New Applications of Boxed Molecular Dynamics: Efficient Simulation of Rare Events

Jonathan James Booth

School of Chemistry

University of Leeds

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

January 2016

The candidate confirms that the work submitted is his/her own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.
Chapter 3 includes work which features in two jointly authored publications:
1) - Booth, J.; Vazquez, S.; Martinez-Nunez, E.; Marks, A.; Rodgers, J.; Glowacki, D.; Shalashilin, D., 'Recent applications of boxed molecular dynamics: a simple multiscale technique for atomistic simulations',*Phil. Trans. Royal Soc. A*, **2014**, 372, 20130384.
2) - Booth, J.; Shalashilin, D., 'Fully Atomistic Simulations of Protein Unfolding in Low Speed Atomic Force Microscope and Force Clamp Experiments with the Help of Boxed Molecular Dynamics', *J. Phys. Chem. B.*, **2016**, 120, 700-708.
WIth publication 1, only the chapter on AFM pulling features in the thesis. The work, figures and writing in this section were done by myself. The remainder of the publication was written by the other authors.
The work in publication 2 was done by myself, as were the figures and all the writing apart from the theory sections. D. Shalashilin wrote the other sections and assisted with the overall editing.

Some of the cyclisation simulations for the 8 and 10 membered peptides featured in chapter 4 were carried out by MChem student Stuart Croft and by summer student Ross Schofield.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.
© 2016 The University of Leeds and Jonathan James Booth.
The right of Jonathan James Booth to be identified as Author of this work has been asserted by him in accordance with the Copyright, Designs and Patents Act 1988.

This thesis is dedicated to the memory of John Booth, who always suspected that I would turn into a boffin.

Acknowledgements

My thanks go to Dmitry Shalashilin for his supervision and guidance during the project, as well as to his family for their excellent Christmas banquets. I would also like to thank David Glowacki, David Brockwell and Lorna Clarke for their patient reading of our manuscripts and their sound advice, as well as Seishi Shimizu for his insights on thermodynamics. The experimentalists in the Jaspars group at Aberdeen should be thanked for providing such a fun challenge in the peptide cyclisation blind test. The assistance in IT related matters from the Shalashilin group as well as the hard work of the CHARMM GUI team made the whole project much easier than it might otherwise have been.

On a personal level I would like to thank the caving clubs of York and Leeds universities for the adventures in potholes and on mountains that allowed me to briefly forget about the bugs in my code. Thanks go to Matt and Lisa Gosling for their excellent cooking. And of course none of this would have been possible without the support of my family who are always in my thoughts.

Abstract

This work presents Boxed Molecular Dynamics (BXD), an efficient simulation tool for studying long time scale processes which are inaccessible to conventional methods of simulation. Boxed Molecular Dynamics is explained and introduced in the context of modelling the dynamics of proteins and peptides. Two major applications of Boxed Molecular Dynamics are reported.

1) - The mechanical unfolding of proteins induced by Atomic Force Microscopy methods is investigated. For the first time, experimental data is reproduced and unfolding pathways are investigated without the use of high artificial pulling forces which makes the simulation less realistic.

2) - A cheap and accurate *in silico* screening tool is developed to aid with the discovery and production of medicinal cyclic peptides. Enzymatic peptide cyclization is investigated by BXD and the ability of amino acid sequences to cyclize is predicted with an accuracy of 76 %.

Abbreviations

AFM	Atomic Force Microscopy
API	Active Pharmaceutical Ingredient
AXD	Accelerated Classical Dynamics
BXD	Boxed Molecular Dynamics
EEF1	Effective Energy Function 1
FC	Force Clamp
FFS	Forward Flux Sampling
FPT	First Passage Time
MCMM	Monte Carlo Multiple Minimisation
MD	Molecular Dynamics
PCC	Pre Cyclization Conformation
SMD	Steered Molecular Dynamics
TST	Transition State Theory
VC	Velocity Clamp
WHAM	Weighted Histogram Analysis Method

Contents

List of Figures

Chapter 1

Introduction

1.1 Proteins

Proteins are large biopolymers which perform a vast array of tasks in living organisms. Examples include forming structures, catalysing reactions, DNA replication, molecule and ion transport as well as enabling movement.[1] Proteins are made up of chains of amino acids linked together. There are 20 naturally occurring amino acids all sharing a common structure of a central carbon atom flanked by COOH and NH_2 groups. The general structure of an amino acid is shown in figure 1.1.

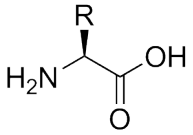

Figure 1.1: general structure of an amino acid. Note the left handed stereochemistry which is present in all naturally occurring amino acids.

Amino acids differ in the identity of the R group known as the side chain, which defines the property of the amino acid. The twenty natural amino acids are shown in figure 1.2 and can be divided into those with polar, non-polar and ionic side chains. Amino acids with polar or ionic side chains are hydrophillic and are generally found on the outside of a protein where they can interact with water. Ionic side

chains often form salt bridges where a cationic side chain interacts with an anionic side chain, stabilising the structure of the protein. Amino acids with non-polar side chains are hydrophobic and are usually found in the water free core of a protein. These competing effects underly many of the structural and functional properties of proteins and other biomolecular machinary.[2]

