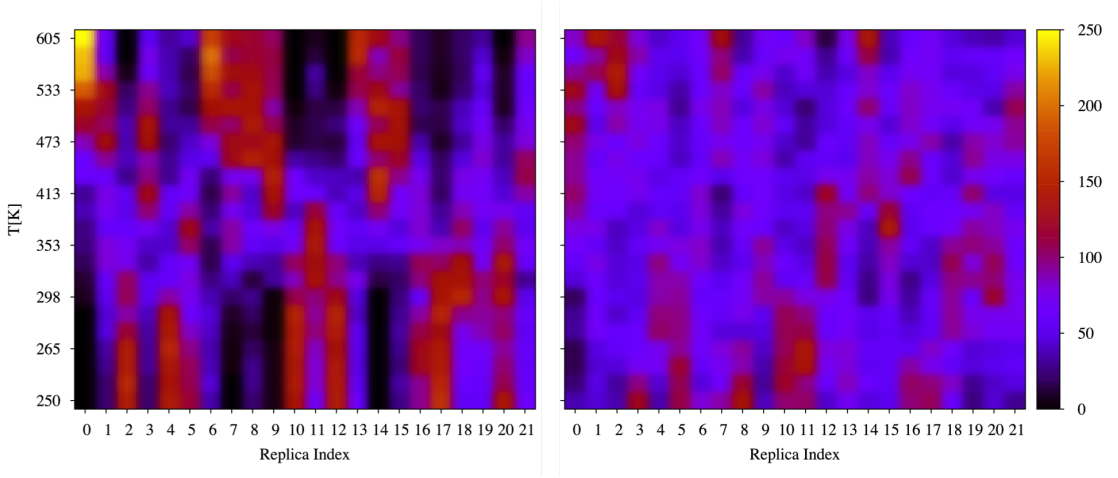

Figure 5.3: Mixing of replicas in temperature space, as observed in regular REMD (left) and in our RET approach (right).

that RET leads indeed to the correct ensemble and therefore allows one to calculate thermal averages atlas with the same efficiency as regular REMD.

We demonstrate now that the increased ability to walk along temperature space can be used to decrease the number of replicas in a simulation. For this purpose, we have simulated the same system with only twelve replicas, spread over the same temperature interval as in the previous case: 250, 273, 298, 315, 333, 353, 393, 433, 473, 513, 555, and 605 K. In Table 5.2 we list the acceptance rates for exchange moves between neighboring temperatures. These rates are always larger for RET than for regular REMD, but note also the extreme small rates for REMD in the temperature range of 350 K-470 K, indicating a bottleneck for replicas walking through temperature space. As a consequence of this bottleneck we observe no tunneling event during the full 300 ns of the trajectories generated by regular REMD, while we find 53 events in the trajectories generated by RET sampling. This correspond to an average time of 25.1 (2.9) ns to traverse the whole temperature range with RET sampling. Assuming a simple diffusion process through temperature, the

T[K]	RET-MD	REMD
250-273	49.20	15.20
273-298	49.60	13.87
298-315	49.20	36.67
315-333	52.13	35.87
333-353	53.47	33.73
353-393	49.07	2.40
393-433	52.40	5.73
433-473	50.80	8.53
473-513	50.53	14.80
513-555	49.87	16.14
555-605	48.80	13.20

Table 5.2: Acceptance rate for RET and REMD using 12 replicas.

time needed to move between the lowest temperature and the highest temperature increases with the square of the number of replicas, i.e., the number of tunneling events decreases accordingly with number of replicas. Our measure times for the RET sampling are consistent with such diffusive behavior, and consequently, we find for the RET simulations a much larger number of tunneling events for a system of twelve replicas than for the system with 22 replicas. However, this argument is only valid if there is a sufficiently high probability that replicas will exchange between neighboring temperatures. This is the case for our RET approach, however, for REMD the exchange probabilities are significantly reduced for temperature between 350 K and 470 K, making it difficult for replicas to pass this bottleneck and walk along the whole temperature range. As a consequence, the sampling efficiency of regular REMD is greatly reduced when lowering the number of replicas from 22 replicas to 12, while the reduction in sampling efficiency is less for our new RET approach.

This can be seen again in Figure 5.4 where we also show the frequencies of native-like configurations (root-mean-square deviation smaller than 2.7 Å) as obtained in

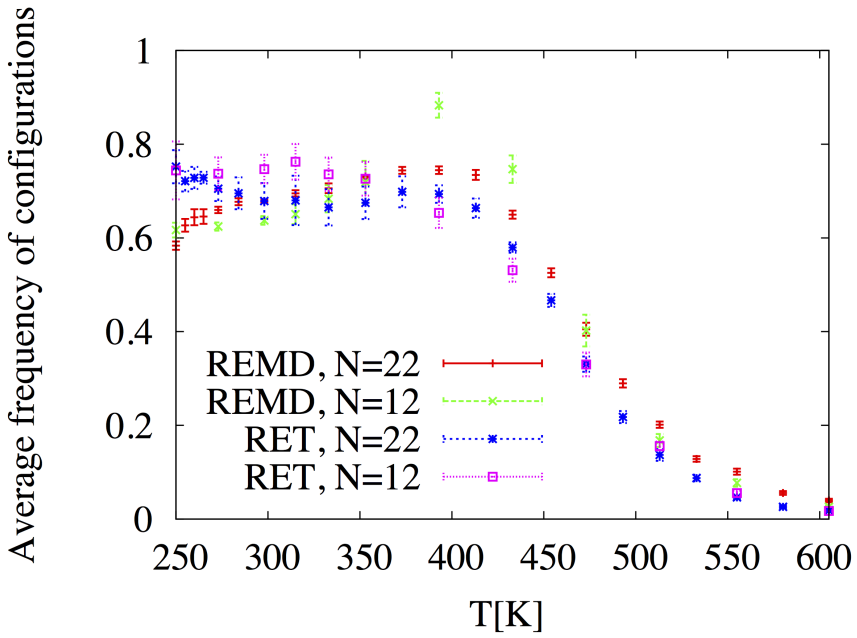

Figure 5.4: Average fraction of configurations with a backbone rmsd with a root-mean-square-deviation smaller than 2.7 Å to the native structure as deposited in the PDB (1L2Y). Shown are data from regular REMD and such of RET simulations with the same distribution of 22 replicas, and from simulations with both methods using only twelve replicas distributed over the same temperature range.

replica-exchange molecular dynamics and RET simulations with twelve temperatures. When taken over the whole trajectories of 300 ns length, our results overlap for both methods strongly with the ones obtained in the corresponding simulations for 22 replicas. The exception is the melting temperature where we expect large fluctuations in energy and the system to be most sensitive to sampling difficulties. At $T \approx 350 - 450K$, the values obtained with regular replica-exchange molecular dynamics deviate strongly between the runs with twelve replicas and those run with 22 replicas. These strong deviations are expected as there is little exchange between neighboring temperatures in this range (see Table 5.2). Around the melting temperature this low exchange rate leads to large sampling errors. In RET, the exchange

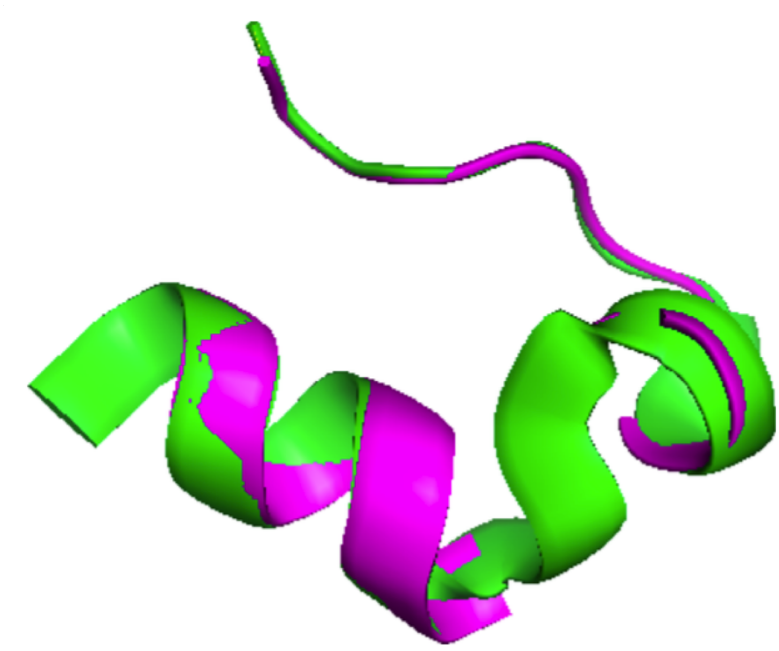

Figure 5.5: The lowest RMSD Trp-cage configuration (in magenta) superimposed on the NMR structure (green).

rates between neighboring temperatures are not reduced in this temperature range when reducing the number of replicas from 22 to 12. Since we do not have the bottlenecks observed for REMD, we do not find such deviations for RET sampling. The lack of bottleneck demonstrates the improved sampling by our new method, which even in simulations with this small number of replicas (twelve as opposed to the 22 replicas in earlier work[173]) allows us to find structures that differ by less than 1 Å from the experimentally determined one, see Figure 5.5.

5.4 Conclusions

In summary, we have introduced a variant of replica-exchange molecular dynamics that increases the flow of replicas through temperature by allowing the system to

"tunnel" through unfavorable "transition states" generated by the exchange move. We have tested this idea by simulating the Trp-cage protein in an implicit solvent, an often used toy-model for evaluating new sampling techniques. We show that both methods lead to the same thermodynamic averages; but thermalization is faster for RET when a too large spacing in temperature leads for regular REMD to very low acceptance rates. This is a persistent problem in replica-exchange molecular dynamics of proteins in an explicit solvent where the large number of water molecules results in a huge number of degrees of freedom which in turn leads to the need for very small spacing in temperature (and therefore a large number of replicas). We remark that explicit solvent simulations are not the only example where low acceptance rates limit the use of regular REMD. Another example is resolution exchange[91] which suffers from diminishing acceptance rates for even small differences in the graining of the models. We are currently evaluating how RET can be used in this context to study the folding and free energy landscapes of the A and B domain of protein G (in preparation).

5.5 Acknowledgments

We acknowledge support from the National Science Foundation (research Grant No. CHE-1266256) and from Hacettepe University Scientific Research Fund under Project No. FHD.2015.6939. The simulations were done on the BOOMER cluster of the University of Oklahoma. We thank to Dr. Jiang Ping for help at an early stage of this project. F.Y. also thanks the Department of Chemistry and Biochemistry for kind hospitality during his sabbatical stay at University of Oklahoma.

Chapter 6: Simulating Protein Fold Switching by Replica-Exchange-with-Tunneling

The following chapter was published in The Journal of Chemical Theory and Computation by the author of this dissertation as the following article: Nathan A. Bernhardt, Wenhui Xi, Wei Wang, and Ulrich H. E. Hansmann. Simulating protein fold switching by replica exchange with tunneling. J. Chem. Theory. Comput., 12(11):5656–66, 2016. All text and figures are taken with the permission of the publisher.

Author Contributions: Dr. Wenhui Xi is credited for providing data used in validating the method, specifically from simulations of AFP and BFP. Likewise, Dr. Wei Wang is credited for his contribution of the the serum amyloid A section. The remaining portion of this chapter pertaining to GA98 and GB98 was contributed by the author of this dissertation. Implementation of the Go-model feeding method was also performed by the author.

6.1 Introduction

Detailed knowledge of the processes by which proteins fold or aggregate is crucial for understanding disease pathways and the working of drugs at the cellular level. This is because a protein's function depends on its specific three-dimensional structure. In the common picture of folding, the sequence of amino acids encodes a funnel-like energy landscape that guides the folding pathways into a single and distinct native state.[5,6] However, protein landscapes are often more complex. For instance, mutation experiments by Orban, Bryan, and co-workers[21,50,51] led to mutants of

the A and B domains of protein G that have over 90% sequence identity but still preserve their distinct structures and functions. These experiments and other observations[178–180] suggest that the sequence of a protein often encodes not only the native fold but also an ensemble of structures that are essential for the function or are important during folding or association. The various forms populate a multi-funnel folding and association landscape where mutations, changes in environment, or interaction with other molecules switch between the encoded folds.

It is a challenge to probe such conformational transitions in experiments or computer simulations. The latter suffer from the problem that the time scales of folding and assembly of proteins are difficult to cover in atomistic simulations. This is because the computational efforts increase exponentially with system size in constant temperature molecular dynamics simulations. While replacing the exponential growth in computational cost by a power law, enhanced sampling techniques such as replica exchange molecular dynamics[85–87,160,163,181] and other generalized ensemble techniques.[158,159] are often still not efficient enough for exploring the landscape of proteins to a degree that would allow one to determine the relative weight of the various competing structures and the pathways connecting them.

There exists a large and diverse ensemble of enhanced sampling techniques, including accelerated molecular dynamics, metadynamics, and multiple Markov chain approaches, to name only a few. For recent reviews, see, for instance refs.[182–186] However, the most commonly used enhanced sampling technique is replica exchange molecular dynamics as the underlying idea is easy to grasp and the technique is simple to implement into a standard program package without breaking its scaling on massively parallel computer systems. We have recently proposed a method to overcome some of the shortcomings that hold back replica exchange molecular dynamics

through the use of replica exchange with tunneling (RET).[187] In the present article, we extend RET in a way that enables simulations of proteins and protein aggregates that can switch between distinct forms. We demonstrate first that our approach does not introduce sampling biases by simulating two small peptides, the designed 11-residue α-helical tendency peptide (AHTP)[18] and the β-hairpin forming C-terminal fragment (residues 41-56) of the B1 domain of protein G.[19] We then look into a 13-residue long fragment of serum amyloid A[20] that is central for the role of this protein in colonic amyloidosis,[45] and we use this fragment to study the role of certain mutations in shifting the equilibrium between helical and β-hairpin configurations. As a more taxing application, we finally compare the landscapes of two mutants of the A and B domains of protein G[21,50,51] that have over 90% sequence identity but still preserve their distinct structures and functions.

6.2 Materials and Methods

Replica exchange sampling aims to achieve faster convergence at a (low) target temperature by enforcing a random walk through temperature space that allows escape out of local minima. For this purpose, protein conformations are exchanged between neighboring temperatures T_i and $T_{j=i+1}$ with a probability

$$w(\ \mathbf{C}^{old} \rightarrow\ \mathbf{C}^{new}) = \min(1, \exp(-\beta_i E(C_j) - \beta_j E(C_i) + \beta_i E(C_i) + \beta_j E(C_j)))\ . \quad (6.1)$$

Despite its successful application in many folding studies, replica exchange sampling is often restricted because the exchange move leads to a proposed state C^{new} of the multiple-replica system that is exponentially suppressed but is transient in the sense that if they are accepted the two replicas would quickly evolve to a state with a

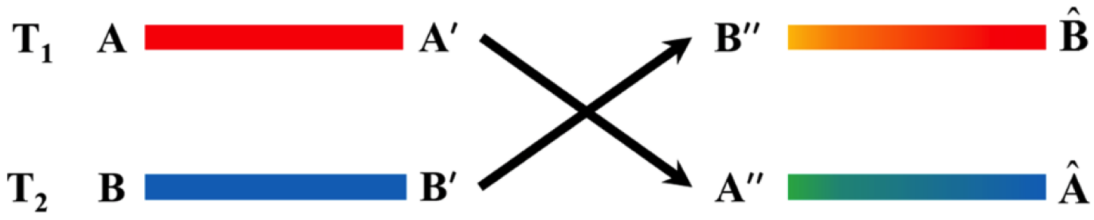

Figure 6.1: Sketch of the replica exchange with tunneling (RET) move.

weight comparable to the one before the exchange move. Recently, we have proposed the tackling of this problem of "tunneling" through the unfavorable "transition state" by a RET approach.[187] RET aims to replace configuration A by configuration $\hat{B}$ at temperature T_1 and configuration B by configuration $\hat{A}$ at temperature T_2; see the sketch in Figure 6.1. Here, $A = (q_A, v_A)$ denotes a state characterized by coordinates q_A of all of its atoms and the associated velocities v_A. The crossing arrows mark the exchange step, and the coloring of the bars indicates the average velocities in the respective systems during the microcanonical segments. Hence, the RET move consists of four parts.

1. In the first part, configuration A (B) evolves by a short microcanonical molecular dynamics run to configuration $A' = (q'_A, v'_A)$ (B') , without which the total energy $E_{tot} = E_{Pot} + E_{kin} = E_1$ (E_2) changes. Similarly, the other replica evolves from state B to state $B' = (q'_B, v'_B)$ while the total energy E_2 stays again constant.

2. The two replicas are now exchanged, and at the same time their velocities are rescaled, such that $E_{tot}(B'') = E_{tot}(A') = E_1$ and $E_{tot}(A'') = E_{tot}(B') = E_2$.

Here, $A'' = (q'_A, v''_A)$, and the velocities are rescaled by

$$v''_A = v'_A \sqrt{\frac{E_2 - E_{pot}(q'_A)}{E_{kin}(v'_A)}} \quad \text{and} \quad v''_B = v'_B \sqrt{\frac{E_1 - E_{pot}(q'_B)}{E_{kin}(v'_B)}} \; . \tag{6.2}$$

Hence, the exchange move generates a "transition state" in the multiple replica system where unfavorable potential energies at the two temperatures are compensated by kinetic energies resulting from the rescaled velocities.

3. In the third step, configurations $A''(B'')$ evolve again by a short microcanonical molecular dynamics run, in which kinetic energy is transformed into potential energy and vice versa, until at temperature T_1 the final configuration $\hat{B} = (\hat{q}_B, \hat{v}_B)$ has a comparable potential energy to configuration A and a velocity distribution typical for T_1. In a similar way, at temperature T_2, configuration $\hat{A} = (\hat{q}_A, \hat{v}_A)$ will have a comparable potential energy as configuration B and a velocity distribution corresponding to T_2. The color coding in Figure 6.1 emphasizes the exchange between potential and kinetic energies in this segment.

4. Finally, the set of configurations $\hat{A}$ and $\hat{B}$ is compared with the set of initial configurations A and B and accepted by a Metropolis step with probability

$$\exp\left(-\beta_1(E_{pot}(\hat{q}_B) - E_{pot}(q_A)) - \beta_2(E_{pot}(\hat{q}_A) - E_{pot}(q_B))\right), \tag{6.3}$$

with $\beta = 1/k_B T$. If rejected, the molecular dynamics simulations will continue at $T_1(T_2)$ with configuration $A(B)$. In both cases, the velocities of the configurations at $T_1(T_2)$ are newly drawn from a distribution corresponding to the respective temperatures.

The acceptance criterium of eq 6.3 in the final step of the RET move is derived by writing the probability to find configurations with potential energy $E_{pot}(q_A)$ and total energy E_1 as

$$P(E_{pot}(q_A), E_1) \propto \Omega(E_{pot}(q_A)) \times E_{kin}^{3N/2}(v_A) = \Omega(E_{pot}(q_A)) \times (E_1 - E_{pot}(q_A))^{3N/2} \tag{6.4}$$

with N being the number of particles and $\Omega(E_{pot}(q_A))$ being the density of states with potential energy E_{pot}. As the total energy at T_1 and T_2 is conserved, the acceptance probability for the RET move is 1. However, the Metropolis-Hastings algorithm, which ensures convergence to the correct distribution, requires the product of acceptance and proposal probability. The latter is the probability to start at temperature $T_1(T_2)$ in a configuration with coordinates $q_A(q_B)$ and picking a configuration with coordinates $\hat{q}_B(\hat{q}_A)$ and is given by

$$\left(\frac{E_1 - E_{pot}(\hat{q}_B)}{E_1 - E_{pot}(q_A)}\right)^{3N/2} \times \left(\frac{E_2 - E_{pot}(\hat{q}_A)}{E_2 - E_{pot}(q_B)}\right)^{3N/2} \tag{6.5}$$

Hence, the Metropolis-Hastings criterium for accepting or rejecting the RET move is in general given by

$$w(C^{old} \rightarrow C^{new}) = min\left(1, \left(\frac{E_1 - E_{pot}(\hat{q}_B)}{E_1 - E_{pot}(q_A)}\right)^{3N/2} \times \left(\frac{E_2 - E_{pot}(\hat{q}_A)}{E_2 - E_{pot}(q_B)}\right)^{3N/2}\right) \tag{6.6}$$

This equation is cumbersome to evaluate. However, as both functions on the right side of eq 6.4 grow strongly with their arguments, the distribution of potential energies $P(E_{pot}, E)$ is a sharply peaked function for large N, and a saddle-point

expansion will lead to

$$P(E_{pot}, E) \propto \Omega(E_{pot}) \exp\left\{-\beta_E E_{pot} - \frac{3N}{2}\left(\frac{E_{pot}-\hat{E}_{pot}}{E-\hat{E}_{pot}}\right)^2 + \mathcal{O}\left[\left(\frac{E_{pot}-\hat{E}_{pot}}{E-\hat{E}_{pot}}\right)^3\right]\right\} \tag{6.7}$$

with the inverse microcanonical temperature $\beta_E = 1/k_B T_E = dln\Omega(E)/dE$ and where $\hat{E}_{pot}$ is the most probable potential energy. Hence, for sufficiently large N and long enough trajectories, the RET acceptance criterion of eq 6.6 reduces to eq 6.3, which can be evaluated more easily.[187]

For the purpose of simulating conformational transitions in systems with competing attractors, we combine RET with exchange moves between "physical" models and models relying on Go-type force fields that bias toward distinct configurational states of a protein. A typical situation would be a switch between two competing structures, α and β. Assuming one knows both states, it is possible to define Go-type force fields biasing toward either state α or β and to introduce exchange moves as sketched in Figure 6.2: Such "feeding" of physical systems by Go-models has been previously proposed by us[188] and others[105,189] as a way to avoid the intrinsic bias in Go-model simulations, but it is combined here with our improved sampling technique. This allows us to overcome the main obstacle that has held back Hamilton replica exchange simulations of hybrid physical/Go model simulations, namely, the low acceptance rates of such moves that exchange configurations between models with substantially varying energy functions. In fact, we will show in our simulations of mutants of the 56-residue A and B domains of protein G that the "feeding" of physical systems by Go-models is impractical without the RET move. Note that while the above sketch assumes use of two single-basin Go-models, one can also

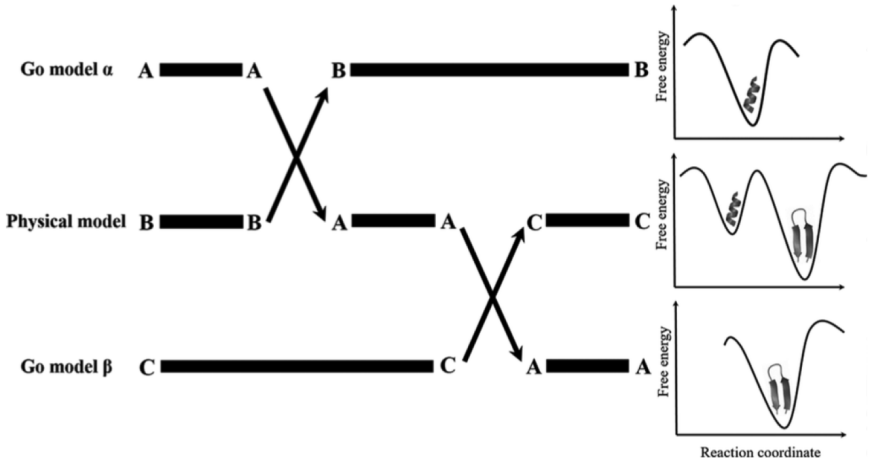

Figure 6.2: RET move "feeding" physical systems by suitable Go-models.

use multibasin Go-models and there is no fundamental restriction in the number of Go-models.

