King's College London

Department of Natural and Mathematical Science

Unfolding Nature: A Simulation Study into Protein Dynamics

Author:
Student 1870087

Supervisor:
Prof. Franca Fraternali

A report submitted for the degree of

MSc in Non-Equilibrium Systems

September 2, 2019

Abstract

Herein we have determined the potential of mean force (PMF) for a mechanical unfolding simulation of two small protein structures, an α-helix and β-hairpin, using umbrella sampling (US) and the weighted histogram analysis method (WHAM). Results show a free energy change of 12.2 ± 2.8 and 23.8 ± 2.1 kcal/mol for the helix and hairpin respectively, with the uncertainty in the calculations obtained through various bootstrapping methods. Structural analysis in conjunction with force-distance plots indicate no major transitions or energy barriers throughout the dynamics, nor does the PMF reveal any intermediate structures for either protein; though high degrees of uncertainty may mask the results. We propose these finding elucidate the difficulty of predicting protein secondary structure from chemical principles.

Acknowledgements

To my friends in the CANES office, what a year it's been.

Contents

List of Figures

Chapter 1

Introduction

The function of a protein is fully dependent on the structure it exhibits at equilibrium. For more than half a century, the prediction, formation, function, and denaturation of protein structures has been actively studied in the fields of structural biology and physics[6]. As of current, more than 80,000 protein structures have been studied experimentally and theoretically[2]; yet even with this monstrous number of cases at our disposal, we are still at a loss for a general theory of protein folding.

The transition from a 1-D amino acid sequence to a 3-D structure is an inherently non-equilibrium process, only in unique molecules and extreme conditions would a protein oscillate between its folded and unfolded states[46], and even then the intermediate states are transient. For this reason, it has been a challenge to assess the dynamics of protein folding in a meaningful way that goes beyond the laws of fundamental physics. This poses important questions for biological and medical research: how can a molecule with so many degrees of freedom find the unique minimum energy configuration, without becoming stuck in a local minimum? How is the same end state reached by a variety of starting structures? One possible way of gaining insight into these problems is through artificial manipulation of proteins, quantifying their stability and intermediate structures throughout dynamical processes. By considering the effects of various external forces on and constrained trajectories of proteins, such as mechanical unfolding [15], we hope to observe the general processes underlying their behavior.

Herein, we investigate the unfolding of two small protein structures, an α-helix and β-hairpin, and quantify the potential of mean force throughout the unfolding process. We observe structural modifications of the proteins and variations in the PMFs during the simulations, and attempt to identify critical points that occur throughout the dynamics. To this end, we considered various physical and chemical properties of the proteins, quantifying hydrogen bonding and applied force, with a qualitative view of hydrophobicity and intermediate structures. The methods used were all performed in a non-equilibrium framework, stressing the importance of non-stationary rather than fixed values

The report has been structured in the following manner: in Chapter 2 we give a broad overview of the research in protein-folding dynamics, and outline methods for computing the potential of mean force; additionally, we provide a complete description of the protein structures, including acquisition and preprocessing of the data. Next, in Chapter 3, we describe our methods of analysis, including umbrella sampling, WHAM, bootstrapping, and various methods of structural changes. We present the results and discuss their relevance to the protein folding process in Chapter 4 and finally conclude with our final remarks in Chapter 5.

Chapter 2

Background

Here we provide a brief review and the necessary background for the report: first we will discuss the protein folding problem in general, and attempt to 1) review theories as to why and how the protein so quickly conforms to the minimum energy structure, and 2) explain the major physical/chemical factors that dictate the structural changes. Then, we will consider the typical methods to determine free energy and PMFs, and give some background to each. Finally, we introduce the data used throughout the computational investigation.

2.1 Protein Folding

Proteins are the end product of the large sequence of complex and mysterious events: transcription and translation. Complex processes decode DNA base pair sequences into messenger RNA, which is then translated through the ribosomes to construct long chains of amino acids (AAs). These chains then quickly fold into 3-D conformations and hence begin their role as an active protein.

The last paragraph seems innocuous to anyone with even a general knowledge of biology, nonetheless in each sentence we find a profound occurrence of an intricate and highly complex set of events whose function, although explained by the principles of physics, seems nothing short of miraculous. Though these events have been well studied for nearly a century, no universal theory has yet been able to explain or predict the subtleties of the transcription and translation process', nor, as we focus on in this paper, the mechanism of protein folding.

Proteins are incredibly large molecules, sometimes as long as 1000 AAs (see Figure 2.1), and as such have a huge phase space. Clearly then, proteins cannot randomly sample each individual configuration, it would take over a billion years for a protein of only 100 AAs to find its minimum energy structure (MES)[27]. As a result of this complexity, studies over the last half century has left us with three major unanswered questions about protein folding[7]: 1) How do the physico-chemical properties of the AA sequence determine the structure of a protein? 2) How can proteins fold so fast, with such a large configuration space (see Levinthal's paradox [28])? and 3) How can the final 3-D structure be predicted from its 1-D amino acid sequence?

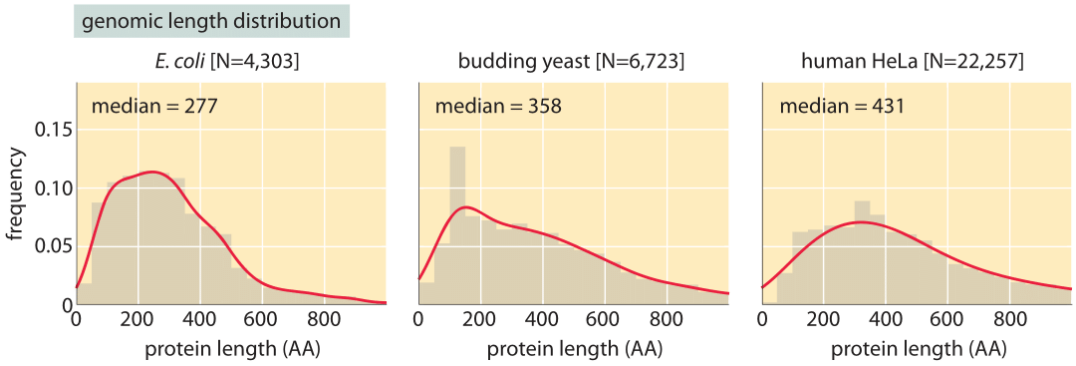


Figure 2.1: Protein length distribution in three species[34].

The prior questions have been studied extensively[12, 8]. For our purposes, the most concerning aspects are the non-equilibrium process that occur: the mechanism of folding itself, and the influence of physico-chemical properties on the mechanism. The first has driven incredible advances in experimental structure determination (see Dill et. al, 2008 : pg 294-295 for a review), though here we focus on the theoretical and computational aspects of the problem.

A common theoretical approach is to use the idea of a protein folding through traversing downhill on a funnel-shaped energy landscape. Though the term 'traversing down an energy landscape' provides no real description of the process; by what means does it 'traverse'? Which paths are restricted or favored? How can thermal fluctuations be included in the picture? Why is it assumed the active structure corresponds to the minimal energy conformation? Regardless, theories of this kind have provided a useful mental image of and given new vocabulary to the process, though it leaves a lot unanswered. *Se non è vero, è ben trovato.*

Recently, at least one aspect of folding pathways in general has been recognized which can provide some insight into Levinthal's paradox. Within the last century, several researchers have recognized that large proteins fold through units called foldons[11, 30, 29], secondary structures of helices, loops, turns, and sheets, rather than the complete AA sequence as a whole. As these units develop, structural stability increases[11] and the parallel local tasks explains how fast folding rates could come about. In addition, the culmination of parallel tasks is able to lead the protein to structures that differ from the global minimum. Though this is a significant result, it cannot explain the small structures that we investigate herein.

Many could claim that the second question, regarding the physico-chemical effects involved in the folding mechanism, is already answered by electrostatic forces. This is in some sense true, though infeasible to use in practice. What we are really concerned about is the higher-level chemical effects: the influence of hydrogen bonding, steric strain, hydrophobicity, and how their interactions dominate one another[5]. The two most important effects in the development of secondary structures are often taken to be hydrogen bonding and the hydrophobic effect, which we now briefly review.

That hydrogen bonds are a significant driver of protein structure was not at first as clear as it may at seem[42]. The first suggestion that the carbonyl and amide groups could hydrogen bond between the AAs was in 1936 by Mirsky and Pauling[35]. Hydrogen bonds, depending on the orientation and electronegativity of the donor and acceptor, are estimated to have a strength of 2 - 10 kcal/mol[38], although, the protein can also form hydrogen bonds with the surrounding solvent. It wasn't until the X-ray crystallography studies by Pauling and Corey in the 1950s[39, 40] that the idea obtained full support (though even still, the significance of the contribution had been debated[48, 24]). Soon after, theoretical studies[45, 54] have shown internal protein hydrogen bonding is more energetically favourable than protein-solvent hydrogen bonding in the helix secondary structure; however, only if a significant configurational entropy barrier is overcome[54]. This significant reduction in energy is due to a repeating chain of hydrogen bonds along the backbone, widely recognized as a key feature in α-helix and β-sheet structures since the spectras of the 1950s. Because the C=O bond is a strong acceptor and the amide a weak donor, any modifications of the solvent (as in [9, 37]) have been shown to exacerbate the amount of inter- or intra- protein h-bond formation and denature the protein. For these reasons, hydrogen bonding is thought to be a key feature of the dynamics and structure.