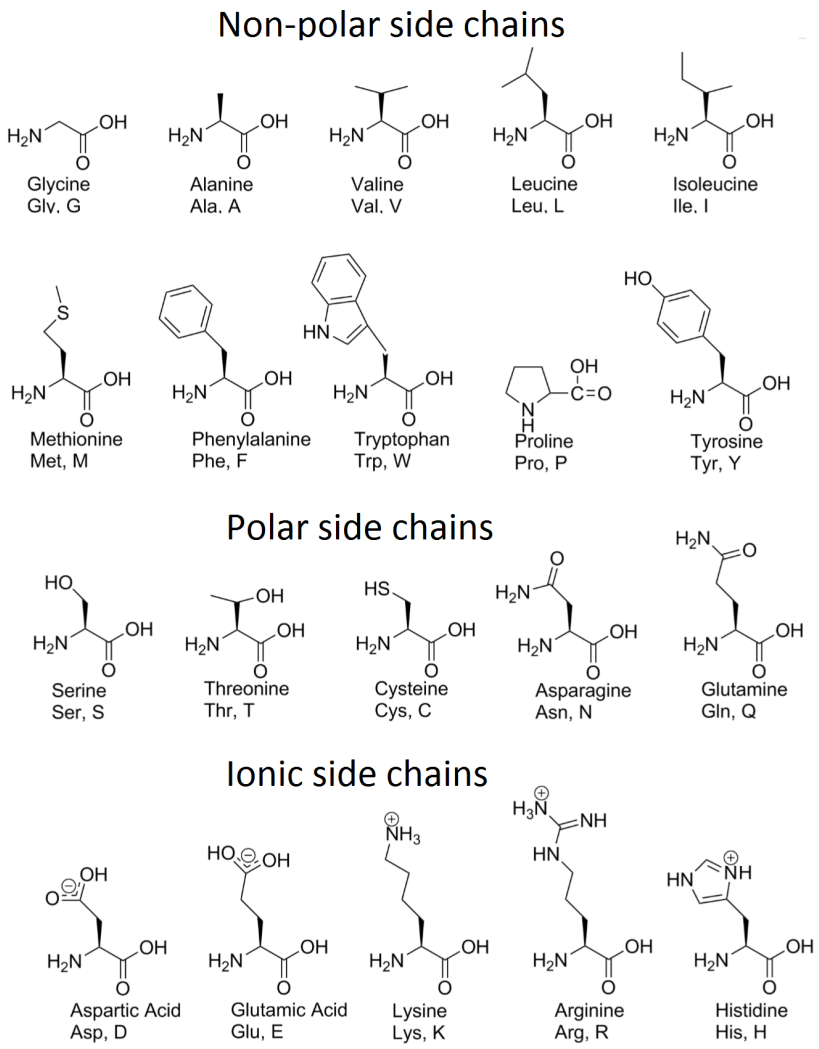

Figure 1.2: the 20 naturally occurring amino acids which can be classified according to the nature of their side chains.

The ribosome is a structure in the cell which builds up a chain of

amino acids into a protein one link at a time. Amino acids join together via the condensation reaction shown in figure 1.3.

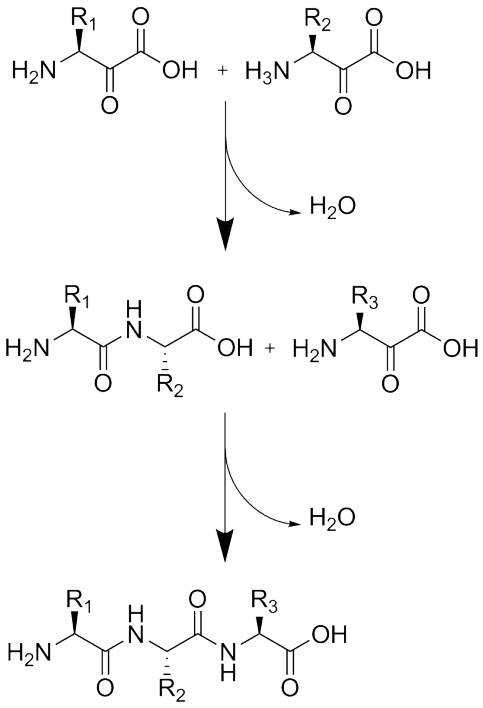

Figure 1.3: amino acids form a chain via the condensation reaction to build up a protein.

Once incorporated into a protein chain the amino acids are known as residues as some of the original atoms are lost in the condensation reaction. The chain of carbon and nitrogen atoms which has been formed is known as the backbone. The smallest proteins known as peptides have less than ten residues while the largest discovered, titin, has around 34000 which equates to over half a million atoms.

Protein structure exists on four levels. Primary structure is the sequence of residues forming the chain. Secondary structure is the local three dimensional form of the chain, held in place by the hydrogen bonding between carbonyl oxygens and amide hydrogens on the protein backbone. The two forms of secondary structure found in nature are

alpha helices and beta sheets, shown in figure 1.4. In an alpha helix the backbone forms a right handed helix with the side chains radiating outwards. In a beta sheet the backbone is layered into parallel strands while the side chains stick out to the sides. Alpha helices and beta sheets are joined by short unstructured sections of chain known as turns or links.

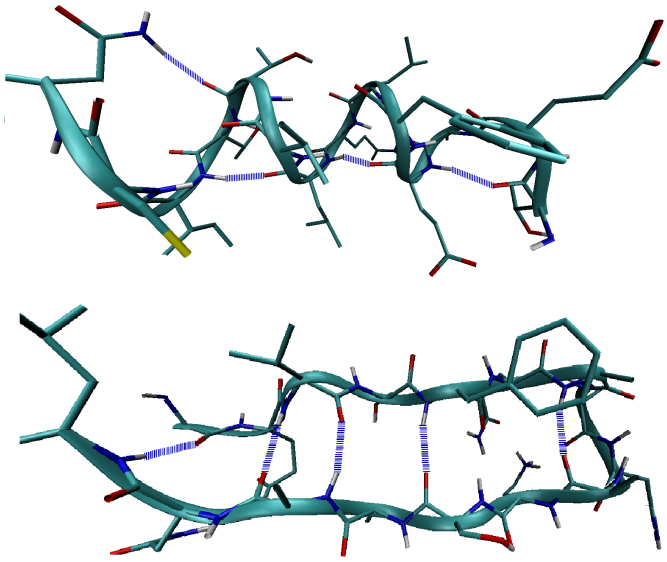

Figure 1.4: a section of alpha helix (top) and beta sheet (bottom). The peptide backbone is shown by the ribbon and the hydrogen bonds, which stabilize the secondary structure are shown by the dashed lines.

Tertiary structure is the arrangement of secondary structure into independent units known as a domains. Unlike secondary structure, which is fixed by hydrogen bonds between backbone atoms, tertiary structure is also determined by interactions between the side chains. As there are 20 amino acids there is a very large number of possible interactions and hence a wide range of possible tertiary structures. Water plays a role in tertiary structure formation; in many proteins the hydrophobic side chains are placed in the water free core of the protein while hydrophilic side chains are exposed to the solvent. A selection of domains is shown below in figure 1.5.