6.3 Implementation and Technical Details

Our simulations rely on an in-house modification of the Gromacs package[174] , version 4.6.5 (available from the authors), and use a potential energy function made of three terms:

$$E_{pot} = E_{phys} + E_{Go} + \lambda E_{\lambda} \tag{6.8}$$

The first term, E_{phys}, is given by the physical interactions between the atoms in the protein as described by Amber 99SB-ILDN[71] or another suitable force field and a generalized Born surface area (GBSA) implicit solvent[177] to approximate the interaction of the protein with the surrounding solvent. As the Go term, we use E_{Go}, the SMOG energy function;[106,107] i.e., given a target structure, we use the SMOG server at http:// smog-server.org to generate a set of parameters for use in GRO-

MACS. The so-generated Go-model is an all-atom model (without hydrogens), uses a van der Waals contact term, defines contacts by a shadow map, and uses no atom charges. In addition, we scale the masses in the Go-model (and the corresponding potential energy terms and forces) such that the temperature scale is the same as in the physical model. The physical model and the Go-model are coupled by a third term, as proposed in ref[104] and tested in ref,[105] with parameter λ describing how strongly the physical model is biased by the Go term. This coupling term has the form

$$E_\lambda = \begin{cases} \frac{1}{2}\left(\Delta^2(i,j)\right) & -ds < \Delta(i,j) < ds \\ A + \frac{B}{\Delta^S(i,j)} + f_{max}\Delta(i,j) & \Delta(i,j) > ds \\ A + \frac{B}{\Delta^S(i,j)}(-1)^S - f_{max}\Delta(i,j) & \Delta(i,j) < -ds \end{cases} \tag{6.9}$$

where $\Delta_{ij} = \delta_{phys}(i,j) - \delta_{go}(i,j)$ and $\delta(i,j)$ is the distance between C_α atoms i and j in the respective model. The two parameters A and B ensure continuity of E_λ at the preselected switching distance $\Delta_{ij} = \pm ds$, f_{max} is the maximum force as $\Delta_{ij} \to \pm\infty$, and parameter S controls how quickly this value is approached. In our simulations, we calculated factors A and B from preset $S = 1, f_{max} = 0$, and ds = 3Å by

$$A = \left(\frac{1}{2} + \frac{1}{S}\right) ds^2 - \left(\frac{1}{S} + 1\right) f_{max} ds \quad and \quad B = \left(\frac{f_{max} - ds}{S}\right) ds^{S+1} \tag{6.10}$$

With this definition, Eq. 8.6 for the rescaling of velocities takes the form

$$v''_A = v'_A \sqrt{\frac{E_2 - E_{phys}(q'_A) - E_{Go}(q'_A) - \lambda_2 E_\lambda(q'_A)}{E_{kin}(v'_A)}}$$

$$v''_B = v'_B \sqrt{\frac{E_1 - E_{phys}(q'_B) - E_{Go}(q'_B) - \lambda_1 E_\lambda(q'_B)}{E_{kin}(v'_B)}} , \tag{6.11}$$

and RET moves are accepted with probability

$$\exp\left(-\beta_1 \left(\Delta E^{(1)}_{phys} + \Delta E^{(1)}_{go} + \lambda_1 \Delta E^{(1)}_{\lambda}\right) - \beta_2 \left(\Delta E^{(2)}_{phys} + \Delta E^{(2)}_{go} + \lambda_2 \Delta E^{(2)}_{\lambda}\right)\right) \tag{6.12}$$

where $\Delta E^{(1)}_{phys} = E_{phys}(\hat{q}_B) - E_{phys}(q_A)$ and $\Delta E^{(2)}_{phys} = E_{phys}(\hat{q}_A) - E_{phys}(q_B)$; $\Delta E^{(i)}_{Go}$ and $\Delta E^{(i)}_{\lambda}$ are defined accordingly.

Because our setup (see the sketch in Figure 6.2) introduces "feeding" of the physical system by two Go-models, for technical reasons we use a setup with two $\lambda = 0$ replicas: one with the physical system and a Go-model leading to structure A and one with the physical system and a Go-model leading to structure B. Physical and Go-model do not interact in both $\lambda = 0$ replicas. Only the configurations of the physical systems are exchanged between the two $\lambda = 0$ replicas. The microcanonical segments in each RET step are 1 ps long to allow for relaxation ("tunneling") before/after the exchange move; i.e., this segment is chosen such that the velocity distribution before and after the RET move is, on average, the same. This is necessary to avoid introducing a bias by the RET move. We chose 50 ps as the time between RET moves, with the temperature controlled by the stochastic v-rescaling method.[132] The cutoff of van der Waals (vdW) and electrostatic interactions is 1.5 nm. Bond lengths of hydrogen atoms are constrained for all residues with the LINCS algorithm[78] in the physical model, where we capped the N- and C-termini with methyl groups. The integration time step in the Leapfrog algorithm is 2 fs.

6.4 Results and Discussions

In a first step, we checked that our new approach does not lead to a bias in protein simulations. As test systems, we have chosen two small peptides with distinct secondary structures. The first one is AFP, a designed 11-residue long peptide with sequence ELLEKLLEKEK[18] that has an experimentally measured helicity of 51% at physiological temperature. The second system is the 16-residue long C-terminus (residues 41-56) of the B domain of protein G,[19] termed BFP here, which is known for form β-hairpins[190,191] with an experimentally determined frequency of 42%.[192] Note that we have simulated the latter system with the OPLS/AA force field[193] instead of AMBER 99SB-ILDN[71] used by us otherwise as this choice allowed us to compare our results for this peptide with earlier work.[191] Solvent contributions are approximated by a generalized Born term.[177]

The two molecules are simulated in a setup where the Go-model contribution biases the system toward either an α-helix or β-hairpin, with a strength that varies with parameter λ for each replica. Physical results are obtained only by analyzing replicas where there is no interaction between physical and Go-model (i.e., $\lambda = 0$). The Go-model energy terms were downloaded from the SMOG server[107] using the two molecules in either an ideal helix or ideal hairpin structure as the input. For the β-hairpin structure of BFP, we extract this structure from the Protein Data Bank structure of the full B domain of protein G (PDB ID: 1GB1). In all other cases, the α-helix or β-hairpin structures are constructed by setting the dihedral angles of the residues to the standard values for the respective models, using the TLEAP module in AmberTools. These structures are then first minimized; afterward, they are heated for 5 ns to 600 K and finally cooled to our target temperature of 310 K

in another 5 ns of explicit solvent molecular dynamics. The so-generated structures define not only the various Go-models but also serve as reference structures for the calculation of root-mean-square deviations.

Go and physical models are coupled by the E_λ term in eq 8.2. Sixteen replicas are used for both peptides. The λ distribution for AFP is (helix) 0.2, 0.1, 0.05, 0.025, 0.0125, 0.00625, 0.003125, 0, 0, 0.003125, 0.00625, 0.0125, 0.025, 0.05, 0.1., and 0.2 (hairpin), and that for BFP is given by (helix) 0.1, 0.05, 0.025, 0.0125, 0.00625, 0.003125, 0.0015625, 0.0, 0.0, 0.0015625, 0.003125, 0.00625, 0.0125, 0.025, 0.05, and 0.1 (hairpin). Our results from the two RET simulations are compared with that of replica exchange molecular dynamics in temperature (T-REMD) using 32 replicas distributed between 300 and 406 K and relying on the same energy functions and setup as described above. The temperature distribution is given by 300.00, 303.05, 306.12, 309.22, 312.34, 315.48, 318.65, 321.84, 325.06, 328.30, 331.56, 334.86, 338.17, 341.52, 344.88, 348.28, 351.70, 355.15, 358.62, 362.12, 365.65, 369.21, 372.79, 376.40, 380.04, 383.71, 387.41, 391.14, 394.89, 398.68, 402.49, and 406.32. Replicas are exchanged here as in the RET simulations every 10 ps. Data are analyzed for the replica closest to T = 310 K and reweighted to that temperature. In all simulations, we follow a replica for 100 ns, but we use only the last 50 ns in our analysis, for which snapshots are stored every 10 ps.

While the two peptides are simulated with a setup in which the peptides walk along a ladder of replicas where on one side the Go-model contribution to the energy biases toward an α-helix and on the opposite side the bias is toward a β-hairpin, the AFP peptide is seen in the physical replica ($\lambda = 0$) with a frequency of 42% in a helical configuration and never as a hairpin. These numbers are comparable to the experimental value of 51% and the one (53%) found in a regular temperature

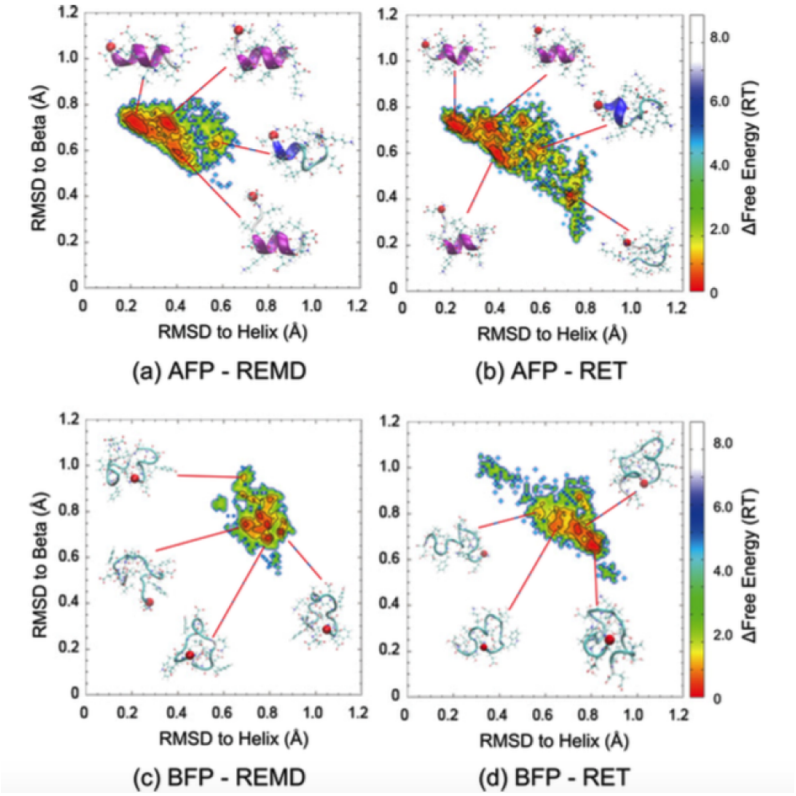

Figure 6.3: Free energy landscape of the 11-residue α-helix forming peptide AFP (a, b) and the 16-residue β-hairpin-forming peptide BFP (c,d) projected on the root-mean-square deviation (RMSD) with respect to the peptide in an ideal helix-configurations and in an ideal β-hairpin configuration. Free energies are given in units of RT. Data were obtained with either regular REMD or our new RET method. Representative configurations are shown for the main basins in the landscape.

replica exchange molecular dynamics simulation of the system. In Figure 6.3a,b, we show the free energy landscape of this peptide projected on the RMSD to the peptide in either an ideal helix or ideal hairpin structure, as obtained with our RET and with regular T-REMD. Similar landscapes for the BFP peptide are shown in Figure 6.3c,d. For this peptide, we find no helical structures at the $\lambda = 0$ replica but hairpin-like configurations with a frequency of 48%, which is also similar to what we find in regular T-REMD simulations (38%) of the peptide. We remark that both

values are also close to the experimentally measured frequency of 42%.[192] In both panels of this figure, we also show typical configurations of the dominating basins in the landscape: in Figure 6.3b, these are helical structures, and in Figure 6.3d, they are β-hairpins.

Hence, we find that despite our setup, which "feeds" the physical replica from replicas biasing to either a helix or a β-turn, we do not find the "wrong" structures in the landscape for either of the two peptides (i.e., we did not find a β-hairpin for AFP or a helix for BFP). This demonstrates that the RET sampling does not introduce biases to the landscape. However, the landscapes derived from RET simulations are broader than the corresponding ones obtained from T-REMD. This broadening reflects the enhanced sampling by our new technique: for such small peptides, we expect broader landscapes than seen in T-REMD simulations.

From a medical point, our second system, namely, the 13-residue long N-terminal fragment of serum amyloid A, is more interesting. This is because the presence of fibrils formed by the first 76 N-terminal residues of serum amyloid A[194] is a hallmark for a common form of colonic amyloidosis.[45,46] Nordling and Abraham-Nordling[20] recently proposed that the N-terminal first 13 residues, usually part of an α-helix, can misfold into a β-hairpin, which in turn may multimerize and act as an anchor for fibril formation. This assumption is consistent with experimental observations that peptides which lack the first 11 residues do not form fibrils.[44] In molecular dynamics simulations of Nordling et al.,[20] the isolated 13-residue fragment lost its helical structure at 300 K after 25 ns and refolded into a β-hairpin during the last 25 ns. We have repeated this simulation at the physiologically more relevant temperature of 310 K, extending it up to 200 ns. The simulation setup is the same as that used above for the AFP and BFP peptides, with the starting configuration

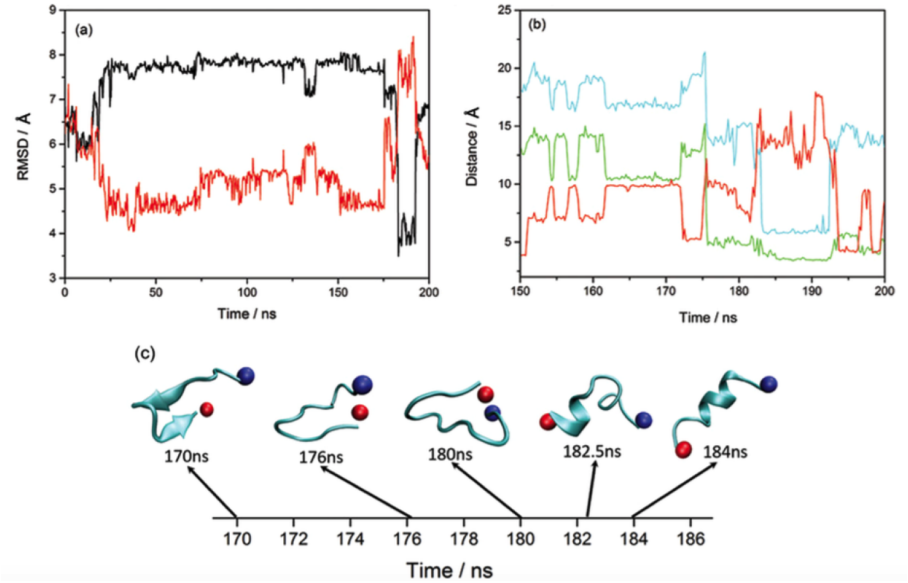

Figure 6.4: Molecular dynamics simulation of a thirteen-residue fragment of the Serum Amyloid A protein at T=310 K, followed over 200 ns. (a) Root-mean-square deviation (RMSD) in Å with respect to the fragment in an ideal helix-configurations (black) and to an ideal β-hairpin (red) as function of time; (b) distances in Å between the positively charged residue 1R and negatively charged residues 5S (green), 9E (blue) and 12D (red) as function of time for an interval where the structure changes; (c) representative configurations during the transition of the molecule from helix to hair pin.

derived from the corresponding fragment of the first entry in the NMR ensemble of the serum amyloid A protein (PDB ID: 4IP8).

In Figure 6.4a, we show a time series of RMSD with respect to the fragment in a helical configuration or measured with respect to the fragment in an ideal hair-pin configuration. Note that, unlike Nordling and Abraham-Nordling, we not only observe the decay of the helix and its reforming as a β-hairpin but also the opposite process, indicating that this peptide can switch between the two forms. These transitions are observed in a time interval between 170 and 200 ns. The transitions between the two structures are associated with the dissolving and reformation of

various salt bridges between the positively charged arginine of the first residue (1R) and the negatively charged serine of residue 5 (S5), glutamic acid of residue 9 (9E), and aspartic acid of residue 12 (12D) (see Figure 6.4b). Going from an α-helix to a β-turn, contacts between residues 1R and 5S and between 1R and 9E dissolve, whereas the appearance of a contact between residues 1R and 12D marks the formation of the hairpin. Representative configurations along the path from a helical configuration to the β-hairpin and back are shown in Figure 6.4c. This evolution of configurations shows how the loss of stabilizing contacts 1R-5S and 1R-9E encourages unwrapping of the α-helix, which in turn allows the two ends to approach each other. Once the peptide finds a U-shaped configuration, contact 1R-12D can be formed, which stabilizes this structure and allows it to evolve into a β-hairpin.

A single transition does not provide sufficient statistics to obtain reliable estimates on the free energy difference between the two forms or on the mechanism of the transition. This is different with our new RET approach, which is designed to enable these transitions. In order to define the Go-model energy term in eq 6.8 by way of the SMOG server,[107] we generate again configurations of the molecule in either an ideal helix or an ideal hairpin structure with the same schedule as described above for the AFP and BFP peptides. Force fields and the setup of the RET simulations are also the same as those for the AFP and BFP molecules studied above, with the exception that replica exchanges are attempted every 50 ps and a total of 18 replicas are used. As for the canonical simulation, we follow a trajectory for 200 ns, but we use only the last 100 ns for our analysis where data are collected every 2 ps. The following distribution of λ parameters is used: (helix) 0.8, 0.4, 0.2, 0.1, 0.05, 0.025, 0.0125, 0.00625, 0.0, 0.0, 0.00625, 0.0125, 0.025, 0.05, 0.1, 0.2, 0.4, 0.8 (hairpin).

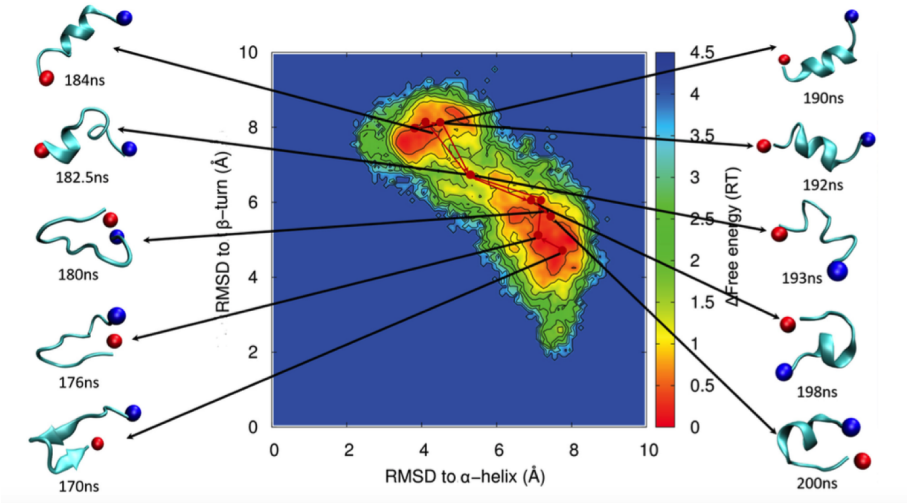

Figure 6.5: Free energy of the Serum Amyloid A fragment as obtained from our RET-simulation projected on the root-mean-square deviation (RMSD) with respect to the fragment in an ideal helix-configurations or in an ideal β-hairpin configuration. Representative configurations are shown for the main basins in the landscape. For comparison, we mark with a red line the transition path from helix to hairpin as seen in our regular molecular dynamics simulation.

The resulting free energy landscape at $\lambda = 0$, i.e., for the physical model without bias by Go-model terms to either a helical or hairpin structure, is shown in Figure 6.5. Here, the landscape is projected on the RMSD with the peptide in either an ideal α-helix (x-axis) or an ideal β-hairpin (y-axis) configuration. The landscape is characterized by two dominant basins, corresponding to the two structures, that are separated by an energy barrier of about 1.5 kcal/mol. About 30% of the configurations are part of the α-helical basin and 40% are part of the β-hairpin (see Table 6.1). These frequencies result from using the RMSD-based default clustering method in GROMACS, applying a cutoff of 3 Å. The α-helical basin is divided in two sub-basins where the smaller ones contain configurations in which the helix lacks the terminal residues and the peptide starts to bend into a U-shape. Note that the transition path seen in the regular molecular dynamics simulation first goes through

		Mutants				
Structure	Wild type	R1A	D12A	S5A	E9A	S5A/E9A
α-helix	30 (4)%	25 (5) %	39 (3)%	30 (4) %	25 (5)%	17 (4) %
β-hairpin	40 (6)%	43 (6) %	31 (6)%	43 (6) %	44 (5)%	53 (4) %

Table 6.1: Frequency of α-Helical and β-Hairpin Configurations of Wild-Type Serum Amyloid A Fragment (1-13) and Five of Its Mutants

this sub-basin. This path and the representative configurations in the two basins and the transition region are also shown in the figure.