Another major force from the physico-chemical perspective that garners some mention is the hydrophobic effect. Much like micelle formation, protein structure is in a large capacity dictated by the aversion to water of its non-polar side chains[22]. Several investigative studies have found clustering of non-polar residues is a dominant feature of protein structure[52, 33]. Further experiments investigated the relationship between non-polar exposed surface area[47], side-chain replacement[32], and truncation[23] with respect to destabilization of the protein structure, demonstrating the effects to be proportional to the change in hydrophobicity of the protein. Unfortunately, there are no absolute measures of hydrophobicity for single, static molecules; and due to the unique dynamics of our simulation, standard techniques such as measuring hydrogen bond lifetime[41] with or orientation relaxation[53] of the nearby solvent molecules could not be accurately assessed.

2.2 Determination of Potential of Mean Force

Here we will discuss two methods of potential of mean force determination; the use of Steered Molecular Dynamics (SMD) with the Jarzynski equality, and the conjunction between US and WHAM. Both can be used to obtain equilibrium results (free energy profiles) from non-equilibrium dynamics, though by very different means. The PMF is in most cases described as the free energy along a particular reaction coordinate associated with the system[4]. SMD is the application of typical molecular dynamics, but with an additional force that changes the Hamiltonian of the system[17]; in this case, the added force is a pulling of the terminal amino acids, as we have performed in our study of the dynamics (Section 4.1). These simulations attempt to study major changes in bio-molecular structures, such as binding/unbinding processes and unfolding, and to do so apply strong forces to drive the molecule of interest out of equilibrium. By determining the work done between transitioning from the folded to unfolded state, one can use the well known Jarzynski equality $\langle \exp -\beta W \rangle = \exp \langle -\beta \Delta F \rangle$ to determine the free energy (or PMF), F, from the non-equilibrium work, W, performed[18].

On the other hand, US takes the opposite approach, and attempts to constrain the system within a range of configurations using a biased potential in order to effectively sample that region of the phase space[49]. This can be used to sample intermediate states that would have been transitory in typical MD/SMD simulations. By sufficiently sampling along the entire reaction coordinate, an accurate PMF can be obtained after un-biasing the complete set of individual histograms via:

$$A(R) = -\frac{1}{\beta} \ln P^b - w_i(R) + F_i \tag{2.1}$$

Where A(R) is the free energy along reaction coordinate R, w_i are the applied biased potentials, and F_i is a residual constant. The above equation is derived fully in Section 3.1. Due to the inter-dependency of the neighbouring windows, the unbiased state occupation probability cannot be obtained analytically[20]. One method of estimating the unbiased probability is WHAM[25], which seeks to determine the unbiased probability distribution through maximizing the log-likelihood of observing the data. A complete description of US and WHAM is given in Section 3.1 and 3.2.

Both methods yield the same end result, though each have their own peculiarities that may influence which is used. First, the SMD + Jarzynski approach is formally exact, no approximation is needed other than those used in the dynamics for the simulation; whereas US + WHAM is has inherent accuracy limitations when using more than one window. However, practically, this is not a concern: the average on the LHS of the equality is dominated by the occurrence of rare and extreme events (for small values of W)[19], hence, accurate convergence relies on taking many runs to reduce uncertainty to an acceptable degree. Second, US + WHAM calculates the state distribution, and then finds the PMF, whereas SMD + Jarzynski directly computes the PMF from work done. In principle, either approach should work fine for simple systems (length 12 and 13 proteins) and goals (calculate PMFs) such as ours, though here we only consider the US + WHAM approach.

There are alternatives to WHAM for un-biasing the resultant histograms in US simulations, which we briefly mention here. Umbrella integration avoids the interdependence between windows by differentiating Eqn. 2.1 and then integrating over the reaction coordinate, assuming the distributions at each window are Guassian with some mean and variance[21]. The dynamic histogram analysis method (DHAM) is similar to WHAM, in that it maximizes the likelihood of the unbiased probability of observing the histograms, though with the additional assumption that the system is Markov[44]. Thus, one seeks to maximize with respect to observing a certain transition matrix for the observations, and can hence analyze systems that may not be fully equilibrated. It can be shown that DHAM reduces to WHAM in the limit of completely uncorrelated trajectories.

2.3 Data

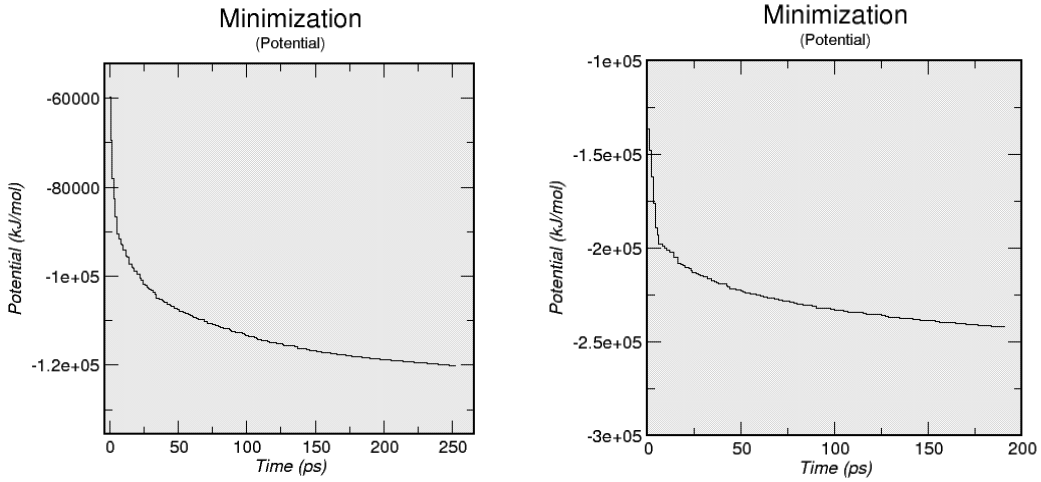
Our simulations presented herein are based on two synthetic protein structures, one α-helix of sequence length 13 and a β-hairpin with length 12 as detailed in Table 2.1. The hairpin structure is available on the RCSB Protein Data Bank[43] as 1LE0, the helix is a structure commonly found in

nesprins[31]. All data, including input, log, structural, and trajectory files were graciously provided by Franca Fraternali and Irene Marzuoli in their original format[13].

Table 2.1: Description of the protein structures used throughout simulations.

	Secondary Structure	Sequence	MW (g/mol)	PDB File
Protein 1	α-helix	L Q K W Q Q F N S D L N S	1606.77	—
Protein 2	β-hairpin	S W T W E G N K W T W K	1607.75	1LE0

After the addition of solvent, the two systems were minimized via gradient descent with a cutoff when force falls below 1000 kJ/mol/nm (approx 250 steps). Electrostatics were computed by the Particle-Mesh Ewald (PME) method with a cutoff of 1.4nm.



(a) System potential for helix

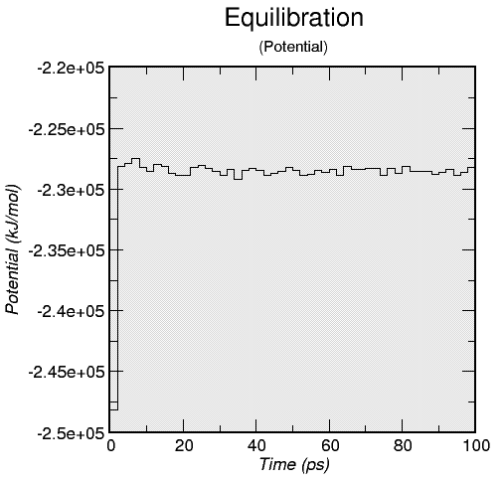
(b) System potential for hairpin

Figure 2.2: The potential energy for (a) helix and (b) hairpin proteins during energy minimization by gradient descent.

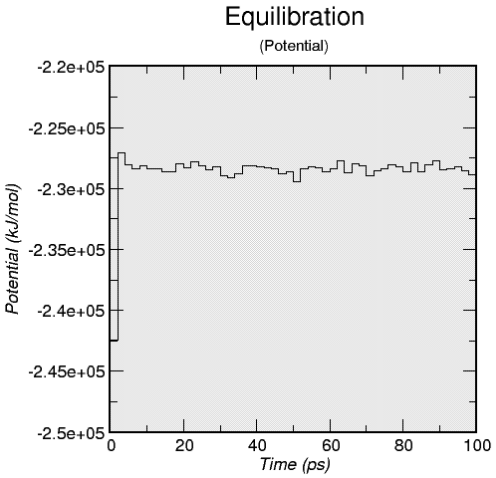
The initial pdb files were pre-processed into GROMACS[51] file format, solvated, and charge-balanced with sodium and chloride ions. Following minimization, the structures were equilibrated for 100 ps (50000 steps of 2 fs) using leapfrog integration and PME electrostatics with a 1.4nm cutoff. The simulation was carried out in an NPT ensemble, with weak (Berendsen) temperature and pressure coupling to an external bath. Figure 2.3 shows the resulting potential energy and pressure for the helix and hairpin structures subsequent to the equilibration.

With the equilibrated structures, five pulling simulations were performed; four for the helix at 0.1 - 0.0001 nm/ps speeds (100 ps - 1 ns simulations), and one for the hairpin at 0.01 nm/ps (660 ps) at each structure's terminal points. The dynamics were constrained by an umbrella (quadratic) potential with a force constant of 1000 ($kJ\ mol^{-1}\ nm^{-2}$).