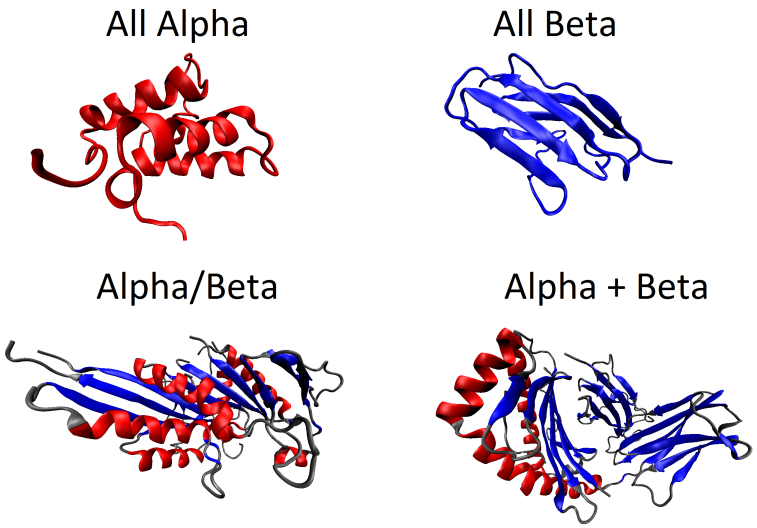

Figure 1.5: A selection of protein domains showing different tertiary structures which are classified as all alpha, all beta, alpha/beta (alternating sections of alpha helix and beta sheet) and alpha+beta (segregated alpha helixes and beta sheets). Alpha helixes are shown in red, beta sheets in blue and the flexible linkers between units of secondary structure are shown in grey. These structures were chosen to show a broad spectrum of the types of secondary structure and their combinations. Structures were downloaded in pdb form from the RCSB databank: 1IMQ (all alpha), 1TIT (all beta), 2ZFL(alpha/beta) and 3BGM (alpha + beta).

Quaternary structure is the highest level of structure and is defined as the arrangement of multiple folded subunits relative to each other. Each folded subunit is a seperate chain of amino acids with distinct tertiary structure. As such a quaternary protein is not a single covalent molecule but a complex of several smaller units. An example of a protein showing quaternary structure is shown below in figure 1.6.

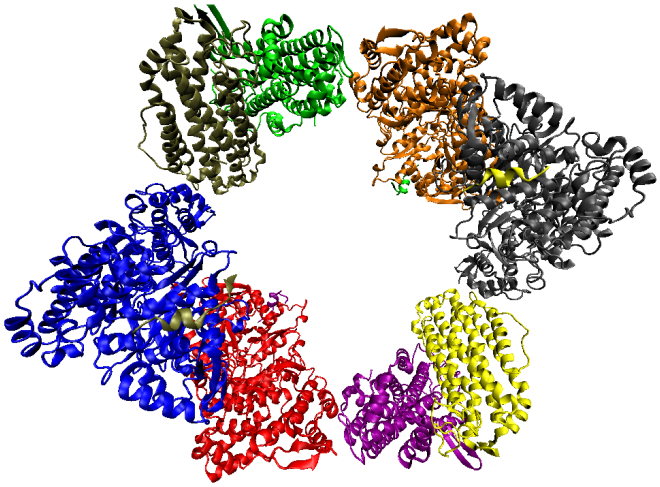

Figure 1.6: a protein showing quaternary structure: the arrangement of folded subunits, shown by different colours, relative to each other. The protein shown is an *E. coli* class Ia ribonucleotide reductase, the structure was downloaded from the RCSB databank (5CNU) and was chosen because it shows a high level structure which is clearly made up of independant folded subunits.

Many proteins show a very high degree of conformational flexibility and do not exhibit any stable structure. Known as intrinsically disordered proteins, these species can form indepentent subunits or occur as flexible linkers between stable protein domains in larger structures.[3]

Protein function is crucial to countless biological processes and understanding it is vital to treating and curing most diseases and conditions. Function comes from structure and dynamics; proteins move and change conformation at room temperature and form labile complexes with other proteins and smaller molecules. These processes are hard to observe in experiment[4–6] so some form of simulation is necessary to investigate protein dynamics. One such method is Molecular Dynamics (MD).

1.2 Molecular Dynamics

Since the first Molecular Dynamics study in 1957[7] MD simulation has become a useful tool for the advancement of chemical and biological understanding.[4] MD allows the elucidation of many structures and processes that cannot be directly observed by experiment[4–6] and is becoming increasingly popular in the study of biological systems since it was first used to study protein structure in 1976.[8] Aspects of biology explored by MD include enzyme catalysis,[9] protein folding,[10–14] ion and small molecule transport[15;16] and ligand binding.[17]

MD is used to observe dynamic processes of biomolecules such as conformational change, or to calculate properties such as structure. The property or process under investigation is modelled as a function of the position and velocity of each atom in the system. Biomolecules undergo significant structural fluctuation at room temperature and a simulation needs to be time resolved to capture this. For example, to find the structure of a protein, the positions of the atoms are calculated over a period of time and averaged, as the structure is not fixed and will vary with the dynamics of the system. This differs from a small molecule quantum mechanical calculation where a structure is obtained from a simple potential energy minimisation.

To simulate the time evolution of a molecule (known as a trajectory) it is necessary to model the interactions between the atoms. MD defines these interactions classically, for example covalent bonds are treated as simple harmonic oscillators while non-covalent interactions are modelled with simple electrostatic and Lennard-Jones potentials. This is done in order to reduce the cost of the calculation; there are too many atoms in biomolecules for electrons to be explicitly included. Each atom is described by a set of parameters such as covalent bond force constants and Lennard-Jones constants. The parameters used to calculate the potential energy, known as a force-field, differ between

models.

Knowing the potential energy, the position and velocity of each atom can be calculated and propagated forwards in time. This is done by applying Newtonian mechanics to each atom using a finite difference method where dynamics are calculated over a series of discrete time steps rather than on a continuous scale in order to make the computation less expensive. MD simulations use numerous finite difference methods, or integrators, to calculate a trajectory. The primary integrator used in this work is the Leapfrog integrator.[18]

1.3 The Long Timescale Problem

The idea of inputing a set of atomic coordinates into a computer and watching a protein fold, or an enzyme catalyse a reaction, sounds too good to be true. Indeed, MD faces a major problem with the timescales involved in most biological processes. This is demonstrated by the example of protein folding - given a sequence of amino acids, what is the lowest free energy structure? In theory this would be answered by starting an MD simulation with a chain of amino acids and watching the trajectory take the molecule into its the lowest free energy conformation. The quantity calculated here would be the average position of the atoms in the conformation corresponding to the native structure. This is unlikely to be found by MD as proteins fold on a timescale of milliseconds to seconds and sometimes even longer,[19] while the time step of an MD simulation is around one femtosecond. This would require propagating the trajectory over a very large number of time steps.