Figure 6.4 and 6.5 indicate that the transition between the two structures (and their corresponding basins in the free energy landscape) is correlated with the dissociation and reformation of certain salt bridges. A salt bridge between residues 1R and 5S and between 1R and 9E stabilizes the α-helical structure, whereas the β-hairpin appears to be stabilized by a salt bridge between the side chains of 1R and 12D. To verify the role of these residues, we studied five mutants of the fragment. The first four mutants are single point mutations where one of the charged residues, 1R, 5S, 9E, or 12D, is replaced by alanine: R1A, S5A, E9A, and D12A. In the fifth mutant, the two helix-stabilizing charged residues 5S and 9E are both replaced by alanine: S5A/9EA. These five mutants are simulated with the same protocol as that used for the wild type. For analyzing these simulations, it is important to find suitable coordinates onto which to project the resulting high-dimensional landscapes. Finding such coordinates is especially important for determine the physics of small and flexible peptides; see, for instance, ref.[195] From the above discussion, it would appear that the distances between residues 1 and 5 or 9 and between residues 1 and 12 would be natural coordinates, and we also present the free energy landscapes projected on these coordinates as Figure S1 of supplementary material.[196] However, we found that it was more suitable to project the landscapes on the RMSD to the

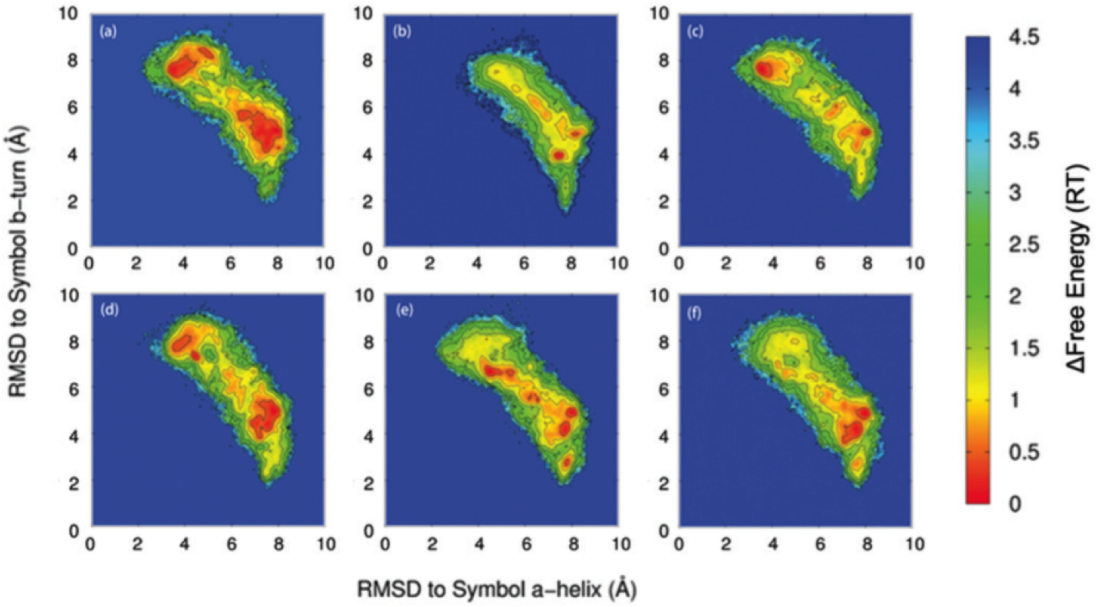

Figure 6.6: Free energy landscape of the wild type Serum Amyloid A fragment (a) and the five mutants R1A (b), D12A (c), S5A (d), E9a (e), and S5a/E9A (f) as obtained from our RET-simulation. The landscapes are projected on the root-mean-square deviation (RMSD) with respect to the fragment in an ideal helix-configurations and to an ideal β-hairpin configuration.

peptide in an ideal helical configuration or to an ideal hairpin configuration. These landscapes are displayed in Figure 6.6b-e. For comparison, we show the wild-type landscape in this figure (Figure 6.6a).

From these six landscapes, we see that the charged residues indeed play an important role in the transition process as the landscapes change dramatically with the mutations. In the wild type, the first residue is an arginine. This is the only positively charged residue in the peptide. Hence, in the R1A mutant, no salt bridge can be formed. While the mutant can still take both forms, there is no clear barrier separating them. The situation is different for the second mutant, where the aspartic acid of residue 12 is mutated into alanine. As we assume that the 1R-12D

contact is stabilizing the β-hairpin configuration, we expect that this mutation leads to a less-populated β-hairpin basin while leaving the α-helix basin in the landscape unchanged. The opposite is expected for the two single mutants, S5A and E9A, where a serine (residue 5) or a glutamic acid (residue 9) is mutated to alanine. Both mutants indeed lead to a less-populated α-helix basin, with the effect being more pronounced for E9A than for S5A. Finally, the reduction in helicity is the highest in the S5A/E9A double mutant, where the β-hairpin configurations are now dominating. While our results from simulations of our small fragment cannot be transferred directly to the full serum amyloid A protein, they suggest two possible mechanisms for potential drug candidates that target colonic amyloidosis: either by stabilizing the α-helix through protecting contacts 1R-5S and 1R-9E or by suppressing formation of the 1R-12D contact, which would destabilize the β-hairpin configuration. However, such drug candidates would have to be studied for the full protein and not only our 11-residue fragment, which is too small to show binding pockets. This goes beyond the purpose of this study, namely, to demonstrate that the RET approach is a suitable tool for the study of switching processes in proteins.

A classical example of fold switching in proteins is the series of mutation experiments by Orban, Bryan, and co-workers,[21,50,51] which led to mutants of the A and B domains of protein G (GA and GB) that have over 90% sequence identity but differ in structure and function. The extreme cases are mutants GA98 and GB98 that vary only by having residue 45L in GA98, which is 45Y in GB98. GB98 assumes the same fold as wild-type GB, and GA98 overwhelmingly possesses the GA fold; however, the competing GB structure is also observed with a low frequency in the experiments on GA98.[50]

Catching the difference between the two mutants in numerical studies is a chal-

lenge. The folding mechanism of the two proteins has been studied in simulations with modified Go-models that introduce a bias to two structures instead of only one,[197,198] but to our knowledge, there exists no all-atom simulations with a physics-based force field that could successfully derive the difference between the two mutants. Additionally, in recent papers,[199,200] it was claimed that present force fields are not accurate enough to explain why these almost identical sequences fold into very distinct structures. The question arises as to whether these previous failures indeed indicate that current force fields do not describe the landscape of these proteins with sufficient accuracy or whether they do lead to a landscape with the correct topology but where the roughness is too large to allow for the necessary broad sampling. If the previous failure of molecular dynamics simulations with physical force fields results from a sampling problem instead of the principle limitations of today's force fields, then it may be overcome, or at least alleviated, by our RET method. We have therefore chosen simulations of the GA98 and GB98 mutants as our last and most taxing application of our new sampling approach.

Randomized start configurations for our simulations are generated from short (100 ps) molecular dynamics runs at an unphysical high-temperature of 3000 K that are relaxed before the E_λ term is turned on and the systems are cooled over 1 ns to the target temperature of 310 K. The RET simulations of the GA98 and GB98 mutants use the same setup as described in the Materials and Methods, with the distribution of λ values describing the coupling between the physical and Go-models, given by (GA-fold) 0.4, 0.2, 0.1, 0.075, 0.055, 0.035, 0.028, 0.015, 0.01, 0.005, 0.0, 0.0, 0.005, 0.01, 0.015, 0.028, 0.035, 0.055, 0.075, 0.1, 0.2, 04 (GB-fold). However, in a slight variation from the previous examples, the two $\lambda = 0$ replicas do not exchange the configurations of the "physical" model. Instead, we have set up a

third replica without any Go-model that is fed by the two $\lambda = 0$ according to a heat bath sampling move. In this way, we can reduce the distance that a replica has to walk in a random process in λ space by a factor of 2, leading to a 4-fold reduction in computational cost. A total of 100 ns is sampled in two independent trajectories, with a total of 60 ns used for analysis. Within these 60 ns, we observed 45 tunneling events in the GA98 simulation and 54 in the corresponding simulation of GB98. Here, a tunneling event is defined as a replica crossing the whole range from $\lambda = 0$ to λ_{max} and back. We remark that we also performed additional short simulations of the two systems that rely on the same energy function of eq 6.8 and setup as the RET simulations but that use a regular exchange move, i.e., configurations are accepted or rejected with a probability $exp(\beta\Delta\lambda\Delta E_\lambda)$. In these simulations, we did not observe a single tunneling event for either of the two systems. The differences in tunneling times is correlated with the much lower average acceptance rates for exchange moves in these simulations: 16% compared with approximately 50% in the RET simulations.

The faster walk through λ space leads to an enhanced sampling of configurations at $\lambda = 0$, where the "physical" model is not biased by Go-model contributions. As a consequence of this improved sampling, we are able to show in Figures 6.7 and 6.8 the free energy landscapes of GA98 and GB98, projected on the RMSD to the GA- and GB-fold as order parameters. We also display in these two figures typical configurations for the main basins in the two landscapes.

In agreement with previous simulations, we find in Figure 6.7 that for GA98 both the GA- and GB-folds are populated and that the GA-fold is favored by roughly $3k_BT$ in free energy; i.e., at T = 310 K, the GA-fold is the dominant structure. The corresponding basin in the landscape is composed of ≈60% of sampled con-

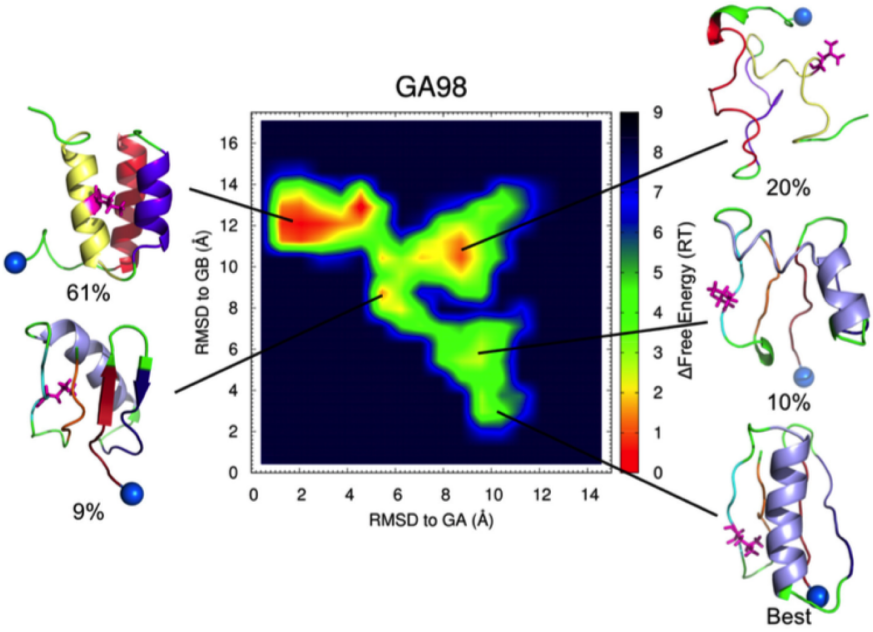

Figure 6.7: Free energy landscape of the GA98 mutant as obtained from our RET-simulation and projected on the root-mean-square deviation (RMSD) with respect to the GA98 crystal structure (x-axis) and GB98 crystal structure (y-axis). Representative structures for the main basins are shown together with the frequency of these structures. We also show the lowest-energy structure found in our simulations and mark its position in the landscape.

figurations, all built by three α-helices packed together in the GA-fold. Residue 45L packs well, with no sign of steric clashes that may prevent helix packing, but it forms various long-range contacts with residues on the central and C-terminal helixes. For instance, 74% of configurations have a 45L-33I contact. Following a pseudotrajectory out of the basin, the central helix dissolves first, followed by the terminal helices. This is similar to what has been observed earlier by Kouza et al.[197] However, the landscape is more rugged than the one observed in previous simulations relying on a modified Go-model that interpolated between a GA- and GB-fold,[197] and as a consequence, we do not see a continuous decrease in free energy

when moving from the GB-fold to the GA-fold. One of the intermediates, a basin populated by about 9% of configurations, resembles a mirror structure of the GB-fold with an incorrectly placed α-helix and only partially formed strands S1 and S2. The two Go-models only slightly favor the correct fold over the mirror structures; therefore, we cannot exclude the possibility that the observed mirror structures at the "physical" $\lambda = 0$ replica (where there are no Go-model contributions) are artifacts of our search procedure. However, we think that this is unlikely as mirror structures of the GA-fold are observed only with a much smaller frequency than that of the GB-fold. The appearance of mirror structures has been also observed previously in protein simulations with physical force fields and is discussed in ref.[201] The GB-fold is populated by about 10% of configurations; however, the β-sheets are often formed poorly and the helix is broken or bent to form a contact with 45L. Unlike in Kouza et al.,[197] we find only an insignificant propensity in the configurations of this cluster to form 45L-23A contacts.

The differences in the landscapes between our RET simulation, shown in Figure 6.8, and the previous Go-model simulations of ref[197] are larger for the GB98 mutant. In our simulations, both the GA- and GB-folds are observed with similar frequencies, and the GB-fold is lower in free energy by less than $0.5k_BT$. This is in contrast to experiments where one observes GB98 only in the GB-fold. In the previous Go-model simulations of Kouza et al.,[197] configurations of GB98 in the GA-fold were observed, but the GB-fold is much more highly populated than the GA-fold and is separated from it by a high barrier of about $6k_BT$. The authors argued that this high barrier, separating disordered configurations with the GB-fold from configurations with the GA-fold, makes it unlikely that the GA-fold can be observed in experiments. We see the same type of barrier in the GB98 landscape obtained in our RET simulations

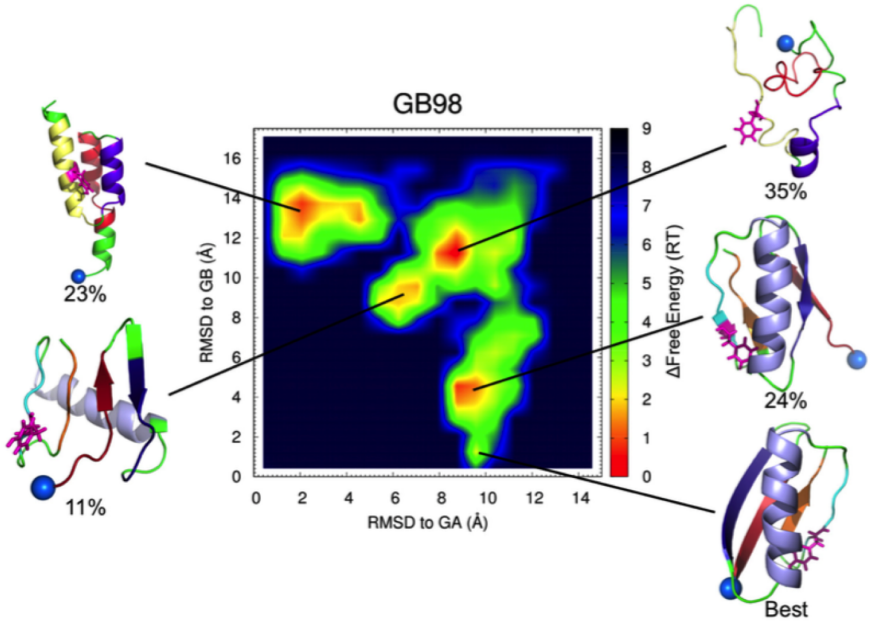

Figure 6.8: Free energy landscape of the GB98 mutant as obtained from our RET-simulation and projected on the root-mean-square deviation (RMSD) with respect to the GA98 crystal structure (x-axis) and GB98 crystal structure (y-axis). Representative structures for the main basins are shown together with the frequency of these structures. We also show the lowest-energy structure found in our simulations and mark its position in the landscape.

and shown in Figure 6.8. While about 23% of configurations have the GA-fold, this basin is separated by barriers of about $6 - 8k_BT$ separated from the reminder of the protein landscape, which again suggests that for the GB98 mutant the GA-fold is thermodynamically possible but kinetically difficult to access. Unlike Kouza et al.,[197] we find that contacts between 45Y and 33I appear with the same frequency of 74% as the 45L-33I contact in GA98 configurations having the GA-fold. On the other hand, the GB-fold is populated by about 24% of configurations, with both the helix and the four strands well-formed and 45Y-23A formed, unlike 45L-23A in GA98 configurations with the GB-fold. Following a pseudotrajectory into this basin, visual inspection of the configurations suggests that the α-helix forms first,

followed by strands S3 and S4 arranging as the C-terminal β-sheet, before at last the N-terminal strand forms and builds a β-sheet. Note that the mirror structure of the GB-fold is again formed and appears with a frequency of 11%; however, far fewer contacts are formed with 45Y than in GA98 with 45L. This is because, unlike leucine, the tyrosine is positioned on the rear face of the β-sheets facing away from the helix. Thus, there are no contacts between tyrosine 45 and the helix, but there are still contacts with β-strands S1, S2, and S4. Another major cluster found in the GB free energy landscape shows the two central β-sheets slightly opened. We speculate a dynamic opening and closing of these sheets in which the α-helix could transition through, thus changing faces. This second GB-fold is found for both GA98 and GB98 but to a much lesser extent in GA98.

6.5 Conclusions

We have introduced RET as a way to increase sampling in simulations of systems with competing attractors. The method works by introducing a random walk along a coordinate λ that describes the strength with which the system is biased toward one of the attractors. Movement along this coordinate is enhanced by the RET move, which allow the system to tunnel through unfavorable transition states generated by the exchange move. As in similar Hamilton replica exchange methods that aim to speed up simulation by combining Go-model-like terms with "physical" energy functions,[105,188] only the $\lambda = 0$ replica, i.e., the one where there is no bias from Go-models on the physical force field, is used for data generation and analysis.

Simulating two small peptides, we have shown that this approach does not introduce a bias into the simulations: while both peptides are simulated with the same setup in which the "physical" replica ($\lambda = 0$) exchanges both with replicas where

the Go term biases toward an α-helix and with replicas where the bias is toward a β-hairpin, the helix-forming peptide is observed in the physical ($\lambda = 0$) replica only in a helix configuration and the hairpin-forming peptide is observed only in hairpin configurations.

We have used our sampling approach to study the role of certain salt bridges on the transition between helix and sheet states in a 13-residue fragment of serum amyloid A that is implicated in colonic amyloidosis. While our results from simulations of a small fragment cannot be transferred directly to the full protein, they suggest that residue 12D is a target for potential drug candidates. Our second application is computationally more challenging. In it we compare the 56-residue long GA98 and GB98 proteins, two proteins that differ in only one residue but take very different structures. Unlike previous physics-based all-atom simulations, which failed to reproduce these differences, we find very different landscapes for these proteins, consistent with the experiments. This is the more astonishing as our simulations approximate the protein-solvent interaction by an implicit solvent model. This suggests that the previous difficulties in simulating these two proteins reported in recent papers[199,200] are not so much due to insufficient accuracy of the force fields (as was claimed) but incomplete sampling. This gives us hope that despite limitations in the present generation of force fields it will be possible with RET to study the conformational changes in switching proteins and intrinsically disordered proteins in order to understand how mutations, changes in environment, or interaction with other molecules result in switching between the encoded structures. As a step in this direction, we have now started RET simulations of two small switching proteins, the transcription factor RfaH[22] and lymphotactin.[179] These systems will also allow a direct comparison with experimental results. Other simulations are under way that

try to quantify the scaling of our approach with system size.

A principal shortcoming of our approach is that it requires knowledge of the structures into which a system folds or between which it transitions. This is because knowledge of these structures is needed to define the biasing Go-models for the replicas with $\lambda \neq 0$. There are quite a number of problems where such structures are known but the transitions between them (or folding pathways to them) are not; therefore, our approach already has sufficient utility to RET to make it a useful tool in simulations of biological molecules. We remark that the need for knowledge of the structures is not a principal limitation of our approach. One possibility to avoid this shortcoming would be the use of (nonstructure-based) coarse-grained models instead of the Go-models utilized in the present version. We have started to test such an approach for predicting conformational changes and new structures in the context of protein design.

6.6 Acknowledgments

The simulations in this work were done using XSEDE resources funded under project MCB160005 and the BOOMER cluster of the University of Oklahoma. We acknowledge financial support from NSF CHE-1266256 and OCAST HR14-129.

Chapter 7: Multi-Funnel Landscape of the Fold-Switching Protein RfaH-CTD

The following chapter was published in The Journal of Physical Chemistry B by the author of this dissertation as the following article: N. A. Bernhardt and U. H. E. Hansmann. Multifunnel landscape of the fold-switching protein rfah-ctd. J Phys Chem B, 122:1600–1607, 2018. All text and figures are taken with the permission of the publisher.

Author Contributions: All work presented in this chapter is credited to the author of this dissertation.

