THE FREE ENERGY LANDSCAPE FOR THE FOLDING OF A SMALL ACID-SOLUBLE PROTEIN FROM MOLECULAR DYNAMICS SIMULATIONS

A Thesis

by

LAUREN MCGREGOR

B.S., The University of Texas at Austin, 2017

Submitted in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in

CHEMISTRY

Texas A&M University-Corpus Christi
Corpus Christi, Texas

May 2020

May 2020

THE FREE ENERGY LANDSCAPE FOR THE FOLDING OF A SMALL ACID-SOLUBLE PROTEIN FROM MOLECULAR DYNAMICS SIMULATIONS

A Thesis

by

LAUREN A. MCGREGOR

This thesis meets the standards for scope and quality of
Texas A&M University-Corpus Christi and is hereby approved.

Dr. Timothy Causgrove, PhD
Chair

Dr. Patrick Larkin, PhD
Co-Chair/Committee Member

Dr. Nicolas Holubowitch, PhD
Committee Member

May 2020

ABSTRACT

Intrinsically disordered proteins (IDPs) have been studied widely due to their abundance in biological systems and most notably, for their important roles in cellular functions. Despite their prevalence in nature, there is still much to be known about these proteins. Due to their dynamic nature, these proteins have posed a problem for scientists in the past who have tried to characterize them using traditional methods. In particular, the α/β-type small acid-soluble protein has been of interest. This IDP has been attributed as the main factor for bacterial spore survival under extreme conditions. Therefore, the determination of the binding mechanism of the α/β-type small, acid-soluble protein to spore DNA is essential for understanding its role in spore resistance.

In order to investigate the protein-DNA binding event, the C chain of the protein was first isolated and used to determine both folding and unfolding properties of the protein, with respect to their energies. This was achieved through the use of the molecular dynamics package Gromacs 2018.6 and the Plumed 2.4.4 plugin. In combination this allowed for a new form of accelerated conformational sampling to be achieved, known as Well-Tempered Bias-Exchange Metadynamics (WT-BEMETA). A total of 400 ns was obtained for an unbiased simulation (UB) of the C chain. Two biased simulations of the initial, folded structure obtained from X-ray crystallography (F1, F2) were simulated with a total of 100 ns collected for each. Lastly, two biased simulations of unfolded structures (UF1, UF2) were simulated with a total of 100 ns collected for each. These simulations incorporated the use of several collective variables for biasing, namely the number of hydrogen bonds by monitoring two separate regions of the helices (N_H), and the distance between the two helices (D_C). For each CV, a replica of the system was produced in order to apply the time-dependent bias potential. An additional, unbiased replica of the system was also incorporated into the exchange process. This combination yielded a total of 4 replicas for each of the 4 biased

systems using WT-BEMETA. Analysis with the Gromacs plug-in Metagui 3 allowed for the identification of prominent folded and unfolded structures, based on their sets of collective variables (CV) and corresponding energies for the F1. Two post-processing CVs (RMSD, root-mean-square deviation from the reference, folded structure and the number of native contacts, Q) were also used to extract more information about the structural states of the systems as well as to confirm the results. For the comparison of the unbiased simulation (UB) against the first, folded biased simulation (F1), the sampling efficiency was determined to be significantly improved by the incorporation of a bias. The sampling of the UB simulation confirmed that it was not capable of accessing unfolded structural states, even with a longer simulation time. These results were confirmed based on the small ranges of values obtained from the post-processing CVs. This outcome illustrated the need for advanced sampling techniques. For the biased systems, F1 and F2 were capable of sampling both folded and unfolded structural states based on the CV range of values that were obtained for the simulations. Meanwhile, UF1 and UF2 only sampled unfolded structures based on their corresponding CV values. These results were confirmed by the assessment of their free energy profiles, as well as with post-processing CVs. Therefore, more sampling was needed to allow the 4 WT-BEMETA systems to sample the same phase space and to improve reliability. The (UF1, UF2) systems were not used for further analysis, since they were not capable of sampling both folded and unfolded structures. Since F1 and F2 displayed similar results in their free energy profiles, only F1 was used for the determination of microstates and the free energy landscape, for simplification.

Ultimately, the use of WT-BEMETA allowed for the identification of prominent microstates of the protein as it unfolded. The free energy landscape was determined for these structures, where folded structures were associated with lower energies (1-3.5 kJ mol^{-1}) and

unfolded structures were associated with higher energies (3.5-12 kJ mol^{-1}). Since the IDP exists in an unfolded, structural state in nature, unfolded structures should correspond to lower energies. This meant that these results had only captured transitioning, unfolded structures of the protein, rather than true unfolded structures. The higher energies obtained for these transitioning structures were attributed to the initial breaking of hydrogen bonds as the protein unfolded. In order to obtain true unfolded structures, as well as to improve the reliability of the results, more sampling should be conducted in the future. With more simulation time, these results can be extended further in order to determine the binding mechanism of this IDP to spore DNA, as well as to understand the effects of an additional protein chain present. Ultimately, this research will aid in the development of identifying novel inhibitors that may prevent this binding process, as well as to avoid the consequences of malfunctioning IDPs.

DEDICATION

I would like to dedicate this thesis to my family, whom have supported me throughout my Masters. Thank you for motivating me every single day by your actions and words.

ACKNOWLEDGEMENTS

First, I would like to acknowledge Dr. Timothy P. Causgrove who served as my committee chair. I would like to thank him for being an excellent mentor throughout this research, as well as for all the knowledge and guidance he has provided. I would also like to thank my committee members, Dr. Patrick Larkin and Dr. Nicolas Holubowitch for their continued support and the additional knowledge they provided. Lastly, I would like to extend my gratitude to Texas A&M University-Corpus Christi, who provided me with their High Performance Computing (HPC) cluster. I would also like to thank Tom Merrick, for all of his help with the cluster.

TABLE OF CONTENTS

CONTENTS PAGE

LIST OF FIGURES

FIGURES PAGE

CHAPTER I – Introduction

1.1 Biological Importance of Intrinsically Disordered Proteins

Intrinsically disordered proteins (IDPs) are defined by their lack of secondary and tertiary structure in their unbound state, which is important for their function. These dynamic proteins sample a range of different conformations over time, giving them the advantage of binding to different targets and therefore producing different consequences. The structural plasticity of IDPs allows them to achieve functional modes that are inaccessible to folded proteins, including folding-upon-binding, the formation of transient complexes via nonspecific interactions, or interactions with rapid dissociation rates where the IDP remains dynamic in a so-called 'fuzzy' complex.[1,2] These proteins also exist based on a range of a disordered continuum, where there is no distinct classification that pertains to the amount of the disorder. The range of disorder can vary from having no defining structure, to random coiling and collapsed globules.[3] Due to their dynamic properties, IDPs are essential for facilitating numerous biological functions within cells of both prokaryotes and eukaryotes. This includes executing protein interaction processes such as molecular recognition, cell regulation, and signal transduction.[4] IDPs have become of growing interest recently for several important reasons. A significant fraction of eukaryotic proteins has been predicted to be composed of unstructured or disordered regions of more than 50 amino acids in length.[4] At first glance, these proteins seem to defy the well-known "form follows function" paradigm. However, for IDPs, this perspective is shifted since their random coiling features are essential to their function. Malfunctions of IDPs have also been linked to various diseases, as seen in the cases of the protein alpha-synuclein in Parkinson's, the tau protein in Alzheimer's, as well as the role of the tumor suppressor p53 in cancer formation.[2]

Intrinsically disordered proteins also serve a role in the spread of infectious bacterial diseases such as anthrax, tetanus, botulism, and food born illnesses. In particular, the protein-DNA interaction between the bacterial DNA and the intrinsically disordered α/β-type small, acid-soluble protein (SASP) is quite significant. When exposed to extreme environments, bacteria such as Bacillus and Clostridium are capable of forming endospores. In their spore form, they are able remain dormant for years, withstanding extreme conditions. Experimental data has confirmed the α/β-type small, acid-soluble protein as the main determinant for DNA resistance to damage caused by UV radiation, heat, and oxidizing agents in spores of Bacillus and Clostridium.[5,6,7] The spore DNA has been investigated under different conditions, both with and without the presence of the SASP in the spore's core. These proteins are crucial for spore DNA protection, as illustrated by the sensitivity of the DNA to many DNA-damaging agents. DNA sensitivity was compared *in vitro* between *B. subtilis* spores that lacked ~85% of their α/β-type SASPs to *B. subtilis* spores that contained higher levels of the bound α/β-type SASP to DNA, under extreme conditions.[6] In Setlow et al, experimental results illustrated that spores with low levels of α/β-type SASPs inside the core experienced significantly more deaths when held at temperatures of 85 °C for 30 minutes. In addition, these surviving spores had also acquired more mutations after being exposed. In contrast, spores with high levels of α/β-type SASPs inside the core did not experience as many auxotrophics or asporogenous mutations after being exposed.[6] In another experiment, Setlow et al illustrated that high levels of α/β-type SASPs also protected the spore's DNA from single-strand breaks caused by UV radiation.[7] The process of how the SASPs prevent this is still unclear, but has been linked to the prevention of DNA depurination. The results of their experiment illustrated that DNA depurination was also reduced in vitro at least 20-fold after exposure to radiation.[7] These survival rates were also similar to those obtained from placing bacterial spores under other extreme

conditions, such as cold temperatures and oxidizing agents. The effects of the α/β-type SASP on the DNA's properties have been investigated as well. Studies *in vivo* and *in vitro* have shown that α/β-type SASP proteins have profound effects on DNA properties.[6] Upon binding, the previously disordered protein adopts a highly alpha-helical structure. In the DNA, stiffening occurs in the helix and bends are eliminated. This interaction not only produces structural changes, but also changes in both the chemical and photochemical reactivity of DNA.[6] These changes have been linked to the preservation of the DNA, especially under UV radiation. Without this protection DNA degradation occurs, which is caused by the formation of cyclobutane-type pyrimidine dimers and photoproducts between adjacent pyrimidine residues.[6] In contrast, SASP-containing bacterial DNA are less susceptible to these degradation pathways under UV irradiation. Instead, a rather a protective intrastrand thyminyl-thymine adduct termed the spore photoproduct is formed in growing DNA cells.[6] This type of photoproduct is also beneficial, since it can be repaired much easier during outgrowth. Another consequence of the SASP-DNA interaction is that the protein exhibits cooperative binding with an additional amino acid, monomer. This phenomenon facilitates the binding of a second monomer at the active site, which is illustrated in the bound X-ray crystallographic structure.[1,8] The mechanism of SASP-DNA cooperative binding should also be investigated, in order to aid in the development of potential inhibitors. The work herein aims to establish a computational system that is capable of monitoring the folding and unfolding process of the isolated monomer, first. The information obtained can then be extended to the monomer-bound protein-DNA complex, as well as the dimer-bound complex. Ultimately, this work will lead to a better understanding of the cooperative binding event and its protective mechanism in bacterial DNA.

1.2 Literature Review of Experimental Methods

Due to its role in bacterial spore survival, the binding mechanism of SASPs is of current interest in drug discovery. However, IDPs in general pose many difficulties due to their interchanging short-lived conformational states in time and space. Traditional methods for studying their unbound state structure, such as X-ray crystallography and NMR, lack the ability to resolve all of their conformational states, therefore only capturing a snapshot of the protein's spectrum of states. Despite these difficulties, some researchers have managed to use advanced NMR techniques in order to acquire more information about their structural states. One example of a newer approach for studying IDPs is the use of in-cell NMR. Smith et al. used in cell-NMR to show that the α-Synuclein IDP retains a level of structural disorder comparable to unstructured peptides when measured within the bacterial cytoplasm both buffer and in *E. coli*.[9] Therefore, in-cell NMR has been illustrated as a suitable approach for studying these proteins. In addition to advanced NMR techniques, combined approaches are also often incorporated. Theillet et al. used a combination of in-cell Nuclear Magnetic Resonance (NMR) and Electron Paramagnetic Resonance Spectroscopy (EPR) techniques to characterize the structure of α-Synuclein as well as its dynamics in neuronal cell lines and non-neuronal cell lines.[10] This experiment has shown that combinations of NMR techniques can also be useful for determining both the structure and dynamics of IDPs as well. However, a major limitation of in-cell NMR is that protein overexpression is required to obtain a detectable signal. This oversaturation may also affect the interactions among the pool of interaction partners effectively rendering the contributions from key cellular interactions a minor fraction of the total detectable signal.[9]

Along with the use of in-cell NMR, solid-state NMR is also another common technique for studying IDPs. This technique recently demonstrated remarkable results for the IDP-related amyloid fibrils in particular, since they usually are insoluble and have larger molecular weights. Rienstra et al. recently reported a high-resolution structure of full length α-synuclein in fibrils, which reveals a β-sheet topology reminiscent of the Greek key motif commonly found in folded, globular, proteins.[11] However, proper sample preparation is extremely crucial in order to achieve reproducible results, since the structures are easily affected by factors such as the concentration of the peptide and the pH. NMR techniques have also been commonly used to study various post-translational modifications of IDPs, such as acetylation and phosphorylation.[11] With its continued application, advanced techniques of NMR can provide more information about these proteins, as well as how these covalent modifications can affect the protein's function. However, there are several difficulties that can affect the efficiency of NMR, as well as its spectroscopic interpretation. This ranges from sample preparation to deconvolution of the spectroscopic results, since dynamics are typically involved. Due to these difficulties, analysis is still an often, difficult process for most researchers today.

1.3 Molecular Dynamics Simulations

Computer simulations have been used in the past to investigate the linked folding and binding mechanism of several different IDPs to DNA.[12,13] The benefits of computational modeling have been unmatched, especially when experiments are quite difficult to perform as well as costly. Molecular dynamics (MD) simulations, in particular, are excellent for studying conformational changes in proteins and DNA study as well as, the binding mechanisms of proteins.[12] In general,

these simulations are often used to study large, complex systems present in biology. Molecular dynamics can be described as a technique where the time-evolution of a set of interacting atoms is followed by numerical integration of Newton's equation of motion.[14] Since molecular interactions are modeled by a given potential energy function, it is important to determine all relevant forces acting on the atoms and incorporate them accordingly. Based on these principles, determining the most appropriate implementation of the set of interatomic forces (force field) allows the most accurate representation of the system *in vivo*, as compared to *in vitro* studies in the laboratory. Force fields are used in simulations to describe the time evolution of bond lengths, bond angles and torsions, non-bonding van der Waals forces, and electrostatic interactions between atoms.[15] During each time step of the simulation, numerical integration of the potential functions produces new positions and velocities, which are defined as trajectories. These trajectory files contain all relevant physical quantities that are used to describe the system. With enough simulation time, typically on a timescale of nanoseconds to microseconds, most molecular interactions can be modeled.

1.4 Statistical Mechanics

Molecular dynamics simulations are capable of observing the microscopic chemical behavior of a system, however, they can also be extrapolated to describe macroscopic properties. This can be achieved if the Ergodic hypothesis is satisfied. Where with a long enough simulation time, the system will eventually sample all possible states.[16] Therefore, this means that the ensemble average is equal to the simulation time average. This connection is made via statistical mechanics, which provides the rigorous mathematical expressions that relate macroscopic

properties to the distribution and motion of the atoms and molecules of the N-body system; molecular dynamics simulations provide the means to solve the equation of motion of the particles and evaluate these mathematical formulas.[17] However, it can be both difficult and time consuming to run simulations long enough to sample all of the likely energy states. Conformational states that have relatively high energies or are separated from the current state by large energy barriers are less likely to be sampled throughout a short simulation time. In contrast, states that are lower in energy and easier to access will be sampled more frequently. This dilemma emphasizes the need for advanced MD approaches.

1.5 Incorporation of Metadynamics

In addition to traditional molecular dynamics approaches, a variant of accelerated conformational sampling will be used in order to improve sampling efficiency and decrease the computational demand. Within accelerated sampling, there are several methods that are typically categorized as either path-based or non-equilibrium approaches. For these simulations, non-equilibrium approaches were used since they do not require significant end-state knowledge prior to the simulation.[13] One of the most common non-equilibrium approaches works through the application of a bias potential known as metadynamics. By applying a bias, the system's potential energy landscape is modified, allowing it to reach more energetically favorable states and prevent the resampling of conformational states. The potential forces the system away from the kinetic traps in the potential energy surface and out into the unexplored parts of the energy landscape.[13] This perturbation is typically achieved by raising the energy levels that fall below a particular threshold. Metadynamics adds a history-dependent biasing potential (collective variable) in the

form of Gaussian functions, commonly referred to as Gaussian kernels. By summing the deposition of Gaussian kernels along the trajectory in the collective variable (CV) space, we can alter the probability distribution of sampling and reconstruct the free energy surface.[18,19] The technique is represented mathematically by the following equation.

$$V(\vec{s},t) = \sum_{k\tau<t} W(k\tau) \exp\left(- \sum_{i=1}^{d} \frac{(s_i - s_i(q(k\tau)))^2}{2\sigma_i^2} \right) \quad \textbf{Eq. 1}$$

Where $W(k\tau)$ is the height of the Gaussian, k is the energy rate, τ is the Gaussian deposition stride, and σ_i is the width of the Gaussian for the ith CV.[18] The exponential term includes the set of atomic positions s_i and the pre-determined CV as a function of the positions, $s_i(q)$.[18] In standard metadynamics, the height of the gaussian kernels are kept constant throughout the simulation. In contrast, a well-known variant of metadynamics known as well-tempered metadynamics decreases the height of the gaussian kernels over time.[18] This method allows for a more conservative approach when adding the bias potential and prevents from overfilling during loner simulation times. This method is described by the following mathematical equation.

$$W(k\tau) = W_0 \exp\left(- \frac{V(\vec{s(q}(k\tau)), k\tau)}{k_B \Delta T} \right), \quad \textbf{Eq. 2}$$

Where W_0 is an initial Gaussian height, ΔT an input parameter with the dimension of temperature, and k_B the Boltzmann constant.[18] In standard metadynamics, the bias potential behaves as a function of the selected CVs, where over a long time scale it will converge to the negative of the

free energy. This is illustrated in Eq. 3 without the presence of the temperature ratio, where $\Delta T=0$ would equal to standard metadynamics. In well-tempered metadynamics, the bias potential converges over a long time scale, but the free energy isn't always fully compensated for. This is due to the incorporation of a new parameter that describes a variation of temperature based on a ratio of the system's temperature and the CV's temperature. This relation can be described by the following equation.

$$V(\vec{s}, t \to \infty) = -\frac{\Delta T}{T + \Delta T} F(\vec{s}) + C$$

Eq.3

Where T is the temperature of the system and, with a long time limit, the CVs sample an ensemble at a temperature $T+\Delta T$, which is higher than the system temperature (T). [18] Using ΔT, the extent of free energy can be manipulated to match the free energy barriers of the system that want to be crossed. The manipulation of this term is described by the bias factor (γ) in present literature, which represents the ratio between the temperature of the system (T) and the temperature of the relevant CVs ($T+\Delta T$), as seen below.[18]

$$\gamma = \frac{T + \Delta T}{T}$$

Eq. 4

The ideal bias factor is scaled to cross relevant energy barriers and allow the system to efficiently escape from local minima throughout the duration of the simulation. Scaling is determined by plotting the time evolution of the selected CVs along with the Gaussian height. If in all simulations a diffusive behavior is observed in the biased CV when the Gaussian height is very small, and very similar free-energy surfaces are obtained, then we can be quite confident that the simulations are converging to an accurate value.[18] This diffusive behavior is illustrated by the fluctuation observed

within the graph, where the system is allowed to sample various energy states rather than being trapped in local energy minima. Scaling was carried out by summing the height for all individual Gaussian hills (which are stored in HILLS files during the simulation). This information correlated to the free energies and therefore, can be plotted for each CV against simulation time in order to determine the system has effectively escaped from its initial local minimum. The bias factor was set at 10 for all biased MD production runs after being scaled until this fluctuation was achieved.