This brute force structure calculation has been shown to be possible in principle by the Shaw group, who used a specially built supercomputer to fold a number of proteins using MD simulations.[20–22] Despite the promise of these simulations, the problem of long time scales remains as the force fields used in MD simulations are parameterised by

reproducing physical quantities of liquids, such as density, from short timescale simulations. It has been shown that these force fields do not remain fully valid when extended over the much longer timescales of protein folding.[23]

The long timescales arise from the fact that the free energy landscape of a protein is very complicated, featuring many wells and barriers with meta-stable states and kinetic traps. This makes the conformation space hard to explore in sufficient detail in a reasonable amount of time.

Given a one dimensional reaction coordinate x which describes a process such as protein folding, the probability of x having a particular value is related to the corresponding free energy and the temperature in the following way:

$$P(x) = e^{\frac{-G(x)}{k_B T}} \tag{1.1}$$

It is clear from equation 1.1 that areas of high free energy will rarely be sampled because of the vanishingly small probability of the trajectory arriving there. This means that the trajectory will move downhill into whatever free energy well it finds first and stay there as the barriers around the well are too high to be crossed in a timescale accessible to simulation. Convergence will be poor as few regions of the conformational space have been sampled. This problem, common to many biological processes, is illustrated below in figure 1.7.

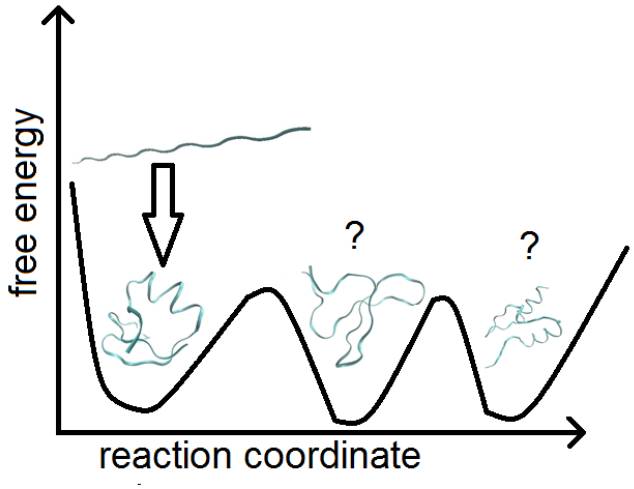

Figure 1.7: in the ideal simulation the trajectory takes the molecule into its lowest free energy conformation. In reality the free energy landscape is complex, featuring many local minima with barriers between them. The trajectory finds a minimum and stays there. The simulator cannot know whether this is a local or a global minimum as only a small fraction of the conformation space will have been sampled.

1.4 Long Timescale Methods

Numerous methods exist which deal with this problem. Known as accelerated sampling methods, these techniques usually fall into one of three categories: biasing methods, reactive flux methods and temperature methods. In this section some of the more common techniques are reviewed.

1.4.1 Biasing Methods

With these methods a reaction coordinate x is defined to describe the process under investigation. The potential energy is modified so that the trajectory moves along x more easily, crossing barriers that would otherwise be too high for an unbiased trajectory to move over.

Umbrella Sampling

Umbrella sampling[24;25] modifies the potential energy function by adding a bias which pushes the trajectory over barriers. The modified potential $U'(r)$ is obtained by adding a biasing term which is a function of the position along the reaction coordinate:

$$U'(r) = U(r) + W(x) \tag{1.2}$$

where $U'(r)$ is the new potential energy function, $U(r)$ is the original potential and $W(x)$ is the biasing potential along the reaction coordinate, usually chosen to be a harmonic potential or umbrella:

$$W(x) = k(x - x_0)^2 \tag{1.3}$$

where k is a constant and x_0 is the centre of the umbrella. In practice many simulations, with biasing potentials centred on different values of x_0, are combined. Between the set of simulations the reaction coordinate is covered by umbrellas, allowing all barriers to be crossed. This is shown in figure 1.8.

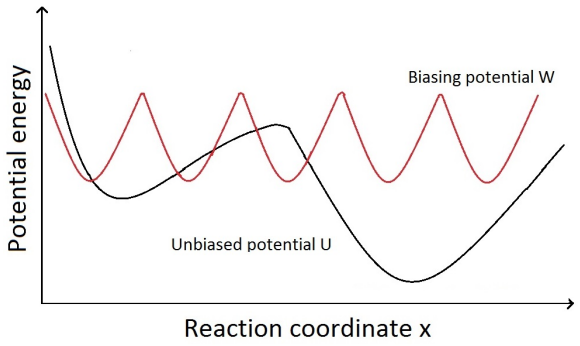

Figure 1.8: umbrella sampling places a series of harmonic biasing potentials along the reaction coordinate.

The unbiased free energy is recovered via the Weighted Histogram Analysis Method or WHAM[26]. From the whole set of simulations, the

biased probability $P_i'(x)$ of observing a state in which the reaction coordinate has a value of x in simulation i is

$$P_i'(x) = \frac{P(x)\, e^{\frac{-W_i(x)}{k_B T}}}{\sum_i e^{-\frac{U(r)+W_i(x)}{k_B T}}} \tag{1.4}$$

where $P(x)$ is the probability of the reaction coordinate having value x with the unbiased free energy (see equation 1.1), $W_i(x)$ is the biasing potential used in simulation i and $U(r)$ is the unbiased potential energy. Recovering $P(x)$ will allow the unbiased free energy along the reaction coordinate to be obtained as, by rearranging equation 1.1,

$$\Delta G(x) = -k_B T ln\left[P(x)\right] \tag{1.5}$$

where $\Delta G(x)$ is the free energy a function of the reaction coordinate. Computing $P(x)$ is non trivial and convoluted.