7.1 Introduction

Proteins play a central role in the biochemistry of cells, participating in transcription, cell signaling, migration and muscle movement to name only a few of their roles. Protein function is correlated with the molecule assuming a specific three-dimensional shape, but the process by that a protein folds into a certain structure is not known in all details, and depends not only on the sequence of amino acids (the chemical composition of a protein) but also on environment and interaction with other molecules. In the standard model of protein folding, one assumes that a protein has a funnel-shaped energy landscape[5,6] that guides a multitude of possible folding pathways into a unique structure where the protein is biologically active. However, such a single-funnel picture cannot describe all aspects of folding for all proteins. For instance, intrinsically disordered proteins[202–204] do not have a defined structure, but may assume one when interacting with other proteins, with the struc-

ture changing with the binding partner. In other cases, proteins exist in an ensemble of different (but defined) structures[178–180] which may allow proteins to have more than one function in the cell. In these cases we would expect a multi-funnel shaped folding landscape.

Take as an example RfaH,[22] a protein that triggers gene expression in Escherichia coli by switching the structure of the C-terminal domain from an α-helical hairpin (PDB-ID: 2OUG) to a β-barrel (PDB-ID: 2LCL), see Figure 7.1. In the first form, stabilized by interaction with the N-terminus, the C-terminal domain masks a RNA polymerase binding site on the N-terminal domain thus regulating transcription. However, when not in contact with the N-terminal domain, or as an isolated protein, the C-terminal domain spontaneously rearranges into a β-barrel (Figure 7.1). In this form, RfaH binds directly to the ribosomal protein S10, thus recruiting the prokaryotic ribosomal 30S subunit to the elongating RNA and promoting translation. Hence, the fold switch of the C-terminal domain alters dramatically the function of the RfaH protein. Both folds are encoded in the sequence of the C-terminal domain, and it is the interaction (or lack of interaction) with the N-terminus of RfaH that selects the fold. Hence, we would expect a double-funneled landscape for the isolated 66-residue large C-terminal domain of RfaH (RfaH-CDT), with one funnel leading to the β-barrel, and the secondary funnel leading to the α-helical hairpin.[57] The rather small size, the experimentally observed fold switching, and the resolved structures of the two folds make RfaH-CTD an ideal model to study the factors that determine protein plasticity and the mechanism of fold switching in proteins.

However, probing such fold switching and mapping their energy landscape by experiments or *in silico* is a challenge.[57,58,60,61,205] Computationally, the problem is that the exploration of the ensemble of possible structures and the conversion

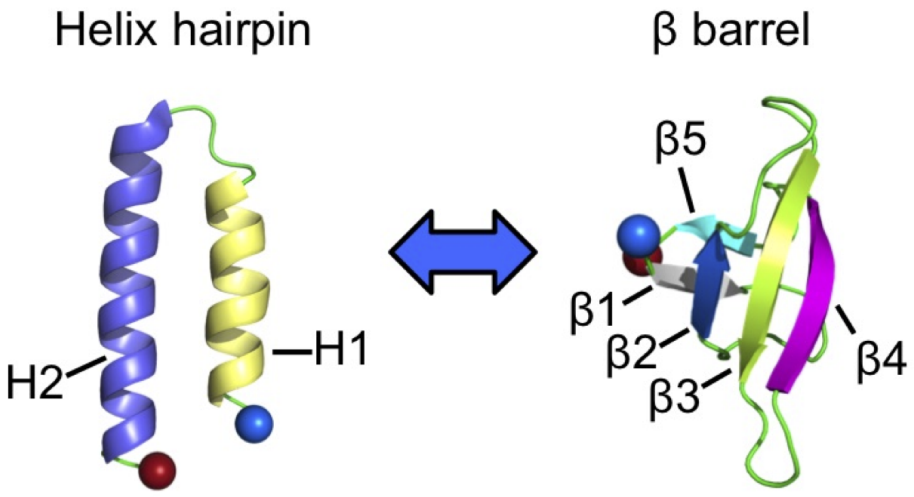

Figure 7.1: The two folds observed for the C-terminal domain (RfaH-CTD) of the transcription factor RfaH. The N-terminus of RfaH-CTD is marked by a blue ball and the C-terminus a red ball.

between these structures happens on timescales that, on general-purpose computers, are not accessible in all-atom molecular dynamics simulations with explicit solvent. Enhanced sampling techniques such as Replica Exchange Molecular Dynamics (REMD)[85–87,160,163,181] promise to overcome this problem by realizing a random walk in temperature which allows the system to escape out of traps and cross barriers by explorations to higher temperatures. However, the sampling efficiency of REMD is often below the theoretical maximum. One problem is that the probability for a replica exchange depends on the temperature spacing which shrinks dramatically with system size. Hence, with the inclusion of water molecules one needs even for small proteins a huge number of replica. This often makes REMD simulations of proteins with explicit solvent impractical, and previous REMD simulations of RfaH[205] had for this reason to rely on an implicit solvent. While these simulations allowed the authors to propose a transition pathway between the two folds, choice of an implicit solvent is not without problems. While the helix-hairpin was found, the

simulation was unable to completely fold the β-barrel. Only when continuing the simulations from the best configurations by including solvent molecules explicitly was the correct β-barrel fold found.[205]

We have recently proposed to overcome some of the limitations that hold back REMD by a Replica-Exchange-with-Tunneling (RET) approach.[187] We have shown that RET in conjunction with a Hamilton-Replica-Exchange[89,206] of systems where the "physical" system is coupled to varying degrees with biasing Go-models, allows efficient simulation of proteins and protein assemblies that can take more than one state.[196,207] For instance, we have used this approach in a a recent study[196] of the two mutants GA98 and GB98, which differ only in a single residue but keep the distinct original folds of the A and B domain of protein G, GA and GB.[21,50,51] We have shown how the mutation leading from GA98 to GB98 alters the energy landscape leading to selection of one fold over the other.[196] While there exist alternative techniques, for instance, double-Lorentzian restraints,[208] to enhance transitions between configurations, we use in the present work the above discussed approach, with which we are more familiar, to explore the folding and switching landscape of the C-terminal domain of RfaH, Rfah-CTD, and propose a conversion process that connects the two forms.

7.2 Materials and Methods

When simulating conformational transitions in systems with competing attractors as in the case of RfaH-CDT (an α-helical hairpin and a β-barrel), one way to enhance transitions between the two attractors is to utilize exchange moves between "physical" models and such relying on Go-type force fields that bias toward one or the other of the competing configurational states. Go-models are defined by energy

functions that depend directly on the similarity to a pre-selected structure, and therefore lead to a smooth energy landscape with a single funnel located around the target fold. As a consequence, Go-models fold proteins quickly, but are by construction unable to capture accurately the energetics of non-native folds. Thus, in an effort to exploit the quick folding of Go-models but remove the associated bias against non-native folds, one can design a Hamilton Replica Exchange Method[89] where at each replica a "physical" model is "fed" by a Go-model, but where the bias differs for each replica.[105,188,207] In the present implementation, the physical and the Go-model are coupled through a potential that depends on the similarity between configurations in the two models[104,105]

$$E_\lambda = \begin{cases} \frac{1}{2}\left(\Delta^2(i,j)\right) & -ds < \Delta(i,j) < ds \\ A + \frac{B}{\Delta^S(i,j)} + f_{max}\Delta(i,j) & \Delta(i,j) > ds \\ A + \frac{B}{\Delta^S(i,j)}(-1)^S - f_{max}\Delta(i,j) & \Delta(i,j) < -ds \end{cases} \tag{7.1}$$

such that $\Delta(i,j)$ is the difference in distances between alpha carbons i and j in the respective models, f_{max} is a parameter that controls the maximum force as $\Delta(i,j) \to \infty$, and S controls how fast this value is reached. The parameters A and B are set so that the potential and its first derivative are continuous at values of $\Delta(i,j) = \pm ds$, and are expressed as

$$A = \left(\frac{1}{2} + \frac{1}{S}\right) ds^2 - \left(\frac{1}{S} + 1\right) f_{max} ds \quad and \quad B = \left(\frac{f_{max} - ds}{S}\right) ds^{S+1} \tag{7.2}$$

Thus the total potential energy of the system is

$$E_{pot} = E_{phy} + E_{go} + \lambda E_\lambda \tag{7.3}$$

where λ controls how strongly the physical and Go-models are coupled. Hamilton Replica Exchange now introduces a random walk in λ-space, with data to be analyzed only from the replica where the bias from the Go-model on the physical model vanishes, i.e., $\lambda = 0$.

However, exchange rates are often low in such an approach. This is a common problem in replica-exchange sampling[85,86,163] which in its standard implementation aims to enforce a random walk in temperature as a way to escape out of local minima in order to achieve faster convergence at a (low) target temperature. Unfortunately, the exchange move between neighboring temperatures often leads to a proposal state that is exponentially suppressed, but if accepted the multiple replica system quickly relaxes to a state of comparable probability to that before the exchange. As a way to overcome this bottleneck and tunnel through the unfavorable transition we have recently introduced Replica-Exchange-with-Tunneling (RET)[187,196,207] by the following four-step-procedure:

1. In the first step, the configurations $A(B)$ evolve on two neighboring replica over a short microcanonical molecular dynamics trajectory to configurations $A'(B')$, without that the total energies E_1 and E_2 change on the two replicas. However, there will be an exchange between potential and kinetic energy on each replica.

2. Next, the configurations A' and B' are exchanged, and the velocities are rescaled according to the following equations such that the energies remains constant before and after the exchange: $E'_1 = E_1$ and $E'_2 = E_2$.

$$v''_A = v'_A \sqrt{\frac{E_2 - E_{pot}(q'_A)}{E_{kin}(v'_A)}} \quad \text{and} \quad v''_B = v'_B \sqrt{\frac{E_1 - E_{pot}(q'_B)}{E_{kin}(v'_B)}} \tag{7.4}$$

3. After the exchange, the two replica evolve again by microcanonical molecular dynamics. While the total energies E_1 and E_2 on the two replica do not change, the exchange between potential and kinetic energy will lead to final states $\hat{B}$ on replica 1 and $\hat{A}$ on replica 2 that have potential energies comparable to the corresponding configurations before the exchange move, and velocity distributions as one would expect for the given temperatures at each replica.

4. The final configurations on each replica are either accepted or rejected according to the following Metropolis criterion

$$\exp\left(-\beta_1(E_{pot}(\hat{q}_B) - E_{pot}(q_A)) - \beta_2(E_{pot}(\hat{q}_A) - E_{pot}(q_B))\right), \qquad (7.5)$$

with $\beta = 1/k_B T$. If rejected, molecular dynamics simulations continue with the original configurations $A(B)$. However, in both cases, new velocity distributions are randomly drawn according to the temperatures on the respective replica.

For more details on this approach and its limitations, see Refs 187 and 196.

In the present study we use RET moves to overcome the problem of low exchange rates in our above described set-up. Velocities are rescaled according to

$$\begin{aligned} v''_A &= v'_A \sqrt{\frac{E_2 - E_{phys}(q'_A) - E_{Go}(q'_A) - \lambda_2 E_\lambda(q'_A)}{E_{kin}(v'_A)}} \\ v''_B &= v'_B \sqrt{\frac{E_1 - E_{phys}(q'_B) - E_{Go}(q'_B) - \lambda_1 E_\lambda(q'_B)}{E_{kin}(v'_B)}}\,, \end{aligned} \qquad (7.6)$$

and RET moves are accepted with probability

$$\exp\left(-\beta_1\left(\Delta E^{(1)}_{phys}+\Delta E^{(1)}_{go}+\lambda_1\Delta E^{(1)}_{\lambda}\right)-\beta_2\left(\Delta E^{(2)}_{phys}+\Delta E^{(2)}_{go}+\lambda_2\Delta E^{(2)}_{\lambda}\right)\right) \tag{7.7}$$

where $\Delta E^{(1)}_{phys}=E_{phys}(\hat{q}_B)-E_{phys}(q_A)$ and $\Delta E^{(2)}_{phys}=E_{phys}(\hat{q}_A)-E_{phys}(q_B)$; $\Delta E^{(i)}_{Go}$ and $\Delta E^{(i)}_{\lambda}$ are defined accordingly.

In the present example we have two ladders of replica, each covering a range from $\lambda=0$ to a value $\lambda=\lambda_{max}$. In the one ladder, replicas walk between a system with no bias on the physical model to one where there is maximal bias toward the α-helix hairpin; in the other the biasing is toward the β-barrel. The two $\lambda=0$ replica do not exchange configurations but serve as reservoirs from that a canonical simulation at the same temperature is "fed" by a heat bath move.

We consider in our simulations not the full-length RfaH protein but only the 48-residue-long C-terminal domain, RfaH-CTD. The RfaH-CTD protein is capped at the N-terminus by an acetyl group, and by a methylamine group at the C-terminus. Prior to running large scale simulations initial structures were randomized by high temperature molecular dynamics (T = 3500 K). Replicas were then brought to their initial lambda values in a short preproduction run. A total of 25 replicas with a lambda distribution of λ = 0.4, 0.2, 0.1, 0.075, 0.055, 0.035, 0.028, 0.023, 0.015, 0.010, 0.005, 0.00, 0.00, 0.00, 0.005, 0.010, 0.015, 0.023, 0.028, 0.035, 0.055, 0.075, 0.1, 0.2 and 0.4 is used. The E_λ energy function of Eq. 8.2 is parametrized with $ds=0.3$Å, $S=1$ and $f_{max}=0$. Data were generated over 100 ns trajectories using an in house version of GROMACS 4.6.5[174] (available from the authors on request), modified to accommodate RET sampling and the Go-model feeding. Potential energy calculations relied on the CHARM36 force field[72] in combination with a GBSA

implicit solvent[177] for the physical model and the smog energy function[106,107] for the Go-model (using the online SMOG server at http:// smog-server.org). Equations of motion are integrated by a leapfrog integrator with a time step of 2 fs, which requires use of the linear constraint solver (LINCS)[78] for constraining hydrogen and heavy atom bond distances. A plain cutoff of 1.5 nm was used for treatment of electrostatics, and the v-rescale thermostat[132] is used to keep the temperature at 310 K.

7.3 Results and Discussions

We start our analysis by first testing that our simulations have converged. For this purpose, we have calculated the later discussed free energy landscape for different intervals of the 100 ns trajectory, see figure S1 of supplemental material.[209] Comparing these landscapes we see that our trajectory has converged after 30 ns, and therefore use the last 70 ns for our analysis. Within this time interval, we observe an average exchange rate between neighboring replicas, with individual rates listed in table S1 (supplemental material[209]), of $26 \pm 3\%$, a value that is similar to the one seen by us in previous RET simulations where we also showed that regular Hamilton Exchange Replica Exchange let to lower rates (especially around $\lambda = 0$) if the same number of replica is used.[187,196,207] As a consequence, replica can walk on both sides of the ladder between replica with $\lambda = \lambda_{max}$ where the physical model is biased strongly by the corresponding Go-model, and $\lambda = 0$ where the physical model is not biased by the Go-model. The number of walks between the two extreme values (called by us tunneling events) are a measure for the quality of simulation. In the present study, we observe a total of 34 tunneling events, with examples shown in Figure 7.2, a value that in our previous work indicated that our simulations had sampled suf-

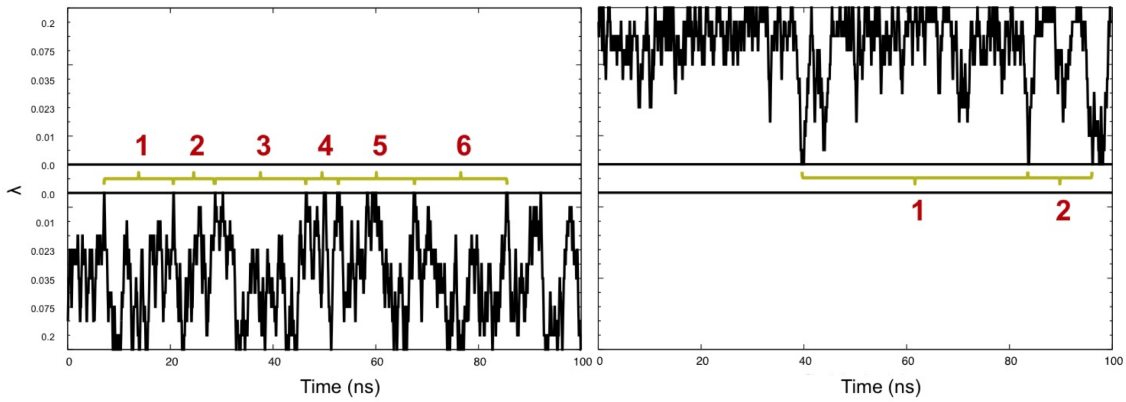

Figure 7.2: Example of replica walking through λ space, from with a Go-model biasing toward the helix hairpin to a replica with $\lambda = 0$ (no bias) (left), and from $\lambda = 0$ toward replica with a Go-model biasing to the β-barrel (right). Tunneling events are numbered in red. Horizontal black lines mark the $\lambda = 0$ replica.

ficient statistics. We remark that we saw in previous simulations[187,196,207] always much higher numbers of tunneling events when using RET exchange moves than in regular Hamilton Exchange Replica Exchange with the same number of replicas and the same λ distribution, reflecting the superior sampling that results from the RET move.

Note that the observed tunneling events cannot be interpreted as folding events leading to either the helix-hairpin or the β-barrel state as our RET simulations rely on an artificial dynamics. This is a common problem in all generalized-ensemble and replica-exchange simulations, but one that can be circumvented by reconstructing the free energy landscape of the system under consideration. We show in Figure 7.3 this landscape projected on the root-mean-square-deviation to either the helix hairpin structure (x-axis) or the β-barrel (y-axis). Bin sizes were chosen as 0.8 angstroms, a value smaller than the maximal root-mean-square deviation between models of the 2LCL NMR ensemble, and the landscape was smoothen to interpolate between bins. Note, that this landscape is derived only from the unbiased replica,

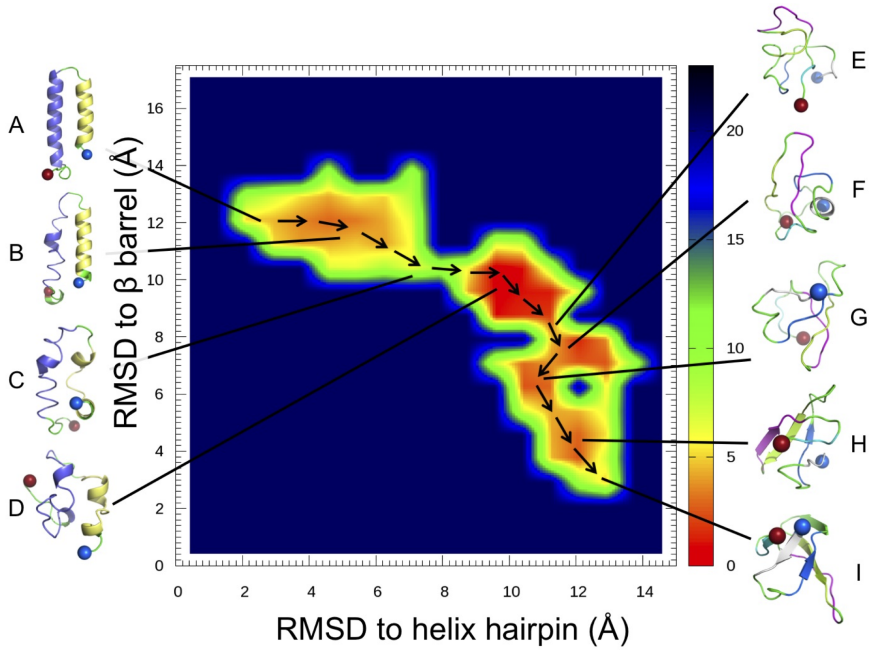

Figure 7.3: Free energy landscape, in units of RT, of the switching protein RfaH-CTD projected on the root-mean-square deviation (RMSD) with respect to the helix hairpin structure of the protein and with respect to the protein in the β-barrel form. Representative configurations are shown for the main basins in the landscape.

i.e. the one that has no contribution from a Go-term but is "fed" by the two sides of the ladder of replica, on one side is the physical model biased by the Go-term with varying degrees toward the helix-hairpin, and on the other side is the bias is toward the β-barrel.

Besides the landscape we show also in Figure 7.3 representative structures for the various regions, labeled A to I. Visual inspection and clustering analysis[210] of the landscape indicates that the β-barrel state (region H and I) is the preferred fold of RfaH-CTD, with about 21% of all configurations in the β-barrel form. However, the bound state state (region A and B) is also significantly populated, with roughly 6% of configurations in the helix hairpin state. Both folds differ by only approximately 2 RT in free energy, but are separated by a barrier of at least 10 RT. The majority of

sampled configurations, 73 %, are either disordered or not representative of either fold. Visual inspection of the free energy landscape in Figure 7.3 suggests that the transition between helix-hairpin and β-barrel state might involve a disordered crossover state, see the chain of arrows in the landscape used to trace a possible transition pathway.

Moving from the region of fully-formed helix hairpin (A), helix 2 of RfaH-CTD begins to deteriorate as seen by visual inspection of the configurations in region (B). Moving further along this region, helix 1 begins also to dissolve. Moving out of this basin requires to dissolve the backbone hydrogen bonds that stabilize the two helices, leading to a free energy barrier of about 10 RT, see region C. This is supported by the hydrogen-bond analysis in Figure 7.4 where we define a hydrogen bond by donor acceptor distances of less than 3.5 angstroms and an α angle of less than 30^o.

Upon crossing the barrier (C), the RfaH-CTD molecule moves through an ensemble of disordered configurations with little or no defined secondary structure. However in this region (D), β-hairpins begin to form in what appears to be a random fashion, and eventually, after crossing a much smaller barrier (E) of about 4 RT, a stable hairpin between $\beta 3$ and $\beta 4$ forms in region F. At this point $\beta 1$ has also begun to make contacts with $\beta 2$. Upon entering region G, $\beta 2$ starts to attach to $\beta 3$ of the β-hairpin structure, bringing $\beta 1$ with it. Interestingly, there exists a small helix, stabilized by several hydrogen bonds, in the linker region connecting $\beta 1$ and $\beta 2$. This helix positions $\beta 1$ higher up than in the ideal fold likely making it difficult for $\beta 5$ to lay on top of it. Only after finally crossing a third much smaller barrier of less than 2 RT, possibly due to loss of hydrogen bonds in the small helix between $\beta 1$ and $\beta 2$, does RfaH-CTD start to assume in region H the β-barrel form.