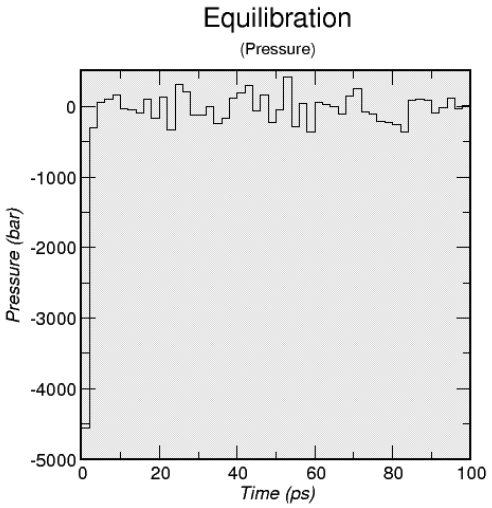
Lastly, to generate the PMF for the unfolding process, umbrella sampling simulations were performed on 25 and 21 intermediate structures for the helix and hairpin proteins respectively. Each sampling was performed for 5 ns collecting 2500 data points at each intermediate stage with an umbrella force constant of 500 ($kJ\ mol^{-1}\ nm^{-2}$). All pulling simulations utilized PME electrostatics with a 1.4nm cutoff, with the NPT ensemble generated by a Nose-Hoover and Parrinello-Rahman thermostat and barostat respectively.



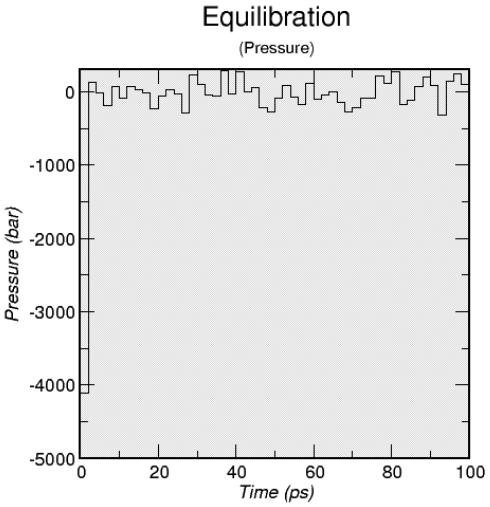
(a) System potential for helix



(b) System potential for hairpin



(c) System pressure for helix



(d) System pressure for hairpin

Figure 2.3: The potential energy and pressure for (a,c) helix and (b,d) hairpin proteins during equilibration.

Chapter 3

Methods

3.1 Pulling and Umbrella Sampling (US)

Our pulling simulations are performed at constant rates varying from 0.1 - 0.0001 nm/ps using an 'umbrella force' method which stretches the protein using a harmonic potential, rather than forcing fixed atomic coordinates. This method allows the protein more structural flexibility as the simulation is carried out, helping to produce more stable intermediates.

One run through of the pulling simulation, however, does not garner nearly enough samples to construct an accurate probability distribution for the PMF. To accurately describe the PMF of our system, we must have a large amount of samples at each point along the reaction coordinate to determine the most likely (or lowest energy) states. One 'potential' solution is to constrain our molecule at various points during the simulation by modifying the system energy itself and collecting samples on the new energy landscape[20]. This is the foundation of a technique dubbed 'Biased Molecular Dynamics' or 'Umbrella Sampling', in which one places an additional force, most often a harmonic potential, to constrain the dynamics around a region of interest without allowing the molecule to evolve towards it's typical equilibrium structure.

In order to determine the PMF for the entirety of our reaction coordinate, we run multiple windows (25 and 21 for helix and hairpin respectively) of the protein pulled to various lengths from its starting structure, held by a biased potential. The bias, w_i, of region i modifies the energy as:

$$E^b(\boldsymbol{r}) = E^u(\boldsymbol{r}) + w_i(R(\boldsymbol{r}))$$

Denoting b and u for biased and unbiased quantities respectively. For our simulations the potential is independent of momentum, thus occupation probabilities, P, along the axis, R, are given by:

$$\begin{aligned}
P^u &= \frac{\int d\mathbf{r}\delta(R^* - R(\mathbf{r}))e^{-\beta H(\mathbf{r})}}{\int d\mathbf{r}\int e^{-\beta H(\mathbf{r})}} \\
P^b &= \frac{\int d\mathbf{r}\delta(R^* - R(\mathbf{r}))e^{-\beta[H(\mathbf{r})+w_i(R(\boldsymbol{r}))]}}{\int d\mathbf{r}\int e^{-\beta[H(\mathbf{r})+w_i(R(\boldsymbol{r}))]}} \\
&= e^{-\beta w_i(R(\boldsymbol{r}))}\frac{\int d\mathbf{r}\delta(R^* - R(\mathbf{r}))e^{-\beta H(\mathbf{r})}}{\int d\mathbf{r}\int e^{-\beta[H(\mathbf{r})+w_i(R(\boldsymbol{r}))]}} \\
\implies P^u &= P^b e^{\beta w_i(R(\boldsymbol{r}))}\langle e^{-\beta w_i(R(\boldsymbol{r}))}\rangle \\
\implies A(R) &= -\frac{1}{\beta}\ln P^b - w_i(R) + F_i
\end{aligned}$$

Where $F_i = -\frac{1}{\beta}\ln\langle e^{-\beta w_i}\rangle$ is a residual term necessary for connecting the free energy, A, of individual windows[20], H is the Hamiltonian of the system, β is the thermodynamic beta, and R^* is a specific location on the reaction coordinate.

3.2 WHAM!

Unfortunately, when dealing with the overlap between multiple simulations, we cannot treat the F_i term as a simple additive constant. There are several [1, 21, 25, 44] techniques available to approximate the contribution from F_i, one of the most common prescriptions being the weighted histogram analysis method (WHAM).

WHAM is able to provide a best estimate for the unbiased probability distribution, and hence the PMF, by providing a maximum likelihood estimation for the global distribution given the biased histograms obtained via umbrella sampling. In our procedure, we simulated the pulling at 21 and 25 states for hairpin and helix respectively, collecting $N_{samp} \approx$ 1500-2500 samples (depending on simulation window) at each location.

Assume our histograms have been constructed using Q bins, in that case, the probability of observing a particular histogram of results in the i'th simulation is given by the multinomial distribution[25]:

$$P(n_{i1}, ..., n_{iQ}) \propto \prod_{k=1}^{Q} (p_{ik}^b)^{n_{ik}}$$

Denoting p_{ik}^b and n_{ik} as the biased equilibrium probability and histogram count for window i and bin k. The joint likelihood is simply the product of our independent trials:

$$P(n_{11}, ..., n_{1Q}, ..., n_{N_{sim}1}, ..., n_{N_{sim}Q}) = \prod_{i=1}^{N_{sim}} \prod_{k=1}^{Q} (p_{ik}^b)^{n_{ik}} \tag{3.1}$$

Let the biased probability be represented by $p_{ik}^b = f_i c_{ik} p_k^u$ with some weight, c_{ik}, and normalization factor $f_i = \frac{1}{\sum_k c_{ik} p_k^u}$. We now wish to maximize the likelihood of Eqn 3.1 for observing our collected data:

$$L = \ln \prod_{i=1}^{N_{sim}} \prod_{k=1}^{Q} (f_i c_{ik} p_k^u)^{n_{ik}}$$

With the addition of a Lagrange multiplier to ensure the probability does not exceed unity:

$$\bar{L} = \ln \prod_{i=1}^{N_{sim}} \prod_{k=1}^{Q} (f_i c_{ik} p_k^u)^{n_{ik}} + \sum_{i=1}^{N_{sim}} \lambda_k \Big(1 - \sum_{i=1}^{Q} f_i c_{ik} p_k^u\Big) \tag{3.2}$$

Differentiating with respect to f and p_k^u we find $\lambda_k = -N_{samp}$ and:

$$p_k^u = \frac{\sum_{i=1}^{N_{sim}} n_{ik}}{\sum_{i=1}^{N_{sim}} N_{samp} f_i c_{ik}}. \tag{3.3}$$

The additional factor, F_i, can then be determined by iterating Eqn. 3.3 with the explicit expression:

$$\begin{aligned} e^{-\beta F_i} &= \langle e^{-\beta w_i} \rangle \\ &= \int P e^{-\beta w_i} dR \end{aligned}$$

3.3 Bootstrapping

The conjunction of US and WHAM results in a PMF that provides the best estimate given our data. However, our simulations sample only a finite number of times at each of the locations chosen along the reaction coordinate, meaning there is uncertainty inherent in the results. Unfortunately, we have no reference distribution to compare to our empirical distribution, making it difficult to determine how accurately our simulations come to the true value. We are, however, able to

determine the statistical uncertainty through bootstrapping to yield some guess as to how well our underlying data describes the PMF.

Bootstrapping is a non-parametric method of statistical inference that makes no prior assumptions about the distribution of the data[36], making it ideal for cases when no analytical estimation is possible. The technique relies on re-sampling observed data to better understand how the underlying data is distributed[10]. We consider two different types, each with two sub-types, of bootstrapping on our data: trajectory based and re-sampling of complete histograms, as performed by the g_wham package[16].

In trajectory based bootstrapping (traj), we draw new data based on the distributions described by the individual umbrella sampled histograms, and create a new set of histograms and PMF for each run. Alternatively, we can model the observed histograms as Gaussians and draw from those instead (trajG). This is the standard method of bootstrapping described in the original paper by Efron[10] and used in most applications[3].