In addition to the well-tempered bias, the application of bias-exchange metadynamics was also incorporated to further increase sampling efficiency.[20] This new technique was developed for complex systems that incorporate several CVs, since this would require a greater computational expense. The CVs used in a metadynamics simulation are not correlated to each other qualitatively and therefore, each one needs to be sampled enough in order to produce its own reconstructed free-energy landscape. The dimensions of the free-energy landscape of a system are also related to the number of biased CVs used in a simulation. The more CVs incorporated, the more the free-energy landscape grows. However, several or more CVs are usually needed to describe a complex system. With bias-exchange metadynamics (BEMETA), each CV is biased in parallel via multiple replicas of the same system at the same temperature, but differ by their time-dependent potentials. The exchange aspect of this method allows for even faster escape from low-energy states. Instead of waiting for each biased replica to converge to a free-energy equilibrium, an exchange is attempted at a given number of steps. This exchange is between randomly selected replica pairs. The probability that each exchange will be accepted is illustrated by a Metropolis rule.

$$P = \min(1, \exp\left[\beta(V_G^a(x^a, t) + V_G^b(x^b, t) - V_G^a(x^b, t) - V_G^b(x^a, t))\right]) \qquad \textbf{Eq. 5}$$

Where x^a and x^b are the coordinates of replicas a and b and $V_G^{a(b)}(x,t)$ is the metadynamics potential acting on the replicas.[21] The introduction of this jump process can be seen as an addition to the application of traditional metadynamics. The time-dependent Gaussian potentials also converge to the negative of the free energy. These jumps greatly increase the capability of each replica to diffuse in the CV space, and hence the accuracy of the free energy reconstruction.[16] Therefore, this method allows for each of the CVs to be sequentially biased using each of the different metadynamics potentials, still acting on one CV at a time. Throughout the simulation, each trajectory can evolve through the high dimensional free energy landscape produced by the multiple CVs, resulting in N one-dimensional projections of the free energy.[18,21] To further extend the computational efficiency of this method, the well-tempered bias will be applied rather than the standard biasing method, as described. This combination results in the emerging technique, known as well-tempered bias-exchange metadynamics (WT-BEMETA).

1.6 Biasing with Collective Variables

The Plumed 2.4.4 plug-in will be used for incorporating several CVs. Plumed is a plug-in that is commonly used in conjunction with Gromacs, since it is capable of working with various MD codes. Plumed can be used simultaneously in MD production runs as well as for post-processing analysis of simulations. CVs are advantageous for defining the processes of interest that occur in larger systems with a few parameters. CVs are defined as explicit functions of the system coordinates and represent the degrees of freedom of the system along which rare events are happening.[18] The two CVs determined to be the most beneficial for studying this IDP are the number of backbone hydrogen bonds (N_H) and the distance between contacts (D_C). The number of

backbone hydrogen bonds were monitored separately within the two helices of the protein that were resolved, designated as region 1 and 2. Both of these 2 regions contained hydrogen bonds that were present in the helical portions of the folded structure. This was determined by the Visual Molecular Dynamics Program (VMD), as described in the Methods section. The N_H CV is be defined by the following rational equation.

$$N_{\mathrm{H}} = \sum_{i \in \mathrm{O}} \sum_{j \in \mathrm{H}} \frac{1-\left(\frac{r-d_0}{r_0}\right)^n}{1-\left(\frac{r-d_0}{r_0}\right)^m} \qquad \textbf{Eq. 6}$$

Where r_{ij} is the distance between atoms i and j (O and H atoms of the backbone, respectively, where O is a carbonyl oxygen and H is an amide hydrogen).[22] In the equation, the allowed difference in distance between the atoms (d_0) was set to 0.25 nm and the initial distance between the atoms (r_0) was 0.3 nm for all of the biased simulations. The default parameters for n and m were also used, where n was 6 and m was 12. The sum was over pairs of atoms that form hydrogen bonds in a helical part of the folded protein. Since the protein is highly helical upon folding, the number of these hydrogen bonds was higher for the initial, folded structure. This allowed for the unfolding process to be monitored by observing decreases in the amount of hydrogen bonds as the simulation progressed. Both regions of the IDP that were investigated will be used to assess the likely structural states of the unfolded protein. For the second biasing CV, the distance between the contacting atoms of the two helices was monitored as D_C. This was achieved by distinguishing which atoms of interest would be monitored, as described in the Methods section. Instead of using direct atomic positions, their center of mass from these groups of atoms were used. This is a common technique used by Plumed to help ease the computational load, without losing important

data. After the contacts between the two helices were determined, the distances between them were then tracked by the D_C as the simulation progressed. In the starting structure, these distances were small since the two helices were folded towards each other. As the simulation progressed, D_C was capable of monitoring the protein's degree of unraveling by means of distance between the two helices. Therefore, the unfolded structures were associated with higher values of D_C. The changes in distance between the two helices were then used, in conjunction with N_H, to determine the likely structures and characteristics of the unfolded protein.

1.7 Post-processing with Collective Variables

Along with the biased CVs that were actively used in the WT-BEMETA simulations, two other CVs were also used in post-processing. Through post-processing techniques, further information can be extracted about the nature of the sampled structures. Nonetheless, these CVs can be used in the same manner despite not playing active roles during the simulation. The first CV used post-simulation was the root-mean-square deviation (RMSD) with respect to a reference structure. This CV is also known as the distance-RMSD and will be referred to as RMSD for simplification. The reference structure used for all calculations was the original, folded structure, exempting atoms that were not resolved by X-ray diffraction. By superimposing a selected structure from the simulation to the reference structure, the RMSD values can determine how much the structure has evolved at that point in time. This calculation is based on the following equation, where $\mathbf{x}_a$ and $\mathbf{x}_b$ refer to atom positions in the two structures.

$$d(\mathbf{X}^A, \mathbf{X}^B) = \sqrt{\frac{1}{N(N-1)} \sum_{i \neq j} [d(\mathbf{x}_i^a, \mathbf{x}_j^a) - d(\mathbf{x}_i^b, \mathbf{x}_j^b)]^2}$$

Eq.7

In the equation, N is the number of atoms and $d(x_i,x_j)$ represents the distance between atoms i and j.[23] Notably, it can be difficult to align structures efficiently due to a couple of complications. Implications with the alignment of structures may lead to falsely, high RMSD values. Since only the internal vibrational motions are of interest, it can be difficult to remove rotational and translational motions from the alignment. However, this issue can be resolved with the application of lower and upper cutoffs for the pairs of atoms that can be considered in the calculations. A high RMSD can have more ambiguity in regards to protein folding and unfolding. For example, the protein may have refolded, but that structure could still register as high RMSD if it did not reform as the exact structure of the reference. Along with this issue, the obtainable values for RMSD do not allow for an explicit interpretation of the structure, aside from having a distance of 0 nm. Where 0 nm, would indicate perfect alignment in regards to the reference structure, while any larger value doesn't necessary default to the unfolded structure. It is also difficult to gauge at what distance in between this range that the protein starts unfolding. Therefore, a high RMSD does not always correspond to an unfolded structure based on these situations. This is mainly due to the RMSD's sensitivity for more longer ranges of motion. This CV is appropriate for indicating change, however, it does not allow for the explicit interpretation of where in the protein that change is occurring. Nonetheless, the RMSD CV can still provide additional insight about the protein structure as well as confirm the results from other CVs.

The second post-processing CV was the fraction of native contacts, which can be referred to as Q. This CV works by using a switching function to calculate and transform the distances

between pairs of atoms, which are referred to as contact maps. The transformed distance is then used in comparison to a reference value (r_0 in the equation below), which allows for the squared distance between the two contact maps to be calculated. The contacts were defined by the selecting pairs of non-hydrogen atoms that were within 4.2 Å of each other, as well as more than 3 residues apart. By allowing residues that were 4 apart, this included contacts within an alpha helix. The alpha helix is also commonly defined as hydrogen bonds between *i* and *i+4* residues, for reference.[24] The switching function used is defined by the following equation, where *s(r)* is a defined minimum.

$$s(r) = \frac{1}{1+\exp(\beta(r_{ij}-\lambda r_{ij}^0))}$$

Eq. 8

In the equation, when $r \leq r_0$ then $s(r)=1.0$ and when $r>r_0$ the function decays smoothly to 0.[25] Default values of β (50 nm^{-1}) and λ (1.8) were used. Plumed offers various switching functions based on how the decay is treated, however, the specific function above was used by Best, et al.[26] for observing protein folding. Q is then defined as the sum of *s(r)* over all *i,j* pairs divided by the number of pairs (237 in our protein). By the use of Q, characteristics of protein folding and unfolding can easily be determined in post-processing. Since a high value for Q (close to 1) would only pertain to these specific pairs of atoms remaining in close contact, there is less ambiguity in high values for this CV. This is also true for low values of Q, since only a lack of contacts would yield a low value for distance. Without the issues caused by structure alignment, Q can be used to help distinguish the results obtained by RMSD as well as confirm them.

1.8 Analysis with METAGUI

Along with Plumed 2.4.4, a new plug-in that provides a graphical user interface (GUI) specifically for determining the thermodynamics, microstates, and kinetic basins of complex simulations was incorporated. This plug-in is known as Metagui 3, which is based on the visual molecular dynamics program 1.9.3 (VMD). Within VMD, access to both the Plumed plug-in and the Metagui 3 plug-in is available. For Plumed, VMD offers a tool that supports the analysis for CVs used in the simulation as well as the capability of appending new CVs used in post-processing techniques. Within Metagui, 3D analysis of the systems was easily achieved for all relevant microstates and their corresponding free energies. Metagui was also used to assess the sets of CV values acquired for each microstate, which was ultimately used to narrow down likely unfolded structures of the IDP. Therefore, Metagui was used to describe the free energy landscape of the unfolded structures that the folded protein could access.

1.9 Literature Review of MD Simulations on IDPS

While first originating in the 1950s with Alder and Wainwright, MD methods have been significantly modified throughout the decades in order to develop the techniques that are most commonly used today.[13] By the 1970s, MD simulations were synonymous for studying complex systems, such as proteins. One of the earliest MD simulations was conducted on a small bovine pancreatic trypsin inhibitor (BPTI) with a total simulation time of 9.2 picoseconds.[27] Today, these simulations typically run on a scale of hundreds of nanoseconds to microseconds and are capable of providing information on explicit, atomistic scale. Using MD simulations, researches are

capable of gaining extensive information about their system of interest such as loop motions, allosteric transitions, helix-coil transitions, rotation of solvent-exposed side chains, as well as binding related conformational changes.[13] Researchers have been particularly interested in studying IDPs related to aggregation, which can lead to diseases such as Alzheimer's and amyotrophic lateral sclerosis (ALS), for example. Recently, researchers have produced atomic-level simulations of the monomeric TDP-43 protein that revealed local helix stabilizations at the C-terminus in the presence of ALS mutations.[28] The identification of these mutations has been essential to understanding their effects on the role of the protein in relation to the progression of the disease. Another simulation illustrated that the known aggregation propensity of four amylin sequences, an implication of type-2 diabetes, can be explained by their local helical propensity differences.[29] Therefore, these small differences in helical propensity have been attributed to the aggregation of this protein. For Alzheimer's disease, researchers have also been interested in studying an intrinsically disordered α/β peptide that has been linked to early onset of the disease based on acquired mutations. Simulations from several groups have shed light onto the structural alterations resulting from those point mutations in the full-length and fragment α/β, particularly within the monomeric and dimeric landscape.[30]

Since simulations occur on short time scales with respect to the events that they are trying to monitor in nature, advanced sampling MD techniques are commonly used to improve sampling efficiency. Therefore, researchers are turning to accelerated approaches in order to better capture these processes of interests for IDPs. Accelerated sampling can be achieved in many ways, but one of the most common techniques has been shown as the modification of bias potentials.[13] Bias potentials are typically added through the use of collective variables (CVs) and the most efficient way to incorporate them is in parallel replicas of the system. One recent example of this method,

was an intrinsically disordered peptide derived from the Neh2 domain of the Nuclear factor erythroid 2-related factor 2 (Nfr2) protein that was monitored using well-tempered metadynamics (WT-META) and bias-exchange metadynamics (BE-META).[12] They acquired over 10 microseconds of simulation that allowed them to produce a free-energy reconstruction of the free protein as well as confirm the presence of the β-hairpin conformation. Most notably, they found that these two methods were very comparable, with the exception of BE-META sampling slightly more clusters.[12] In contrast to replica exchanges where the system's bias potential is modified using various replicas at different temperatures, the bias exchange method allows for this modification to be done through the use of collective variables. In another study, the benefits of BE-META were further highlighted as researches were able to analyze the conduction and selectivity in Na^+ ion channels. Using this method, they were able to calculate free energy profiles in the presence of a variable number and type of permeating ions.[31] Their collective variables of choice also corresponded to the system they were investigating, such as the distance along the channel-axis between an ion and the center of mass of the carbonyl oxygen of atoms of specific residues.[31]

Until 2011, the α/β-type small acid soluble protein in particular, had not been studied. However, Ojeda-May and Pu recently conducted a replica exchange molecular dynamics simulation on the α/β-type small acid soluble protein.[32] However, they conducted their simulations using a different molecular dynamics package called Chemistry at Harvard Macromolecular Mechanics (CHARM) and a different force field associated with the package. Additionally, replica exchanges were conducted using various temperatures of the systems. In contrast, our research performed replica exchanges on systems that were biased by different CVs. They were able to gain information about the melting temperature and overall, their research was able to confirm that only

a small free energy barrier (4.184 kJ/mol) separated the conformational ensembles at high and low temperatures.[32] In addition, they used several different CVs for post-processing such as the radius of gyration and end-to-end distance of the protein chain. Ultimately, by using different CVs, they were able to gain different information about the system, such as the weight of clusters in relation to their melting temperatures. Their results provide essential information about the particular IDP along with the results obtained from this experiment. Ultimately, advanced MD techniques have been widely used to determine the structural propensities of difficult systems, as well as to determine the free energy landscapes for particular structures of interest.

1.10 Research Goals

The purpose of this research was to determine the nature of protein folding and unfolding for the α/β-type small, acid-soluble protein. The information derived from these simulations can also be used in future research to determine the binding mechanism of this protein to bacterial spore DNA. Further investigation can lead to the determination of potential inhibitors for the binding mechanism of the α/β-type SASP as well as other IDPs. IDPs have been studied extensively in the past, both experimentally and through molecular dynamics simulations; however, the α/β-type SASP has only been studied minimally by means of MD simulations, with physical data also being quite limited. Results of this research aims to provide more insight of these proteins, fill in gaps that past research has not yet answered, and ultimately guide the development of novel IDP inhibitors.

CHAPTER II-Methods

2.1 Computational Resources

All MD production runs were computed using Gromacs 2018.6 and the Plumed 2.4.4 plugin. All production runs were done on TAMUCC's Tsunami high performance computing cluster, which contains 1 head node, 44 cpu nodes, and 4 gpu nodes. The cluster contains a total core count of 1288 and 11.7 TB of memory. Each computer node is equipped with two Xenon processors, while each gpu node is equipped with two Xenon processors and two Nvidia Tesla K20XM gpus. The cluster utilizes both SLURM and Bright Cluster Manager for processing all tasks. The HPC is accessible to those affiliated with TAMUCC and is funded by the National Science Foundation.

2.2 Obtaining the PDB file

A PDB file was first obtained from the Protein Data Bank, containing atomic positions that have been experimentally derived from its X-ray crystallography structure. The PDB file essentially contains all atomic positions. Under most circumstances, the PDB file must be edited prior to preparing the system. In the initial X-ray crystallography, there were three proteins present in addition to a section of double-stranded DNA, with only two being bound to the spore DNA. The unrelated protein, labelled as B, was removed from the PDB file prior to further necessary editing. In addition, some atoms could not be resolved by X-ray crystallography and had to be added in VMD. Since only chain C of the protein was used for the simulations, both chain A and

the DNA had to be removed as well. This was done for simplification, while system parameters were first being determined. Due to time limitations, other system combinations were not used for the biased simulations. Therefore, all reported results pertain to chain C of the protein structure. All protein chains share the same amino acid sequence and overall folded structure.

2.3 Preparation of the System

The PDB file was converted into a file that could be read into Gromacs. Along with this, hydrogen atoms are not detected by x-ray crystallography and therefore, were generated. Both of these were done using the Gromacs program, pdb2gmx. In this program, a topology file was produced that contained a molecular description of the complex, including all interactions such as covalent bonds, for example. During this step, the force field was also chosen, which decides what information is written into the topology file. For this experiment, the AMBER ff99SB*-ILDN force field was used since it was optimized for helix-coil transitions as well as side chain torsions. During this step, the explicit solvation model TIP3P was also chosen in order to obtain a full description of the protein-DNA structure. Along with a topology file (containing energy parameters necessary for simulation), a gro file was produced which contains the same information as the pdb file, but now included the velocities.