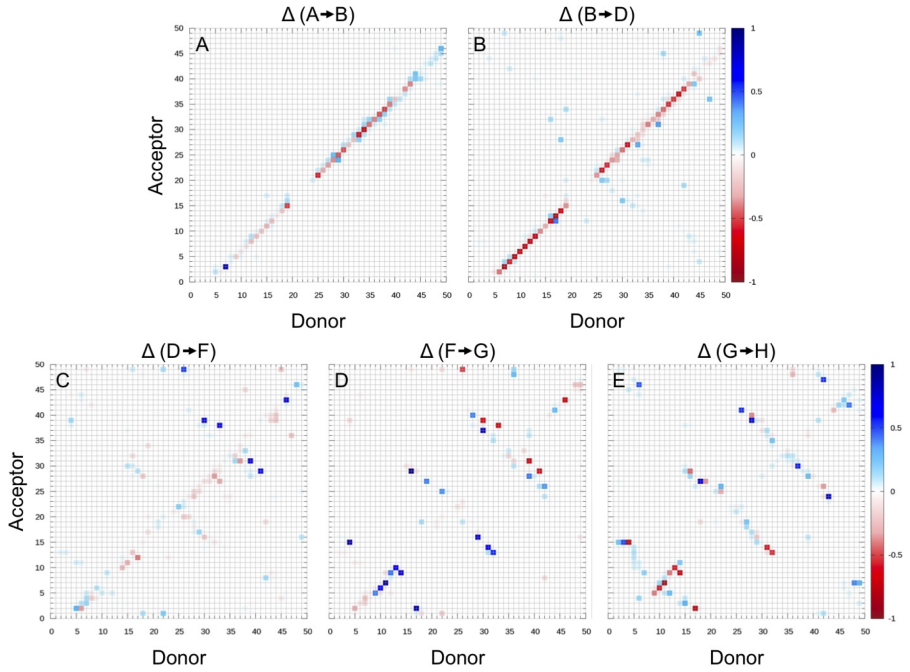

Figure 7.4: Percentage of structures that gained (blue) or lost (red) backbone hydrogen bonds between the residues indicated on x and y axis, when converting from one region of the free energy landscape to another. The involved regions are indicated at the top of each panel, with the indices corresponding to the ones defined in Figure 7.3.

Surprisingly the contacts between $\beta 5$ and $\beta 4$ are maintained in this step where the upper portion of $\beta 4$ bends down slightly allowing $\beta 5$ to lay on top of $\beta 1$ but facing in the wrong direction. In the final step, $\beta 5$ works its way around the N-terminus, thus completing the β-barrel (I). This chain of events is again supported by the hydrogen bond analysis of Figure 7.4. Note that this chain of events is also observed in the tunneling events that we show in Figure 7.5. While such tunneling events do not necessarily represent "true" transition paths (as they rely on an artificial dynamics), they are added here for illustration.

The above conversion process is similar to the one proposed in previous work[205] that relied on regular REMD simulations. One difference is that in this earlier

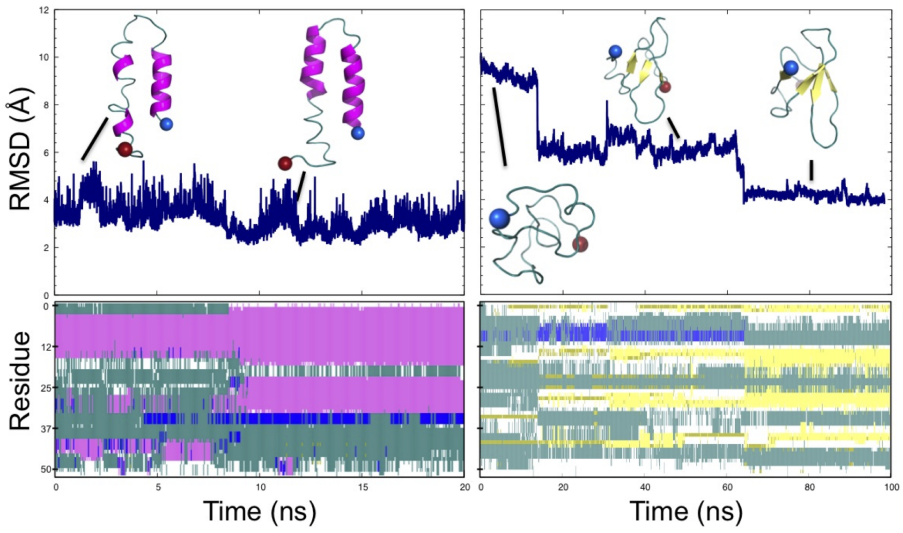

Figure 7.5: Formation of the helix hairpin (left panels) and the β-barrel (right panels). Top panels show the root-mean-square-deviation (RMSD) to the helix hairpin (left) and the β-barrel (right). Configurations from various time points are shown. Secondary structure analysis by the VMD program[211] is shown in the bottom panel. Here, pink represents α helices and yellow β- sheets.

work, helix 1 breaks first as opposed to our simulations where helix 2 is the one that starts dissolving first. However, in the earlier work, a transition pathway was obtained by following a single replica moving through temperature space. As at high temperatures a replica can cross barriers insurmountable at low temperatures, the observed transition pathways result from an artificial dynamics and do not necessarily describe the correct paths. On the other hand, our scenario follows from interpreting the free energy landscape of the protein, not from observed trajectories (which would also result from an artificial dynamics and therefore not necessarily describing the correct pathway). In addition, our scenario is also supported by a comparison of the root-mean-square-fluctuation (RMSF) of residues at the two termini. The C-terminus of RfaH-CTD has a large tail consisting of seven residues following helix 2, while at the N-terminus only three residues precede helix 1. The

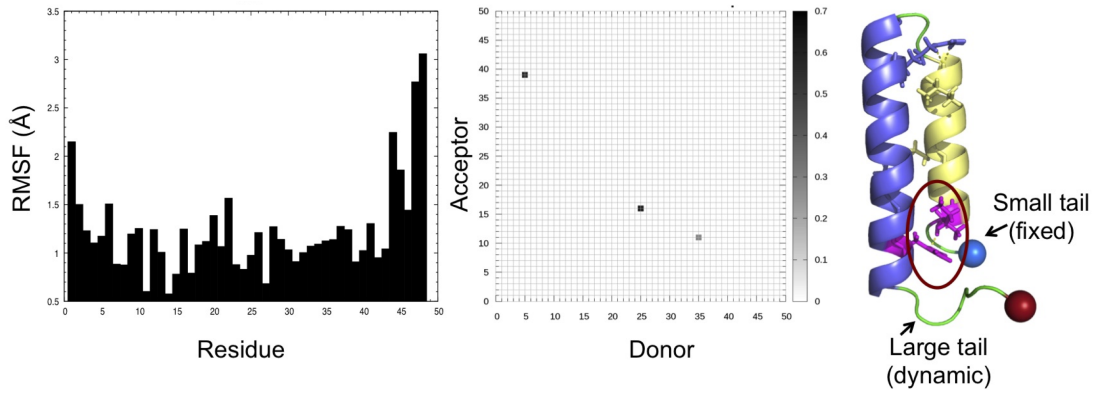

Figure 7.6: Root-mean-square-fluctuation (RMSF) computed for configurations close to the the helix-hairpin state (less than 2.5 angstroms to the ideal fold) (left panel). Hydrogen bonding between helix 1 and helix 2 for these configurations (middle panel). The helix hairpin structure is shown in the right panel, with the tail segments labeled and the N-terminal tail stabilizing hydrogen bond circled in red.

RMSF values in Figure 7.6 indicate that the larger tail at the C-terminus is much more flexible than the short end at the N-terminus whose three residues are stabilized by hydrogen bonds with residues 5 and 39 of helix 1 and helix 2. The increased mobility of the seven C-terminal residues adds extra strain on helix 2, thus disrupting its hydrogen bond pattern as is seen in Figure 7.4, and from visual inspection of clustering data for region (B). Note that this interpretation would not apply to the full-size RfaH protein (instead of only the C-terminal domain RfaH-CTD) which has a much larger linker region preceding helix 1.

The differences are much smaller in the remaining parts of the conversion process. Previous work also found that dissolution of the helices is followed by a disordered interconnecting state which precedes formation of a β-hairpin between $\beta 3$ and $\beta 4$ and then addition of $\beta 2$. The strand $\beta 1$ was suggested to take longer to align with the developing β-sheet due to a larger linker region but once formed would provide a template for addition of $\beta 5$, thus completely folding the barrel. This order of

β-sheet formation is the same as in our scenario, with the caveat that in our picture the process appears to be more dynamic: β-strands continue to grow and re-arrange as additional strands attach to the initial β-hairpin, as seen in the panels C-E of Figure 7.4, and become fully-formed only late in the folding pathway toward the β-barrel.

7.4 Conclusions

Using a variant of Replica-Exchange-with-Tunneling (RET) we have studied the fold switching process of the 66-residue C-domain RfaH-CTD of the transcription factor RfaH. Our enhanced sampling method allows us to calculate the free energy landscape of the protein projected on suitable coordinates. Analyzing this landscape we propose a mechanism for the conversion process between the helix-hairpin form seen when RfaH-CTD is bound to the he N-terminal domain of RfaH and blocks transcription, and the β-barrel form seen in the unbound RfaH-CTD which promotes translation. Consistent with experiments we find that the β-barrel is the preferred fold for the isolated RfaH-CTD. However, its free energy is only marginally lower than the helix-hairpin seen in the bound RfAH, but both folds are separated by large barriers resulting from the main chain hydrogen bonds of the helix hairpin. Upon dissolution of the helix hairpin, RfaH-CTD evolves into a disordered state, before a β-hairpin forms between $\beta 3$ and $\beta 4$. Later $\beta 2$ attaches to $\beta 3$ of this hairpin, with $\beta 1$ being in contact already with $\beta 2$. In the final steps, $\beta 4$ bends slightly trapping temporarily $\beta 5$ on top of $\beta 1$ before this strand rearranges and completes the β-barrel fold. While the overall pathway is similar to earlier work using traditional REMD simulations[205] , our improved sampling method adds important detail, showing a less structured conversion process with secondary structure only forming late in the

process. Together with our earlier work, these results establish the usefulness of our approach for studying switching proteins. We intend now to use our simulation protocol for the simulation of larger switching proteins such as the 93-residue lymphotactin[179] that would be difficult to study with regular REMD.

7.5 Acknowledgments

The simulations in this work were done using the SCHOONER cluster of the University of Oklahoma and XSEDE resources allocated under grant MCB160005 (National Science Foundation). We acknowledge financial support from the National Science Foundation under grant CHE1266256 and the National Institute of Health under grant GM120578 and GM120634.

Chapter 8: Multi-Scale Methods for Fast Exploration of Protein Landscapes.

The following chapter is taken from work that was unpublished at the time this dissertation was written. A manuscript is in preparation.

Author Contributions: All work presented in this chapter is credited to the author of this dissertation.

8.1 Introduction

In the last several decades, molecular dynamics (MD) simulations have been used to study a variety of bimolecular systems adding to our understanding of important processes such as protein folding[8–10,212] and aggregation.[33,35,110,147,155] These simulations approximate the dynamics of real molecular systems by application of atomic force fields.[7] With simulations of fine-grained models, where every atom of the system is represented explicitly, many properties of a protein system may be computed with reasonable accuracy.[71,72,213,214] Unfortunately, there remain a host of biological phenomena, such as the formation of amyloid fibers from monomers,[33,35,38] which occur on a timescale[38,39] currently unobtainable by high-resolution MD simulation. Holding back these simulations is the vibration of hydrogen atoms and the inclusion of an explicit solvent. Together these factors restrict the size of Δt used when updating the system and increase the number of computations required for each update. However, these problems may be reduced by lowering the resolution of the system thus enabling sufficient exploration of the protein landscape as needed for accurate computation of ensemble averages.[75]

By starting with a fine-grained model and removing select degrees of freedom, coarse-grained models[94,95,212,215,216] have been built that require less computation when solving Newton's equations. As a result, simulations using coarse-grained models are able to reach a timescale unobtainable to fine-grained models given the same resources.[75,212] However, removing degrees of freedom from the system results in a reduction of entropy, an effect that must be compensated for by enthalpic contributions.[75] In practice, an exact balance of these two factors is not possible and coarse-grained models tend to be less accurate than their fine-grained counterparts.[75] Still, the efficient exploration of conformational space observed in coarse-grained simulations[92,93] is an attractive feature and in principle, the accuracy of such simulations could be improved by refinement using a fine-grained model.[217]

With the resolution exchange method,[90,91] this is accomplished by performing a replica exchange[86,87] simulation where not the temperature varies but instead the resolution. Exploration of resolution space then results in the faster convergence of simulations at high resolution as compared to standard MD.[90,91] However, the resolution exchange method is held back by the problem of reintroducing missing degrees of freedom needed to construct a fine-grained model from a coarse one. While there have been methods[90,91,101–103] developed with this task in mind many introduce bias to samples[102,103] or result in the proposal of high energy states likely to be rejected.[90,91,101] Thus, in a recently proposed method[105] inspired by multiscale essential sampling (MSES),[104] the problem of reintroducing missing degrees of freedom is circumvented by the introduction of a restraining potential E_λ.[105] Because this and the MSES method rely on hamiltonian replica exchange[89] the method is susceptible to exchange bottleneck[88] problems similar to those observed in replica exchange molecular dynamics simulations (REMD),[86,87] a problem that may be cir-

cumvented by introduction the replica-exchange-with-tunneling (RET)[187] method recently developed by our group. We refer to this combination of MSES with RET as MSES/RET and test its efficiency together with another variant, ResET, that relies on only two replicas, in simulations of the Trp-cage protein.[17,162] This mini-protein has been investigated extensively[165,170,212,218,219] enabling direct comparison between our data and that from past studies.

8.2 Materials and Methods

Because molecular dynamics simulations of low-resolution protein systems are able to rapidly explore configurational space,[92,93] coarse-grained models may be used to enhance sampling in other simulations. Specifically, one can construct a molecular dynamics simulation where both a coarse- and fine-grained model evolve in parallel but that the fine-grained model is fed[196,207,209] configurations by the coarser representation. This "feeding" of states is accomplished by introduction of a restraining potential E_λ,[105] typically a function capable of enforcing a strict agreement between the two models, and a control parameter λ. The potential energy of such a system takes the form

$$E_{pot}(q_{fg}, q_{cg}) = E_{fg}(q_{fg}) + E_{cg}(q_{cg}) + \lambda E_\lambda(q_{fg}, q_{cg}), \tag{8.1}$$

where $E_{pot}(q_{fg}, q_{cg})$ is the total potential energy of the system, $E_{fg}(q_{fg})$ and $E_{cg}(q_{cg})$ are the potential energies of the fine- and coarse-grained models respectively and $E_\lambda(q_{fg}, q_{cg})$ is the restraining potential. One possibility is to use for $E_\lambda(q_{fg}, q_{cg})$ a

function of the form[105]

$$E_\lambda = E_\alpha = \begin{cases} \frac{1}{2}\left(\Delta^2(i,j)\right) & -ds < \Delta(i,j) < ds \\ A + \frac{B}{\Delta^S(i,j)} + f_{max}\Delta(i,j) & \Delta(i,j) > ds \\ A + \frac{B}{\Delta^S(i,j)}(-1)^S - f_{max}\Delta(i,j) & \Delta(i,j) < -ds \end{cases} \tag{8.2}$$

where $\Delta(ij) = \delta_{fg}(ij) - \delta_{cg}(ij)$ and is the difference in distances (δ) between α-carbons i and j of the two models. To reflect the dependence of equation 8.2 on α-carbon distances we have made here the change in notation $E_\lambda \rightarrow E_\alpha$. What's more, the control parameter f_{max} allows fixation of the maximum force as $\Delta(i,j) \rightarrow \infty$ and S sets how fast this value is realized. Because the functional form of 8.2 changes when $\Delta(i,j) = \pm ds$, the parameters A and B are included to ensure continuity of $E_\alpha(q_{fg}, q_{cg})$ and its first derivative at these values. These parameters are thus computed by

$$A = \left(\frac{1}{2} + \frac{1}{S}\right) ds^2 - \left(\frac{1}{S} + 1\right) f_{max} ds \quad and \quad B = \left(\frac{f_{max} - ds}{S}\right) ds^{S+1}. \tag{8.3}$$

What's more, it may be shown that a reflection of atomic coordinates across an arbitrary axis leaves 8.2 invariant. For this reason, a strict agreement between structures is not guaranteed by the inclusion of $E_\alpha(q_{fg}, q_{cg})$ alone, even with large λ values, leaving fine-grained models susceptible to the feeding of mirror structures.[196] Thus in our simulations, we add as a secondary restraint a function based on dihedral angles

$$E_\phi(q_{fg}, q_{cg}) = 1 + cos(\Delta_\phi(ijkl) + \pi), \tag{8.4}$$

such that $\Delta_\phi(ijkl) = \phi_{fg}(ijkl) - \phi_{cg}(ijkl)$ and is the difference in the dihedral angle (ϕ) formed by atoms i,j,k and l within each model. With this modification, the

potential energy of our system takes the final form

$$E_{pot}(q_{fg}, q_{cg}) = E_{fg}(q_{fg}) + E_{cg}(q_{cg}) + \lambda_\alpha E_\alpha(q_{fg}, q_{cg}) + \lambda_\phi E_\phi(q_{fg}, q_{cg}). \tag{8.5}$$

Introduction of hamiltonian replica exchange[89] now results in a random walk through $\lambda_\alpha/\lambda_\phi$ space with data analysis performed solely at the $\lambda_\alpha = \lambda_\phi = 0$ simulation i.e. where all biasing vanishes.

Because exchange moves often lead to a state of the multi-replica system which is exponentially suppressed, exchange rates are often low for the method just described. However, if accepted molecular systems typically evolve, in a short time, to a state with probability similar to before the exchange. Inspired by this observation we have introduced the replica-exchange-with-tunneling (RET)[187,196,207,209] method summarized in the following four-step-procedure:

1. In the first step, configurations $A(B)$ are updated by a short microcanonical simulation to configurations $A'(B')$ such that the energy on each system remains constant.

2. Following step one, the coordinates and velocities of A' and B' are exchanged. The velocities are also rescaled in this step such that the total energy (for each replica) remains unchanged by the exchange. The new velocities are thus computed by

$$v''_A = v'_A \sqrt{\frac{E_2 - E_{pot}(q'_A)}{E_{kin}(v'_A)}} \quad \text{and} \quad v''_B = v'_B \sqrt{\frac{E_1 - E_{pot}(q'_B)}{E_{kin}(v'_B)}}. \tag{8.6}$$

3. Following the exchange move, both replicas evolve by a second microcanonical simulation to configurations $\hat{B}(\hat{A})$ where the total energy is again conserved.

4. In the final step, configurations $\hat{B}(\hat{A})$ are accepted or rejected with probability

$$min\left(1, \exp\left(-\beta_1(E_{pot}(\hat{q}_B) - E_{pot}(q_A)) - \beta_2(E_{pot}(\hat{q}_A) - E_{pot}(q_B))\right)\right), \quad (8.7)$$

where $\beta = 1/k_BT$. If rejected, the system returns to its initial state $A(B)$. However, in either case, new velocities are randomly generated from a Boltzmann distribution according to the system temperature.

For more information regarding the RET method and its application, refer to References 187, 196, 207 and 209.