A more advanced version of Umbrella Sampling is Adaptive Bias Umbrella Sampling[27] where the biasing potential is expanded to cover the whole reaction coordinate, and varied on the fly until the simulation populates all states equally. At this point the combination of the free energy and biasing potential results in a flat energy surface as all points along the reaction coordinate are equally probable. The advantage of this is that the free energy can be recovered as the negative of the biasing potential rather than via more convoluted algorithms.

Metadynamics

Metadynamics fills in free energy wells to flatten out the energy landscape along the reaction coordinate.[28] A biasing potential is constructed by adding Gaussian potentials, until the well is filled up and then crosses into the next well. This is then filled up and the process is repeated until the free energy along the reaction coordinate is flat, allowing the reaction coordinate to be well sampled. This is illustrated in figure 1.9.

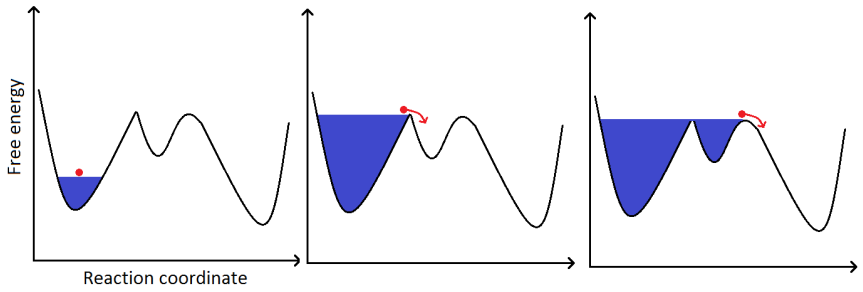

Figure 1.9: metadynamics fills up free energy wells with biasing potentials (blue) and allows the trajectory (red) to travel easily along the reaction coordinate.

As with adaptive bias umbrella sampling the biasing potential will converge to the negative of the unbiased free energy along the reaction coordinate.

In general, biasing methods quickly accelerate the sampling of a reaction coordinate and allow the observation of many processes which would be inaccessible to conventional MD. However, unbiasing the results to obtain equilibrium free energies and kinetics is often complicated. Also for methods such as metadynamics, prior knowledge of the free energy profile along the reaction coordinate is needed which limits the number of applications.

1.4.2 Reactive Flux Methods

Reactive flux methods are based on Transition State Theory[32;33] (TST) in which the phase space is partitioned into two sections separating states A and B with a dividing surface between them. The rate constant for transition from A to B is defined as the flux through this dividing surface:

$$K_{AB}^{TST} = \kappa \frac{e^{-\frac{W(x^*)}{k_B T}}}{\int_{-\infty}^{x^*} e^{-\frac{W(x^*)}{k_B T}} dx} \tag{1.6}$$

where x is the reaction coordinate describing the path from states A to B, x^* is value of x at which the dividing surface is located and $W(x^*)$ is the reversible work needed to move the system from state A to x^*. The fraction represents the equilibrium probability of finding the system at x^* compared to the probability of being at A. The constant κ represents the fraction of trajectories which, on reaching the transition state x^*, cross over and reach state B.

Expressed informally, the TST equation states that the rate constant for a transition from A to B is equal to the chance of finding the system at the transition state, multiplied by the probability of the trajectory then crossing over to state B and remaining there. Equation 1.6 forms the basis of reactive flux methods. The dividing surface x^* is formally the transition state between A and B however this does not have to be the case; equation 1.6 is valid for arbitrary states A and B which do not have to represent stable states. The position of the surface x^* can also be arbitrary.

In general these methods work by dividing up the phase space with many such surfaces and calculating the rate constants between them. A selection of reactive flux methods is reviewed below.

Milestoning

Milestoning[29;30] places a series of planes or milestones in phase space along a reaction coordinate. A set of conventional MD simulations are initiated on and restricted to each plane, generating an ensemble of conformations for each plane. Each of these conformations is then run as a separate unbiased trajectory until it reaches the next plane, at which point it is terminated. A trajectory initiated at plane H_n has a lifetime τ^+ if it ends at plane H_{n+1}, and τ^- if it ends at plane H_{n-1}. This is shown below in figure 1.10.

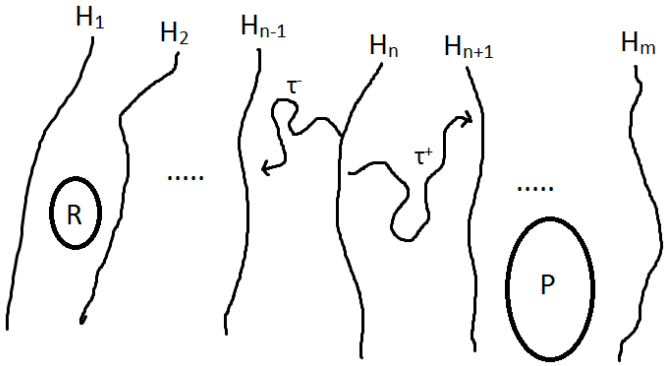

Figure 1.10: in milestoning planes are placed along a reaction coordinate between reactants and products. Conformations are released from a plane until they hit the next one and the time taken from release to collision is recorded.

The distribution of trajectory lifetimes going from plane H_n to H_{n+1} is written as ${K_n}^+(\tau)$, and as ${K_n}^-(\tau)$ for those going from plane H_n to H_{n-1}. ${K_0}^-$ and ${K_m}^+$ are zero as below plane H_0 and above plane H_m there are no other planes to hit. If the equilibrium probability $P^{eq}(n)$ of a trajectory being found on milestone H_n were known then the free energy along the reaction coordinate could be found from equation 1.5, as such a distribution would represent the probability of finding a trajectory along the reaction coordinate. The probability of finding a trajectory at H_n at time t is

$$P_n(t) = \int_0^t \left[1 - \int_0^{t-t'} K_n(t-t')\right] Q_n(t')\,dt' \tag{1.7}$$

where the integrand is the probability of arriving at H_n at time t' and not leaving before time t.

$K_n(t-t')$ is defined as ${K_n}^+ + {K_n}^-$ and $Q_n(t')$ is defined as

$$Q_n(t') = 2\delta(t) P_n(0) + \int_0^t Q_{n\pm1}(t'') K^{\pm}_{n\pm1}(t-t'')\,dt''. \tag{1.8}$$

Equation 1.8 is the probability of a transition to H_n expressed as a sum over the initial conditions of the starting trajectories $P_n(0)$, and over previous transitions to H_{n+1} followed by a transition to H_n. Solving equations 1.7 and 1.8 gives the equilibrium probability of finding a trajectory at H_n at time t:

$$P^{eq}(n) = \lim_{t\to\infty} P_n(t) \tag{1.9}$$

where $\lim_{t\to\infty} P_n(t)$ is found by running the trajectories until the distributions of lifetimes converges. Now the free energy along the reaction coordinate is obtained from equation 1.5.