The next step of preparation for the system was defining the periodic boundary conditions. This step included defining the simulation box and filling it with the appropriate amount of water molecules. Gromacs used the programs editconf and solvate to accomplish this. A dodecahedral box was used in order to cut down on the integration time that would have been attributed to water molecules located at the corners of the box. The box size contained a 1.8 nm buffer in order to

sufficiently enclose the protein system and allow for it to unfold during the simulation time. After generating the box, the system was then solvated using the chosen solvent model. The output of both of these functions were new gro files and an updated topology file.

After the system was solvated, the next step was to obtain a zero net-charge of the system, since biological systems do not exist with a net charge. This was done by using the genion program on Gromacs, which replaced solvent molecules with ions in the topology file. Prior to genion, the grompp module on Gromacs was used in order to produce a run input file for genion. Grompp created a new tpr file by processing the topology file, the coordinate file, and the molecular dynamics file (mdp). The new run input was then used for genion. Along with neutralization, the system was also modified to contain a concentration of 100 mM NaCl, simulating a biological ionic environment.

Once the box has been defined, solvated, and ions added, the next step was to conduct an energy minimization. This step is carried out prior to the production run in order to reduce the number of bad contacts and to identify local energy minimums that the system would likely be in. This is also carried out similarly to the ionization step, using grompp to create the concatenated tpr file and mdrun to carry out the simulation.

After completing the energy minimization step, the system was then equilibrated in order to optimize both the solvent and solute. By applying restraints, the system was then brought to the correct temperature, followed by the correct pressure needed to carry out the simulation. In the first phase of equilibration, the number of particles, volume, and temperature (NVT) was held constant. After the temperature was stabilized, the next ensemble held the number of particles, pressure, and temperature (NPT) constant. Once the pressure had also been stabilized, the system was then ready data collection.

This method was used for creating five systems, each to be used as the starting point for a simulation production run. The first system created was used to run the unbiased simulation of chain C, which was referred to as UB. This system was ultimately used in order to demonstrate the need for metadynamics as well as determine the appropriate parameters for CVs. After preparation until this point, UB was ready for production runs on the cluster.

2.4 Setting up WT-BEMETA Simulations

The next two systems created were referred to as F1 and F2, for folded system 1 and folded system 2. The two identical systems were both used for WT-BEMETA runs. Having two systems with the same starting structures served to assess potential discrepancies between the WT-BEMETA runs, as well as to confirm reproducibility.

The last two systems prepared were the two unfolded protein structure systems, referred to as unfolded system 1 (UF1) and unfolded system 2 (UF2). These two systems were used for WT-BEMETA runs as well. The only difference in preparation for the unfolded systems was that the starting structures were two non-identical, unfolded structures of the protein. The two unfolded structures were acquired by a random selection of unfolded structures from past trial runs, which began from randomly structured extended chains. The identities of their unfolded states were also confirmed in VMD. By using the same starting configurations on different unfolded structures, the results of the simulations were able to capture the reverse mechanism with more reliability. If the protein was capable of sampling the same folded and unfolded structures, regardless of the starting structure, this would be indicative of sufficient sampling, as well as convergence between the systems. Therefore, the results would be regarded as more reliable and reproducible.

As mentioned, essentially 2 CVs were used for biasing in the WT-BEMETA simulations labeled F1, F2, UF1, and UF2. These 2 CVs correspond to N_H and D_C; however, N_H was divided into two regions. This yielded a total of 3 CVs that were incorporated within the simulations. By nature of WT-BEMETA, this technique involved the exchanges of replicas. Notably, only 1 CV actively biased a system at a time, which equated to 3 replicas (1-3) for each simulation. An additional replica that was not biased by any CV, but still participated in the random exchanges, was used to serve as a baseline. This was referred to as replica 0 for all simulations. This combination produced a total of 4 replicas (0-3) that would be ran simultaneously on the cluster, while conducting random exchanges. In post-processing analysis, more CVs were used to extract data, however, they were not used in the molecular dynamics simulations.

For replicas 1 and 2, these systems were biased by the N_H CV, but were separated by the two distinct regions they were monitoring. As seen in Figure 1, the helical portions of the folded protein can be clearly seen in VMD. This was completed by the "Sequence Viewer" analysis tool under Extensions with the dry version (all waters removed) of the folded structure.

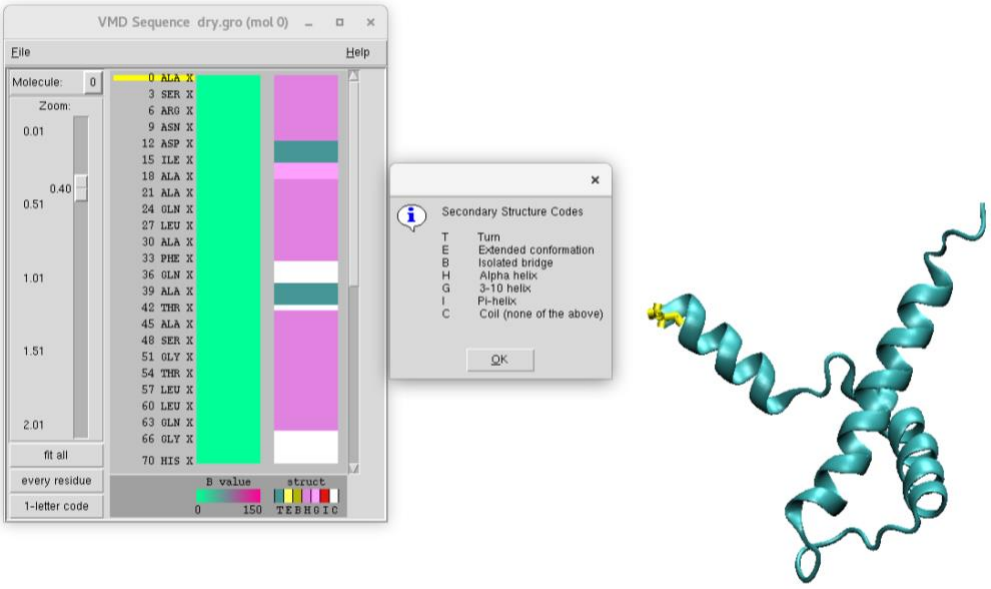

Figure 1. Sequence Viewer tool within VMD to illustrate the helical portions of the protein based on amino acid residue content. The alpha helical structure code is represented by H and purple, as denoted in the secondary window. The two closest, visible helices were used for tracking by means of the N_H CVs.

There were three separate portions of helical content within the protein that can be seen within the VMD tool, which were denoted by the code H and purple shade. The first region pertained to the N-terminal region, denoted by the appearance of yellow in both the sequence viewer and on the protein strand. This helical portion was not included, since the residues were not resolved in the X-ray structure. Its helical structure is an artifact of adding these residues back to the protein, as this was the most compact structure to append the residues. This section was not part of any biasing and was also ignored in analyses such as RMSD and fraction of native contacts. However, it should be noted that experiments with this protein[33] indicate that these N-terminal residues may play an active role in binding to DNA and are therefore relevant.

Based on the other two distinct helical regions, two separate CVs were made for tracking the number of hydrogen bonds that lie within them. The first region (residues 19 to 37) corresponds to the term hbonds1 and "HB1" for short, while the second region (residues 43 to 68) corresponds

to the term hbonds2 and “HB2” for short. Since these two regions were separate from each other, the number of maximum hydrogen bonds present in the folded state differed. In turn, this affected the maximum value that the CV could attain for these regions. This corresponds to their associated grids, where the grid defines to the range of values that the CVs were allowed to sample within the simulation. Therefore, the grid maximums for HB1 and HB2 were 15 and 22 hydrogen bonds, respectively. This was due to a longer helix and therefore more backbone hydrogen bonds within the region that was monitored by HB2. Since the lowest number of hydrogen bonds that may detected by the CV was 0, the grid minimums were both set at 0. Essentially, one tool is being used to monitor the changes in two, separate areas of the protein that are helical in nature upon folding. Furthermore, replicas 1 and 2 for all of the simulations correspond to the systems that were biased by the HB1 CV and the HB2 CV, respectively.

For replica 3, the systems were biased by the distance CV, D_c, which was referred to as D1. The D1 CV corresponded to the distance between the center of mass (COM) of atoms, where the two helices were in contact with each other in the initial folded structure. This selection is illustrated by the following figure.

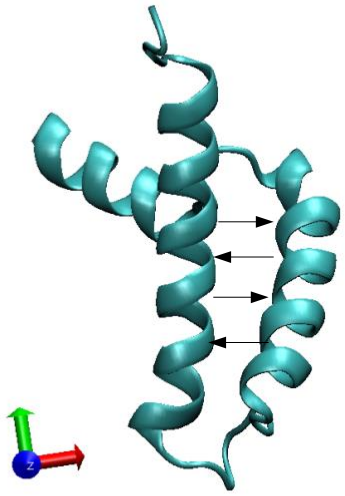

Figure 2. The (dry) folded protein structure illustrated by VMD, where the arrows indicate the selection of atoms used for calculating the COM to track with the D1 CV. These atoms reside within the two main helices of the protein and were monitored for changes in contact distance.

The grids for D1 were also set at 0 and 5.5 Å, based on the folded structure in VMD. The use of the D1 CV allows for another method of tracking protein folding that is also capable of providing definitive results. After each of the CVs were allocated to their corresponding replicas for biasing, all 4 systems were ready for MD production runs.

2.5 Production Runs

After the 5 systems (1 unbiased, 4 biased) had been produced and equilibrated, they were ready for data collection. During this step, the system was transferred to TAMUCC's HPC cluster for the rest of the simulations. The HPC cluster allowed for significantly faster computing and it can store data on a terabyte scale, as mentioned. For optimal computing efficiency, the HPC cluster's gpu nodes were used to conduct the simulations. For the four systems that were biased, the incorporation of Plumed variables was achieved through the assimilation of Plumed input files,

referred to as dat files. These files were prepared in the same manner as previous MD files mentioned. For the WT-BEMETA runs, this included information about sigma, the bias factor, and the height of the Gaussian deposit. The height for all biased replicas (1-3) was set at 1.2 and the bias factor was set to 10.0. For replicas 0-3, their sigma values were set at 1.0, 1.0, 0.5, and 0.2, respectively. These dat files were then read by the activation of Plumed in Gromacs and produced colvar and hills files throughout the duration of the MD production simulation. The colvar file contained the values of all CVs as a function of simulation time. The hills file contained additional information about the Gaussian kernels that were deposited throughout the simulation. The hills files were especially useful for determining free energies, testing for convergence, and for producing plots of the simulation's behavior. As mentioned a total of 400 ns was obtained for the unbiased simulation and 100 ns was obtained for each of the biased simulations (F1, F2, UF1, FU2). Because the biased simulations included four replicas each, these five simulations were comparable in computational effort.

2.6 Post-Processing

Along with the incorporation of the CVs to access the alpha-helical propensities of the IDP structure throughout the duration of the experiment, post-processing analysis was conducted for further analysis. This was completed in VMD 1.9.3 along with the additional Metagui 3 plug-in mentioned. The program VMD has support for traditional MD simulations and was sufficient for analysis of the unbiased system. For the biased systems, all analysis was conducted with the use of Metagui. The Plumed 2.4.4 plug-in for VMD was used for calculating the two post-simulation CVs for all 5 systems as well.

Since each replica for the biased simulations participated in multiple exchanges throughout the entirety of the simulation, by nature the files had to demuxed first in order to obtain a continuous trajectory for each replica. Demuxing files corresponds to a method of reconstructing these trajectories from all of the individual, replica files produced. In simulations involving exchange, one must be careful and examine the time evolution of each replica diffusing in CV space, before conclusions can be made that the simulation is converged.[34] The demuxed files were then loaded within the *define inputs* tab of Metagui. Aside from the unbiased simulation, the frames of the systems were all set to begin at 20,000 rather than 1. This was done to remove earlier frames that were associated to the system equilibrating, without running the risk of removing critical frames. For all biased systems the CV were all set as active and 15 grid points were assigned to each one. Their grid minimum and maximums all corresponded to their initial Plumed input files that contained these values.

After defining the inputs, the *Analyze* tab of Metagui was then used to find the microstates and assess their thermodynamic properties. The k-medoids method by Kaufman and Rousseuw[35] was also used to assign the microstates. For assessing the thermodynamics, ΔG was computed with a delta=2kT, kT=2.4943 kJ mol^{-1}, and an equilibration time of 1 frame. The output was used to produce the 1D free energy profiles that were used to assess for convergence. Within the *Visualization* tab, all relevant information about the systems were sorted based on their assignment of structures to these microstates. This was done automatically in Metagui based on the sets of CV values that were computed for each of the sampled structural states.

This information was then used to assess the thermodynamics and microstates of the 4 biased simulations, in terms of their folded and unfolded structures. Within Metagui, all relevant

microstates were illustrated in conjunction with their corresponding CV values to produce 2D projections.

2.7 Additional CVs

During post-processing, two additional CVs were appended for all 5 simulations. For the unbiased simulation, these were the only CVs used to obtain data since it was not actively biased. Meanwhile, these CVs were used for additional analysis for the biased simulations, resulting in a total of 6 CVs that were used to obtain information about the simulations.

The first post-processing CV was the distance RMSD of the reference structure as described. This reference structure that was used for the calculation was the original, folded structure of the protein. The structure that was used as the reference was also illustrated in Figure 2. This was the only additional input that was required for the CV as well. Based on this structure, the RMSD was calculated for each of the microstates based on their differences to this respective structure. This was done in order to assess protein folding/unfolding, as well as to monitor sampling fluctuation. Implications of this CV were discussed further in regards to alignments of structure.

Similar in nature, the second CV was used to determine the fraction of native contacts (Q). This CV required the input of a set of atom pairs, which were referred to as contact maps. These pairs of non-hydrogen atoms were selected in VMD based on the criteria of being within 4.2 Å of each other, as well as more than 3 residues apart. This was done to assure that contacts of the alpha helices were incorporated within the CV. After these selections were made, they were read by Plumed, in order to carry out the necessary switching function. From these results, the distance

between these contacts was tracked throughout the progression of the simulations. This allowed for the assessment of sampling fluctuation and folded and unfolded states without issues related to alignment.

CHAPTER III-Results and Discussion

3.1 Simulation Results

A total of 400 ns of simulation time was collected for the unbiased folded-protein system (UB) using the setup as described in the Methods section. For the two starting configurations (F1, F2), a total of 100 ns of simulation time was collected for each. A total of 100 ns of simulation time was also collected for the two unfolded-protein systems (UF1, UF2).

3.2 Comparison of Unbiased Simulations

The unbiased simulation of the folded-protein (UB) was used for several reasons. This system was first prepared and then used to determine the appropriate parameters for running metadynamics. This included determining the appropriate CVs and their corresponding parameters. The UB system was also used to illustrate the need for metadynamics, and in particular, WT-BEMETA. This need was illustrated by the lack of sampling areas of phase space. Even with a substantially longer simulation time at 400 ns, the results of UB did not compare to those of F1and F2 at 100 ns. This result was due to the unbiased system not being able to escape local minima as easily during the time period. This ultimately resulted in the lack of sampling of higher energy states as well as states that were separated by a high energy barrier. Therefore, without the help of the energy bias, the system was unable to sample all likely states for the folded and unfolded structures. Since the simulations occur only on a nanosecond scale, one cannot draw

any conclusions about what is occurring in nature without sufficient sampling. For example, an event that occurs instantaneously in nature may actually take hours to mimic in a simulated environment.[25] Therefore, due to these limitations, it is critical to be able to achieve high sampling efficiency within a short period of time.

In order to assess if the simulations were sampling both unfolded and folded structures, post-processing CVs were used to track variations in regions linked to folding. Ideally, sampling both the unfolded and folded structures should lead to large fluctuations in the CVs. Since UB was not biased by CVs during the simulation, only post-processing CVs were able to be used as comparison. As previously described, these two CVs were the RMSD relative to the reference structure (RMSD) and the fraction of native contacts (Q). First looking at the RMSD from the reference structure, Figures 3 and 4, illustrate the different outcomes of UB and the F1 simulations, where F1 incorporated WT-BEMETA.

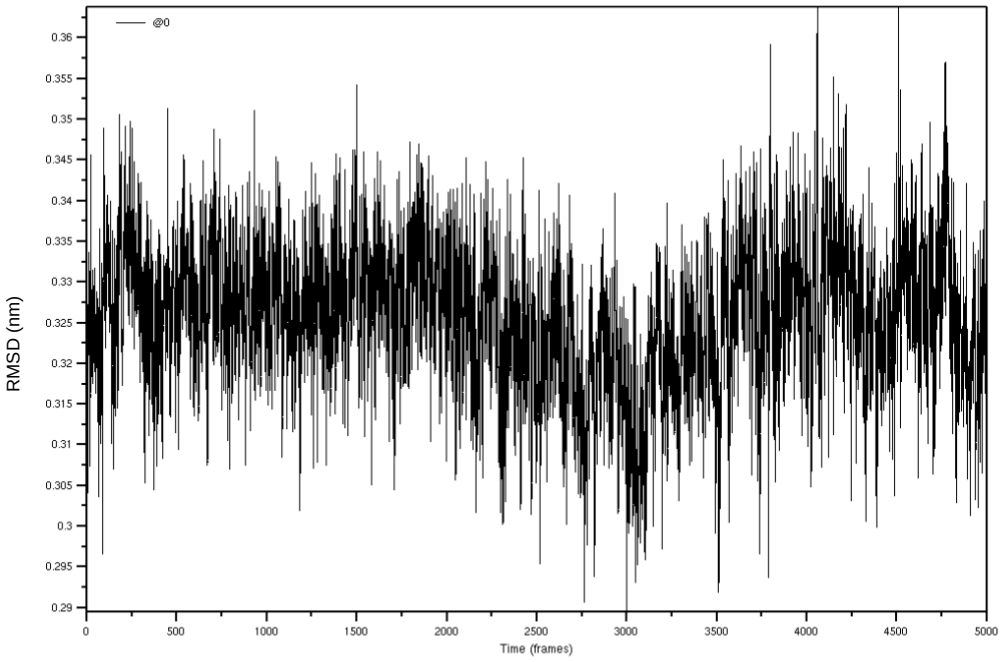

Figure 3. Plot of the post-processing, RMSD CV against simulation time for the unbiased simulation of the folded structure (UB). Where the RMSD is in nm and time is represented by the number of frames (standard in Metagui), which totaled 5,000. The total simulation time was 400 ns. The range of values for RMSD fluctuated from 0.29-0.37nm.