However, trajectory based approaches are often not feasible when considering systems with a large phase space. Because we are only sampling along one reaction coordinate, we must ensure that, for proper convergence, we sufficiently explore regions perpendicular to it. Without fully exploring orthogonal coordinates, we cannot assume sampling has taken into account all (or most) trajectories. Hence, if our histograms are not fully converged, then we severely underestimate our error. For this reason, extracting individual trajectories from the histograms for bootstrap analysis is expected to yield inaccurate error estimates, for both the typical and Gaussian case.

The proposed solution by Hub, de Groot, and van der Spoel[16] is to instead bootstrap complete histograms themselves, this avoids the issue of generating new (also unconverged) histograms and PMFs from biased trajectories. Two methods are used: the first groups histograms by proximity, and then selects entire histograms with replacement from the individual groups (hist). Bayesian complete histogram bootstrapping (bhist) skips the partitioning into groups and instead assigns random weights to each histogram. A general algorithm is as follows:

Algorithm 1 Complete Histogram Bootstrapping

1: **procedure** BS
2: Hists = Set of all histograms $h_i : 1 \leq i \leq N$
3: Assign weights $P(h_i) = w_i \quad \forall i : \sum_i w_i = 1$
4: Draw N samples from Hists by $P(h_i)$
5: Compute PMF
6: **end procedure**

Figure 3.1: Algorithm for the complete histogram bootstrap procedure described by Hub et al.[16]

GROMACS g_ wham ensures that the sampling procedure does not leave any gaps along the reaction coordinate, so that the bootstrapped samples can be used to calculate new PMFs[16]. Having generated several ($N_b = 2000$) profiles, we can then calculate the uncertainty from our bootstrapped PMFs, Θ_b as:

$$\sigma^2_{PMF} = \frac{1}{N_b - 1} \sum_{k=1}^{N_b} (\Theta_{b,k} - \langle \Theta_b \rangle)^2 \tag{3.4}$$

3.4 Structural Changes

From our simulations we are provided with both the variation of force over time and the PMF along the reaction coordinate. We predict that these variables will be strongly correlated to structural changes within the protein, as purported in Section 2.2. Because our simulations don't allow for easy determination of the hydrophobic effect, we have decided to simply measure the variations hydrogen bonds during the simulation. To do so, two measures were considered: hydrogen bond auto-correlation, and average (N-H)–O distance. Hydrogen bonds were determined

by hand through the Visual Molecular Dynamics (VMD) software as the atom pairs presented in Figure 4.1. From our initial selection, we monitored the auto-correlation of the population as[14]:

$$C(t) = \left\langle \frac{\sum \theta_{ij}(0)\theta_{ij}(t)}{\sum \theta_{ij}(0)^2} \right\rangle \tag{3.5}$$

Where $\theta_{ij}(t)$ is a discrete variable with value one if hydrogen bonding occurs between pairs $(N-H)_i$ and O_j, and zero otherwise. The trajectory is considered at multiple starting points to determine the average for each pulling simulation. Hydrogen bonds are defined to have an angle greater than 130 degrees and no further than 3.0 Å separation between the (N-H)–O pair. Weak hydrogen bonding, which can occur from angles greater than 90 degrees and 4 Å separation[26], was excluded from the calculation.

Our secondary measure, average (N-H)–O bond length, simply computed the euclidean distance of the hydrogen bonded pairs and took the average between them.

Beyond hydrogen bonding, we also considered the radius of gyration (RoG) and end-to-end distance (E2E) of the protein along the trajectory, which provides an indicator of progress for the pulling simulation. The E2E was determined by taking the distance between the two terminal alpha carbons. RoG is defined as:

$$R_g = \sqrt{\frac{1}{M}\sum_{i=1}^{N} m_i(\boldsymbol{r}_i - \boldsymbol{R})^2} \tag{3.6}$$

where M is the total mass, $\boldsymbol{r}_i$ and m_i the atomic coordinates and mass of atom i, and $\boldsymbol{R}$ is the mean position of the protein.

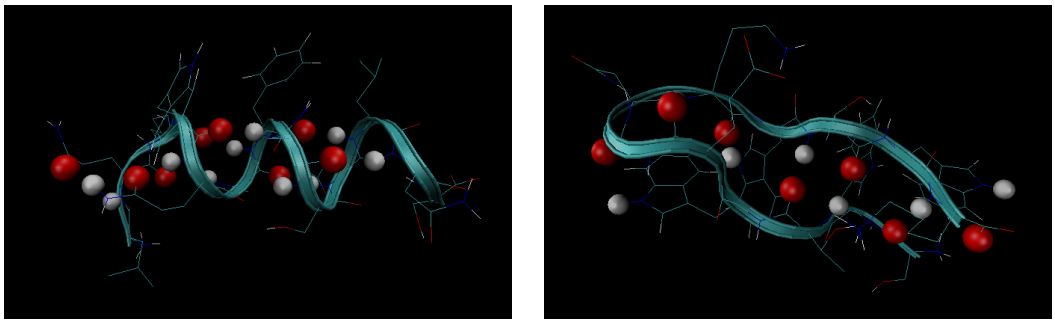
Chapter 4

Results

Results will be presented in the following format: first, we explore the relationship between the variation of force applied and structural changes in the protein. Second, we analyze the histograms obtained via umbrella sampling routines and discuss the overlap between windows. Third, we present the PMF as obtained by the WHAM algorithm on the umbrella sampled data, and consider the statistical uncertainty associated with the results. Finally, we observe the protein structure along the PMF, provide a meaningful interpretation, and give some concluding remarks.

4.1 Pulling Force and Dynamics

Because the pulling simulations were carried out at a constant rate, the force applied to the protein must vary as the conformation changes. The number of hydrogen bonds, torsional strain along the backbone, steric strain among the side-chains, and van der Waals forces all vary with the structure, requiring an increasing force to drive the protein away from its equilibrium state. We have studied the relationship between force, hydrogen bonding, and radius of gyration, to discover how these factors contribute to the stability of the folded state.

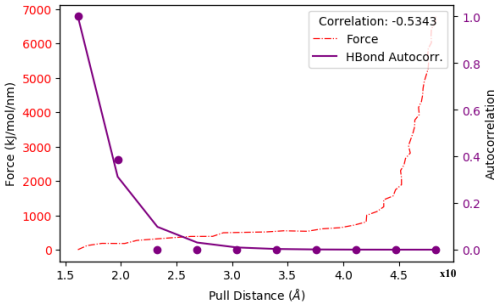


(a) (b)

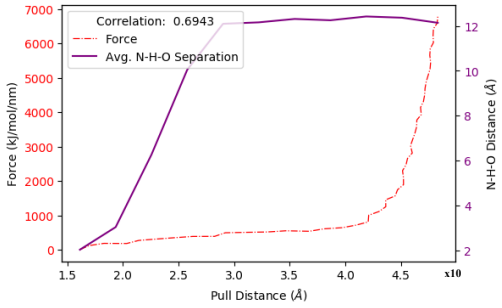
Figure 4.1: Hydrogens and oxygens involved in hydrogen bonding for the (a) helix and (b) hairpin proteins.

To quantify the correlation between applied force and hydrogen bonding, we considered two variables: the hydrogen bond auto-correlation and mean (N-H)–O bond distance for the atoms shown in Figure 4.1, following the methods described in Section 3.4. The results are displayed in Figures 4.2 and 4.3. For fast pulling speeds (0.1 - 0.01 nm/ps) we find a moderately high correlation between force and our two hydrogen bond variables ($|>0.5|$), though, there does not appear to be any clear indication that the breaking itself is directly contributing to the increased force. Had the existence and subsequent cleavage of H-bonds been a major contributor, we would expect to see step-wise increases in force at each hydrogen bond separation to overcome the associated bond enthalpy. For slower simulations, results were inconclusive, due to the very small distance covered by the pulling simulation and large fluctuations in force. The hairpin structure likewise

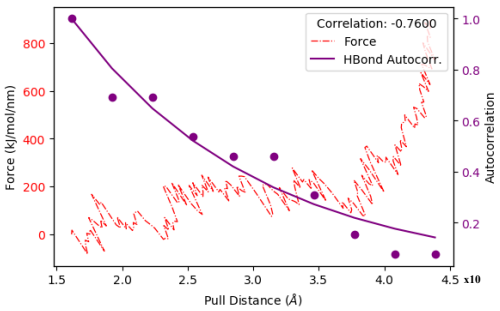
had a relatively fast pulling speed (0.01 nm/ps) and showed similar results to those for the helix; although there is a moderately strong correlation, there is no indication that hydrogen bonding is the *cause* of the increased force.