A recent modification[31] of milestoning replaces the dividing surfaces with polygons rather than simple planes. The polygons, known as Voronoi polyhedra, divide the phase space into a network of tessellating cells. Trajectories are initiated in each cell and confined within it. The frequency at which the trajectory hits a boundary between polyhedra is used to calculate the rate constant for passage between them. In this way the phase space can be sampled in multiple dimensions rather than along a one dimensional reaction coordinate.

Forward Flux Sampling

In forward flux sampling[35] (FFS) a series of planes (denoted $\lambda_0, \lambda_1, ...\lambda_n$) are defined in the phase space between the product and reactant regions (A and B). Trajectories are initiated on λ_0 which lies on the initial state A. If a trajectory reaches λ_1 then its conformation is stored and used to restart multiple trial trajectories. If a trial trajectory reaches the next plane λ_2 then it is counted as a success, and as a failure if it returns to λ_1. Successful trajectories are again stored and used as starting points for a fresh set of trial trajectories until the final state B is reached. This is illustrated below in figure 1.11.

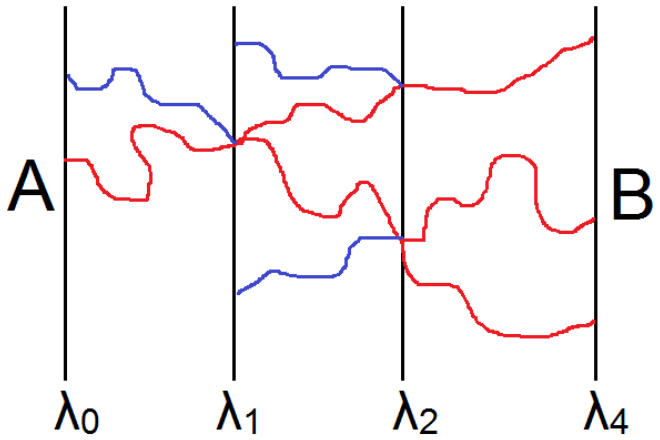

Figure 1.11: in forward flux sampling, planes are placed along the phase space between initial state A and final state B. A trajectory is released from plane λ_0 at state A and the conformation is stored if it reaches the next plane λ_1. This stored conformation is used to seed multiple trial trajectories from λ_1. The ratio of successful trajectories that reach the next plane (red) to trajectories which return to the previous plane (blue) is used to calculate the probability of passage between planes. Once this is known for each plane the free energy from states A to B can be calculated.

The ratio of successful to failed trajectories at plane λ_m gives the probability of transition from λ_m to λ_{m+1}. Once this is known for all the planes then the free energy for the entire process is obtained by using equation 1.5. FFS has the added advantage that a sequence of successful trial trajectories that make it all the way to state B can be joined together to form an unbiased trajectory representing the overall transition from A to B.

In addition to the two methods described above there are many other reactive flux methods which work in a similar way. These include transition path sampling,[36] transition interface sampling,[37] the weighted ensemble method[38] and the finite temperature string method.[39] These methods accelerate the rate of sampling because a trajectory must only reach the next boundary. This allows energy barriers to be crossed as the boundaries act as a ratchet, allowing the process to be simulated

in a series of small hops rather than one large unlikely jump from A to B.

The advantage of these techniques is that no biasing of the potential energy is needed which simplifies the generation of equilibrium free energies and kinetics. However these methods often require multiple trajectories to be run at different points in the phase space, complicating the procedure and requiring prior knowledge of the free energy landscape.

One of the major limitations of both biasing and reactive flux methods is the need to define a reaction coordinate. This is often difficult as the motion of biomolecules is highly dimensional, with many processes requiring movement accross a large proportion of the $3N$ degrees of freedom available to a molecule of N atoms. Because of this it can be hard to find a one dimensional reaction coordinate which differentiates between the states of the system and describes the pathways between them.[40]

1.4.3 Temperature Methods

Temperature methods vary the temperature of the simulation in order to push the trajectory over free energy barriers. The most basic of these techniques is simulated annealing.

Simulated Annealing

In simulated annealing[41] the temperature is raised which lifts the system over free energy barriers. The temperature is then reduced so that the molecule falls into an energy minimum and the process is repeated, exploring multiple minima which would not be accessed by a conventional trajectory, see figure 1.12.

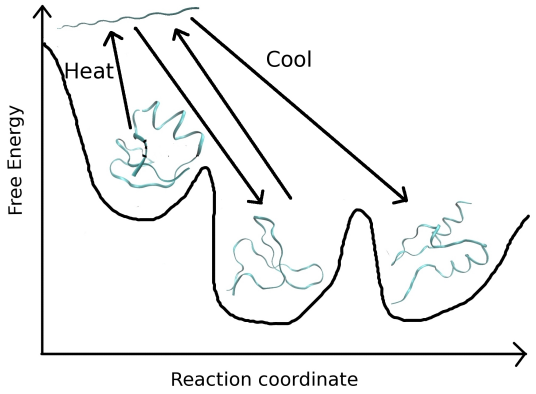

Figure 1.12: simulated annealing heats and cools the system to explore multiple energy minima.

Replica Exchange

Simulated annealing has developed into the replica exchange method[42] where separate simulations are carried out at different temperatures. These simulations, known as replicas, do not interact but can swap temperatures with their neighbours. Because of this the trajectories take a random walk in temperature space which allows them to cross barriers easily and cool down into minima that would otherwise not be sampled.

The advantage of temperature methods is that the phase space can be efficiently sampled without modifying the potential energy or dividing up the phase space with boundaries. This requires no prior knowledge of the process under investigation. Also there is no need to define a reaction coordinate to describe the process of interest, which is a major limitation of many accelerated sampling methods as a suitable coordinate does not always exist. However, simulated annealing does not allow equilibrium dynamics to be reconstructed and replica exchange is a complicated process to implement as many trajectories

must run in parallel and swap temperatures with each other.