Since F1 and F2 illustrated similar results, only F1 is shown for comparison against UB to avoid complication. For F1, replica 0 was used as a direct comparison. As described, replica 0 for all of the simulations involving WT-BEMETA corresponds to the system that was not biased by a CV. However, the replica still participated in random exchanges during the simulation. Therefore, this replica served as a baseline for the rest of replicas in the WT-META simulations.

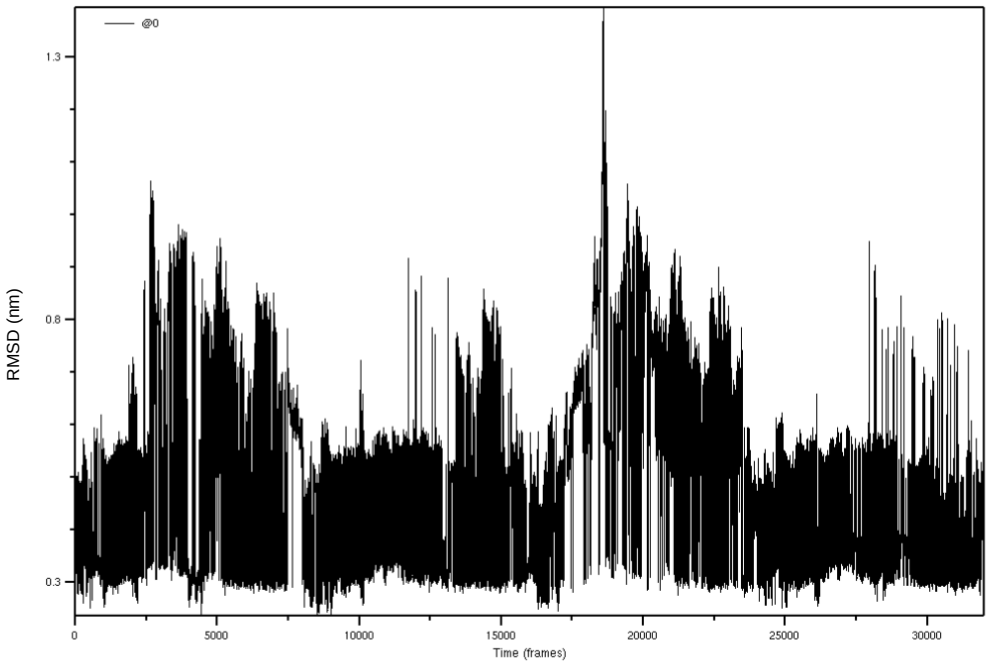

Figure 4. Plot of the post-processing, RMSD CV against simulation time for the biased simulation of the first, folded structure (F1). Where the RMSD is in nm and time is represented by the number of frames (standard in Metagui), which totaled 30,000. The total simulation time was 100 ns. The range of values for RMSD fluctuated from <0.3 nm to >1.3 nm.

The RMSD CV alone proved that there were substantial differences in sampling ranges. In Figure 3, the results of the post-processing collective variable RMSD illustrated fluctuation that ranged from 0.29-0.37 nm. Despite the initial appearance of the graph, these values were only fluctuating by a small in range in comparison to F1. Meanwhile, Figure 4 demonstrated that F1 had sampled much larger ranges with the smallest value at less than 0.3nm and the largest value at greater than 1.3nm. Therefore, resulting in a sampling range of 1nm. This result was expected for F1, since the bias allowed it to sample states more easily. The fluctuation of the RMSD CV suggests that the biased system was capable of experiencing both high and low RMSD, with respect to the reference structure. The reference structure used was also the same for both simulations. Since the application of the bias was the only difference between the two simulation setups, it can be concluded that the bias caused the drastic difference in sampling efficiency.

The reference structure was determined by selecting atoms that were present in the helical portions of the protein when it was folded. Based on findings from literature[1,2,5] and the X-ray crystallography structure[26], the protein has been known to become highly alpha helical upon folding and binding to DNA. Since both simulations began with the folded structure as the starting point, a low RMSD value indicates that the protein hasn't moved much from the starting point. In contrast, a high RMSD value indicates that the protein has undergone conformational changes that differ from both the starting point of the simulation and the reference structure. Since the atoms used to construct the reference structure play a role in defining the protein's helical structure, this result would indicate a lack of helical structure.

A high RMSD indicates that the protein became unfolded during the simulation. Ultimately, the simulation should be capable of fluctuating between high and low RMSD, to ensure efficient sampling of both structural forms. However, large fluctuations of the RMSD CV were only seen in F1, despite its shorter sampling time. The lack of significant fluctuations, in UB, indicates that the simulation was not able to sample folded and unfolded structures within the simulation time. As expected for UB, the protein experienced large energy barriers that did not allow it to access higher energy states or escape from local energy minima. Therefore, one cannot easily sample all plausible states or determine the free energy landscape, without the use of a bias within this time frame.

The RMSD findings above were corroborated by measurements within the native contacts CV, which was represented as Q in Figures 5 and 6. Where given the nature of Q, this CV should also illustrate sufficient fluctuation if the system is sampling both folded and unfolded structures. This collective variable was developed using the fraction of native contacts between the two helical

strands in the folded, starting structure. This domain includes residues 22 to 64 since they define the critical portion of the folded structure (Figure 5).

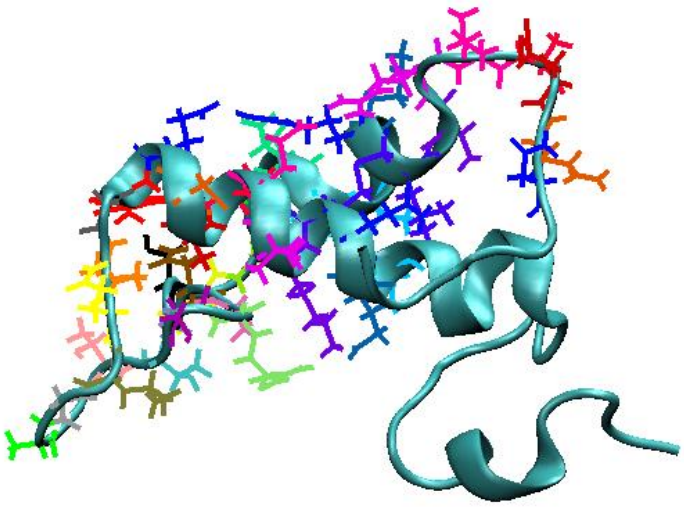

Figure 5. The (dry) folded protein structure illustrated by VMD, where residues 22-64 are represented as lines in contrast to ribbons. The coloring method for these residues pertains to their residue ID. These residues represent the regions that were critical for defining the folded structure of the IDP.

Within these residues, all non-hydrogen atoms that were within 4.2 Å of each other but more than 3 residues apart were incorporated, giving a total of 237 contact pairs that were monitored. Since these strands will only be in close proximity when they are folded and the protein is known to be alpha helical upon folded, this CV is capable of confirming both properties. Thus, monitoring the native contacts is highly advantageous for the detection of protein folding and unfolding. Obtaining a high value for the native contacts would indicate that the protein has stayed within its original structure, since those initial contacts remained close. A low value obtained for the native contacts would indicate the protein has moved away from its initial structure and the selected atoms were no longer close enough to remain in contact. A fluctuation in the fraction of native contacts is also indicative of the protein's capability to effectively transition between folded

and unfolded structures. In agreement with the results from the RMSD CV, fluctuations were also shown within F1 only (Figure 6). In contrast, UB showed minimal fluctuation (Figure 6).

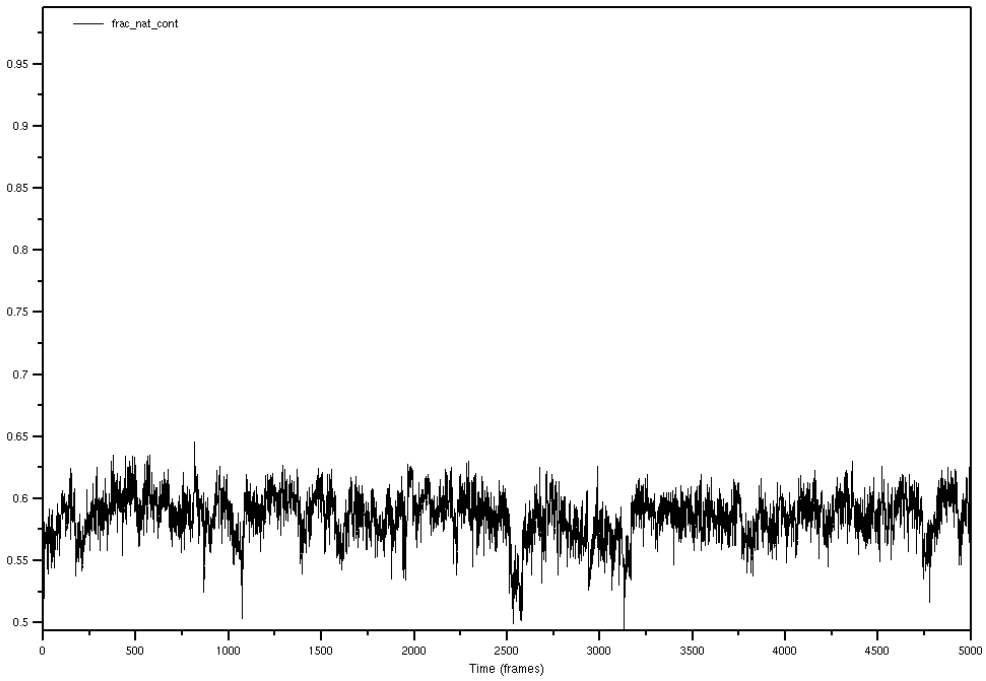

Figure 6. Plot of Q against simulation time for the unbiased simulation of the folded structure (UB). Where Q is the fraction of native contacts present and time is represented by the number of frames, which totaled 5,000. The total simulation time was 400 ns. The range of values for Q fluctuated from 0.50-0.65 native contacts.

The fluctuation of distance between the native contacts indicated that the protein remained relatively folded throughout the duration of the simulation. As illustrated, the range was very minimal between 0.50-0.65 and did not approach 1. The difference between UB and F1 was much larger for Q as seen in the following figure for F1.

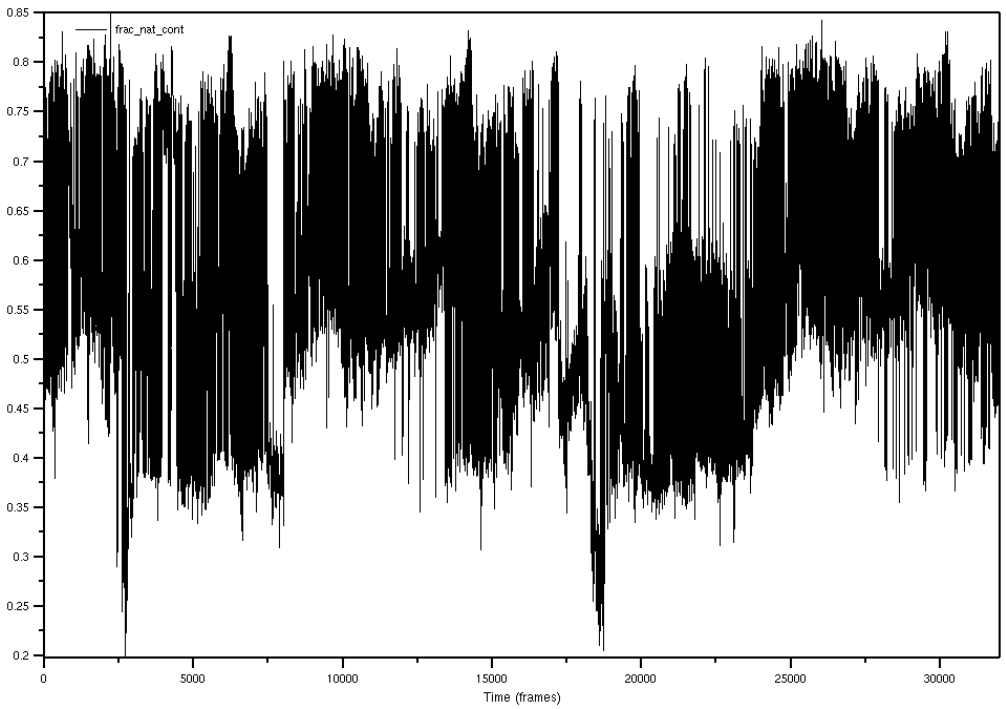

Figure 7. Plot of Q against simulation time for the biased simulation of the first, folded structure (F1). Where Q is the fraction of native contacts present and time is represented by the number of frames, which totaled 30,000. The total simulation time was 100 ns. The range of values for Q fluctuated from 0.2-0.85 native contacts.

In contrast to UB, the sampling efficiency was much larger. The number of native contacts ranged from 0.2-0.85, meaning more states were sampled when a bias was incorporated. This simulation also obtained values closer to 1. Additionally, the use of the native contacts provides a better interpretation of what is occurring, relative to RMSD analysis. The RMSD CV likely has more ambiguity in the values due to a couple of scenarios potentially occurring. Since the RMSD is only measuring the root mean square deviation of the structural differences from the reference structure, a high number may correlate to the protein's current structure not aligning well to the reference structure. Therefore, despite the protein becoming folded once again or even remaining folded, if these positions do not align, a high RMSD may still be recorded. Another assumption that cannot be made with RMSD, is that a low RMSD always corresponds to the protein becoming

folded. In some scenarios, the protein may become unfolded (in the sense of the D1 CV), but still remain roughly helical. In contrast, the high and low numbers obtained from the use of Q are less ambiguous. Only small distances between these native contacts would yield a high number for the CV, where a value of 1 would be the highest. In contrast, obtaining a low number would only be caused by a loss of the native contacts in the sampled structure. Since contacts are defined based on the folded state, the degree of contacts provides direct correlation to degree of folding. For these reasons, more than one collective variable should always be used to analyze the behavior of complex systems.

Overall, UB illustrated a poor sampling efficiency, despite it having a longer simulation time and having started with the same structure as F1. This result was seen in the values obtained from the two, post-processing CVs and was illustrated by their corresponding graphs. These results emphasize the need for metadynamics as well as the need for multiple collective variables for analysis. Therefore, the application of WT-BEMETA is needed to obtain reliable results with a nanosecond time frame.

3.3 Well-Tempered Bias Exchange Metadynamics (WT-BEMETA)

A. Comparison of the Free Energy Profiles for Determining Convergence of the Folded Systems

The First Folded System (F1)

For the two folded protein systems that underwent WT-BEMETA, the same starting structures and system configurations were used. For both systems, they each contained a total of 4

replicas corresponding to 3 replicas that were biased by their respective CV. Where, the 0_{th} replica that participated in the exchange process of the simulation, but was not biased. The three replicas for each system (1-3) corresponded to the HB1, HB2, and D1 CVs. By running identical systems separately, the reliability of results is increased and potential artifacts of simulation error are avoided. The direct comparison of these results on multiple scales also provides better resolution of the overall sampling and energy landscape during the folding and unfolding process. Ideally, sufficient sampling should be obtained of the starting, folded structure going through stages of unfolding and folding. However, since this process is relatively large and complex, it can be computationally exhaustive to track the entire system. Therefore, the use of CVs to monitor smaller processes of interest is highly beneficial for monitoring this process. As described in the setup, several collective variables were selected based on their efficiency for monitoring protein folding and were used to extract information about the system. The selected CVs were also chosen based on past successes from the literature.[12,13,31] As seen in the unbiased simulations, one CV was not sufficient for describing this process. On the other hand, too many CVs can be both computationally exhaustive and difficult to analyze. Therefore, the number of hydrogen bonds and distance between contacts within the helical portions of the folded protein were chosen for tracking. Since the protein is helical upon folding, tracking the number of hydrogen bonds present between helices can provide very specific information about the protein's structure. This CV provides a similar purpose as the RMSD, but with less ambiguity. For this reason, hydrogen bonds in region 1 (HB1), see Methods, was one of the CVs used for biasing and the RMSD CV was only used in post-processing analysis.

First, both of the simulations were assessed for characteristics of convergence and then later compared with each other. System convergence was investigated by means of the free energy

profiles generated for each CV. Over the course of the simulation, CVs are capable of filling up these free energy surfaces quite rapidly as they explore larger regions throughout time. Since metadynamics is based on principles of biasing, the accuracy of the free energies is directly related to how reliable the biased potentials are. If a simulation converges, then the bias potential serves as an estimator of the free energy, namely that it fluctuates around an average value, with ripples whose size is determined by the metadynamic parameters.[34] Therefore, throughout the duration of the simulation there should be an average, constant value with only minimal fluctuation from beginning to end. In order to assess the simulation from beginning to end, Metagui divides the simulation into the first and second half. These halves are labeled as the "first part" and "second part" in Figures 8 and 9 and are denoted by red and blue lines, respectively. The values from start to end should be consistent with each other in order to ensure reliability, therefore the red and blue lines produced should remain in alignment with each other throughout the simulation. As seen in Figure 8, there are characteristics of convergence within the free energy profiles generated. F1 and F2 are depicted in the left- and right-hands sides of Figure 8. The energy profiles correspond to the biased CVs of HB1, HB2, and D1 from top to bottom. Each of the free energy profiles for F1 and F2 were placed side-by-side in order to provide a direct comparison.