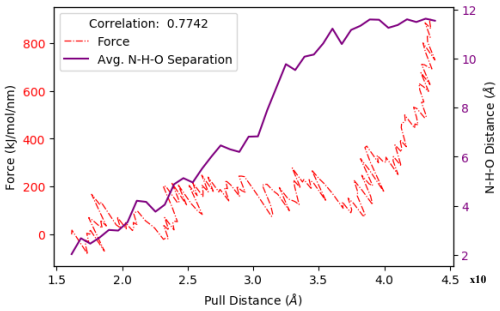
(a) HBAC, speed = 0.1 nm/ps



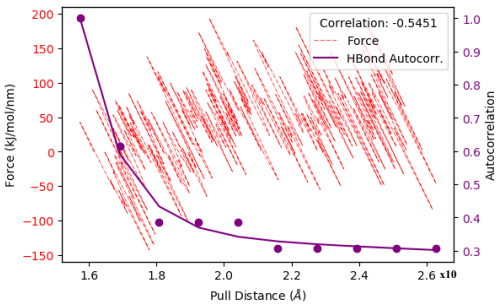
(b) HBD, speed = 0.1 nm/ps



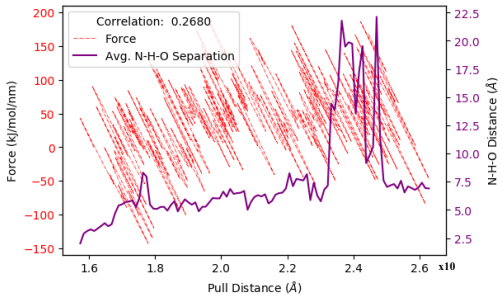
(c) HBAC, speed = 0.01 nm/ps



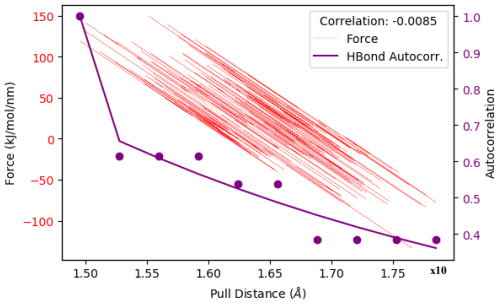
(d) HBD, speed = 0.01 nm/ps



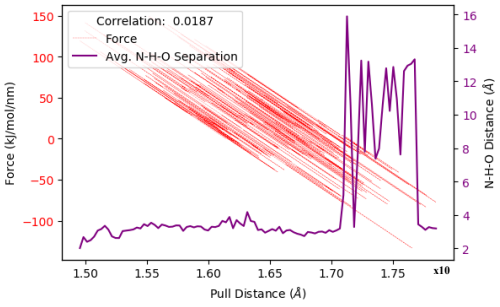
(e) HBAC, speed = 0.001 nm/ps



(f) HBD, speed = 0.001 nm/ps

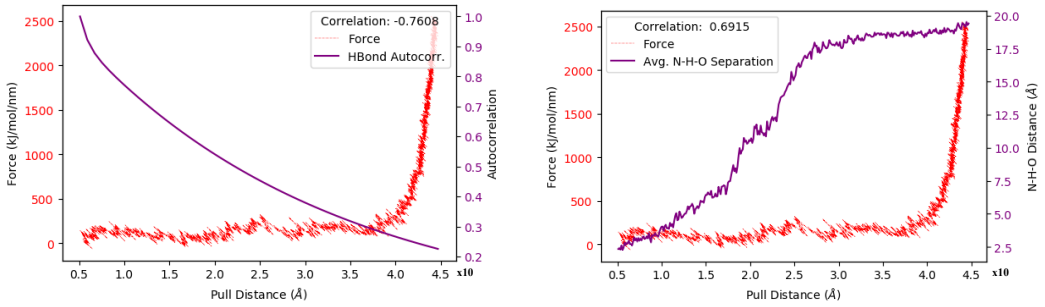


(g) HBAC, speed = 0.0001 nm/ps



(h) HBD, speed = 0.0001 nm/ps

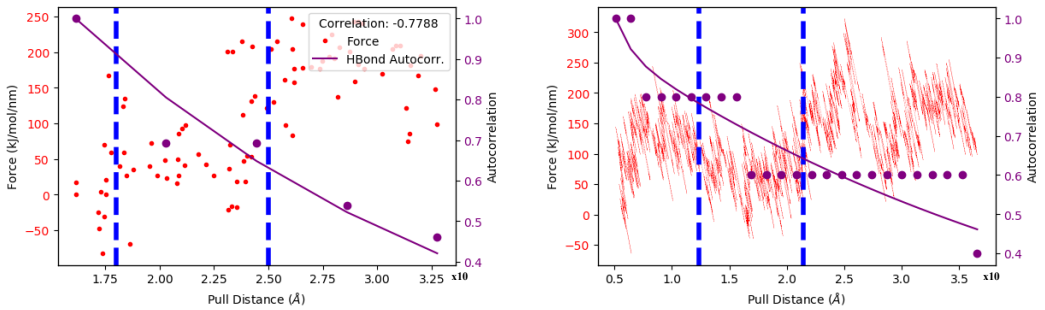
Figure 4.2: Hydrogen bond auto-correlation (Eqn. 3.5) and average (N-H)–O distance for the helix structure plotted with the applied force along the reaction coordinate. The legend displays the Pearson correlation coefficient between the two curves.



(a) HBAC (b) HBD

Figure 4.3: Hydrogen bond auto-correlation (Eqn. 3.5) and average (N-H)–O distance for the hairpin structure plotted with the applied force along the reaction coordinate. The legend displays the Pearson correlation coefficient between the two curves.

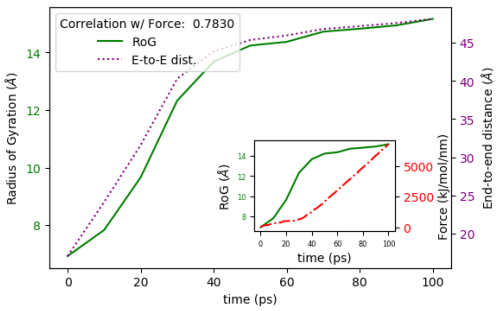
Originally, it was assumed that the sharp increase in force at the end of the pulling simulations were overshadowing the effects of hydrogen bonding. To test this assumption, we considered only the first quarter of the pulling simulation as in Figure 4.4. In it we find some interesting results; upon zooming into the reaction coordinate, there seems to be change-point behaviour in the distribution of applied force. Using the python ruptures package[50] we utilized binary segmentation with a Gaussian cost function to identify change-points in the applied force for helix and hairpin 0.01 nm/ps pulling simulations. In the helix structure, possible change-points were detected at 1.8 and 2.5 nm, and for the hairpin at 1.2 and 2.1 nm as indicated by the dashed blue lines on Figure 4.4. For the former, these points occur right before a set of hydrogen bonds are cleaved, indicating the step-wise behaviour we had predicted in the prior paragraph. However, for the latter, we see no clear relationship between the suspected change-points and number of hydrogen bonds.



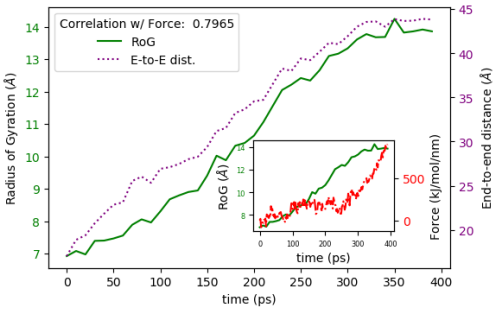
(a) First quarter HBAC, speed = 0.01 nm/ps, helix (b) First quarter HBAC, hairpin

Figure 4.4: The hydrogen bond auto-correlation for the first 1/4 of data points, to obtain a more accurate observation of the h-bond cleavage/force relationship. Blue lines represent change-points detected by binary segmentation[50].

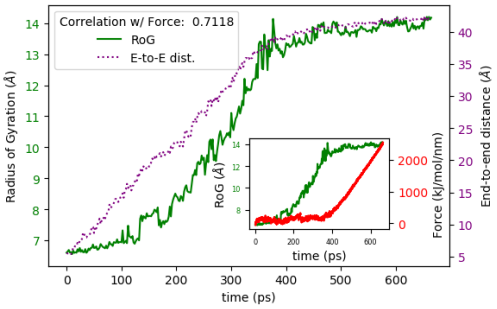
Our secondary structural comparison was how the length of the molecule varies over time, measured by radius of gyration (RoG) and end-to-end distance (E2E) between the terminal alpha-carbons. The results are presented in Figure 4.5 showing the RoG and E2E curves along with the correlation to the applied force for 0.1-0.01 nm/ps helix and hairpin simulations, and the distribution of E2E for the four helix pulling speeds. Clearly, we can see strong correlation between the force and distance pulled, though again, more-so than in the case above, we should not be quick to draw a causal inference. There is, however, one aspect of the force-distance relationship that seems undeniable in light of the plots. At some point (especially visible in Figures 4.5(a) and 4.5(c)) the RoG and E2E distance appears to level out, and the force begins to rapidly increase. We suspect that at this point, because the structure is completely unfolded, the covalent bonds along the backbone are now being torn apart. Severing covalent C-C bonds along the backbone requires an incredible amount of energy, thus explaining the sudden increase once RoG and E2E



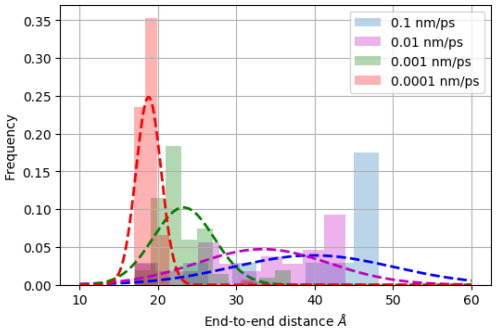
(a) speed = 0.1 nm/ps, helix



(b) speed = 0.01 nm/ps, helix



(c) hairpin



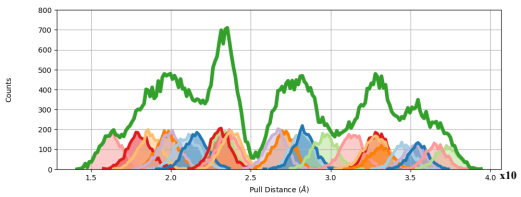
(d) E2E values for all helix pulling speeds

Figure 4.5: RoG and E2E for (a,b) helix and (c) hairpin structures. The subplot shows the RoG plotted against the applied force, with the legend displaying the Pearson correlation between the two curves. (d) Histograms of E2E values for all helix pulling speeds, fitted to Gaussians.