Now that MD has been introduced along with a selection of accelerated sampling methods, the method used in this work will be explained.

Chapter 2

Method

Boxed Molecular Dynamics (BXD) is a method of calculating the free energy and kinetics of a slow process. It is a simple method which does not require any modification of the potential energy or any prior knowledge of the process under investigation. In this chapter the foundations of BXD will be introduced along with the assumptions on which it relies and the conditions under which these are valid.

2.1 Accelerated Classical Dynamics (AXD)

BXD is based on Transition State Theory[32] and is similar to several reactive flux methods. Boundaries are placed along a reaction coordinate which partition the phase space into a series of boxes. The boundaries are reflective and confine the trajectory within a box. This is done by monitoring the value of the reaction coordinate as the simulation progresses. If the value crosses a threshold at which a boundary is located then the velocities of the atoms are inverted with respect to the reaction coordinate. This is illustrated in figure 2.1.

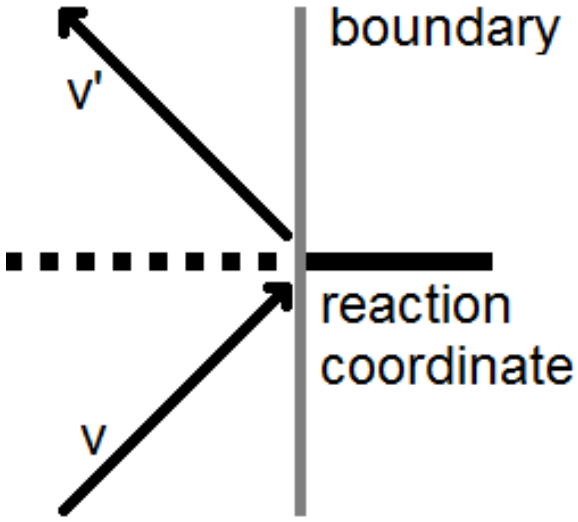

Figure 2.1: the velocity of an atom is reflected with respect to the reaction coordinate when a boundary is hit. If the velocity before the collision is v with a component $v_{parallel}$ along the reaction coordinate then the velocity after the reflection is given as $v' = v - 2v_{parallel}$.

To understand the link with TST consider a simple system of two reflecting boundaries shown below in figure 2.2. This set-up is an early version of BXD known as Accelerated Classical Dynamics (AXD).

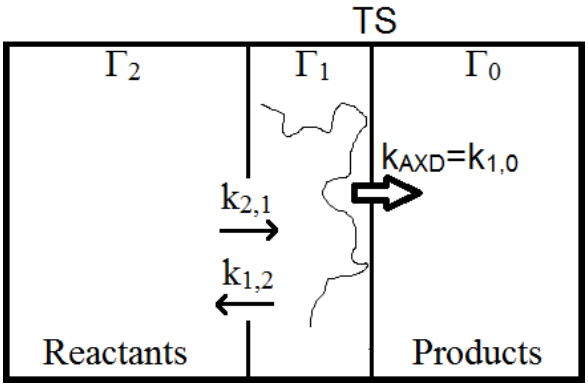

Figure 2.2: the AXD set-up. Reflective boundaries lock the trajectory in the region near the transition state. The accelerated rate constant k_{AXD} is quickly converged as the important region of phase space near the transition state is well sampled.

One boundary is placed on the transition state while another is

placed nearby on the reactants side, locking the trajectory in the area of phase space Γ1. According to TST the reaction rate will be the frequency at which a trajectory hits the boundary between Γ0 and Γ1. TST assumes that trajectories cross the boundary at the transition state, while in BXD and AXD the trajectory is reflected from the boundary. However these two cases are equivalent providing the boxes are in equilibrium. This is because a reflected trajectory is the same as a trajectory which leaves and is replaced by an incoming one on the reflected path. [43]

By locking the trajectory in Γ1 the region around the transition state is sampled more often leading to an accelerated rate constant k_{AXD}. This is related to the actual rate constant k_{TST} in the following way:

$$k_{TST} = k_{AXD} P_{CORR} \tag{2.1}$$

where k_{TST} is the actual rate constant for the reaction, k_{AXD} is the accelerated rate and P_{CORR} is a correction factor equal to the probability of finding the system within Γ1, calculated as

$$P_{corr} = \frac{\Gamma 1}{\Gamma 1 + \Gamma 2} \tag{2.2}$$

which is the fraction of reactant phase space enclosed in region Γ1. P_{corr} can be reearanged into a function of the rate constants for diffusion between Γ1 and Γ2:

$$\frac{\Gamma 1}{\Gamma 1 + \Gamma 2} = \frac{1}{1 + \frac{\Gamma 2}{\Gamma 1}} = \frac{1}{1 + \frac{k_{1,2}}{k_{2,1}}}. \tag{2.3}$$

Equation 2.2 can be derived from the principles of TST, as the TST rate constant for the reaction shown in figure 2.2 is defeined as

$$k_{TST} = \frac{\langle |\mu| \delta(q,p) \Theta(q,p) \rangle}{\Gamma_R} \tag{2.4}$$

where k_{TST} is the TST rate constant for diffusion from the reactant state to product space, $|\mu|$ is the magnitude of the velocity vector normal to the dividing surface in phase space, $\Theta(q,p)$ is a function of the position q and momenta p of the system which is unity when the system is in the reactant region R and zero otherwise, $\delta(q,p)$ is a Dirac delta function which is unity at the dividing surface and Γ_R is the phase space volume of region R. As the reactant space is divided by a reflecting boundary into $\Gamma 1$ and $\Gamma 2$ equation 2.4 can be rewritten as

$$\begin{aligned} k_{TST} &= \frac{\langle |\mu| \delta(q,p) \Theta(q,p) \rangle}{\Gamma_1 + \Gamma_2} \\ &= \frac{\langle |\mu| \delta(q,p) \Theta(q,p) \rangle}{\Gamma_1} \frac{\Gamma_1}{\Gamma_1 + \Gamma_2} \\ &= k_{AXD} k_{CORR} \end{aligned} \tag{2.5}$$

which gives equation 2.2, as the term $\frac{\langle |\mu| \delta(q,p) \Theta(q,p) \rangle}{\Gamma_1}$ is directly calculated from AXD by locking the trajectory in $\Gamma 1$. The rate constants k_{12} and k_{21} necessary to calculate P_{CORR} can be obtained by locking the trajectory in box $\Gamma 1 + \Gamma 2$. The speed-up gained from AXD is due to the fact that it is much faster to converge k_{AXD} and P_{CORR} separately than it is to converge k_{TST}. This is because an unconstrained trajectory would rarely visit the region near the transition state if there was a barrier between reactants and products.