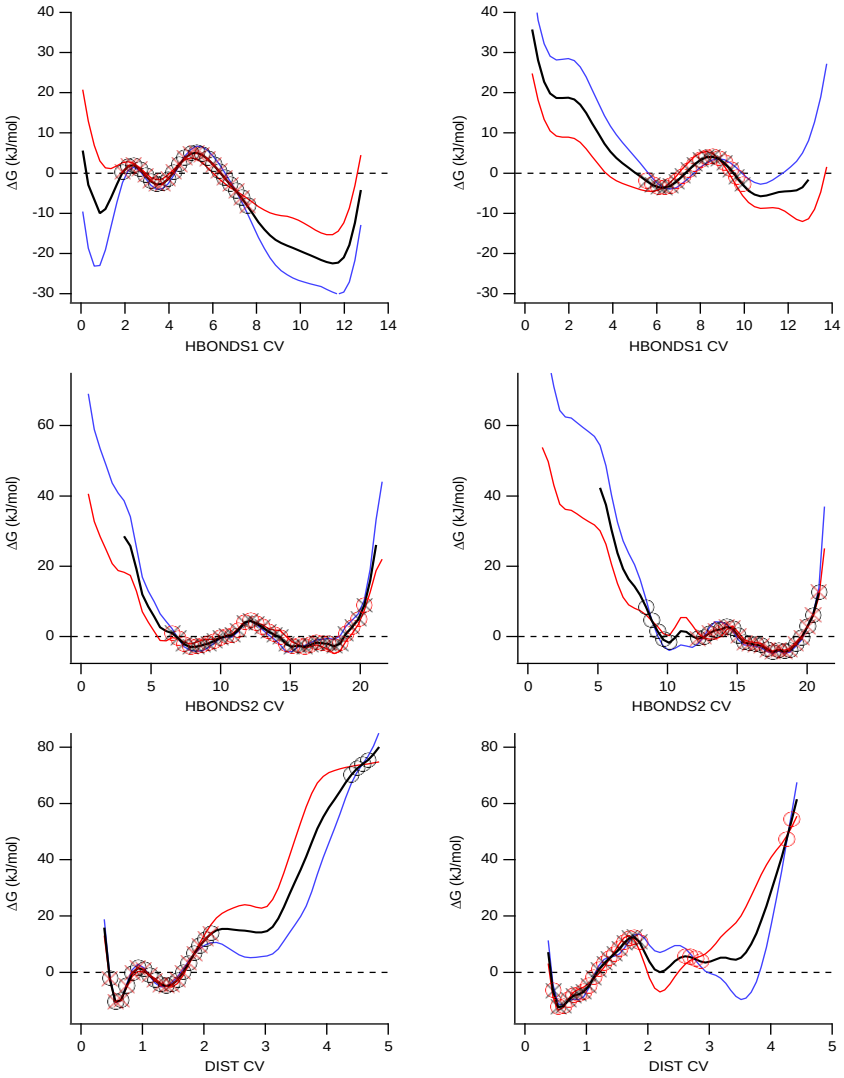

Figure 8. The 1D free energy profiles for systems F1 (left-hand side) and F2 (right-hand side). In each, ΔG in kJ/mol was plotted against the value of the CV that was used for biasing. The first row pertains to the replicas biased by HB1 and were referred to as "HBONDS1 CV" in the diagram. The second row pertains to the replicas biased by HB2 and were referred to as "HBONDS2 CV". The last row pertains to the replicas biased by D1 and were referred to as "DIST CV" in the diagram. The red and blue lines correspond to the first halves of the simulations and the second halves of the simulations, respectively. The black lines represent the average of the two simulation halves. The red crosses and black circles correspond to connectivity regions and within-range values, respectively.

In the beginning of the simulation, it was apparent that the system was equilibrating based on the initial low free energies observed until about 20,000 frames. For this reason, the first 20,000

frames were removed. This was sufficient enough to remove non-significant frames, but without running the risk of cutting out frames that held important information about the simulation. Therefore, it is still partially apparent where the simulation rises from those initial low energy frames. However, the first and second simulation halves for all of F1's replicas remained in partial agreement with each other as seen in the above figure.

In the energy plot corresponding to HB1, the average between the first and second halves of the simulation (black line) is less aligned as compared to the other two individual CV data sets F1. The presence of the average line in HB2 is also lacking when the number of hydrogen bonds is lower. This is likely attributed to a lack of sufficient sampling at low numbers of hydrogen bonds, as well as the differences between the 1st and 2nd half having been too large to compute a reliable average value. Despite the appearance of some alignment within the CV energy profiles, these results indicate that the simulation would benefit from more sampling and longer simulation time. There is also an apparent lack of connectivity and values within range, as denoted by the red crosses and black circles, respectively. The lack of within range values corresponds to the measured differences between simulation halves 1 and 2, not falling close enough within range of the delta parameter (set to $2kT$). This parameter allows for specifying the maximum allowed difference that two different free energy estimates can take in order to be considered reliable.[34] Metagui uses an algorithm that is capable of aligning the first and second halves of the simulation, by which the maximum size of the region they are allowed to differ by is set by delta.[34] By assigning delta to be equal to 2, this means that any deviations between the two halves that are greater than 2 will not be plotted as "within range", as illustrated by the lack of black circles. Therefore, by only allowing small deviations, the appearance of black circles is more authentic and reliable for determining convergence. For HB1, only minimal within range circles were

obtained, where the sampling was between 1 and 8 hydrogen bonds. This meant that only a small amount of sampling ranges had acquired deviations that were smaller than delta. This is another indication that the simulation would benefit from more sampling and longer periods of time. A longer simulation time would decrease the differences by filling in the CV free energy space. In HB2, more within range values were obtained at 5-20 hydrogen bonds. This indicated that the results of HB2 were more in agreement as different values of the CV were sampled. In contrast, D1 also indicated a lack of within range sampling for most of its CV values. Only values between 0.5 and 2 nm fell within range. A lack of black circles and red crosses also occurs shortly after, as well as the divergence of lines related to the first and second halves of the simulation. In order to assess what is occurring at these distances, more simulation time would be needed. Notably, at larger values for D1, the simulation is likely to sample states with larger variations in energies. Since the distance is based on the initial contacts between the helical portions of the folded protein, a large distance indicates that the protein has moved away from these contacts. In its unfolded form, the structural states also vary widely and therefore, divergence is likely seen at high values of D1.

Lastly, the reliability of the simulation convergence can be assessed based on the amount of connectivity, which is illustrated by the red crosses. As a further constraint, the region on which the profiles are aligned must be continuous, as free energy estimators are reliable only within connected regions.[34] For this reason, only the frames that fall within these connected regions were used for further analysis in Metagui. Connectivity regions were obtained in all three energy profiles for F1, however, they were not present throughout all values that were sampled. In the free energy profile for HB1, the sampling results are only considered reliable within 2-8 hydrogen bonds. Meanwhile, for HB2, reliable sampling results are obtained within the range of 12-21 hydrogen

bonds. In D1, sample results are only considered reliable within 0.25-2.25 nm. There also is the appearance of "within range" values that begin at the highest distance recorded in D1. In order to assess what happens at this point, more sampling within this range is needed.

Overall, the sampling ranges with the most reliability are in agreement with what was expected for this set of CVs. However, there are still some ranges of CV values for F1 that need to be sampled more, since they still fall within the grid. The grid for each CV pertains to all values that are possible for sampling within the system, as determined prior in VMD. Although there are characteristics of convergence, these results are still indicative of inadequate statistics. In order to make extrapolations outside of the sampling regions that are considered reliable, more simulation time should be obtained. However, the free energy results that fall within connectivity regions for each CV is reliable. Therefore, only these regions will be used for the determination of microstates during the folding and unfolding process.

The Second Folded System (F2)

Along with F1, the second folded system (F2) was incorporated in order to determine if these results were reproducible. In Figure 8, the free energy profiles obtained from Metagui were also illustrated for the second system. By investigation of F2's free energy profile for HB1, there were similar amounts of connectivity and within ranges values as seen in F1. However, in F2, the divergence between the first and second half of the simulation is greater in energy prior to where the sampling connectivity appears. The appearance of connectivity crosses began at 5 hydrogen bonds, versus 2 hydrogen bonds in F1's energy profile for HB1. The range where the sampling connectivity ends is also at 10 hydrogen bonds, rather 8 hydrogen bonds as seen in F1. These small

differences are likely due to the nature of molecular dynamics and its method of random sampling. Along with their notable similarities, the energy profile for HB1 indicates that more sampling should be acquired to obtain a higher degree of convergence. The free energy profile for HB2, also mimics that of the first folded simulation. Both of the free energy profiles displayed similar sampling curves. In contrast to F1, the range of values that were considered reliable were about 6-21 hydrogen bonds, for HB2. For D1, the energy profile varied widely at larger distances as expected. The range of distance values that were connected also mirrored F1's values. However, the appearance of "within range" values at higher distances is also an indication that more simulation time is needed to determine what is occurring beyond that range. Another aspect to consider is the difference in the scaling of the free energies between the two folded systems. Overall, there were only minimal differences between the two folded simulations. These differences were likely attributed to the inherent, random sampling in MD simulations. These differences were mainly found in the free energy profiles for HB1 and D1, where the free energy differed much greater between the ranges of CV values. However, the connected regions were much closer in comparison between the two energy profiles. Since only the connected regions were used for further analysis, both systems were considered quite similar based on the results for these regions. These results also confirmed the reliability of our methods, since both systems were produced using the same starting structure and simulation configurations.

B. Comparison of the Free Energy Profiles for Determining Convergence of the Unfolded Systems (UF1, UF2)

Following the comparison of the two, folded-protein systems, the free energy profiles of the two, unfolded-protein systems were also analyzed for convergence. As described in the setup, two different unfolded structures of the protein were obtained, but the same system configurations for biasing were applied. The first system starting from unfolded protein was labeled as UF1 and the second was labeled as UF2 for convenience. This was done to determine if the reverse mechanism, going from unfolded to folded, would be capable of obtaining within the same time frame. Ideally, the reverse mechanism of protein folding and unfolding should yield the same results. The protein should also be capable of transitioning between structured and unstructured forms without becoming trapped in local energy minima. If both the folded and unfolded protein systems are capable of sampling the same microstates and pathways, then the results would be highly indicative of convergence. However, if the protein is trapped by high energy barriers or is not given sufficient time to sample all states, this can lead to inadequate sampling. Therefore, the results would not be indicative of what is occurring nature. For both UF1 and UF2, the free energy profiles illustrated characteristics of convergence (Figure 9), with the exception of one outlier that will be discussed later. The free energy profiles produced by Metagui for UF1 can be seen in the following figure.

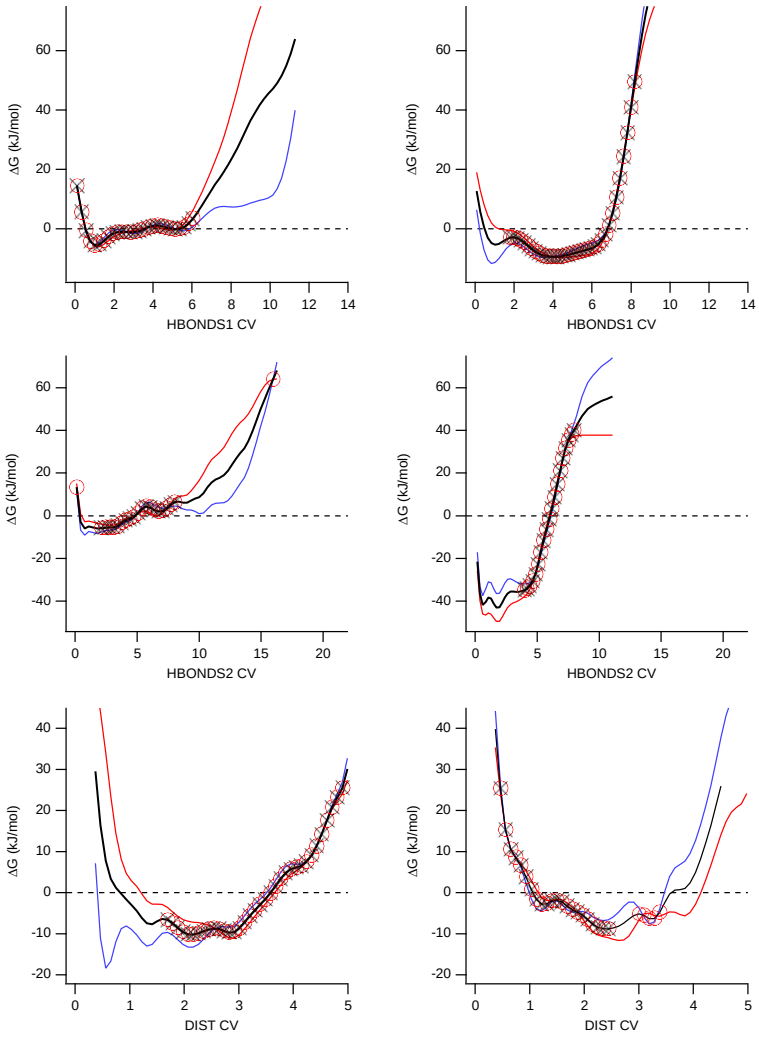

Figure 9. The 1D free energy profiles for systems UF1 (left-hand side) and UF2 (right-hand side). Each panel shows Δ*G* in kJ/mol plotted against the value of the CV that was used for biasing. The first row pertains to the replicas biased by HB1 and were referred to as "HBONDS1 CV" in the diagram. The second row pertains to the replicas biased by HB2 and were referred to as "HBONDS2 CV". The last row pertains to the replicas biased by D1 and were referred to as "DIST CV" in the diagram. The black lines represent the average of the two simulation halves. The red crosses and black circles correspond to connectivity regions and within-range values, respectively.

In the free energy profile for HB1 of UF1, there is more evidence of agreement with the first and second halves of the simulation when the number of hydrogen bonds is lower. At around 7 hydrogen bonds and higher, the simulation showed greater differences, which can be attributed

to the lack of sampling beyond that point. This was also seen in the folded simulations, however, UF1 has a larger, linear increases in energy. The region where the two simulation halves show connectivity lie within 1 and 6 hydrogen bonds. Only within this range are the results considered reliable for HB1, which corresponds to sampling of the unfolded structure.

In comparison, the free energy profiles of the second, unfolded system (UF2) are illustrated in Figure 9. For UF2, the connectivity range for HB1 is between 2 and 8 hydrogen bonds. Along with this comparison, more regions of connectivity were apparent throughout the energy plot. Overall, UF2's free energy profile for HB1 showed more characteristics of system convergence based on its ability to capture sampling, even at higher numbers of hydrogen bonds.

In comparison to HB1, there were more noticeable differences between the free energy profiles for HB2. For UF1, the range of connectivity is very minimal, spanning between 2-8 hydrogen bonds. The two halves of the system also seem to deviate past this point, which is where more folded structures would be present within sampling. There is also the occurrence of a within range circle near the end. This may indicate agreeance in sampling starting to appear around this point, however, more sampling would be needed before making any extrapolations. In reference to the outlier mentioned earlier, the HB2 profile for UF2 scales on a drastically different range. The range of connectivity for HB2 is between 4-8 hydrogen bonds in contrast, as well. The odd "s" shaped graph produced can be mainly attributed to the large, negative free energy values the system was sampling at the beginning of the simulation. Along with this, the sampling did not include higher numbers of hydrogen bonding, which is uncharacteristic of HB2. With these factors, there appears to have been an inherent conflict that may have occurred. It is possible that the system experiences topological frustration[28], in which the path toward unsampled CV space is blocked by the protein itself. However, two systems were prepared just in case an outcome, such as this one,

occurred. The regions where connectivity was present for UF2's HB2 energy profile were still able to be extracted for further analysis in Metagui.

In a larger comparison between the two unfolded systems and folded systems, HB2 varied the most in all of the CV's energy profiles. This effect was likely attributed to the region of the protein from where the hydrogen bonds were selected to construct HB2. That particular selection of the protein likely had more variability and activity during the simulation. The HB2 CV illustrates how different CVs are capable of extracting different information about the system. This was evident when the same CV was used but with different selections. This was also evident when the same CV was used, but with different selections. The combination of multiple CVs can provide an even wider spectrum of information about complex systems and processes. In particular, the region monitored by HB2 might explain more about the variations in structural states that occur in the unfolded conformation of the protein. This portion of the protein may vary in structure based on the presence of the other CV values, along with the possibility that it may not be affected by them at all. However, the effects of the CV values on each other and the overall structures are explored further in their CV projections.

Lastly, the free energy profiles for D1 were also assessed for convergence in Figure 9. Looking at UF1 first, the energy plot displayed a period of system equilibration prior to sampling a distance at about 1.5 nm. From 1.5 nm to roughly 4.75 nm, the free energy profile displayed connectivity between the first and second halves of the simulation. In comparison, the D1 energy plot from UF2 experienced connectivity from roughly 0.5 nm to 2.5 nm. After this point, the simulation seemed to diverge, which indicated insufficient sampling at higher D1 values. In comparison to each other, UF1 had reliable sampling at large distances and UF2 had reliable sampling at smaller distances only. In comparison to the folded systems, UF2 had similar

connectivity ranges and divergence at higher D1 values. Meanwhile, UF2 lacked sampling in lower ranges, but had more total connectivity in sampling. Overall, the free energy profiles for UF1 and UF2 illustrated characteristics of convergence within regions that corresponded to unfolded structures; low numbers of hydrogen bonds and high values of distance. Meanwhile in UF1 and UF2, CV values that corresponded to folded structures where indicative of a lack of sampling based on the gaps in sampling and divergence. The two unfolded systems likely did sample folded structures based on their free energy profiles. Regions of sampling that were considered reliable based on these profiles were extracted for further analysis in Metagui to construct microstates.

3.4 Determination of the Microstates

After assessing the reliability of the simulation results based on the convergence of their individual free energy profiles, the microstates were then identified using F1. Since each of the simulations produced similar results in terms of reliability and sampling most of the CV ranges, F1 was used for the rest of the analysis for simplification as well as due to time constraints. The F1 system was also capable of sampling unfolded, structural states as described below. The microstates represent a group of structures containing similar sets of CV values. The configurations are grouped together in microstates simply by dividing the 3-dimensional CV-space into a grid of small 3-dimensional cubes.[34] The cubes are also defined by the size of their sides running in each direction. These factors determine how far the center of the cube is as well as the distance that is allowed between neighboring microstates. Each frame of the trajectory is assigned to the cube to which it belongs and the set of frames contained in a cube defines a microstate.[34] These parameters were modified by the distinctions of the grid minimum and maximum for each

CV. Based on this information, highly populated microstates that fit the criteria were able to be extracted and visualized in Metagui. This allowed for further analysis of all relevant structures in terms of their individual binding energies, ultimately determining their roles in the pathway for the folding mechanism of the intrinsically disordered protein.