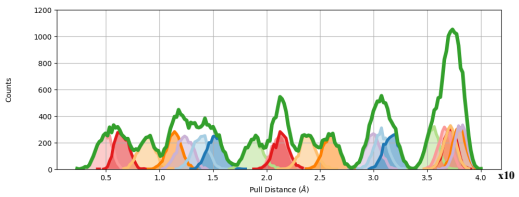
approach 40 nm.

4.2 Sampling Results

Throughout the pulling simulation, the protein was constrained to various regions along the reaction coordinate as described in Sections 2.2 and 3.1 and position was sampled. The histogram of values obtained for hairpin and helix structures are shown in Figure 4.6.



(a) US histograms for helix

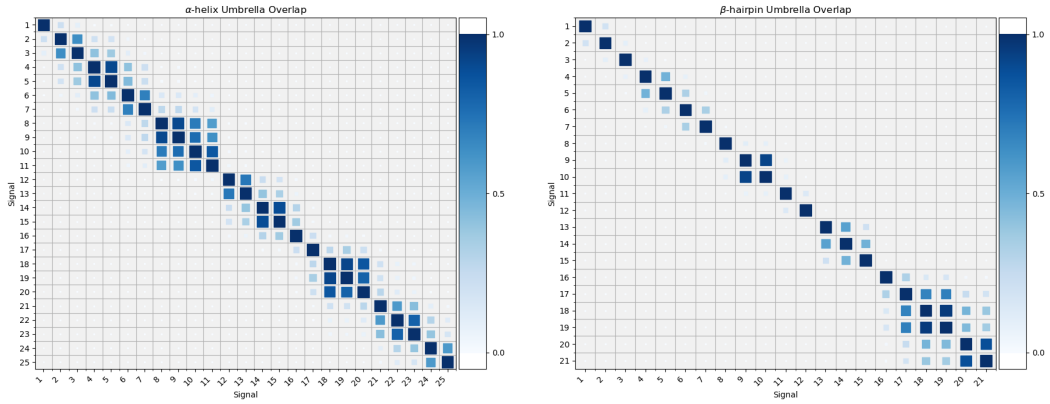


(b) US histograms for hairpin

Figure 4.6: Histograms obtained via umbrella sampling along the reaction coordinate for (a) helix and (b) hairpin structures. The green curve represents the cumulative count of all samples taken at each point.

To ensure accurate occupation probabilities, we must obtain sufficient samples along the entire reaction coordinate, which requires umbrella sampling windows to have strong overlap. This means that we must have sampling without any holes or gaps, the ideal case being uniform sampling across the whole coordinate, which guarantees that we avoid discontinuities when calculating the PMF.

To determine the efficacy of our sampling, the cumulative count of all samples across all windows at each pulling distance was determined as in Figure 4.6. Additionally, we computed a percentage-overlap matrix of all windows (sorted by position along the coordinate) presented by Figure 4.7. We were careful to calculate individual rather than total overlap, for the following argument: consider four histograms, two centered at 1.0 nm and the other two at 5.0 nm, with no overlap between the two groups; the area overlap is in this case 100%, even though a large portion of the reaction coordinate has not been sampled. From Figure 4.6 we can see there are poorly sampled regions around 2.5 nm and 4 nm in the helix structure, and at 1.75, 2.8, and 3.4 nm for the hairpin, corresponding to low overlap regions in Figure 4.7.



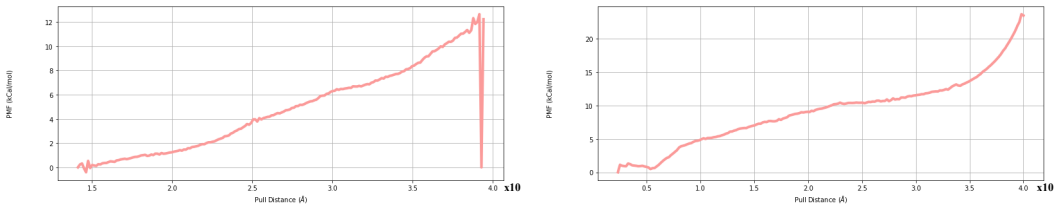
(a) US histogram overlap for helix

(b) US histogram overlap for hairpin

Figure 4.7: Percentage overlap matrix for umbrella sampled histograms, sorted by position along the reaction coordinate. The size and colour of each square represents the degree of overlap between two histograms.

4.3 PMF and WHAM

From the umbrella sampled data, the GROMACS g_ wham package performs the WHAM algorithm to obtain the PMF for the pulling process, as shown in Figure 4.8. We find, in accordance with our observations in the last section, minor discontinuities at 2.5 nm in the helix structure, and 2.25, 2.75, and 3.4 nm in the hairpin; precisely where we had poor sampling and overlap. We further observe a major disruption around 4 nm in the helix structure, in an area where we had no observations.



(a) PMF for helix unfolding

(b) PMF for hairpin unfolding

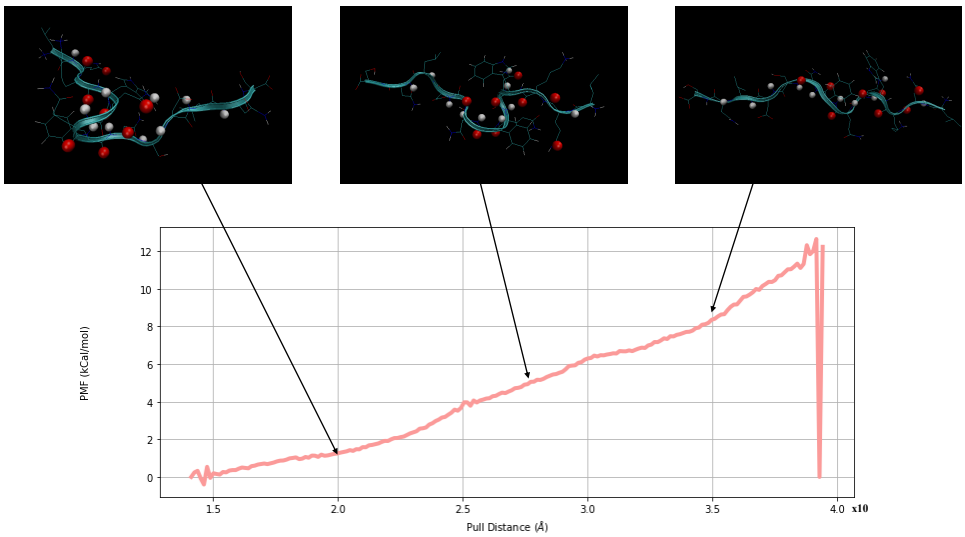
Figure 4.8: PMFs of the unfolding process for (a) helix and (b) hairpin structures. Sharp discontinuities arise from poor overlap of the umbrella windows in Figure 4.6

As described in Section 3.3, the statistical uncertainty was determined via bootstrapping both trajectories and complete histograms. We found that at 2000 bootstrapping runs the change in uncertainty leveled off (i.e. $d\sigma^2/dN_b = 0$) and hence took the standard deviation of the individual runs to be accurate to the true value of uncertainty. For the helix structure, trajectory-based methods reported a far lower level of uncertainty than the opposing complete histogram procedure, 1.2 compared to 2.8 kcal/mol. This does not mean that the trajectory based methods are more

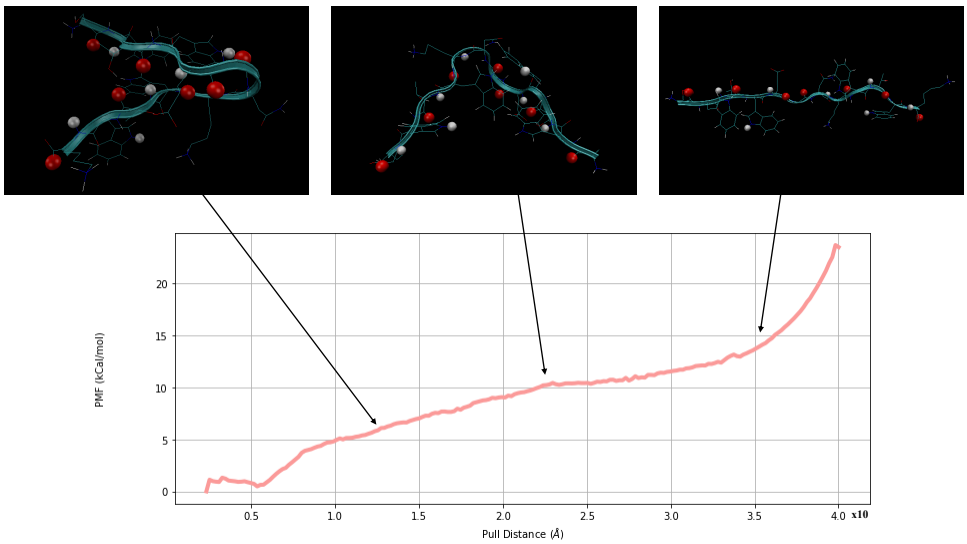
accurate, quite the contrary, it means that the true uncertainty has been severely underestimated, for the reasons given in Section 3.3. In the hairpin, we see the opposite results, with trajectory based methods yielding a higher uncertainty (2.1 vs 1.6 kcal/mol). According to the procedures, this could occur if the following criteria is met: 1) we have effectively sampled the regions orthogonal to our reaction coordinate and 2) have poor histogram overlap. The first increases the accuracy of the trajectory based methods. The second criteria limits the variation of sampled histograms in complete histogram methods; since the algorithm ensures a complete sampling of the collective variable, it must choose the same set of histograms again and again if there are no other overlapping that region. Our deviations of 2.8 and 2.1 kcal/mol for helix and hairpin structures respectively corresponds to percent errors of ± 22.9 and $\pm 8.8\%$.