By including RET moves in the method described above the rescaled velocities of equation 8.6 become

$$\begin{aligned} v''_A &= v'_A\sqrt{\frac{E_2 - E_{fg}(q'_A) - E_{cg}(q'_A) - \lambda_{\alpha 2}E_\alpha(q'_A) - \lambda_{\phi 2}E_\phi(q'_A)}{E_{kin}(v'_A)}} \\ v''_B &= v'_B\sqrt{\frac{E_1 - E_{fg}(q'_B) - E_{cg}(q'_B) - \lambda_{\alpha 1}E_\alpha(q'_B) - \lambda_{\phi 1}E_\phi(q'_B)}{E_{kin}(v'_B)}} \end{aligned} \quad (8.8)$$

and the acceptance criterion of 8.7 changes to

$$\begin{aligned} & \exp\left(-\beta_1\left(\Delta E^{(1)}_{fg} + \Delta E^{(1)}_{cg} + \lambda_{\alpha 1}\Delta E^{(1)}_{\alpha} + \lambda_{\phi 1}\Delta E^{(1)}_{\phi}\right)\right) \\ + & \exp\left(-\beta_2\left(\Delta E^{(2)}_{fg} + \Delta E^{(2)}_{cg} + \lambda_{\alpha 2}\Delta E^{(2)}_{\alpha} + \lambda_{\phi 2}\Delta E^{(2)}_{\phi}\right)\right), \end{aligned} \quad (8.9)$$

where $\Delta E^{(1)}_{fg} = E_{fg}(\hat{q}_B) - E_{fg}(q_A)$ and $\Delta E^{(2)}_{fg} = E_{fg}(\hat{q}_A) - E_{fg}(q_B)$; $\Delta E^{(i)}_{cg}$, $\Delta E^{(i)}_{\alpha}$ and $\Delta E^{(i)}_{\phi}$ are defined accordingly. With the inclusion of RET moves in the MSES method replicas are able to mix even when the spacing of λ values become somewhat large.[187,196,207,209] As a consequence, MSES/RET simulations, even with the inclusion of an explicit solvent, require only a modest number of replicas. For

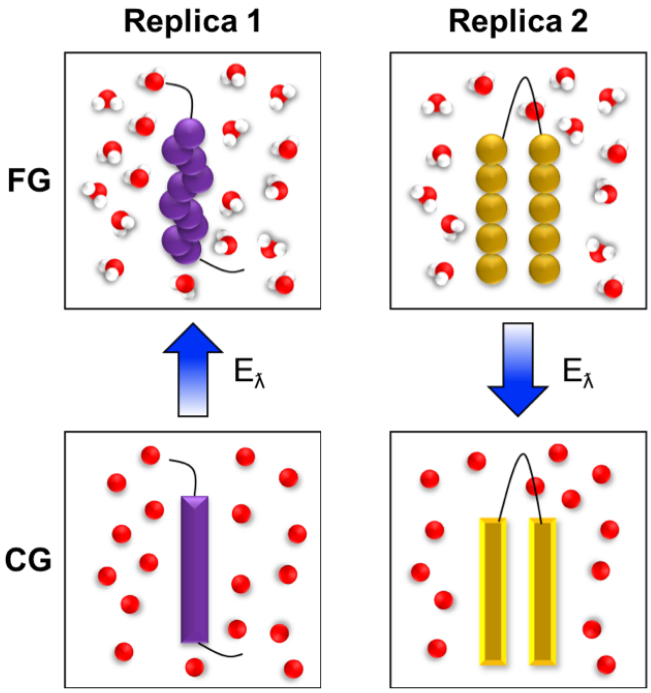

Figure 8.1: Setup for the ResET method.

our simulations of the Trp-cage mini-protein, we use 8. However, one can construct an alternative setup, introduced here as Resolution-Exchange-with-Tunneling (ResET), enabling exploration of protein landscapes using only two replica. The ResET simulation is performed in the following way. On the first replica lives not only a *coarse-grained model* with configuration A_{CG} (described by an energy $E_{CG}(A_{CG})$) but also an *auxiliary fine-grained model* whose configuration A_{FG} depends on the coarse-grained model by an energy $E_{FG}(A_{FG}) + \lambda_1 E_\lambda(A_{CG}, A_{FG})$, that favors configurations of the fine-grained model which resemble the coarse-grained model configuration (8.1). Similarly, we have on the second replica a *fine-grained model* with configuration B_{FG} and energy $E_{FG}(B_{FG})$ and an *auxiliary coarse-grained model* whose configuration B_{CG} depends on the fine-grained model by an energy

$E_{CG}(B_{CG}) + \lambda_2 E_\lambda(B_{CG}, B_{FG})$, ensuring that on this replica the coarse-grained configuration resembles that of the fine-grained model (8.1).

During the exchange move, the coarse-grained configuration A_{CG} on replica 1 is replaced by the configuration B_{CG} of the auxiliary coarse-grained model on replica 2, and the fine-grained configuration B_{FG} of replica 2 by A_{FG} of the auxiliary fine-grained model on replica 1:

$$\begin{aligned} \mathrm{R}_1 :& A_{CG}[E_{CG}(A_{CG})] \rightarrow A_{FG}\left[E_{FG}(A_{FG}) + \lambda_1 E_\lambda(A_{FG}, A_{CG})\right] \searrow\nearrow B_{CG}[E_{CG}(B_{CG})] \\ \mathrm{R}_2 :& B_{FG}[E_{FG}(B_{FG})] \rightarrow B_{CG}\left[E_{CG}(B_{CG}) + \lambda_2 E_\lambda(B_{FG}, B_{CG})\right] \nearrow\searrow A_{FG}[E_{FG}(A_{FG})] \end{aligned} \tag{8.10}$$

Here, the horizontal arrows point to the auxiliary models, and the energy terms for the corresponding models are listed in brackets. Note, that unlike in resolution exchange this move cannot be accepted or rejected according to $min\left(1, exp\left(-\beta\Delta E\right)\right)$ as this would destroy detailed balance. This is because the proposal configurations A_{FG} and B_{CG} are generated by a biased process. Hence, in order to ensure detailed balance, one has to account for the probabilities by that these proposal configurations are generated through a Metropolis-Hastings acceptance criterium:

$$\begin{aligned} w(A \rightarrow B) \;=\; & \min\left(1, \exp(-\beta(E_{CG}(A_{CG}) - E_{FG}(A_{FG}) - E_{CG}(B_{BG}) + E_{FG}(B_{FG})\right. \\ & \left. +\lambda_1 E_\lambda(A_{FG}, A_{CG}) + \lambda_2 E_\lambda(B_{FG}, B_{CG}))))\right). \end{aligned} \tag{8.11}$$

To enable direct comparison between our simulations and those performed in previous studies we follow closely the set up used by Kouza et al[170] for fine-grained models and Han et al[219] for coarse-grained. All fine-grained models are therefore capped at the N-terminus by an acetyl group and at the C-terminus by a methylamine whereas coarse-grained models are left uncapped and the system buffered by 0.15M

Na^+ and Cl^- ions. Both models are solvated with the physical model containing the same box size and water molecule count as in Kouza et al[170] and the coarse-grained model a cubic box of length 5.18 nm and 1118 water molecules. Prior to running NVT simulations, random structures were generated by high temperature (T = 3000 K for the coarse-grained model and 4000 K the fine-grained model) molecular dynamics simulations performed over three nanoseconds. Starting structures were taken periodically following the first nanosecond of melting. Input parameters to the E_λ energy function[105] of Eq. 8.2 were provided so that $ds = 0.3$Å, $S = 1$ and $f_{max} = 0$. Trajectories were then produced using an in-house version of GROMACS 4.6.5[174] (available upon request) which has been modified by the authors to support RET sampling, MSES and the ResET method. Calculations of the potential energy were dependent upon the AMBER94 force field[220] (thus enabling comparison to previous work) and the TIP3P[80] water model for the fine-grained model and the PACE[94,212,219,221–223] energy function for the coarse-grained model. Equations of motion were integrated using a velocity verlet algorithm and a time step of 2 fs. Hydrogen and heavy atom bond distances were constrained using the linear constraint solver (LINCS),[78] for the solvent the settle[79] algorithm was used. Treatment of electrostatic interactions depended on the PME set up[130,131] used in GROMACS for the fine-grained model and a cutoff of 1.2 nm was used for the coarse model. The v-rescale thermostat[132] was used to maintain the temperature. And finally, MD runs were also performed (for both fine- and coarse-grained models) at NVT along with REMD simulations (fine-grained only) enabling direct comparison of MSES/RET and ResET simulations with proven methodologies. Individual temperatures and lambda values are listed for each simulation in table A.1.

8.3 Results and Discussion

To begin our analysis we check folding of the fine-grained model in 3 independent 200 ns canonical simulations, each started from random configurations and momentum distributions. To test for folding the α-carbon root mean squared deviation (RMSD) of a configuration must be less than a cutoff value δ from the native fold. In previous work[165,170] a value of 2.2 Å was used for δ; however, different values have been suggested[218] and we find little difference between 2.2 Å and 3.0 Å. Table 8.1 lists the frequency of folded configurations during different intervals of the simulation using cutoff values of 2.2 Å and a slightly softer 2.4 Å. We note that because we find residue 20 to be highly dynamic in the folded state (data not shown) we exclude this residue from all RMSD calculations. This is different from previous studies[165,170,218] where all 20 residues were included in RMSD calculations. From this analysis, we find approximately 30% of structures to be in the folded state after 200 ns of simulation time. This is compared to 80 % reported in previous REMD simulations[170] of Trp-cage and 70 % from experiment.[162] However, the number of folded structures continues to grow throughout the runs indicating simulations have not converged. For the coarse-grained model, 8 independent simulations are performed and the fraction of folded configurations again monitored. However, for these simulations, RMSD calculations rely on residues $3-19$ following Han et al.[212] Similar to their work we find roughly 50% of structures folded after 1 μs. This number drops to approximately 10% if considering residues $1-19$ for RMSD calculations (table 8.1). As with the fine-grained model, convergence is not achieved in the time simulated.

This is different for our REMD runs where we find the simulation converges after only 50ns (table 8.1). For this reason, only the last 50 ns is used for further analysis.

	PACE 1-19		PACE 3-19			REMD		MSES/RET		ResET			MD	
Time (ns)	2.2	2.4	2.2	2.4	Time (ns)	2.2	2.4	2.2	2.4	2.2	2.4	Time (ns)	2.2	2.4
$0-100$	0	2	11	17	$0-20$	19	24	2	5	3	. 5	$0-50$	0	0
$100-200$	10	14	38	44	$20-40$	42	54	4	8	0	0	$50-100$	0	0
$200-300$	11	15	52	57	$40-60$	51	72	3	6	12	21	$100-150$	8	8
$300-400$	5	8	42	47	$60-80$	64	86	15	21	25	36	$150-200$	30	31 .
$400-500$	4	6	35	38	$80-100$	61	85	13	19	60	68			
$500-600$	5	6	39	43	$100-120$	—	—	13	20	58	71			
$600-700$	4	5	50	52										
$700-800$	8	11	49	55										
$800-900$	8	10	57	60										
$900-1000$	9	12	62	65										

Table 8.1: Percentage of structures in different time intervals of the trajectory that have an RMSD less than the indicated cutoff value (2.2 Å or 2.4 Å).

Within this time interval, approximately 85 % of structures are in the folded state, similar to that found by Kouza et al.[170] Figure 8.2 shows the most populated cluster as obtained using the GROMOS clustering method[210] with a cutoff of 3Å as well as the lowest RMSD and lowest energy structures. These configurations have an RMSD from the native state of 2.0 Å, 0.5 Å and 2.1 Å respectively. Figure 8.3 shows the free energy landscape projected onto the RMSD relative to the native state. This landscape has a minimum at 2.2 Å with a large barrier at 4.8 Å.

Next, we check MSES/RET simulations for convergence (table 8.1). Because coupling coarse- and fine-grained models reduces the speed at which a coarse-grained model may be simulated, we let the coarse representation update multiple steps for each fine-grained update. In the present study, we use an updating ratio of 1:5. Despite showing stable frequencies after 40ns, our MSES/RET simulations are not found to converge. In support of this claim, multiple coarse-grained models are observed to fold late in the simulation. Because information flows between replicas only at large lambda values, it is expected to take some time before this information becomes visible in the fine-grained models. Therefore, MSES/RET simulations likely require more time to converge than standard REMD simulations. To test this, longer trajectories are currently being collected. Still, we speculate on the free

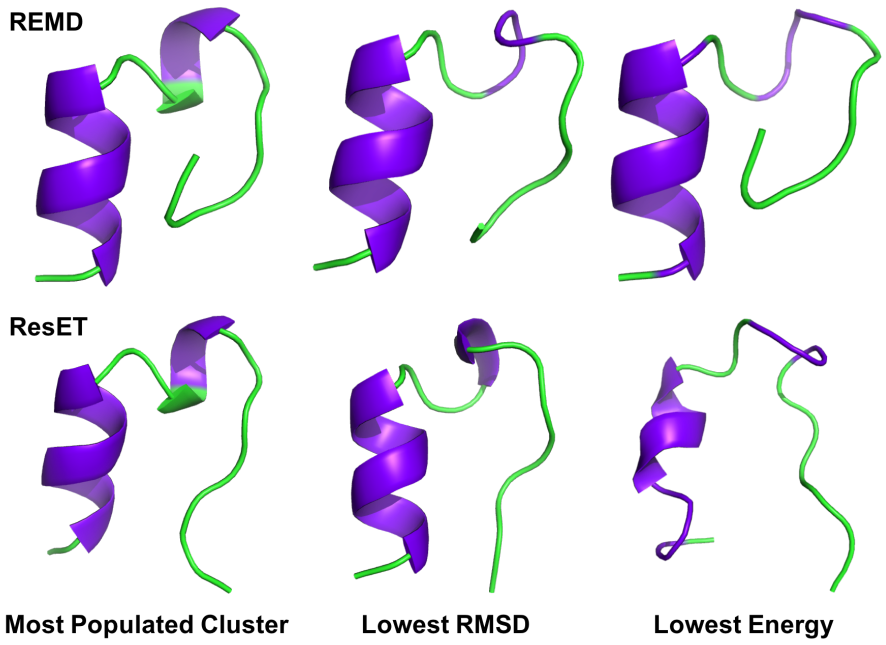

Figure 8.2: Shown are the most populated, lowest energy and lowest RMSD structures from REMD (top panel) and ResET (bottom panel) simulations. Residues known to participate in helices are color coded purpleblue.

energy landscape by using for analysis the last 20ns of the MSES/RET simulation. As shown in figure 8.3, this landscape contains some similarities to that taken from REMD simulations with the minimum lying within 3 Å of the native fold. However, the non-native folds make a much larger percentage of structures.

For the ResET method, a multistep updating scheme is also used with a 1:5 updating ratio. The resulting trajectory is found to converge late in the simulation, after about 80ns. For this reason, we extend this simulation to 120ns with the last 40ns taken for data analysis. In this time interval, about 70 % of structures occupy the native fold. This result is comparable to both experimental results[162] of 70 % and those from both present and past REMD simulations.[165,170,218] Shown in figure 8.2 is the most populated cluster as well as the lowest RMSD and lowest energy structures.

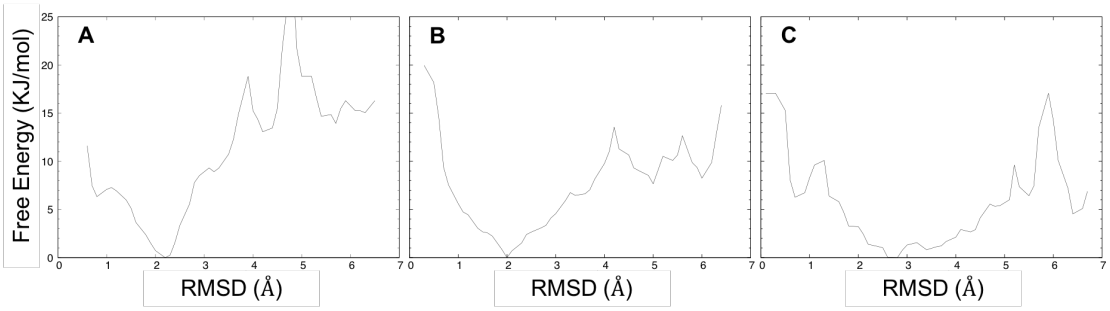

Figure 8.3: Free energy landscape projected onto the root mean square deviation to the NMR structure for REMD (A), ResET (B) and MSES/RET (C) simulations.

These configurations are similar to those found in our REMD simulations with RMSD values to the native fold of 1.5 Å, 0.52 Å and 1.1 Å respectively. Likewise, the free energy landscape (figure 8.3) obtained by the ResET method is similar to that from our REMD simulations. Subtle differences include a broader basin around the native fold and slightly more populated non-native folds. Additionally, the barrier at 4.8 Å is not present in the ResET simulation.

In comparing methodologies, it is found that ResET simulations can give results comparable to REMD simulations but using far fewer resources (only 2 replicas in the present study compared to 40 for REMD). In the case of MSES/RET, similar results may be achievable but larger convergence times may be needed. Still, convergence times may not be so great and, while longer than that required for ResET simulations, could prove to be shorter than for REMD in some cases. For both types of simulation, the occurrence of mirror structures was not observed. This is in contrast to previous simulations[196,209] which used only the lambda energy function of 8.2 as well as preliminary simulations of Trp-cage which did not contain 8.4. Still, it is difficult to say how much this result was influenced by the inclusion of 8.4 or the choice of forcefield. Furthermore, the choice of a coarse-grained model may play

an important role in dictating the behavior of MSES/RET and ResET simulations. However, despite slow convergence times observed in our PACE simulations and an underestimation of the folded state of Trp-cage, ResET simulations employing this coarse-grained model were able to accurately fold Trp-cage with the correct frequencies. This may be the case for the MSES/RET method as well. However, longer trajectories will be needed, the collection of which is currently underway.

8.4 Conclusions

We have introduced a variant of the replica-exchange-with-tunneling and multiscale essential sampling methods as a multi-scale method referred to here as MSES/RET. We have also introduced the resolution-exchange-with-tunneling method (ResET). These methods are able to produce better results than standard high-resolution molecular dynamics simulations and, in the case of ResET, similar results to REMD simulations. This was confirmed by simulations of the Trp-cage mini-protein. Complicating both methodologies is the choice of a coarse-grained model. However, this choice likely affects convergence times only and not the equilibrium distribution as the ResET simulations seem to be uninfluenced by the PACE forcefield underestimating the folded state. Furthermore, the MSES/RET method may be capable of producing results comparable to REMD simulations as well. However, longer trajectories will be needed to confirm this, the acquisition of which is in progress.

Chapter 9: Closing Remarks

9.1 Future Outlook

With the introduction of RET, protein systems of increasing size may be studied with moderate sized computing clusters. The method, however promising, is not expected to act as a silver bullet to the problems holding back REMD and other general ensemble methods but does expand the capabilities of these methods. This was demonstrated in this text by simulations of protein systems both in an implicit and explicit solvent while using fewer resources than REMD and HREMD simulations. By directing RET toward the MSES method, we were able to develop a strategy that enables investigation of proteins with competing attractors such as GA98, GB98 and RfaH-CTD. For GA98, our simulations accurately predict the GA fold to be dominant with GB being only marginally populated. For GB98, the case was more complicated but a picture did emerge that is consistent with experimental observations. These results are an improvement upon past computer simulations of these systems and lend confidence in the correctness of the current generation of force fields. Our simulations of the isolated C-terminal domain of RfaH were also insightful and add important detail to the transition pathway leading from the helix hairpin to the β barrel fold. What's more, the method enabled the identification of important residues that promote fibrilization of the 13-residue fragment of serum amyloid A. Together, these simulations provide important details about the energy landscape for each system and serve as a reference for future studies of proteins that exhibit a dual funnel energy landscape.

Switching gears, we also applied RET to standard MSES simulations that use instead of a Go-model a coarse-grained one. These simulations, tested on the Trp-

cage protein, outperformed standard MD simulations and may even produce results comparable to REMD simulations given long enough trajectories. To test this assumption, we are in the process of extending current trajectories. Moreover, convergence times could likely be shortened upon selection of an appropriate coarse-grained model. Additionally, the ResET method was also introduced and demonstrates how a highly efficient multi-scale simulation may be constructed. With ResET, the free energy landscape of Trp-cage was constructed with remarkable detail provided these simulations used only two replicas. Together, these results are encouraging and suggest that multi-scale simulations can be effective at simulating real protein systems. Still, it is likely that more work will be required before these methods can be applied to more challenging problems such as protein folding and aggregation.

References

[1] A. L. Lehninger, D. L. 1. Nelson, and M. M. Cox. *Lehninger Principles of Biochemistry (6th ed.)*. W.H. Freeman, 2013.

[2] H. Lodish, A. Berk, C. Kaiser, M. Krieger, M. Scott, A. Bretscher, H. Ploegh, and M. Paul. *Molecular Cell Biology*. Macmillan, 6th edition, 2008.

[3] Julie S. Valastyan and Susan Lindquist. Mechanisms of protein-folding diseases at a glance. *Disease Models & Mechanisms*, 7(1):9–14, 2014.

[4] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins*, 21:167–95, 1995.

[5] J. N. Onuchic and P. G. Wolynes. Theory of protein folding. *Curr. Opin. Struc. Biol.*, 14(1):70–5, Feb 2004.

[6] P. E. Leopold, M. Montal, and J. N. Onuchic. Protein folding funnels: A kinetic approach to the sequence-structure relationship. *Proc. Natl. Acad. Sci. (USA)*, 89(18):8721–5, Sep 1992.

[7] Andrew Leach. *Molecular Modelling: Principles and Applications*. Pearson, 2 edition, 2001.

[8] J. L. Dill, K. A. MacCallum. The protein-folding problem, 50 years on. *Science*, 338:1042–6, 2012.

[9] Ken A. Dill, S. Banu Ozkan, M. Scott Shell, and Thomas R. Weikl. The protein folding problem. *Annual Review of Biophysics*, 37(1):289–316, 2008. PMID: 18573083.

[10] C. M. Dobson. Protein folding and misfolding. *Nature*, 426:884–90, 2003.

[11] Cyrus Levinthal. How to Fold Graciously. In J. T. P. Debrunnder and E. Munck, editors, *Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois*, pages 22–24. University of Illinois Press, 1969.

[12] D. Fraser-Pitt, D.; O'Neil. Cystic fibrosis - a multiorgan protein misfolding disease. *Future Sci OA*, 1:FSO57, 2015.

[13] G. R. Cutting. Cystic fibrosis genetics: from molecular understanding to clinical application. *Nat Rev Genet*, 16:45–56, 2015.

[14] J. M. Rommens, M. C. Iannuzzi, B. Kerem, M. L. Drumm, G. Melmer, M. Dean, R. Rozmahel, J. L. Cole, D. Kennedy, and N. et al. Hidaka. Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science*, 245:1059–65, 1989.

[15] J. R. Riordan, J. M. Rommens, B. Kerem, N. Alon, R. Rozmahel, Z. Grzelczak, J. Zielenski, S. Lok, N. Plavsic, and J. L. et al. Chou. Identification of the cystic fibrosis gene: cloning and characterization of complementary dna. *Science*, 245:1066–73, 1989.

[16] M. J. Welsh and A. E. Smith. Molecular mechanisms of cftr chloride channel dysfunction in cystic fibrosis. *Cell*, 73:1251–4, 1993.

[17] C. Simmerling, B. Strockbine, and A. E. Roitberg. All-atom structure prediction and folding simulations of a stable protein. *J Am Chem Soc*, 124:11258–9, 2002.

[18] J. L. Krstenansky, T. J. Owen, K. A. Hagaman, and L. R. McLean. Short model peptides having a high alpha-helical tendency - design and solution properties. *Febs Letters*, 242(2):409–413, Jan 1989.

[19] A. M. Gronenborn, D. R. Filpula, N. Z. Essig, A. Achari, M. Whitlow, P. T. Wingfield, and G. M. Clore. A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein-g. *Science*, 253(5020):657–661, Aug 1991.