3.5 Free Energy of the Microstates

A total of 100 microstates were obtained using the K-medoids method[35] in Metagui with a maximum distance matrix dimension at 5000 along with sieving. After determining the population of microstates, the free energy was analyzed. Since the simulation incorporated a bias, the effect of an external potential had to be normalized. In metadynamics the history-dependent potential provides an estimate of the low-dimensional projections of the free energy.[34] In order to produce the free energy of the microstates, multidimensional projections were derived using the weighted histogram analysis method (WHAM). In the WHAM method, the equilibrium probability of microstate α is given by the following equation.

$$p_\alpha^i = n_\alpha^i e^{\beta(V_\alpha^i - f^i)} \qquad \textbf{Eq. 9}$$

Here *i* represents the replica index, f_i is a shift constant fixing the normalization, n_α^i is the number of times state α is observed in replica *i*, and V_α^i is the bias potential evaluated on the microstate α.[34] Using WHAM, the unbiased probabilities were then used produce normalized quantities of the simulation that can then be used for direct comparison. From these computed free energies, the

microstates were then visualized using various projections. Several combinations, which included 2 CVs at a time, were used to predict the free energy surface of the microstates from F1.

3.6 Visualization of Microstates obtained for F1

The plot for F1 that contained the microstates of HB1 and D1 along the free energy axis can be seen below. As seen in Figure 10, the value of each CV can affect the allowable values of the other CV, as well as their corresponding energies. This information was used in order to determine the sets of CV values that produced highly populated microstates as well as for visualizing what these populated centers looked like.

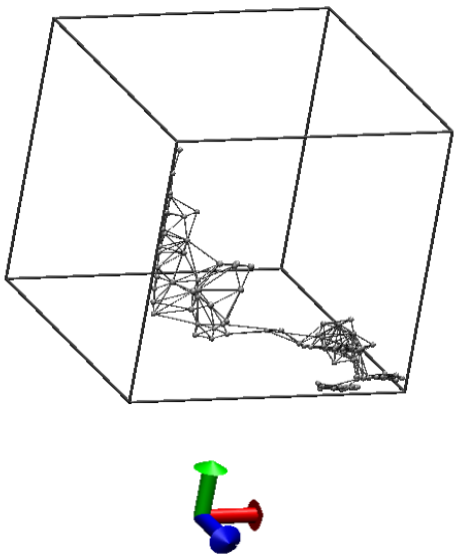

Figure 10. The projections of the microstates obtained for the HB1 CV and the D1 CV against the free energy for the first folded system (F1). In the diagram, the *x*-axis (red arrow) corresponds to the HB1 CV. The *y*-axis (green arrow) corresponds to the D1 CV. The *z*-axis (blue arrow), which was projected towards the viewer, corresponds to the free energy. The values increase in correspondence to the arrow directions. Therefore, the bottom of the *y*-axis equates to small values of D1. For example, the top of the *y*-axis (in the direction of the arrow) equates to small values of D1.

The presence of microstates was prominent at high values of HB1 and small values of D1. Meanwhile, there were no microstates at small values of HB1 and small values of D1. There was also an appearance of microstates at high values of D1 and small values of HB1. As understood, a high number of hydrogen bonds correlates to the protein in its folded, helical form. With both HB1 and HB2 large (both regions helical), then there is likely an upper limit of D1 that is smaller than if the regions are not helical. This first scenario was seen in the plot given by the cluster of microstates with a large number of hydrogen bonds and a small distance between contacts. The first scenario confirmed that the structural state was highly folded. Notably, the free energy was also relatively lower than the other cluster of microstates. This was illustrated in the free energy surface (FES) plot with HB1 and D1, seen below in Figure 11.

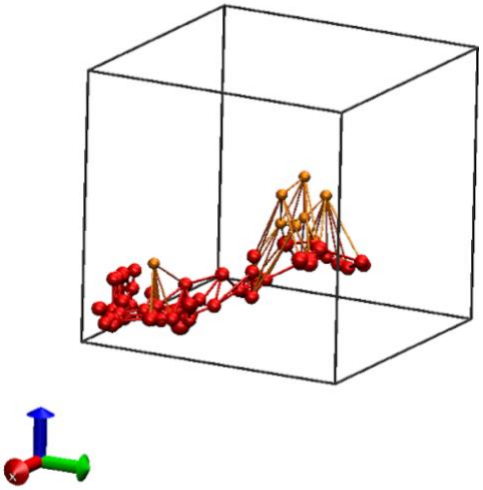

Figure 11. The projections of the free energy surface for the HB1 CV and the D1 CV of the first folded system, F1. The *x*-axis (red arrow) corresponds to the free energies of the HB1 CV, the *y*-axis (green arrow) to the free energies of the D1 CV and the z-axis (blue arrow) to the free energy, which was set at a maximum of 50 kJ mol^{-1}. As illustrated, the free energy increased as both CV values went up. The red and orange beads correspond to different populations of microstates.

The differences in energy between the microstate clusters was likely attributed to the starting structure having been folded and therefore, the protein may have likely resisted change or

experienced more stability in this structural state. A lower energy may have also been attributed to the folded structure being more stable over the transitioning, unfolded structure. There was also a noticeable jump in energy between the two clusters of microstates. This indicated a relatively large energy barrier required to transition from the folded to unfolded state. In the second scenario, a small number of hydrogen bonds present and a large distance between the contacts meant that the protein had unraveled and the two helices had moved away from each other. Therefore, this second outcome represented the protein in its unfolded form. The energy for this cluster of microstates was also comparably higher, which was attributed to the unfolded form of the protein experiencing the initial breaking of hydrogen bonds. The unfolded structure is also capable of taking on various conformations, which was why there was more variability within the cluster of microstates within this corresponding region in Figure 11. Due to the associated high energy and instability, the protein likely experienced many more short-lived microstates. This characteristic is a hallmark of the intrinsically disordered protein. Lastly, the lack of microstates at a small number of hydrogen bonds and a small distance confirmed an unfavorable combination of CVs would not occur. It would have been highly unlikely for the atoms of the D1 CV to have remained in close contact while the protein had unraveled. Overall, the results of the 3D microplot for HB1 and D1, confirmed the predicted behavior of the protein. Sampling for both the folded and unfolded states of the protein was also obtained, based on the locations of the microstates with respect to their CV values.

We next analyze the other helical region of the protein by plotting the HB2 CV along with the D1 CV against the free energy (Figure 12).

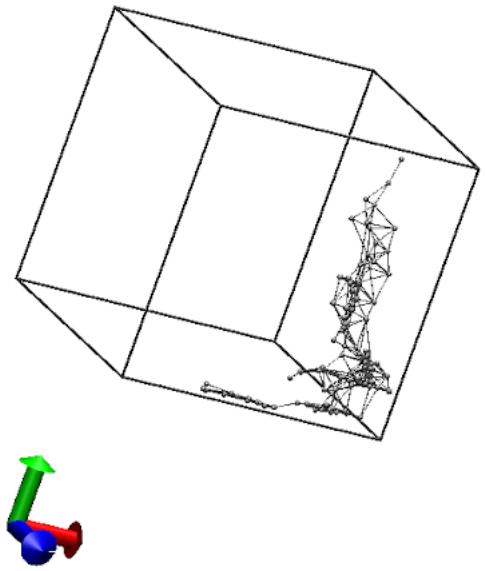

Figure 12. The projections of the microstates obtained for the HB2 CV and the D1 CV against the free energy for the first folded system (F1). In the diagram, the x-axis (red arrow) corresponds to the HB2 CV. The y-axis (green arrow) corresponds to the D1 CV. The z-axis (blue arrow), which was projected towards the viewer, corresponds to the free energy. The values increase in correspondence to the arrow directions.

These data allow for information to be obtained about another region of the protein during the folding and unfolding process, which helped provide a broader illustration of what occurred during the simulation. In comparison to the first region of hydrogen bonds that was tracked, the HB2 and D1 produced several differences, illustrated by the density of microstates present in different regions. When D1 was small and HB2 was within an intermediate range of values, the presence of microstates was small with a narrow distribution. This depicted a scenario where the protein remained roughly helical within that region and that the two helices were in close contact with each other. In addition, there was a strong presence of microstates throughout all ranges of D1 and high values of HB2. This meant that the protein was capable of maintaining a higher number of hydrogen bonds within that region, regardless of whether the two helices were in close contact or not. When D1 was small and HB2 was large, there was a more concentrated portion of these

microstates. This meant that the variation between the clusters of the microstates was very low. With these sets of values, the protein should have been in its most compact, helical form. For this reason, there shouldn't be a large amount of variation between the structures and the populations of microstates within the small section of the diagram. When HB2 is large and D1 starts increasing in value, the distribution of microstates spreads out. These results corresponded to the protein maintaining its helical structure within that region, while the two helices were still moving away from each other. The microstates were also spread out based on the various structures that the protein could undertake as the two helices were separating. This allowed for more independence and variation in the structures that the protein could undertake. The second region of hydrogen bonds that pertained to HB2 was thus capable of maintaining structural integrity despite an increasing value of D1. Therefore, HB2 was likely independent from changes to D1 or at least more resistant to changes than the region that used for HB1. This was also confirmed by the lack of a single microstate present when HB2 was low. However, more sampling would need to be done to confirm that the landscape of possible microstates was sufficiently sampled.

The FES plot for F1 (Figure 13) shows the free energy of the microstates was noticeably lower when the protein was in its folded form.

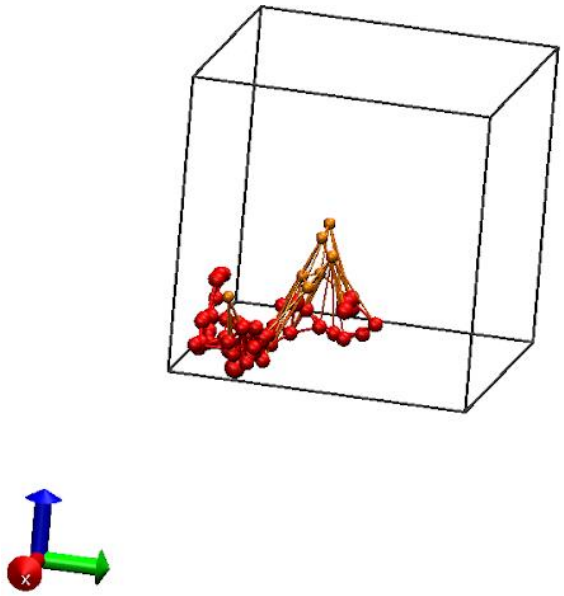

Figure 13. The free energy surface profile for the HB2 CV and the D1 CV of the first folded system, F1. The *x*-axis (red arrow) corresponds to the free energies of the HB2 CV, the *y*-axis (green arrow) to the free energies of the D1 CV, and the *z*-axis (blue arrow) to the total free energy, which was set at a maximum of 50 kJ mol^{-1}. As illustrated, the CV energies increased as they went up and decreased as they went down in diagram. The red and orange beads correspond to different clusters of microstates.

The figure also demonstrated that the free energy was higher when the protein sampled variations of unfolded structure. This result was consistent with the free energy profile of HB1 and D1, since the unfolded structures were associated with higher energies.

Lastly, the presence of all 3 CVs and their corresponding effects on microstate populations were analyzed. The results were illustrated in the figure below containing the 3D plot of HB1, HB2, and D1.

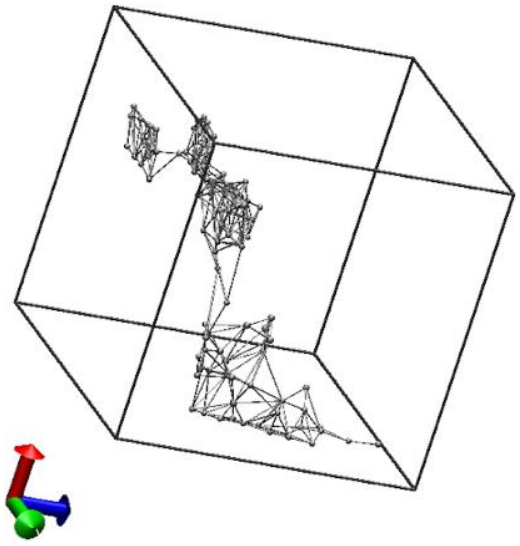

Figure 14. The 3D projections of the microstates obtained for HB1, HB2, and D1 for the first folded system (F1). In the diagram, the *x*-axis (red arrow) corresponds to the HB1 CV. The *y*-axis (green arrow) corresponds to the HB2 CV. The *z*-axis (blue arrow) corresponds to the D1 CV. The values increase in correspondence to the arrow directions.

Notably, there were no populations of microstates when all 3 CVs were at large values. This meant that two helices couldn't have acquired a large distance while both regions of HB1 and HB2 were able to remain helical and therefore, contain a large number of hydrogen bonds. Since the scenario is physically impossible, the lack of microstates at least confirmed the simulation didn't produce any artifacts in that region. A lack of microstates was also seen with all 3 CVs at small values. Microstates did not occur in this region because the protein wasn't able to be both compact and helical without the two helices having folded together as well. This was another impossible scenario, which also confirmed there were no artifacts produced in that region as well. Both of scenarios confirmed the reliability of the results.

In contrast, other sets of CV values were indicative of different structural states that the protein underwent. There were approximately 4 distinctive populations of clusters as seen in Figure 14. The population of microstates that were the most spread out at the bottom of the cube,

likely pertained to unfolded structures based on their corresponding sets of CV values. These populations were used to extract the most, predominant structural states associated with these microstates, as well as to determine their corresponding energies.

3.7 Determining the Structure and Free Energies of the Microstates

One set of values that produced a cluster of microstates pertained to a structure with CV values that corresponded to an unfavorable conformation (Figure 15). Where the total number of associated structures was 49. The structure would be hard for the protein to obtain, since the helices are close and compact, despite the larger region of HB2 having lost its helical content.

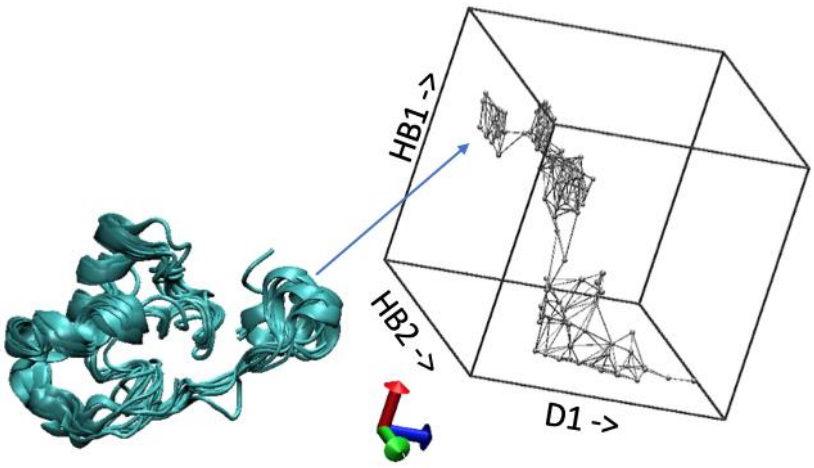

Figure 15. Microstate with a population of 49 structures for F1. The D1 value obtained for this microstate was small at 0.633 nm. The microstate also had 9 hydrogen bonds present for the region monitored by HB1 and only 6 hydrogen bonds present for HB2. This was also evident in the illustration; where only the region for HB1 remained helical, while HB2 did not as the helices moved away from each other. The energy associated with this structure could not be resolved.

This microstate included a relatively large number of hydrogen bonds for HB1 at 9 and a small number of hydrogen bonds for HB2 at 6. The D1 value for this population state was also small in comparison at 0.633 nm. This meant that the structure had lost its helical content within the region monitored by the HB2 CV, while the region monitored by the HB1 CV had still remained helical. Due to the helices being compact at this small distance, it is unfavorable for either HB1 and HB2 to not be helical as well. Additionally, the HB2 CV has also been associated with retaining its helical content better than the HB1 CV. Due to these reasons, the microstate was not sampled enough to compute a reliable energy as well.

The second cluster of microstate populations pertained to the most sampled microstate contained a set of CV values that related to the protein's folded structure. In total there were 1019 structures associated with this microstate, which produced a quite dense image of the microstate (Figure 16).

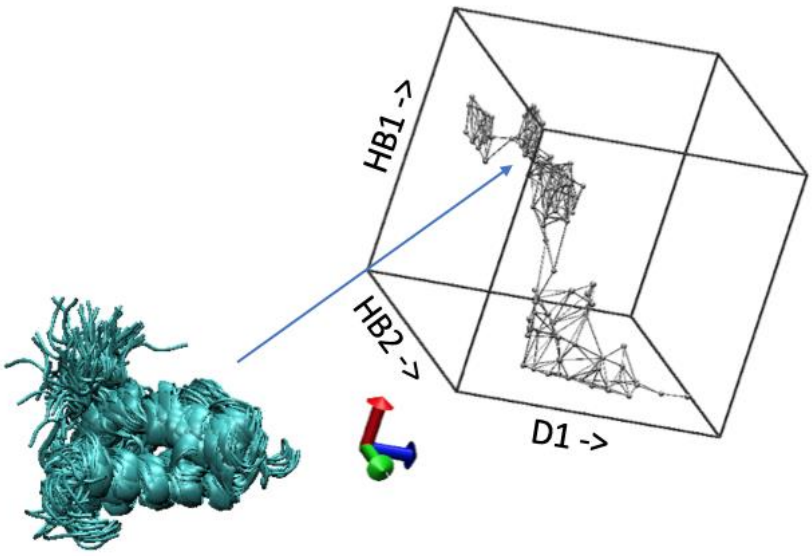

Figure 16. Microstate with the highest association of structures at 1019, which represented the folded structure of the protein from F1. The content of hydrogen bonds was high for both HB1 and HB2, at 11 and 17, respectively. The distance for this structure was also small at 0.642 nm. The energy associated with this structure was also 3.12 kJ mol-1, which was comparably lower than for unfolded structures.

The number of hydrogen bonds within the structure was also high for both regions of HB1 and HB2, with 11 and 17 hydrogen bonds, respectively. The distance between the two helices was relatively, small at 0.642 nm for D1. Therefore, these values related to a folded structure that contained high amounts of helical content. The associated structure of the microstate also further confirmed the presence of the folded protein (Figure 16). The energy associated with this microstate was determined to be 3.12 kJ mol^{-1}. This energy was comparably smaller than the rest of the sampled microstate populations.

The middle population of clusters pertained to a population of microstates with an association of 63 structures (Figure 17).