4.4 PMF Structures

Finally, we take a look at the PMF and corresponding structures for both species. By Figure 4.9 we gain a full view of the unfolding process at various points along the reaction coordinate. The diagram shows there is a 12.2 and 23.8 kcal/mol difference between the folded and unfolded states for the helix and hairpin structures respectively. We find there are no stable intermediates between the folded and unfolded structure. Thus, if the dynamics only follow potential gradient, the protein will return to the folded state. However, here we are making a massive over-simplification; this PMF only corresponds to the singular reaction coordinate we considered, pulling along the Z axis. There is no saying whether the unfolded state finds a different minimum energy structure along a different reaction coordinate by re-folding along another trajectory. The results do however make this scenario unlikely, as the presence of a more stable state should cause a visible dip along some point of the PMF, indicating a low-energy structure somewhere perpendicular to the point on the reaction coordinate. In addition, there appears to be no significant discontinuities or sharp transitions, indicating no major energy barriers are traversed during the simulation.

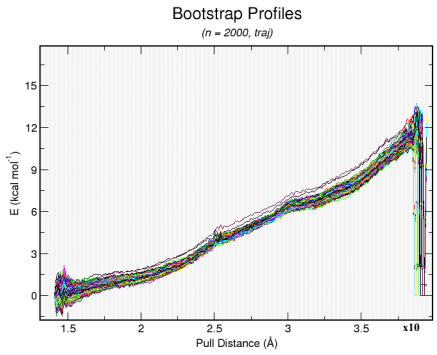


(a) Helix PMF and intermediate species

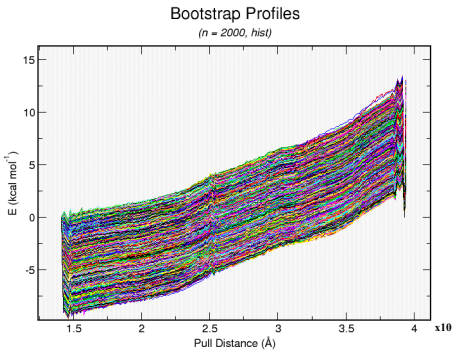


(b) Hairpin PMF and intermediate species

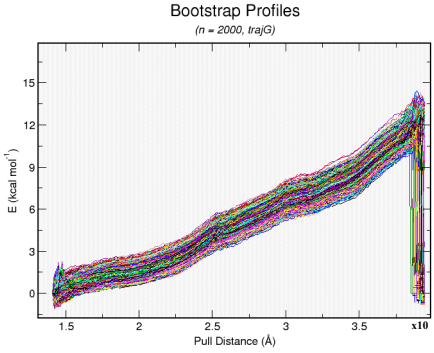
Figure 4.9: The PMF for (a) helix and (b) hairpin structures displayed with intermediate configurations. The energy difference between the two structures is described by the PMF. No stable intermediates were found along the entire reaction coordinate.



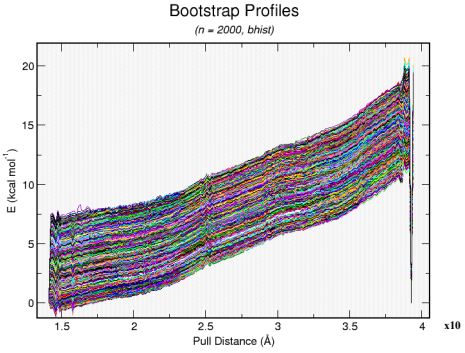
(a) Trajectory



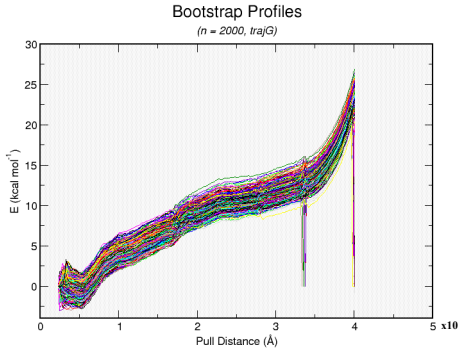
(b) Complete histogram



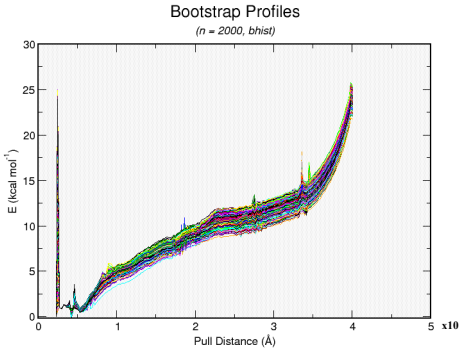
(c) Gaussian-derived trajectory



(d) Bayesian complete histogram



(e)



(f)

Figure 4.10: Bootstrapped PMF profiles for the (a-d) helix and (e,f) hairpin structures using the methods described in Section 3.3 taken over 2000 runs.

Chapter 5

Conclusion

In summary, we have identified how the dynamic changes in structure during the pulling simulation related to the applied force, and visualized our umbrella sampling data and estimated how the overlap affects our computation of the PMF. The PMF was then computed with uncertainty of 8.8 and 22.9% for helix and hairpin structures respectively, as obtained from bootstrapping the data. Lastly, we considered what the shape of the PMF inferred about the dynamics of the folding process, looking at various stages of the protein along the trajectory. From these considerations, we have found inconclusive evidence that hydrogen bonding is a main driver of folding dynamics. Some plausible results have been found through change-point analysis of the force-distance profile, though far more accurate simulations are needed to assess their accuracy. Additionally, we were unable to find any sharp discontinuities in the PMF that did not arise from poor sampling, indicating no major energy barriers or critical points occur throughout the process. The only critical point we were able to discern is the exponential increase in force after the protein has reached its maximum end-to-end distance, corresponding to the pulling of covalent bonds. All of this information provides us with the picture that protein folding of secondary structures relies on a subtle and delicate balance between force and stability. Our PMF indicates that this structure is not one of many, but it truely is an energy minimum, at least along our chosen reaction coordinate. Though, the significant uncertainty in our results may be masking the presence of stable intermediates or critical points in the transition. Future studies will need to obtain far more accurate PMF and force calculations to draw conclusive results.

Bibliography

[1] Christian Bartels. Analyzing biased monte carlo and molecular dynamics simulations. *Chemical Physics Letters*, 331(5-6):446–454, 2000.

[2] Helen M Berman, Tammy Battistuz, Talapady N Bhat, Wolfgang F Bluhm, Philip E Bourne, Kyle Burkhardt, Zukang Feng, Gary L Gilliland, Lisa Iype, Shri Jain, et al. The protein data bank. *Acta Crystallographica Section D: Biological Crystallography*, 58(6):899–907, 2002.

[3] Michael R Chernick. *Bootstrap methods: A guide for practitioners and researchers*, volume 619. John Wiley & Sons, 2011.

[4] Wouter K den Otter. Revisiting the exact relation between potential of mean force and free-energy profile. *Journal of chemical theory and computation*, 9(9):3861–3865, 2013.

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

[6] Ken A Dill and Justin L MacCallum. The protein-folding problem, 50 years on. *science*, 338(6110):1042–1046, 2012.

[7] Ken A Dill and Justin L MacCallum. The protein-folding problem, 50 years on. *science*, 338(6110):1042–1046, 2012.

[8] Christopher M Dobson. Protein folding and misfolding. *Nature*, 426(6968):884, 2003.

[9] Paul Doty and Jen Tsi Yang. Polypeptides. vii. poly-γ-benzyl-l-glutamate: The helix-coil transition in solution1. *Journal of the American Chemical Society*, 78(2):498–500, 1956.

[10] Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer, 1992.

[11] S Walter Englander, Leland Mayne, and Mallela MG Krishna. Protein folding and misfolding: mechanism and principles. *Quarterly reviews of biophysics*, 40(4):1–41, 2007.

[12] Alan R Fersht. From the first protein structures to our current knowledge of protein folding: delights and scepticisms. *Nature Reviews Molecular Cell Biology*, 9(8):650, 2008.

[13] Franca Fraternali and Irene Marzuoli. Personal correspondence. *Email*, July-September 2019.

[14] Richard J Gowers and Paola Carbone. A multiscale approach to model hydrogen bonding: The case of polyamide. *The Journal of chemical physics*, 142(22):224907, 2015.

[15] Nolan C Harris, Yang Song, and Ching-Hwa Kiang. Experimental free energy surface reconstruction from single-molecule force spectroscopy using jarzynski's equality. *Physical review letters*, 99(6):068101, 2007.

[16] Jochen S Hub, Bert L De Groot, and David Van Der Spoel. g_wham[U+E5F8] a free weighted histogram analysis implementation including robust error and autocorrelation estimates. *Journal of chemical theory and computation*, 6(12):3713–3720, 2010.

[17] Barry Isralewitz, Mu Gao, and Klaus Schulten. Steered molecular dynamics and mechanical functions of proteins. *Current opinion in structural biology*, 11(2):224–230, 2001.

[18] Christopher Jarzynski. Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78(14):2690, 1997.