[20] Erik Nordling and Mirna Abraham-Nordling. Colonic amyloidosis, computational analysis of the major amyloidogenic species, serum amyloid a. *Computational Biology and Chemistry*, 39:29–34, Aug 2012.

[21] Patrick A. Alexander, Yanan He, Yihong Chen, John Orban, and Philip N. Bryan. A minimal sequence code for switching protein structure and function. *Proc. Natl. Acad. Sci. (USA)*, 106(50):21149–54, Dec 2009.

[22] Bjoern M. Burmann, Stefan H. Knauer, Anastasia Sevostyanova, Kristian Schweimer, Rachel A. Mooney, Robert Landick, Irina Artsimovitch, and Paul Roesch. An alpha helix to beta barrel domain switch transforms the transcription factor rfah into a translation factor. *Cell*, 150(2):291–303, Jul 2012.

[23] Catherine S. Mocny and Vincent L. Pecoraro. De novo protein design as a methodology for synthetic bioinorganic chemistry. *Accounts of Chemical Research*, 48(8):2388–2396, 2015.

[24] H. Christopher Fry, Andreas Lehmann, Louise E. Sinks, Inge Asselberghs, Andrey Tronin, Venkata Krishnan, J. Kent Blasie, Koen Clays, William F. DeGrado, Jeffery G. Saven, and Michael J. Therien. Computational de novo design and characterization of a protein that selectively binds a highly hyperpolarizable abiological chromophore. *Journal of the American Chemical Society*, 135(37):13914–13926, 2013.

[25] B. I. Dahiyat and S. L. Mayo. De novo protein design: fully automated sequence selection. *Science*, 278:82–7, 1997.

[26] T. Gallagher, P. Alexander, P. Bryan, and G. L. Gilliland. Two crystal structures of the b1 immunoglobulin-binding domain of streptococcal protein g and comparison with nmr. *Biochemistry*, 33:4721–9, 1994.

[27] C.B. Anfinsen and H.A. Scheraga. Experimental and theoretical aspects of protein folding. volume 29 of *Advances in Protein Chemistry*, pages 205 – 300. Academic Press, 1975.

[28] Ken A. Dill. Dominant forces in protein folding. *Biochemistry*, 29(31):7133–7155, 1990. PMID: 2207096.

[29] Ken A. Dill. Polymer principles and protein folding. *Protein Science*, 8(6):1166–1180, 1999.

[30] Theodore T. Herskovits, Barbara Gadegbeku, and Helen Jaillet. On the structural stability and solvent denaturation of proteins: I. denaturation by the alcohols and glycols. *Journal of Biological Chemistry*, 245:2588–2598, 1970.

[31] Jingjing Guo and Huan-Xiang Zhou. Protein allostery and conformational dynamics. *Chemical Reviews*, 116(11):6503–6515, 2016. PMID: 26876046.

[32] Ruth Nussinov. Introduction to protein ensembles and allostery. *Chemical Reviews*, 116(11):6263–6266, 2016. PMID: 27268255.

[33] M. Eisenberg, D. Jucker. The amyloid state of proteins in human diseases. *Cell*, 148:1188–203, 2012.

[34] R. P. Buxbaum, J. N. Linke. A molecular history of the amyloidoses. *J Mol Biol*, 421:142–59, 2012.

[35] C. M. Chiti, F. Dobson. Protein misfolding, functional amyloid, and human disease. *Annu Rev Biochem*, 75:333–66, 2006.

[36] Jed J.W. Wiltzius, Stuart A. Sievers, Michael R. Sawaya, Duilio Cascio, Dmitriy Popov, Christian Riekel, and David Eisenberg. Atomic structure of the cross-β spine of islet amyloid polypeptide (amylin). *Protein Science*, 17(9):1467–1474, 2009.

[37] Pin-Nan Cheng, Johnny D. Pham, and James S. Nowick. The supramolecular chemistry of β-sheets. *Journal of the American Chemical Society*, 135(15):5477–5492, 2013. PMID: 23548073.

[38] Y. S. Eisele. From soluble abeta to progressive abeta aggregation: could prion-like templated misfolding play a role? *Brain Pathol*, 23:333–41, 2013.

[39] K. Weise, D. Radovan, A. Gohlke, N. Opitz, and R. Winter. Interaction of hiapp with model raft membranes and pancreatic beta-cells: cytotoxicity of hiapp oligomers. *Chembiochem*, 11:1280–90, 2010.

[40] J. D. Harper and Jr. Lansbury, P. T. Models of amyloid seeding in alzheimer's disease and scrapie: mechanistic truths and physiological consequences of the time-dependent solubility of amyloid proteins. *Annu Rev Biochem*, 66:385–407, 1997.

[41] K. Jackson, G. A. Barisone, E. Diaz, L. W. Jin, C. DeCarli, and F. Despa. Amylin deposition in the brain: A second amyloid in alzheimer disease? *Ann Neurol*, 74:517–26, 2013.

[42] W. M. Berhanu and A. E. Masunov. Full length amylin oligomer aggregation: insights from molecular dynamics simulations and implications for design of aggregation inhibitors. *J Biomol Struct Dyn*, 32:1651–69, 2014.

[43] J. Zhao, X. Yu, G. Liang, and J. Zheng. Heterogeneous triangular structures of human islet amyloid polypeptide (amylin) with internal hydrophobic cavity and external wrapping morphology reveal the polymorphic nature of amyloid fibrils. *Biomacromolecules*, 12:1781–94, 2011.

[44] Masashi Egashira, Hiroka Takase, Izumi Yamamoto, Masafumi Tanaka, and Hiroyuki Saito. Identification of regions responsible for heparin-induced amyloidogenesis of human serum amyloid a using its fragment peptides. *Archives of Biochemistry and Biophysics*, 511(1-2):101–106, Jul 2011.

[45] C. M. Uhlar and A. S. Whitehead. Serum amyloid a, the major vertebrate acute-phase reactant. *European Journal of Biochemistry*, 265(2):501–523, Oct 1999.

[46] F.C. De Beer, R.K. Mallya, E.A. Fagan, J.G. Lanham, G.R. Hughes, and M.B. Pepys. Serum amyloid-a protein concentration in inflammatory diseases and its relationship to the incidence of reactive systemic amyloidosis. *Lancet*, 2(8292):231–234, 1982.

[47] Bruno J. Strasser. Sickle cell anemia, a molecular disease. *Science*, 286:1488–1490, 1999.

[48] C. B. Anfinsen, E. Haber, M. Sela, and Jr. White, F. H. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc Natl Acad Sci U S A*, 47:1309–14, 1961.

[49] Cyrus Levinthal. Are There Pathways For Protein Folding? *Extrait du Journal de Chimie Physique*, 65, 1968.

[50] Philip N. Bryan and John Orban. Proteins that switch folds. *Curr. Opin. Struct. Biol.*, 20(4):482–8, Aug 2010.

[51] 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. *Proc. Natl. Acad. Sci. (USA)*, 104(29):11963–8, Jul 2007.

[52] S. R. Fahnestock, P. Alexander, J. Nagle, and D. Filpula. Gene for an immunoglobulin-binding protein from a group g streptococcus. *J Bacteriol*, 167:870–80, 1986.

[53] C. Falkenberg, L. Bjorck, and B. Akerstrom. Localization of the binding site for streptococcal protein g on human serum albumin. identification of a 5.5-kilodalton protein g binding albumin fragment. *Biochemistry*, 31:1451–7, 1992.

[54] E. B. Myhre and G. Kronvall. Heterogeneity of nonimmune immunoglobulin fc reactivity among gram-positive cocci: description of three major types of receptors for human immunoglobulin g. *Infect Immun*, 17:475–82, 1977.

[55] Y. He, D. A. Rozak, N. Sari, Y. Chen, P. Bryan, and J. Orban. Structure, dynamics, and stability variation in bacterial albumin binding modules: implications for species specificity. *Biochemistry*, 45:10102–9, 2006.

[56] Y. He, Y. Chen, P. A. Alexander, P. N. Bryan, and J. Orban. Mutational tipping points for switching protein folds and functions. *Structure*, 20:283–91, 2012.

[57] S. K. Tomar, S. H. Knauer, M. Nandymazumdar, P. Rosch, and I. Artsimovitch. Interdomain contacts control folding of transcription factor rfah. *Nucleic Acids Res.*, 41(22):10077–85, 2013.

[58] 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-alpha to all-beta conformational transition of rfah- ctd: Molecular dynamics simulation and markov state model. *J. Chem. Theory. Comput.*, 10(8):2255–64, 2014.

[59] 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, 2007.

[60] Jeevan B. Gc, Bernard S. Gerstman, and Prem P. Chapagain. The role of the interdomain interactions on rfah dynamics and conformational transformation. *J. Phys. Chem. B.*, 119(40):12750–9, 2015.

[61] Nicole Balasco, Daniela Barone, and Luigi Vitagliano. Structural conversion of the transformer protein rfah: New insights derived from protein structure prediction and molecular dynamics simulations. *J. Biomol. Struct. Dyn.*, 33(10):2173–9, 2015.

[62] M. J. Bailey, C. Hughes, and V. Koronakis. Rfah and the ops element, components of a novel system controlling bacterial transcription elongation. *Mol Microbiol*, 26:845–51, 1997.

[63] M. Sofia Ciampi. Rho-dependent terminators and transcription termination. *Microbiology*, 152(9):2515–2528, 2006.

[64] Irina Artsimovitch and Robert Landick. *The Transcriptional Regulator RfaH Stimulates RNA Chain Synthesis after Recruitment to Elongation Complexes by the Exposed Nontemplate DNA Strand*. cell, 2002.

[65] Thomas Englel and Philip Reid. *Physical Chemistry*. Pearson, 2013.

[66] R. Shankar. *Principles of Quantum Mechanics*. Plenum Press, 1994.

[67] Richard P. Feynman and Albert R. Hibbs. *Quantum Mechanics and Path Integrals Emended Edition*. Dover, 2005.

[68] Leonard Susskind and Art Friedman. *Quantum Mechanics The Theoretical Minimum*. Basic Books, 2014.

[69] Keith J. Laidler, John H. Meiser, and Bryan C. Sanctuary. *Physical Chemistry.* Houghton Mifflin Company, 2003.

[70] E. Schrodinger. An undulatory theory of the mechanics of atoms and molecules. *Phys. Rev.*, 28:1049–1070, 1926.

[71] Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, Paul Maragakis, John L. Klepeis, Ron O. Dror, and David E. shaw. Improved side-chain torsion potentials for the amber ff99sb protein force field. *Proteins-Structure Function And Bioinformatics*, 78(8):1950–1958, Jun 2010.

[72] R. B. Best, X. Zhu, J. Shim, P. E. Lopes, J. Mittal, M. Feig, and A. D. Jr. Mackerell. Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone phi, psi and side-chain chi(1) and chi(2) dihedral angles. *Chem Theory Comput*, 8 (9):3257–73, 2012.

[73] Ponder J.W. and Case DA. Force fields for protein simulations. *Adv. Protein Chem.*, pages 27–85, 2003.

[74] Tieleman D.P. Monticelli L. Force fields for classical molecular dynamics. *Biomolecular Simulations. Methods in Molecular Biology (Methods and Protocols)*, 294:197–213, 2013.

[75] Sebastian Kmiecik, Dominik Gront, Michal Kolinski, Lukasz Wieteska, Aleksandra Elzbieta Dawid, and Andrzej Kolinski. Coarse-grained protein models and their applications. *Chemical Reviews*, 116(14):7898–7936, 2016. PMID: 27333362.

[76] J.G. Kirkwood. Statistical mechanics of fluid mixtures. *The Journal of Chem. Phys.*, 3:300–313, 1953.

[77] Benoît Roux. The calculation of the potential of mean force using computer-simulations. 91:275–282, 09 1995.

[78] Berk Hess. P-lincs: A parallel linear constraint solver for molecular simulation. *J. Chem. Theory Comput.*, 4(1):116–122, 2008.

[79] Shuichi Miyamoto and Peter A. Kollman. Settle: An analytical version of the shake and rattle algorithm for rigid water models. *Journal of Computational Chemistry*, 13(8):952–962, 1992.

[80] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, pages 926–935, 1983.

[81] Bernd A. Berg. *Markov Chain Monte Carlo Simulations and Their Statistical Analysis*. World Scientific, 2004.

[82] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Oxford University Press*, 57(1):97–109, 1970.

[83] 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.

[84] Charles J. Geyer. Markov chain monte carlo maximum likelihood. *Interface Proceedings*, 1991.

[85] K. Hukushima and K. Nemoto. Exchange monte carlo method and application to spin glass simulations. *J. Phys. Soc. (Japan)*, 65(6):1604–8, Jun 1996.

[86] U. H. E. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.*, 281(1-3):140–50, Dec 1997.

[87] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314(1-2):141–51, Nov 1999.

[88] W. Nadler and U. H. Hansmann. Dynamics and optimal number of replicas in parallel tempering simulations. *Phys Rev E Stat Nonlin Soft Matter Phys*, 76:065701, 2007.

[89] H Fukunishi, O Watanabe, and S Takada. On the hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *J. Chem. Phys.*, 116(20):9058–67, 2002.

[90] E. Lyman and D. M. Zuckerman. Resolution exchange simulation with incremental coarsening. *J Chem Theory Comput*, 2:656–66, 2006.

[91] E. Lyman, F. M. Ytreberg, and D. M. Zuckerman. Resolution exchange simulation. *Phys Rev Lett*, 96:028105, 2006.

[92] Michael Levitt and Arieh Warshel. Computer simulation of protein folding. *Nature*, 253:694 EP –, 02 1975.

[93] Philip Bradley, Kira M. S. Misura, and David Baker. Toward high-resolution de novo structure prediction for small proteins. *Science*, 309(5742):1868–1871, 2005.

[94] Wei Han and Yun-Dong Wu. Coarse-grained protein model coupled with a coarse-grained water model: A molecular dynamics study of polyalanine-based peptides. *Journal of Chemical Theory and Computation*, 3(6):2146–2161, 2007. PMID: 26636208.

[95] Carol A. Rohl, Charlie E.M. Strauss, Kira M.S. Misura, and David Baker. Protein structure prediction using rosetta. In *Numerical Computer Methods, Part D*, volume 383 of *Methods in Enzymology*, pages 66 – 93. Academic Press, 2004.

[96] Siewert J. Marrink, Alex H. de Vries, and Alan E. Mark. Coarse grained model for semiquantitative lipid simulations. *The Journal of Physical Chemistry B*, 108(2):750–760, 2004.

[97] M. Christen and W. F. van Gunsteren. Multigraining: an algorithm for simultaneous fine-grained and coarse-grained simulation of molecular systems. *J Chem Phys*, 124:154106, 2006.

[98] P. J. Miao, Y. Ortoleva. Molecular dynamics/order parameter extrapolation for bionanosystem simulations. *J Comput Chem*, 30:423–37, 2009.

[99] Y. Chen and B. Roux. Enhanced sampling of an atomic model with hybrid nonequilibrium molecular dynamics-monte carlo simulations guided by a coarse-grained model. *J Chem Theory Comput*, 11:3572–83, 2015.

[100] H. Fujisaki, M. Shiga, K. Moritsugu, and A. Kidera. Multiscale enhanced path sampling based on the onsager-machlup action: application to a model polymer. *J Chem Phys*, 139:054117, 2013.

[101] P. Liu, Q. Shi, E. Lyman, and G. A. Voth. Reconstructing atomistic detail for coarse-grained models with resolution exchange. *J Chem Phys*, 129:114103, 2008.

[102] P. Liu and G. A. Voth. Smart resolution replica exchange: an efficient algorithm for exploring complex energy landscapes. *J Chem Phys*, 126:045106, 2007.

[103] R. Lwin, T. Z. Luo. Overcoming entropic barrier with coupled sampling at dual resolutions. *J Chem Phys*, 123:194904, 2005.

[104] Kei Moritsugu, Tohru Terada, and Akinori Kidera. Scalable free energy calculation of proteins via multiscale essential sampling. *J. Chem. Phys.*, 133(22):224105, Dec 2010.

[105] Weihong Zhang and Jianhan Chen. Accelerate sampling in atomistic energy landscapes using topology-based coarse-grained models. *J. Chem. Theor. Comp.*, 10(3):918–23, Mar 2014.

[106] Paul C. Whitford, Jeffrey K. Noel, Shachi Gosavi, Alexander Schug, Kevin Y. Sanbonmatsu, and Jose N. Onuchic. An all-atom structure-based potential for proteins: Bridging minimal models with all-atom empirical forcefields. *Proteins*, 75(2):430–41, May 2009.

[107] Jeffrey K. Noel, Paul C. Whitford, Karissa Y. Sanbonmatsu, and Jose N. Onuchic. Smog@ctbp: Simplified deployment of structure-based models in gromacs. *Nucleic Acids Res.*, 38:W657–61, Jul 2010.

[108] Joseph T. et al. Jarrett. Seeding "one-dimensional crystallization" of amyloid: A pathogenic mechanism in alzheimer's disease and scrapie? *Cell*, 73:1055 – 1058, 1993.

[109] A. A. Profit, V. Felsen, J. Chinwong, E. R. Mojica, and R. Z. Desamero. Evidence of pi-stacking interactions in the self-assembly of hiapp(22-29). *Proteins*, 81:690–703, 2013.

[110] Todd M. Doran, Alissa J. Kamens, Nadia K. Byrnes, and Bradley L. Nilsson. Role of amino acid hydrophobicity, aromaticity, and molecular volume on iapp(20–29) amyloid self-assembly. *Proteins: Structure, Function, and Bioinformatics*, 80(4):1053–1065, 2011.

[111] W. M. Berhanu and U. H. Hansmann. Side-chain hydrophobicity and the stability of aβ_{16-22} aggregates. *Protein Sci*, 21:1837–48, 2012.

[112] M. F. M. Sciacca, D. Milardi, G. M. L. Messina, G. Marletta, J. R. Brender, and C. Ramamoorthy, A.and La Rosa. Cations as switches of amyloid-mediated membrane disruption mechanisms: Calcium and iapp. *Biophys. J.*, 104:173 – 184, 2013.

[113] S. Bedrood, Y. Li, J. M. Isas, B. G. Hegde, U. Baxa, I. S. Haworth, and R. Langen. Fibril structure of human islet amyloid polypeptide. *J Biol Chem*, 287:5235–41, 2012.

[114] Jakob T. Nielsen, Morten Bjerring, Martin D. Jeppesen, Ronnie O. Pedersen, Jan M. Pedersen, Kim L. Hein, Thomas Vosegaard, Troels Skrydstrup, Daniel E. Otzen, and Niels C. Nielsen. Unique identification of supramolecular structures in amyloid fibrils by solid-state nmr spectroscopy. *Angewandte Chemie International Edition*, 48(12):2118–2121, 2009.

[115] Jillian Madine, Edward Jack, Peter G. Stockley, Sheena E. Radford, Louise C. Serpell, and David A. Middleton. Structural insights into the polymorphism of amyloid-like fibrils formed by region $20-29$of amylin revealed by solid-state nmr and x-ray fiber diffraction. *Journal of the American Chemical Society*, 130(45):14990–15001, 2008. PMID: 18937465.

[116] Sorin Luca, Wai-Ming Yau, Richard Leapman, and Robert Tycko. Peptide conformation and supramolecular organization in amylin fibrils: Constraints from solid-state nmr. *Biochemistry*, 46(47):13505–13522, 2007. PMID: 17979302.

[117] W. Xu, H. Su, J. Z. Zhang, and Y. Mu. Molecular dynamics simulation study on the molecular structures of the amylin fibril models. *J Phys Chem B*, 116:13991–9, 2012.

[118] J. Zhao, X. Yu, G. Liang, and J. Zheng. Structural polymorphism of human islet amyloid polypeptide (hiapp) oligomers highlights the importance of interfacial residue interactions. *Biomacromolecules*, 12:210–20, 2011.

[119] L. H. Tu and D. P. Raleigh. Role of aromatic interactions in amyloid formation by islet amyloid polypeptide. *Biochemistry*, 52:333–42, 2013.

[120] L. Jiang, C. Liu, D. Leibly, M. Landau, M. Zhao, M. P. Hughes, and D. S. Eisenberg. Structure-based discovery of fiber-binding compounds that reduce the cytotoxicity of amyloid beta. *Elife*, 2, 2013.

[121] A. Kahler, H. Sticht, and A. H. Horn. Conformational stability of fibrillar amyloid-beta oligomers via protofilament pair formation - a systematic computational study. *PLoS One*, 8:e70521, 2013.

[122] I. Autiero, M. Saviano, and E. Langella. In silico investigation and targeting of amyloid beta oligomers of different size. *Mol Biosyst*, 9:2118–24, 2013.

[123] A. E. Berhanu, W. M. Masunov. Alternative packing modes as basis for amyloid polymorphismin five fragments. *Polypeptide. Biochemistry*, pages 131–144, 2013.

[124] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins*, 65:712–25, 2006.

[125] U. Zachariae, R. Schneider, R. Briones, Z. Gattin, J. P. Demers, K. Giller, E. Maier, M. Zweckstetter, C. Griesinger, S. Becker, R. Benz, B. L. de Groot,

and A. Lange. beta-barrel mobility underlies closure of the voltage-dependent anion channel. *Structure*, 20:1540–9, 2012.

[126] C. Kutzner, H. Grubmuller, B. L. de Groot, and U. Zachariae. Computational electrophysiology: the molecular dynamics of ion channel permeation and selectivity in atomistic detail. *Biophys J*, 101:809–17, 2011.

[127] W. M. Berhanu and U. H. Hansmann. Structure and dynamics of amyloid-beta segmental polymorphisms. *PLoS One*, 7:e41479, 2012.

[128] H. Ndlovu, A. E. Ashcroft, S. E. Radford, and S. A. Harris. Effect of sequence variation on the mechanical response of amyloid fibrils probed by steered molecular dynamics simulation. *Biophys J*, 102:587–96, 2012.