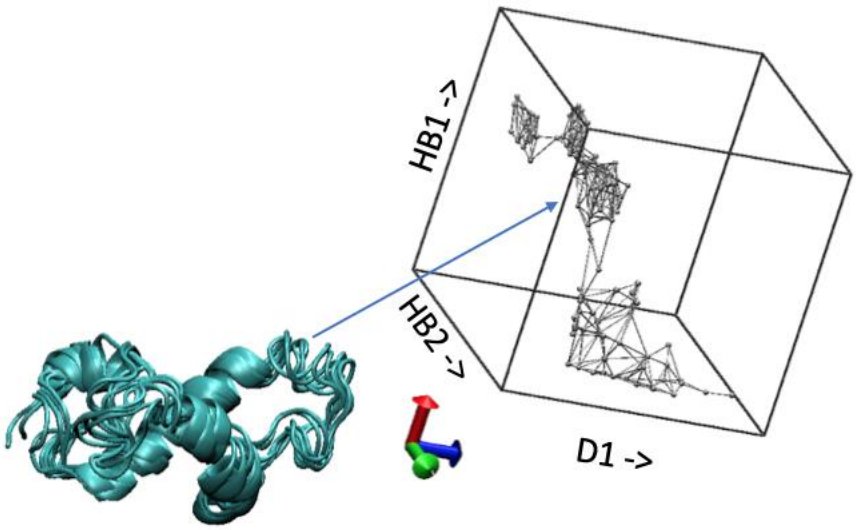

Figure 17. Microstate with a population of 63 structures for F1. In contrast to the folded structure, this microstate sampled a larger distance for D1. The D1 value obtained for this microstate was 1.295 nm. The microstate also had 11 hydrogen bonds present for the region monitored by HB1 and 19 hydrogen bonds present for HB2. This was also evident in the illustration; where the protein remained helical, while the two helices had begun to move away from each other. The energy associated with this structure was 3.7 kJ mol-1, which was slightly higher than the first folded structure.

This microstate had comparably higher values for HB1 and HB2 at 11 and 19 hydrogen bonds, respectively. Meanwhile the value for D1 was large at 1.295 nm. This combination of CVs pertained to an intermediate structural state of the protein as it started to unfold. The associated free energy is also higher than the first folded structure at 3.7 kJ mol-1. Based on the structure (Figure 17) and the CV values, one of the first steps in the unfolding process is likely associated with the two helices moving away from each other.

For the cluster of microstates that were spread out at the bottom of Figure 18, the highest associated number of structures was 161. This region also pertained to unfolded structures of the protein based on the sets of CV values.

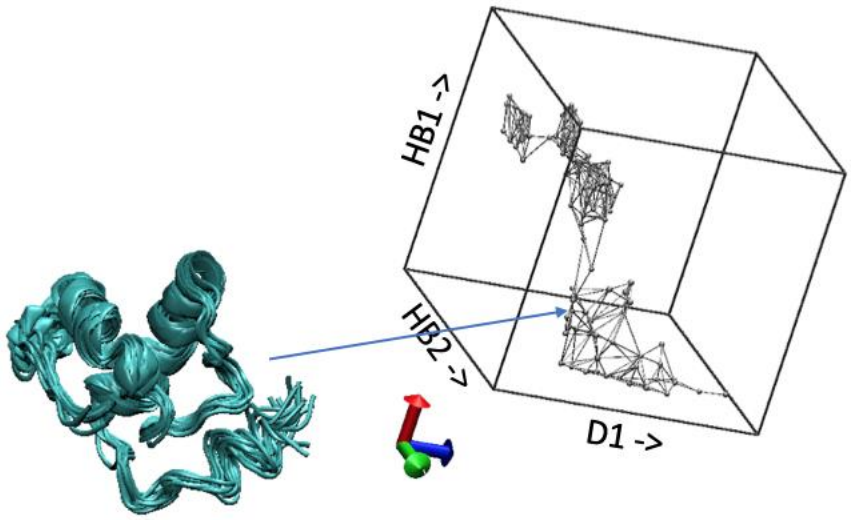

Figure 18. Microstate with an association of 161 structures, which was relatively high for the sampled unfolded structures in F1. In contrast to last structures, this microstate sampled lower values of HB1 and HB2 at 8 and 10 hydrogen bonds, respectively. The D1 value was also relatively large at 0.742 nm. This was evident by the lack of helicity in the illustration, as well as by the larger distance that accumulated between the two strands. This microstate is highly indicative of what the unfolded protein looks like in nature. The energy associated was also much higher at 12.5 kJ mol^{-1}.

This microstate contained much lower values for HB1 and HB2 in contrast, at 8 and 10 hydrogen bonds, respectively. The D1 value was also relatively larger than the overall, folded structures at

0.742 nm. This microstate pertained to a structural state of the protein, where the helices have begun to move away and the helical content of the protein has extensively decreased. The associated energy for this microstate was higher than the folded microstate at 12.5 kJ mol^{-1}, which was expected.

Focusing on the bottom region (Figure 18), other highly populated structures were acquired and used to map the proposed intermediate structural states of the unfolding IDP. Since the IDP is dynamic in nature, it samples many short-lived microstates versus remaining at one. The likely structures that the protein undertakes during this process were determined based on their free energies and CV values. Ultimately, the use of WT-BEMETA allowed for these populated structures to be acquired and used to map the proposed intermediate and unfolded structural states of the intrinsically disordered protein (Figure 19).

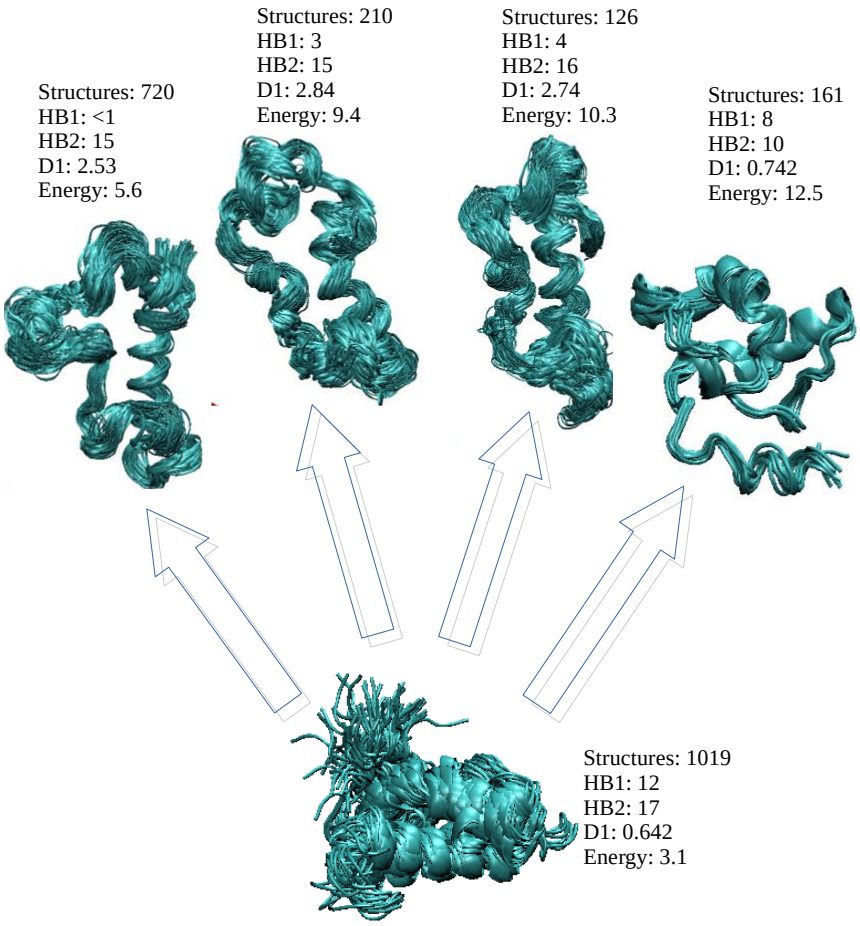

Figure 19. Proposed structures of the IDP from F1, where the bottom microstate corresponds to the folded structure and lowest energy. Navigating from left to right are the variations of the folded protein becoming unfolded, based on the free energy values. Where the first microstate corresponds to structures that have no hydrogen bonds present for HB1, as D1 increases. The second and third microstates correspond to structures where HB2 and D1 are relatively high and HB1 is moderate. The fourth microstate corresponds to structures where D1 and HB2 are high, meanwhile HB1 is very low. This microstate also had the highest sampled and reliable energy during the simulation.

Starting from its folded structure, the protein seems to lose one portion of helicity much faster as the two helices began to move away from each other. This was made evident by the region of the protein tracked with HB1, as D1 increased. Meanwhile, the portion of the protein tracked by HB2 was the last to lose its helical content as the two strands moved away from each other. Despite the presence of either low or high values of D1, the HB2 region of the protein remained

roughly helical during the simulation. This indicated that the region monitored by HB2 was either more resistant to changes in structure or was indifferent to changes in other regions of the protein such as, D1 increasing and HB1 decreasing. Based on these findings, this region should be investigated further in order to determine if it plays a larger role in the mechanistic pathways of the protein.

In this work, the free energies were determined to be much higher for the unfolded structures. This was likely attributed to the energy required to break the hydrogen bonds, initially. However, energy differences were also smaller in comparison to when the protein transitioned between short-lived unfolded states, making the protein more amenable to them. Hence, the unfolded protein is dynamic in nature, which is the hallmark of IDPs. The transitions observed were able to be identified despite occurring on only a nanosecond timescale, thanks to WT-BEMETA. Most notably, several combinations of CVs yielded higher amounts of associated structures in correlation to their microstates. Different microstates were observed based on their differences in structure density, energies, and values for the CVs. A summary of the findings in Figure 19 shows the starting structure and its corresponding values. At the top of Figure 19, were some of the most prominent microstates with their distinct sets of CV combinations. These identified structures were all attributed as likely intermediate states during the unfolding process of the IDP.

3.8 Post-processing CVs

Two CVs were incorporated during post-processing to confirm the results obtained and to extract additional information. As seen in the earlier comparison of the unbiased simulation (UB)

and the biased simulation of the folded structure (F1), the RMSD CV was used to determine sampling fluctuation within the number of simulation frames. Now focusing on the biased simulations, it is important to assess the values of the RMSD CV in relation to the free energies. As seen in the following figure for F1 and F2, the RMSD is plotted against the free energies for the simulations.

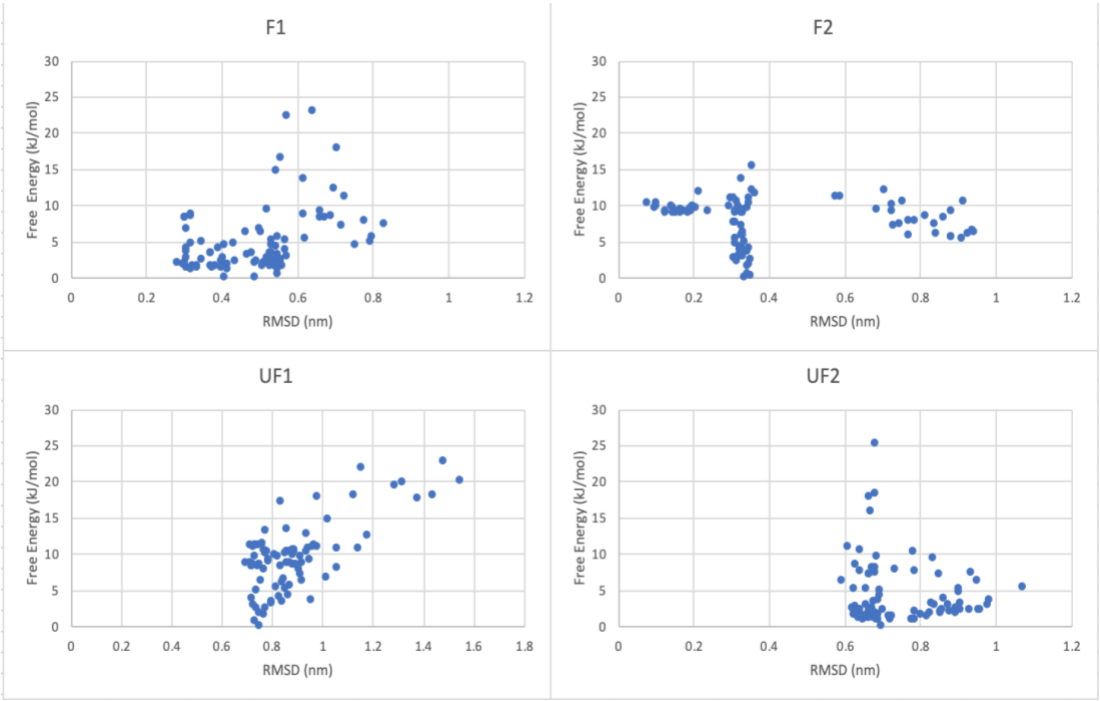

Figure 20. Free energy plots for the RMSD of the biased simulations (F1, F2, UF1, UF2). Where the RMSD was computed based on the differences from the reference structure. The reference structure was obtained from the folded structure of the protein obtained from X-ray crystallography. The unit of RMSD was measured in nm and energy was measured in units of kJ mol^{-1}.

For F1, the RMSD ranges roughly from 0.2 to 1.0 nm. Higher energies, such as 25 kJ mol^{-1} were associated with RMSD values near the higher end of the range. Notably, there were no values lower than 0.2 nm obtained. This meant that there were no structures sampled that were perfectly aligned to the folded, reference structure. This result was likely due to implications in the

alignment of structures and frame-fitting as well as changes in structure during the first 20 ns of simulation, which was not used in the analysis. Therefore, lower values obtained that were not zero, were likely folded structural states. This was supported by the lower energies obtained, in the range of 0-10 kJ mol^{-1}. Upon reaching a RMSD of 0.6 nm, there is a noticeable difference in the distribution of data. From this point on, there were a variety of energy ranges associated with higher RMSDs. Based on the energies obtained from the microstates as well as in the 2D projections, at this point the protein began unfolding. The unfolded structure was determined to be much larger in energy, demonstrated here by the abrupt increase in energy. However, as the RMSD approaches a value of 0.8 nm, there is a noticeable drop in energy. After crossing this high energy barrier associated with reaching the unfolded state, the energy of the protein levels off as it sampled different, unfolded confirmations.

In the diagram for F2, a different pattern was observed. Within 0-0.4 nm, the sampling density was much higher and the energies ranged from about 0-15 kJ mol^{-1}. After this point, there appears to be an absence of sampling until about 0.6 nm and then about 0.8-1.0 nm; there is no large jump in energy as observed in F1. The incongruent data may be attributable to a lack of sampling for this simulation. However, the results indicate that the protein remained roughly similar in structure based on the RMSD for the first half. After the jump in sampling data, it appears the structures were unfolded and therefore, less aligned with reference structure. From the first point to the second point of sampling, the protein transition from a folded state (low RMSD) to an unfolded state (high RMSD). Meanwhile, the energies were roughly the same for both low and high RMSD values, with no jumps in energy present. Ultimately, the system is indicative of requiring more simulation time, due to the apparent gap in sampling data.

In the first unfolded simulation, UF1, the RMSD does not sample any values lower than 0.6 nm. This meant that there were no structures that were similar to the folded reference structure, as expected. This could have resulted in the lack of sampling these unfolded states as well as difficulties with the alignment of the RMSD. Further simulations should be conducted in order to assess this. Lastly, the RMSD values also span from 1.0 to 1.8nm, meaning that simulation likely sampled more unfolded structures and remained unfolded, rather than accessing folding structures. Within this range of higher RMSD values, the energies also peaked to larger values. This concurred with the nature of the unfolded structure and the necessary, breaking of hydrogen bonds. Along with the lack of folded structures, the unfolded structures should have also begun to sample lower energies after the hydrogen bonds were broken. Overall, these results were indicative that more sampling should be done for UF1.

Lastly, the diagram for UF2 was analyzed. Similar to UF1, there was no sampling below 0.6 nm for the RMSD. The fact that both unfolded structures did not sample values that were associated with the folded structure, confirmed prior findings. The structure likely remained unfolded throughout the duration of the simulation, based on these results. However, within RMSD ranges of 0.6 nm and 1.2 nm, there was sufficient sampling. Notably, at 0.7 nm there was a jump in energy at 25 kJ mol^{-1}, which then leveled off to roughly 2-10 kJ mol^{-1}. One assumption that can be confirmed by more sampling is if the structure before the energy jump was folded. Since energy jumps have only been observed amongst these larger structural transitions, this was likely the case. However, more sampling should be done for UF2, as well as for UF1 before any definitive conclusions can be made.

The second post-processing CV incorporated in the biased simulations was Q. The fraction of native contacts works in a similar fashion as RMSD, but provides less ambiguity for discerning

its values. As seen in Figure 21, the plots for Q against the free energy for the 4 simulations was illustrated for comparison.

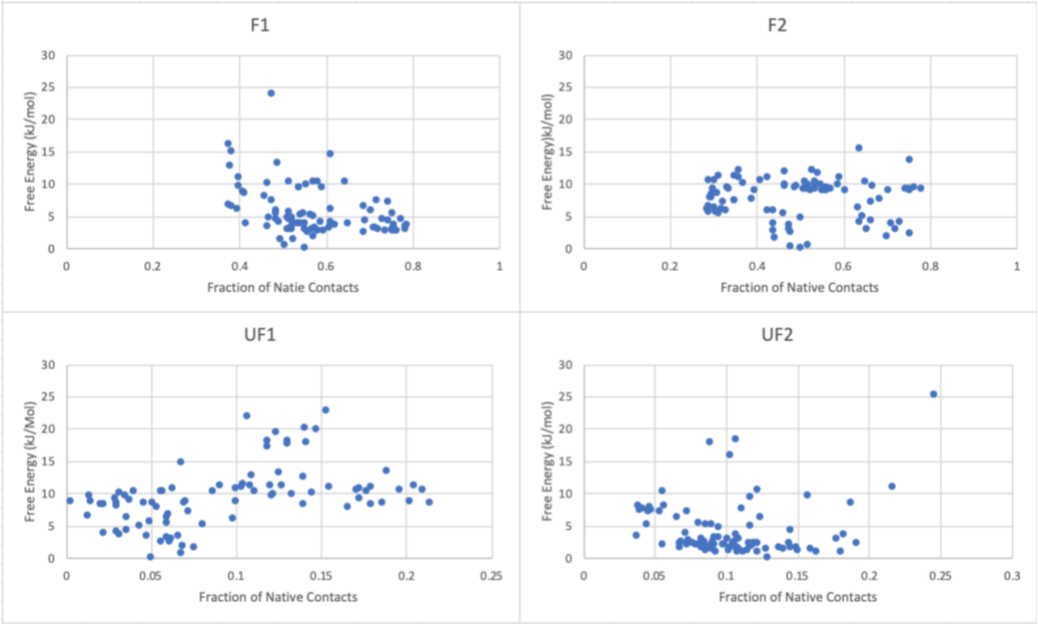

Figure 21. Free energy plots for the distance between the fraction of native contacts (Q) of the biased simulations (F1, F2, UF1, UF2). The fraction of native contacts corresponded to the select pairs of non-hydrogen atoms within the helices. These contacts were chosen based on the criteria of being within 4.2 Å apart, as well as more than 3 residues. The energy was measured in units of kJ mol^{-1}.