[19] Christopher Jarzynski. Rare events and the convergence of exponentially averaged work values. *Physical Review E*, 73(4):046105, 2006.

[20] Johannes Kästner. Umbrella sampling. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(6):932–942, 2011.

[21] Johannes Kästner and Walter Thiel. Bridging the gap between thermodynamic integration and umbrella sampling provides a novel analysis method:"umbrella integration". *The Journal of chemical physics*, 123(14):144104, 2005.

[22] Walter Kauzmann. Some factors in the interpretation of protein denaturation. In *Advances in protein chemistry*, volume 14, pages 1–63. Elsevier, 1959.

[23] James T Kellis, Kerstin Nyberg, Alan R Fersht, et al. Contribution of hydrophobic interactions to protein stability. *Nature*, 333(6175):784, 1988.

[24] Irving M Klotz and James S Franzen. Hydrogen bonds between model peptide groups in solution. *Journal of the American Chemical Society*, 84(18):3461–3466, 1962.

[25] Shankar Kumar, John M Rosenberg, Djamal Bouzida, Robert H Swendsen, and Peter A Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *Journal of computational chemistry*, 13(8):1011–1021, 1992.

[26] Eugene Kwan. *An Introduction to Hydrogen Bonding*. Evans Group, Sep 2009.

[27] Cyrus Levinthal. Are there pathways for protein folding? *Journal de chimie physique*, 65:44–45, 1968.

[28] Cyrus Levinthal. How to fold graciously. *Mossbauer spectroscopy in biological systems*, 67:22–24, 1969.

[29] Haripada Maity, Mita Maity, and S Walter Englander. How cytochrome c folds, and why: submolecular foldon units and their stepwise sequential stabilization. *Journal of molecular biology*, 343(1):223–233, 2004.

[30] Haripada Maity, Mita Maity, Mallela MG Krishna, Leland Mayne, and S Walter Englander. Protein folding: the stepwise assembly of foldon units. *Proceedings of the National Academy of Sciences*, 102(13):4741–4746, 2005.

[31] Irene Marzuoli. Personal correspondence through jozsef mak. *Email*, September 2019.

[32] Brian W Matthews. Studies on protein stability with t4 lysozyme. In *Advances in protein chemistry*, volume 46, pages 249–278. Elsevier, 1995.

[33] H Meirovitch and HA Scheraga. Empirical studies of hydrophobicity. 2. distribution of the hydrophobic, hydrophilic, neutral, and ambivalent amino acids in the interior and exterior layers of native proteins. *Macromolecules*, 13(6):1406–1414, 1980.

[34] Ron Milo and Rob Phillips. *Cell biology by the numbers*. Garland Science, 2015.

[35] Alfred E Mirsky and Linus Pauling. On the structure of native, denatured, and coagulated proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 22(7):439, 1936.

[36] Christopher F Mooney, Christopher L Mooney, Christopher Z Mooney, Robert D Duval, and Robert Duvall. *Bootstrapping: A nonparametric approach to statistical inference*. Number 95. Sage, 1993.

[37] George Nemethy, William J Peer, and Harold A Scheraga. Effect of protein-solvent interactions on protein conformation. *Annual review of biophysics and bioengineering*, 10(1):459–497, 1981.

[38] Linus Pauling. *The Nature of the Chemical Bond...*, volume 260. Cornell university press Ithaca, NY, 1960.

[39] Linus Pauling and Robert B Corey. The pleated sheet, a new layer configuration of polypeptide chains. *Proceedings of the National Academy of Sciences of the United States of America*, 37(5):251, 1951.

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

[41] DC Rapaport. Hydrogen bonds in water: Network organization and lifetimes. *Molecular Physics*, 50(5):1151–1162, 1983.

[42] George D Rose and Richard Wolfenden. Hydrogen bonding, hydrophobicity, packing, and protein folding. *Annual review of biophysics and biomolecular structure*, 22(1):381–415, 1993.

[43] Peter W Rose, Bojan Beran, Chunxiao Bi, Wolfgang F Bluhm, Dimitris Dimitropoulos, David S Goodsell, Andreas Prlić, Martha Quesada, Gregory B Quinn, John D Westbrook, et al. The rcsb protein data bank: redesigned web site and web services. *Nucleic acids research*, 39(suppl_1):D392–D401, 2010.

[44] Edina Rosta and Gerhard Hummer. Free energies from dynamic weighted histogram analysis using unbiased markov state model. *Journal of chemical theory and computation*, 11(1):276–285, 2014.

[45] John A Schellman. The factors affecting the stability of hydrogen-bonded polypeptide structures in solution. *The Journal of Physical Chemistry*, 62(12):1485–1494, 1958.

[46] Thomas Schindler and Franz X Schmid. Thermodynamic properties of an extremely rapid protein folding reaction. *Biochemistry*, 35(51):16833–16842, 1996.

[47] Ruth S Spolar, Jeung-Hoi Ha, and M Thomas Record. Hydrophobic effect in protein folding and other noncovalent processes involving proteins. *Proceedings of the National Academy of Sciences*, 86(21):8382–8385, 1989.

[48] H Susi, SN Timasheff, and JS Ard. Near infrared investigation of interamide hydrogen bonding in aqueous solution. *J. Biol. Chem*, 239:3051–3054, 1964.

[49] Glenn M Torrie and John P Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, 1977.

[50] Charles Truong, Laurent Oudre, and Nicolas Vayatis. ruptures: change point detection in python. *arXiv preprint arXiv:1801.00826*, 2018.

[51] David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E Mark, and Herman JC Berendsen. Gromacs: fast, flexible, and free. *Journal of computational chemistry*, 26(16):1701–1718, 2005.

[52] David H Wertz and Harold A Scheraga. Influence of water on protein structure. an analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule. *Macromolecules*, 11(1):9–15, 1978.

[53] Yu-ling Yeh and Chung-Yuan Mou. Orientational relaxation dynamics of liquid water studied by molecular dynamics simulation. *The Journal of Physical Chemistry B*, 103(18):3699–3705, 1999.

[54] Bruno H Zimm and JK Bragg. Theory of the phase transition between helix and random coil in polypeptide chains. *The journal of chemical physics*, 31(2):526–535, 1959.

All code and bootstrapping outputs can be found on https://github.com/Studen1870087/SimMethods
For all queries, please email K1898719@kcl.ac.uk

Word count
Words in text: 5619
Words in headers: 45
Words outside text (captions, etc.): 507
Number of headers: 20
Number of floats/tables/figures: 16
Number of math inlines: 62
Number of math displayed: 13
Subcounts:
text+headers+captions (#headers/#floats/#inlines/#displayed)
130+10+0 (3/0/4/0) _top_
438+1+0 (1/0/3/0) Chapter: Introduction}\label{chap:Introduction
86+1+0 (1/0/1/0) Chapter: Background}\label{chap:Background
1003+2+6 (1/1/3/0) Section: Protein Folding}\label{sec:bkg_folding
619+6+0 (1/0/4/1) Section: Determination of Potential of Mean Force}\label{sec:bkg_pmf
318+1+64 (1/3/11/0) Section: Data
0+1+0 (1/0/0/0) Chapter: Methods}\label{chap:Methods
297+5+0 (1/0/5/3) Section: Pulling and Umbrella Sampling (US)}\label{sec:US
227+1+0 (1/0/13/6) Section: WHAM!}\label{sec:WHAM
457+1+12 (1/1/3/1) Section: Bootstrapping}\label{sec:DasBoot
287+2+0 (1/0/7/2) Section: Structural Changes}\label{sec:struct_changes
84+1+0 (1/0/1/0) Chapter: Results}\label{chap:Results
630+4+230 (1/5/1/0) Section: Pulling Force and Dynamics}\label{sec:pull_dynamics
231+2+77 (1/2/1/0) Section: Sampling Results
316+3+32 (1/1/4/0) Section: PMF and WHAM
196+2+75 (1/2/1/0) Section: PMF Structures
280+1+0 (1/0/0/0) Chapter: Conclusion}\label{chap:conclusion
20+1+11 (1/1/0/0) Chapter: Appendix}\label{chap:appendix

Figure 1: Word count for the report obtained via https://app.uio.no/ifi/texcount/

Office use only: Date/Time stamp

KING'S College LONDON

King's College London

Department of Mathematics

Plagiarism Statement for Coursework

The following statement must be signed by all students submitting dissertations, projects or essays as a part of their formal assessments for any degree.

Student Name: Jacob Delveaux

Module Code: 7CCMNE08

Title of Project/Dissertation: Unfolding Nature: A Simulation Study into Protein Dynamics

PLAGIARISM -

You are reminded that all work submitted as part of the requirements for any examination of the College or University of London must be expressed in your own words and incorporate your own ideas and judgements.

Plagiarism, that is the presentation of another person's thoughts or words as though they were your own, must be avoided, with particular care in coursework and essays and reports written in your own time.

Direct quotations from the published or unpublished work of others must always be clearly identified as such by being placed inside quotation marks, and a full reference to their source must be provided in the proper form. Failure to observe these rules may result in an allegation of cheating.

This includes repetition of your own work, if the fact that the work has been or is to be presented elsewhere (especially if it has already been presented for assessment) is not clearly stated.

DECLARATION:

I have read and understood the above statement and the College's statement on Academic Honesty and Integrity. I declare that the content of this submission is my own work.

I understand that plagiarism is a serious examination offence, an allegation of which can result in action being taken under the College's Misconduct regulations.

Signature: 

Date: 3 September 2019

Please submit <u>one</u> copy of the plagiarism declaration with your project/essay