[129] S. Pronk, S. Pall, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl. Gromacs 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29:845–54, 2013.

[130] T. Darden, D. York, and L. Pedersen. Particle mesh ewald $\sim$ an $n.\ log(n)$ method for ewald sums in large systems. *J. Chem. Phys.*, 98:10089–10092, 1993.

[131] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen. A smooth particle mesh ewald method. *J. Chem. Phys.*, 103:8577–8593, 1995.

[132] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126(1):014101, Jan 2007.

[133] G. Bussi, T. Zykova-Timan, and M. Parrinello. Isothermal-isobaric molecular dynamics using stochastic velocity rescaling. *J. Chem. Phys.*, 130:074101, 2009.

[134] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12):7182–7190, 1981.

[135] W. L. DeLano. Pymol molecular graphics system, v 1.3.0.4. schrödinger, LLC, 2002.

[136] G. Liang, J. Zhao, X. Yu, and J. Zheng. Comparative molecular dynamics study of human islet amyloid polypeptide (iapp) and rat iapp oligomers. *Biochemistry*, 52:1089–100, 2013.

[137] R. Nelson, M. R. Sawaya, M. Balbirnie, A. O. Madsen, C. Riekel, R. Grothe, and D. Eisenberg. Structure of the cross-beta spine of amyloid-like fibrils. *Nature*, 435:773–8, 2005.

[138] W. M. Berhanu and U. H. Hansmann. The stability of cylindrin beta-barrel amyloid oligomer models-a molecular dynamics study. *Proteins*, 81:1542–55, 2013.

[139] A. W. Fitzpatrick, G. T. Debelouchina, M. J. Bayro, D. K. Clare, M. A. Caporini, V. S. Bajaj, C. P. Jaroniec, L. Wang, V. Ladizhansky, S. A. Muller, C. E. MacPhee, C. A. Waudby, H. R. Mott, A. De Simone, T. P. Knowles, H. R. Saibil, M. Vendruscolo, E. V. Orlova, R. G. Griffin, and C. M. Dobson. Atomic structure and hierarchical assembly of a cross-beta amyloid fibril. *Proc Natl Acad Sci U S A*, 110:5468–73, 2013.

[140] C. Wu and J. E. Shea. Structural similarities and differences between amyloidogenic and non-amyloidogenic islet amyloid polypeptide (iapp) sequences and implications for the dual physiological and pathological activities of these peptides. *PLoS Comput Biol*, 9:e1003211, 2013.

[141] P. Cao, P. Marek, H. Noor, V. Patsalo, L. H. Tu, H. Wang, A. Abedini, and D. P. Raleigh. Islet amyloid: from fundamental biophysics to mechanisms of cytotoxicity. *FEBS Lett*, 587:1106–18, 2013.

[142] M. R. Sawaya, S. Sambashivan, R. Nelson, M. I. Ivanova, S. A. Sievers, M. I. Apostol, M. J. Thompson, M. Balbirnie, J. J. Wiltzius, H. T. McFarlane, A. O. Madsen, C. Riekel, and D. Eisenberg. Atomic structures of amyloid cross-beta spines reveal varied steric zippers. *Nature*, 447:453–7, 2007.

[143] M. Shahnawaz and C. Soto. Microcin amyloid fibrils a are reservoir of toxic oligomeric species. *J Biol Chem*, 287:11665–76, 2012.

[144] C. E. Bulawa, S. Connelly, M. Devit, L. Wang, C. Weigel, J. A. Fleming, J. Packman, E. T. Powers, R. L. Wiseman, T. R. Foss, I. A. Wilson, J. W. Kelly, and R. Labaudiniere. Tafamidis, a potent and selective transthyretin kinetic stabilizer that inhibits the amyloid cascade. *Proc Natl Acad Sci U S A*, 109:9629–34, 2012.

[145] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–637, 1983.

[146] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case, and 3rd Cheatham, T. E. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc Chem Res*, 33:889–97, 2000.

[147] W. M. Berhanu, F. Yasar, and U. H. Hansmann. In silico cross seeding of abeta and amylin fibril-like oligomers. *ACS Chem Neurosci*, 4:1488–500, 2013.

[148] J. Park, B. Kahng, and W. Hwang. Thermodynamic selection of steric zipper patterns in the amyloid cross-beta spine. *PLoS Comput Biol*, 5:e1000492, 2009.

[149] Jesper Sørensen, David S. Palmer, and Birgit Schiøtt. Hot-spot mapping of the interactions between chymosin and bovine k-casein. *Journal of Agricultural and Food Chemistry*, 61(33):7949–7959, 2013. PMID: 23834716.

[150] Sara M. Butterfield and Hilal A. Lashuel. Amyloidogenic protein–membrane interactions: Mechanistic insight from model systems. *Angewandte Chemie International Edition*, 49(33):5628–5654, 2010.

[151] C. A. Lasagna-Reeves, C. G. Glabe, and R. Kayed. Amyloid-beta annular protofibrils evade fibrillar fate in alzheimer disease brain. *J Biol Chem*, 286:22122–30, 2011.

[152] J. D. Pham, N. Chim, C. W. Goulding, and J. S. Nowick. Structures of oligomers of a peptide from beta-amyloid. *J Am Chem Soc*, 135:12460–7, 2013.

[153] Ravi Prakash Reddy Nanga, Jeffrey R. Brender, Subramanian Vivekanandan, and Ayyalusamy Ramamoorthy. Structure and membrane orientation of iapp in its natively amidated form at physiological ph in a membrane environment. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1808(10):2337 – 2342, 2011.

[154] Hyunbum Jang, Laura Connelly, Fernando Teran Arce, Srinivasan Ramachandran, Bruce L. Kagan, Ratnesh Lal, and Ruth Nussinov. Mechanisms for the insertion of toxic, fibril-like β-amyloid oligomers into the membrane. *Journal of Chemical Theory and Computation*, 9(1):822–833, 2013. PMID: 23316126.

[155] James C. Stroud, Cong Liu, Poh K. Teng, and David Eisenberg. Toxic fibrillar oligomers of amyloid-β have cross-β structure. *Proceedings of the National Academy of Sciences*, 109(20):7717–7722, 2012.

[156] Pavan K. GhattyVenkataKrishna and Barmak Mostofian. Dynamics of water in the amphiphilic pore of amyloid β fibrils. *Chemical Physics Letters*, 582:1 – 5, 2013.

[157] K. K. Skeby, J. Sorensen, and B. Schiott. Identification of a common binding mode for imaging agents to amyloid fibrils from molecular dynamics simulations. *J Am Chem Soc*, 135:15114–28, 2013.

[158] B. Berg and T.Neuhaus. Multicanonical algorithms for first order phase transitions. *Phys. Lett. B*, 267(2):249–253, Sep 1991.

[159] U.H.E. Hansmann and Y. Okamoto. Prediction of peptide conformation by multicanonical algorithm: A new approach to the multiple-minima problem. *J. Comp. Chem*, 14(11):1333–1338, Nov 1993.

[160] Walter Nadler and Ulrich H. E. Hansmann. Generalized ensemble and tempering simulations: A unified view. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.*, 75(2):026109, Feb 2007.

[161] S. Duane, A.D. Kennedy, B.J. Pendleton, and D. Roweth. Hybrid monte carlo. *Phys. Lett.*, 195(2):216–222, Sep 1987.

[162] J. W. Neidigh, R. M. Fesinmeyer, and N. H. Andersen. Designing a 20-residue protein. *Nature Structural Biology*, 9(6):425–430, Jun 2002.

[163] G. J. Geyer and E. A. Thompson. Annealing markov chain monte carlo with applications to ancestral inference. *J. Am. Stat. Assn.*, 90(431):909–20, Sep 1995.

[164] A. Schug, W. Wenzel, and U. H. Hansmann. Energy landscape paving simulations of the trp-cage protein. *J Chem Phys*, 122:194711, 2005.

[165] D. Paschek, H. Nymeyer, and A. E. Garcia. Replica exchange simulation of reversible folding/unfolding of the trp-cage miniprotein in explicit solvent: on the structure and possible role of internal water. *J Struct Biol*, 157:524–33, 2007.

[166] D. Paschek, S. Hempel, and A. E. Garcia. Computing the stability diagram of the trp-cage miniprotein. *Proc Natl Acad Sci U S A*, 105:17754–9, 2008.

[167] A. Okur, L. Wickstrom, M. Layten, R. Geney, K. Song, V. Hornak, and C. Simmerling. Improved efficiency of replica exchange simulations through use of a hybrid explicit/implicit solvation model. *J Chem Theory Comput*, 2:420–33, 2006.

[168] X. Huang, M. Hagen, B. Kim, R. A. Friesner, R. Zhou, and B. J. Berne. Replica exchange with solute tempering: efficiency in large scale systems. *J Phys Chem B*, 111:5405–10, 2007.

[169] P. Kar, W. Nadler, and U. H. Hansmann. Microcanonical replica exchange molecular dynamics simulation of proteins. *Phys Rev E Stat Nonlin Soft Matter Phys*, 80:056703, 2009.

[170] M. Kouza and U. H. Hansmann. Velocity scaling for optimizing replica exchange molecular dynamics. *J Chem Phys*, 134:044124, 2011.

[171] A. Schug, T. Herges, and W. Wenzel. Reproducible protein folding with the stochastic tunneling method. *Phys. Rev. Lett.*, 91:158102, 2003.

[172] Y. He, Y. Xiao, A. Liwo, and H. A. Scheraga. Exploring the parameter space of the coarse-grained unres force field by random search: selecting a transferable medium-resolution force field. *J Comput Chem*, 30:2127–35, 2009.

[173] J. W. Pitera and W. Swope. Understanding folding and design: replica-exchange simulations of "trp-cage" miniproteins. *Proc Natl Acad Sci U S A*, 100:7587–92, 2003.

[174] Berk Hess, Carsten Kutzner, David van der Spoel, and Erik Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, 4(3):435–47, Mar 2008.

[175] See supplementary material at http://dx.doi.org/10.1063/1.4936968 for the modified program.

[176] Wendy D. Cornell, Piotr Cieplak, Christopher I. Bayly, Ian R. Gould, Kenneth M. Merz, David M. Ferguson, David C. Spellmeyer, Thomas Fox, James W. Caldwell, and Peter A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995.

[177] A. Onufriev, D. Bashford, and D. A. Case. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins*, 55(2):383–94, May 2004.

[178] Sandipan Mohanty, Jan H. Meinke, Olav Zimmermann, and Ulrich H. E. Hansmann. Simulation of top7-cfr: A transient helix extension guides folding. *Proc. Natl. Acad. Sci. (USA)*, 105(23):8004–7, Jun 2008.

[179] Robbyn L. Tuinstra, Francis C. Peterson, Snjezana Kutlesa, E. Sonay Elgin, Michael A. Kron, and Brian F. Volkman. Interconversion between two unrelated protein folds in the lymphotactin native state. *Proc. Natl. Acad. Sci. (USA)*, 105(13):5057–62, Apr 2008.

[180] Ian M. Sander, Julie L. Chaney, and Patricia L. Clark. Expanding anfinsen's principle: Contributions of synonymous codon selection to rational protein design. *J. Am. Chem. Soc.*, 136(3):858–61, Jan 2014.

[181] M. C. Tesi, E. J. J. van Rensburg, E. Orlandini, and S. G. Whittington. Monte carlo study of the interacting self-avoiding walk model in three dimensions. *J. Stat. Phys.*, 82(1-2):155–81, Jan 1996.

[182] J. D. Chodera and F. Noe. Markov state models of biomolecular conformational dynamics. *Current Opinion in Structural Biology*, 25:135–144, Apr 2014.

[183] C. Sinner, B. Lutz, S. John, I. Reinartz, A. Verma, and A. Schug. Simulating biomolecular folding and function by native-structure-based/go-type models. *Israel Journal of Chemistry*, 54(8-9):1165–1175, Aug 2014.

[184] I. G. Rodriguez-Bussey, U. Doshi, and D. Hamelberg. Enhanced molecular dynamics sampling of drug target conformations. *Biopolymers*, 105(1):35–42, Jan 2016.

[185] C. Abrams and G. Bussi. Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration. *Entropy*, 16(1):163–199, Jan 2014.

[186] D. Shukla, C. X. Hernandez, J. K. Weber, and V. S. Pande. Markov state models provide insights into dynamic modulation of protein function. *Accounts of Chemical Research*, 48(2):414–422, Feb 2015.

[187] Fatih Yasar, Nathan A. Bernhardt, and Ulrich H. E. Hansmann. Replica-exchange-with-tunneling for fast exploration of protein landscapes. *J. Chem. Phys.*, 143(22):224102, Dec 2015.

[188] Jan H. Meinke and Ulrich H. E. Hansmann. Protein simulations combining an all-atom force field with a go-term. *J. Phys.:Cond. Mat.*, 19(28):285215, Jul 2007.

[189] Cheng Zhang and Jianpeng Ma. Folding helical proteins in explicit solvent using dihedral-biased tempering (vol 109, pg 8139, 2012). *Proc. Natl. Acad. SCi. U.S.A.*, 109(40):16392–16392, Oct 2012.

[190] B. Zagrovic, E. J. Sorin, and V. Pande. beta-hairpin folding simulations in atomistic detail using an implicit solvent model. *J Mol Biol*, 313(1):151–169, Oct 2001.

[191] M. Andrec, A. K. Felts, E. Gallicchio, and R. M. Levy. Protein folding pathways from replica exchange simulations and a kinetic network model. *Proc Natl Acad Sci U S A*, 102(19):6801–6806, May 2005.

[192] Francisco J. Blanco, German Rivas, and Luis Serrano. A short linear peptide that folds into a native stable bold beta-hairpin in aqueous solution. *Nature Structural Biology*, 1:584–590, 1994.

[193] W.L. Jorgensen and J. Tirado-Rives. The opls force field for proteins. energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.*, 110(6):1657–1666, Mar 1988.

[194] A. Husebekk, B. Skogen, G. Husby, and G. Marhaug. Transformation of amyloid precursor saa to protein aa and incorporation in amyloid fibrils in vivo. *Scandinavian Journal of Immunology*, 21(3):283–287, 1985.

[195] A. Ardevol, G. A. Tribello, M. Ceriotti, and M. Parrinello. Probing the unfolded configurations of a beta-hairpin using sketch-map. *J. Chem. Theory Comput.*, 11:1086, 2015.

[196] Nathan A. Bernhardt, Wenhui Xi, Wei Wang, and Ulrich H. E. Hansmann. Simulating protein fold switching by replica exchange with tunneling. *J. Chem. Theory. Comput.*, 12(11):5656–66, 2016.

[197] Maksim Kouza and Ulrich H. E. Hansmann. Folding simulations of the a and b domains of protein g. *Journal of Physical Chemistry B*, 116(23):6645–6653, Jun 2012.

[198] Ludovico Sutto and Carlo Camilloni. From a to b: A ride in the free energy surfaces of protein g domains suggests how new folds arise. *J Chem Phys.*, 136(18):185101, May 2012.

[199] Jane R. Allison, Maike Bergeler, Niels Hansen, and Wilfred F. van Gunsteren. Current computer modeling cannot explain why two highly similar sequences fold into different structures. *Biochemistry*, 50(50):10965–10973, Dec 2011.

[200] Niels Hansen, Jane R. Allison, Florian H. Hodel, and Wilfred F. van Gunsteren. Relative free enthalpies for point mutations in two proteins with highly similar sequences but different folds. *Biochemistry*, 52(29):4962–4970, Jul 2013.

[201] K. Kachilshvili, G.G. Maisuradze, O.A. Martin, A. Liwo, J.A. Vila, and H.A. Scheraga. Accounting for mirror-image conformation as a subtle effect in protein folding. *Proc. Natl. Acad. Sci. USA*, 111:8458–8463, 2014.

[202] A Keith Dunker, Israel Silman, Vladimir N Uversky, and Joel L Sussman. Function and structure of inherently disordered proteins. *Curr. Opin. Struct. Biol.*, 18(6):756–64, 2008.

[203] A. Keith Dunker, J. David Lawson, Celeste J. Brown, Ryan M. Williams, Pedro Romero, Jeong S. Oh, Christopher J. Oldfield, Andrew M. Campen, Catherine M. Ratliff, Kerry W. Hipps, Juan Ausio, Mark S. Nissen, Raymond Reeves, ChulHee Kang, Charles R. Kissinger, Robert W. Bailey, Michael D. Griswold, Wah Chiu, Ethan C. Garner, and Zoran Obradovic. Intrinsically disordered protein. *J. Mol. Graph. Model.*, 19(1):26–59, 2001.

[204] H. Jane Dyson and Peter E. Wright. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell. Biol.*, 6(3):197–208, 2005.

[205] Jeevan B. Gc, Yuba R. Bhandari, Bernard S. Gerstman, and Prem P. Chapagain. Molecular dynamics investigations of the alpha-helix to beta-barrel conformational transformation in the rfah transcription factor. *J. Phys. Chem. B.*, 118(19):5101–8, 2014.

[206] W.; Kwak and Ulrich H. E. Hansmann. Efficient sampling of protein structures by model hopping. *Phys. Rev. Lett.*, 95(13):138102, 2005.

[207] Huiling Zhang, Wenhui Xi, Ulrich H. E. Hansmann, and Yanjie Wei. Fibril-barrel transitions in cylindrin amyloids. *J. Chem. Theory. Comput.*, 13(8):3936–44, 2017.

[208] J. Lee, K. Joo, B. R. Brooks, and J. Lee. The atomistic mechanism of conformational transition of adenylate kinase investigated by lorentzian structure-based potential. *J Chem Theory Comput*, 11 (7):3211–24, 2015.

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

[210] X. Daura, K. Gademann, B. Jaun, D. Seebach, W.F.van Gunsteren, and A.E. Mark. Peptide folding: When simulation meets experiment. *Angew. Chem. Int. Ed.*, 38(1):236–40, 1999.

[211] W Humphrey, A Dalke, and K. Schulten. Vmd: Visual molecular dynamics. *J. Mol. Graph.*, 14(1):33–8, 27–8, 1996.

[212] W. Han, C. K. Wan, and Y. D. Wu. Pace force field for protein simulations. 2. folding simulations of peptides. *J Chem Theory Comput*, 6:3390–402, 2010.

[213] V. S. Pande and D. S. Rokhsar. Molecular dynamics simulations of unfolding and refolding of a beta-hairpin fragment of protein g. *Proc Natl Acad Sci U S A*, 96:9062–7, 1999.

[214] Stefano Piana, John L Klepeis, and David 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. Folding and binding / Nucleic acids and their protein complexes.

[215] A. Kolinski. Protein modeling and structure prediction with a reduced representation. *Acta Biochim Pol*, 51:349–71, 2004.

[216] A. Liwo, M. Baranowski, C. Czaplewski, E. Golas, Y. He, D. Jagiela, P. Krupa, M. Maciejczyk, M. Makowski, M. A. Mozolewska, A. Niadzvedtski, S. Oldziej, H. A. Scheraga, A. K. Sieradzan, R. Slusarz, Y. Wirecki, T. Yin, and B. Zaborowski. A unified coarse-grained model of biological macromolecules based on mean-field multipole-multipole interactions. *J Mol Model*, 20:2306, 2014.

[217] A. Warshel and M. Levitt. Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J Mol Biol*, 103:227–49, 1976.

[218] Deepak R. Canchi, Dietmar Paschek, and Angel E. García. Equilibrium study of protein denaturation by urea. *Journal of the American Chemical Society*, 132(7):2338–2344, 2010. PMID: 20121105.

[219] Wei Han and Klaus Schulten. Characterization of folding mechanisms of trp-cage and ww-domain by network analysis of simulations with a hybrid-resolution model. *The Journal of Physical Chemistry B*, 117(42):13367–13377, 2013. PMID: 23915394.

[220] Wendy D. Cornell, Piotr Cieplak, Christopher I. Bayly, Ian R. Gould, Kenneth M. Merz, David M. Ferguson, David C. Spellmeyer, Thomas Fox, James W. Caldwell, and Peter A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995.

[221] Wei Han, Cheuk-Kin Wan, and Yun-Dong Wu. Toward a coarse-grained protein model coupled with a coarse-grained solvent model: Solvation free energies of amino acid side chains. *Journal of Chemical Theory and Computation*, 4(11):1891–1901, 2008. PMID: 26620333.

[222] W. Han, C. K. Wan, F. Jiang, and Y. D. Wu. Pace force field for protein simulations. 1. full parameterization of version 1 and verification. *J Chem Theory Comput*, 6:3373–89, 2010.

[223] Wei Han and Klaus Schulten. Further optimization of a hybrid united-atom and coarse-grained force field for folding simulations: Improved backbone hydration and interactions between charged side chains. *Journal of Chemical Theory and Computation*, 8(11):4413–4424, 2012. PMID: 23204949.

Appendix

Provided in this appendix are supplementary material to chapter 8. Specifically, table A.1 lists the temperatures and lambda values used in REMD and MSES/RET simulations.

REMD				**MSES/RET**		
Rep	Temp	Rep	Temp	Rep	λ_λ	λ_α
1	280.0	26	415.1	1	0.0	0.0
2	284.1	27	422.5	2	0.0075	0.5
3	288.2	28	430.1	3	0.020	1.0
4	292.4	29	438.0	4	0.060	1.5
5	296.7	30	446.0	5	0.18	2.0
6	301.1	31	454.3	6	0.4	3.0
7	305.6	32	462.8	7	1.0	4.0
8	310.2	33	471.6	8	2.5	5.0
9	314.9	34	480.6			
10	319.7	35	489.8			
11	324.6	36	499.3			
12	329.6	37	509.0			
13	334.7	38	519.0			
14	340.0	39	529.2			
15	345.4	40	539.7			
16	351.0					
17	356.6					
18	362.5					
19	368.4					
20	374.6					
21	380.9					
22	387.3					
23	394.0					
24	400.8					
25	407.8					

Table A.1: Temperature distribution used in REMD simulations and λ distribution used in MSES/RET simulations.