For F1, the values of Q ranged from 0.3 to 0.8. The smaller values of Q corresponded to smaller numbers obtained for the fraction of native contacts. Therefore, variations of unfolded structures were present at these lower values. At the value of 0.4, the free energies appear to peak, with the highest energy recorded around 25 kJ mol^{-1}. This point likely signified where the unfolded structure began to fold, since the values of Q were closer to 1. After this peak in energy, the free energies appear to drop off and the distances between the native contacts increased even more. Denser sampling was also present between the values of 0.5 and 0.8. These results were highly

indicative of protein folding and suggest that there were two structural states that were highly favorable during protein folding, since they were at local energy minima as well. This information can be used for determining which microstates correspond to these values and ultimately, which microstates would be worthwhile to investigate further.

In F2, the results were also very similar. The sampling ranges also correspond to values of 0.3-0.8 of Q, with matching energies that ranged from 0-15 kJ mol^{-1}. In contrast, there was no distinct energy jump identified in F2. This may have been due to poor sampling efficiency as well as the need for a longer simulation. However, the characteristic clustering within the range of the unfolded structure at values of 0.4 and the two likely folded structures at values of 0.5 and 0.75, were present. Similarly, the last clusters of data points also pertained to local energy minima. However, more sampling should be done in order to assess if there was an energy jump present between likely unfolded and folded structures.

Looking at the first unfolded system, UF1, the plot was significantly different from the folded simulations. There was notably denser sampling within the ranges of 0-0.225 of Q, as expected for an unfolded protein. The folded systems, in contrast, did not sample lower than values of 0.2 for Q. Meanwhile, a value of 0.2 was on the higher end for Q in UF1. The smaller values obtained for Q (<0.25) were consistent with the presence of unfolded structural states, since they had less than 25% of the native contacts present. These values of Q also presided at lower energies. Notably, the energy started to increase to roughly 20 kJ mol^{-1} around 0.1 to 0.15. These higher free energies were likely associated with structures that were starting to fold or that were less favorable. After 0.15, the energies started to level off and decreased down to 10 kJ mol^{-1}. The lack of sampling past 0.25 for Q indicated that the simulation was unable to efficiently sample folded structures.

Based on both RMSD and Q for UF1 the simulation should be extended further to improve sampling.

Lastly, UF2 had a sampling range of 0.02-0.25 for Q. In comparison to UF1, there were more dense populations of sampling and the free energy was lower, on average by 5 kJ/mol. The free energy ranged roughly from 0-10 kJ/mol. Similar to UF1, there was also a jump in energy at a Q of 0.1. The energy associated with this jump was also roughly the same, around 20 kJ/mol. The last notable difference was associated with the sampling of Q at 0.25, along with a free energy around 26 kJ/mol. There were no other populations of sampling nearby, however, there was another isolated sampling point nearby with a smaller energy around 11 kJ/mol. This structure may have served as an intermediate step in between accessing the last, highest energy state that was sampled. In order to fill in the gaps between sampling and to acquire sampling of the folded structure, the UF2 simulation should be extended further.

Overall, the results of both post-processing CVs provided more information about the phases spaces that were being sampled for both the folded and unfolded systems. For the folded systems, the RMSD CV indicated that both the folded and unfolded structures were being sampled. In comparison, the Q CV indicated a greater need for sampling of the unfolded structures. For the unfolded systems, both CVs indicated that the unfolded structures were predominantly sampled. Meanwhile, sampling of the folded structures for these systems was poor. Based on the results, there was no overlap between the regions of phase space that were sampled in the unfolded simulations and the folded simulations. Therefore, there was not enough sampling on both ends of the simulations to meet in the same sampling space. Lastly, the unfolded structure should inherently have less energy, since this is the preferred state of the protein when DNA is not present. The initial peak in energy between the transitions was also expected due to the energy requirement

needed for breaking the hydrogen bonds. After that step, the protein should begin to sample favorable unfolded structural states. The results of these higher energy states correlated to unfolded states therefore meant that the simulation did not simulate much past the point of it initially becoming unfolded. However, more stable structures of the unfolded protein should be determined in order to begin to characterize the structure of the IDP in nature.

CHAPTER IV- CONCLUSION

Despite IDPs abundance in nature, there is still a general lack of knowledge about the dynamics and structural propensities of these proteins. This is mainly attributed to the inherent difficulties of studying these proteins when using traditional methods. For these reasons, researchers have turned to advanced computational approaches in order to acquire more information about the conformational states of IDPs. The initial findings of our unbiased simulation demonstrated the need for a new approach to accelerate the conformational sampling of IDPs. The results of our 4 WT-BEMETA simulations showed a substantial increase in the IDP conformation sampling efficiency and required less time. The convergence of each system, in terms of reliability, was assessed based on their one-dimensional free energy profiles. Despite the need for further sampling, information about the systems was still able to be extracted for analysis within the sampling regions that were deemed reliable. The results obtained for the identical, initially folded systems (F1 and F2) illustrated only small differences between each other, which was indicative that the results were reproducible. Since the results of the two systems were similar, F1 was used for the majority of the analysis for simplification. As for the initially, unique unfolded systems (UF1 and UF2), these results were indicative that more sampling was required. Despite the initial results, the simulations should be extended further, in order to confirm that both unfolded and folded structures were effectively captured during the simulation. Additionally, a simulation using the same, initial unfolded structures should be conducted as well. This would provide information about the reproducibility of our findings for the initial, unfolded structure systems (UF1 and UF2).

Within F1, several 3D projections were used to demonstrate the populations of microstates and their corresponding free energies, based on their sets of CV values. The most prominent sets of CV values were analyzed further for their structural properties. Ultimately, the difference in their CV values helped predict their structural states as well as the stability of the structural state based on its free energy. This research found that the helical region of the folded structure, monitored by the hydrogen bonding in region 2 (HB2), was more resistant to changes in terms of unfolding. Regardless of the other CV values, this region remained highly helical throughout the sampling of structures. Only at extreme distances between the two helices (high D1 values) did this region began to unfold. The HB1 CV was more affected by higher D1 values and was also the first region to lose its helical properties. In some structures, the region of HB1 remained roughly helical as distance increased, however, the associated energy was much higher. Following the general trend, the folded structure was associated with lower energy as compared to the unfolded structural states. There was a clear energy spike for the folded structure transitioning to the unfolded state, attributed to the breakage of hydrogen bonds. This was based on the results obtained for the 3D projections. After this peak in energy, variations of the unfolded structural states were more spread out in terms of their sets of CV values. This meant that there were more structural variations of the unfolded state, which was expected based on the nature of the IDP. Correspondingly, the energy levels off after reaching of the many possible unfolded states. Most notably, sampling of lower energy states was not acquired for the unfolded protein. This meant that the sampling likely ended past the initial transition, but more stable structures of the unfolded protein were not yet captured by the simulation.

Ultimately, the results of this research were used to identify likely transitional state structures that the IDP underwent as it unfolds. This information was based on the population of

prominent microstates and their associated energies. The results of this simulation can be extended further by the analysis of the IDP in the presence of DNA, as well as another monomer. The communication between these protein-protein interactions is only minimal, with the most recent research having been conducted in 2001 by Hayes et al. Their findings proved that the n-terminal amino acid residues were critical for communication between monomers. They also determined that the binding affinity relied heavily on the charges of only the first few N-terminal residues, based on a series of experiments that included N-terminal residue deletions.[33] However, their results were only based on Circular Dichroism (CD) spectroscopy and have yet to be confirmed by multidimensional NMR spectroscopic techniques. Notably, advanced NMR techniques have been the main method for studying these types of proteins, since they were first developed. A vast majority of research has been done on IDPs using NMR, however, currently the α/β-type small acid soluble protein has remained uninvestigated. However, with the application of WT-BEMETA and appropriate CVs, more information can be provided about the nature of protein folding and unfolding for this IDP. In turn, this data can be used to help deconvolute NMR spectra as well as to help determine NMR parameters. For example, the number of native carbon contacts is especially beneficial for both confirming and deconvoluting C-NMR related spectra.

The use of the new MD simulation technique, WT-BEMETA, can be extended to researching other IDPs along with other biological systems. Numerous IDPs have been researched using various molecular dynamics simulation techniques that range from unbiased methods to different types of accelerated conformational sampling.[12,13,36] Notably, the α/β-type small acid soluble protein was first investigated using computational methods in 2011 by Ojeda-May and Pu[32]. They used a similar technique called replica exchange metadynamics (REMETA), where the main difference was that their replicas were biased using different temperatures rather than CVs.

They were ultimately able to determine that only a small free energy barrier (4.184 kJ/mol) separated the conformational ensembles at high and low temperatures.[32] This number coincided with the ranges of free energy obtained from our simulation as folded and unfolded states were sampled (1-13 kJ/mol). Their results provide essential information about this particular IDP along with the results obtained from this experiment. Through the use of different biasing approaches and CVs, we can begin to understand the nature of this IDP substantially more. Researchers that are interested can investigate the folding and binding process further by conducting these simulations on longer time scales, as well as by manipulating biasing parameters. By fine-tuning this approach, substantially more sampling may be obtained for this protein system. When working with biased approaches, it is also critical to obtain a high order of system convergence, in order to confirm the reliability of the results.

Ultimately, this work can be used to help determine the binding mechanism of this protein to spore DNA and can be extended to investigating other IDPs. With more research, potential inhibitors can also be discovered to prevent the various binding mechanisms of IDPS from occurring. This is typically the main goal for researchers when studying these proteins, since malfunctions of IDPs are linked to various diseases. The use of WT-BEMETA in particular, serves as one of the most powerful tools used in molecular modeling today and can be used to aid in the discovery process.

REFERENCES

[1] Setlow, P., I will Survive: DNA Protection in Bacterial Spores. *Trends in Microbiology* **2007**, 15 (4), 172-180.

[2] Kosol, S.; Contreras-Martos, S.; Cedeno, C.; Tompa, P., Structural Characterization of Intrinsically Disordered Proteins by NMR Spectroscopy. *Molecules* **2013**, 18 (9), 10802–10828.

[3] Uversky, V.N., Natively Unfolded Proteins: A Point where Biology Waits for Physics. *Protein Science* **2002**, 11 (4), 739–756.

[4] Wright, P.E.; Dyson, H.J., Intrinsically Unstructured Proteins and their Functions. *Nature Reviews Molecular Cell Biology* **2015**, 16 (1), 18–29.

[5] Lee, J.K.; Movahedi, S.; Harding, S.; Mackey, B.M.; Waites, W.W., Effect of Small, Acid-Soluble Proteins on Spore Resistance and Germination under a Combination of Pressure and Heat Treatment. *Journal of Food Protection* **2007**, 70 (9), 2168-2171.

[6] Fairhead, H.; Setlow, B.; Setlow, P., Prevention of DNA Damage in Spores and in Vitro by Small, Acid-Soluble Proteins from *Bacillus Species*. *Journal of Bacteriology* **1993**, 175 (5), 1367-1374.

[7] Moeller R.; Setlow, P.; Reitz, G.; Nicholson, W.L., Roles of Small, Acid-Soluble Spore Proteins and Core Water Content in Survival of *Bacillus subtilis* Spores Exposed to Environmental Solar UV Radiation *Applied and Environmental Microbiology* **2009**, 75 (16), 5202-5208.

[8] Setlow, P.; Hayes, C. S.; Peng, Z.-Y., Equilibrium and Kinetic Binding Interactions between DNA and a Group of Novel, Nonspecific DNA-binding Proteins from Spores of Bacillus and Clostridium Species. *The Journal of Biological Chemistry* **2000,** 275 (45), 35040–35050.

[9] Van der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R.J.; Daughdrill, G.W.; Dunker, K.A.; et. al., Classification of Intrinsically Disordered Regions and Proteins. *Chemical Reviews* **2014,** 114 (13), 6589–6631.

[10] Theillet, F.X.; Binolfi, A.; Bekei, B.; Martorana A.; Rose, H.M.; Stuiver, M.; et al., Structural Disorder of Monomeric α-Synuclein Persists in Mammalian Cells. *Nature* **2016,** 530, 45-50.

[11] Tuttle, M.D.; Comellas, G.; Nieuwkoop, A.J.; Covell, D.J.; Berthold, D.A.; Kloepper, K.D.; et al., Solid-State NMR Structure of a Pathogenic Fibril of Full-Length Human α-Synuclein. *Nature Structure and Molecular Biology* **2016**, 23 (5), 409–415

[12] Do, T.N.; Karttunen, M.; Choy, W.Y., Accelerating the Conformational Sampling of Intrinsically Disordered Proteins. *Journal of Chemical Theory and Computation* **2014,** 10, 5081–5094.

[13] Adcock, S. A.; McCammon, J. A., Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins. *Chemical Reviews* **2006,** 106 (5), 1589–1615.

[14] Lopes, P.E.M.; Guvench, O.; MacKerell, A.D., Current Status of Protein Force Fields for Molecular Dynamics. *Methods in Molecular Biology* **2016,** 1215, 47-71.

[15] Theory of Molecular Dynamics Simulations. https://embnet.vital-it.ch/MD_tutorial/pages/MD.Part1.html (accessed **2019**).

[16] De Oliveira, C.R.; Werlang, T, Ergodic Hypothesis in Classical Statistical Mechanics. *Revista Brasileira de Ensino de Física* **2007,** 29 (2), 189-201.

[17] Hospital, A.; Goni, J.R.; Ordozco, M.; Gelpi, J.L., Molecular Dynamics Simulations: Advances and Applications. *Advances and Applications in Bioinformatics and Chemistry* **2015,** 8, 37-47.

[18] Belfast tutorial: Metadynamics. https://plumed.github.io/doc-v2.5/user-doc/html/belfast-6.html (accessed **2019**).

[19] Palazzesi, F.; Barducci, A.; Tollinger, M.; Parrinello, M., The Allosteric Communication Pathways in KIX Domain of CBP. *Proceedings of the National Academy of Sciences of the United States of America* **2013**, 110 (35), 14237-14242.

[20] Bussi, G.; Gervasio, F.; Laio, A.; Parrinello, M., Free-Energy Landscape for β Hairpin Folding from Combined Parallel Tempering and Metadynamics. *Journal of the American Chemical Society* **2006**, 128 (41), 13435-13441.

[21] Marinelli, F.; Pietrucci, F.; Laio, A.; Piana, S., A Kinetic Model of Trp-Cage Folding from Multiple Biased Molecular Dynamics Simulations. *PLoS Computational Biology* **2009**, 5 (8), e1000452.

[22] Piana, S.; Laio, A., A Bias-Exchange Approach to Protein Folding. *The Journal of Physical Chemistry B* **2007**, 111 (17), 4553-4559.

[23] Alpha RMSD. https://plumed.github.io/doc-v2.5/user-doc/html/_a_l_p_h_a_r_m_s_d.html (accessed **2019**).

[24] Eisenberg, D., The Discovery of the α--Helix and β -Sheet, the Principal Structural Features of Proteins. *Proceedings of the National Academy of Sciences* **2003,** 100 (20), 11207-11210.

[25] Switching Function. https://www.plumed.org/doc-v2.5/user-doc/html/switchingfunction.html (accessed **2020**).

[26] Best, R.; Hummer, G.; Eaton, W.A., Native Contacts Determine Protein Folding Mechanisms in Atomistic Simulations. *Proceedings of the National Academy of Sciences* **2013,** 110 (44), 17874-17879.

[27] Duan, Y.; Kollman, P., Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution. *Science* **1998,** 282 (5389) 740-744.

[28] Conicella A. E.; Zerze G. H.; Mittal J.; Fawzi N. L., ALS Mutations Disrupt Phase Separation Mediated by Helical Structure in the TDP-43 Low Complexity C-Terminal Domain. *Structure* **2016**, 24 (9), 1537–1549.

[29] Miller C.; Zerze G. l. H.; Mittal J., Molecular Simulations Indicate Marked Differences in the Structure of Amylin Mutants, Correlated with Known Aggregation Propensity. *The Journal of Physical Chemistry B* **2013**, 117 (50), 16066–16075.

[30] Di Fede G.; Catania, M.; Morbin, M.; Rossi, G.; Suardi, S.; Merlin, M.; et al., A Recessive Mutation in the APP Gene with Dominant-Negative Effect on Amyloidogenesis. *Science* **2009,** 323 (5920), 1473–1477.

[31] Domene, C.; Barbini, P.; Furini, S., Bias-Exchange Metadynamics Simulations: An Efficient Strategy for the Analysis of Conduction and Selectivity in Ion Channels. *Journal of Chemical Theory and Computation* **2015,** 11 (4), 1896-1906.

[32] Ojeda-May, P.; Pu, J., Replica Exchange Molecular Dynamics Simulations of an α/β-type Small Acid Soluble Protein (SASP). *Biophysical Chemistry* **2013,** 184 (31), 17-21.

[33] Hayes, C.S.; Alarcon-Hernandez, E.; Setlow, P., N-terminal Amino Acid Residues Mediate Protein-Protein Interactions between DNA-bound α/β-Type Small, Acid-soluble Spore Proteins from Bacillus Species. *The Journal of Biological Chemistry* **2001,** 276 (3), 2267-2275.

[34] Biarnes, X.; Pietrucci, F.; Marinelli, F.; Laio, A., METAGUI. A VMD Interface for Analyzing Metadynamics and Molecular Dynamics Simulations. *Computer Physics Communications* **2012,** 183 (12), 203-211.

[35] Kaufman, L.; Rousseuw, P., Clustering by Means of Medoids. *Delft University of Technology: Reports of the Faculty of Mathematics and Informatics* **1987,** 87 (3), 1-28.

[36] Laio, A.; Gervasio, F.L., Metadynamics: A Method to Simulate Rare Events and Reconstruct the Free Energy in Biophysics, Chemistry and Material Science. *Delft University of Technology: Reports on Progress in Physics* **2008,** 71 (12), 6601.