2.2 Main Idea of BXD

BXD works in the same way as AXD but with multiple boundaries placed along the whole reaction coordinate. The trajectory starts in a box and after a certain number of collisions with the next boundary it is allowed through into the next box. This process is then repeated and the trajectory passes into the next box, and so on until the final box has been reached and the direction is reversed. The location of the boxes does not affect the result as TST is still valid if the boundary does not

lie on a transition state.[33]

Placing these boundaries along the reaction coordinate allows free energy barriers to be crossed quickly. This is due to the fact that, after the trajectory enters a new box, it cannot go back into the previous one. In this way the boxes act as a ratchet, preventing the trajectory from rolling back downhill. This is illustrated in figure 2.3. When the final box has been reached the direction of travel is reversed and the process continues until the entire reaction coordinate has been sampled multiple times in both directions.

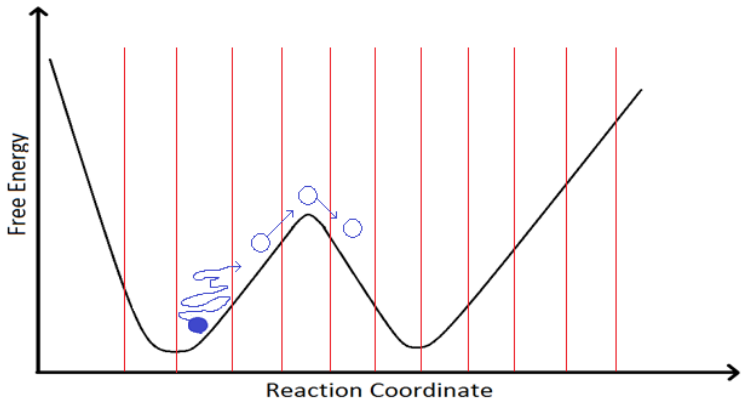

Figure 2.3: reflecting boundaries (red lines) along the reaction coordinate confine the trajectory (blue ball) into boxes allowing free energy barriers to be crossed. The boxes act like a ratchet and stop the trajectory rolling back downhill.

Figure 2.4 shows a plot of the reaction coordinate value against time for a typical BXD simulation, illustrating how the trajectory moves through the boxes.

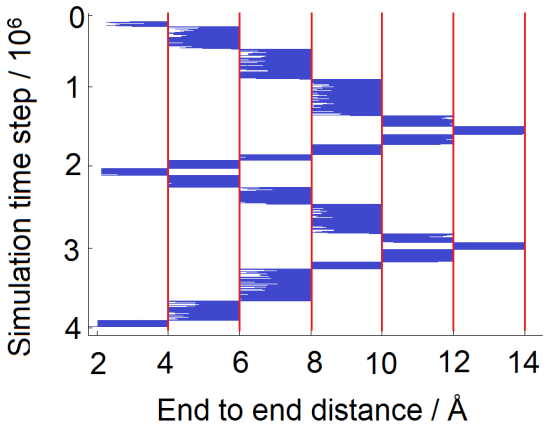

Figure 2.4: a plot of the reaction coordinate value against simulation time from a BXD simulation shows how the trajectory (blue) moves through the boxes and samples the phase space. The reflective boundaries are shown by red lines.

In this way the entire reaction coordinate is scanned until the sampling converges. Splitting the phase space into boxes not only allows free energy barriers to be crossed, it also makes it possible to calculate the rate constants and free energy along the reaction coordinate. This is shown in figure 2.5.

Increasing the resolution in this way reveals many fine details of the free enrgy landscape. These details are generally meaningful as the uncertainty in the free energy, resulting from the distribution of first passage times used to calculate the rate constant, is usually less than one percent. This figure is very low because on each boundary there are tpyically over 100000 first passage times. It should be noted that in this work the free energies presented are the result of averaging 10 to 20 individual free energies and the uncertainty is calculated as the standard deviation of the set of free energies used to calculate the average. This error is larger than the error coming from the distribution of first passage times, which suggests that independent trajectories extensivly sample slightly different pathways.

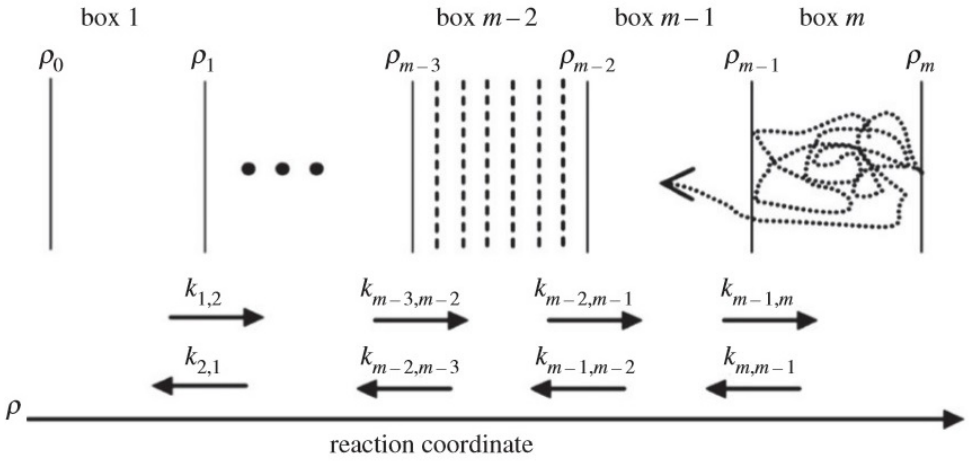

Figure 2.5: illustration of BXD showing boundaries $(\rho_m, \rho_{m-1}, ...\rho_0)$ placed along the reaction coordinate, dividing the phase space into boxes. The rate constants between each box and its neighbours are quickly calculated and are used to obtain the free energy along the reaction coordinate. Dotted lines show smaller bins within the boxes which increase the resolution of the free energy.

BXD is similar to umbrella sampling in that the trajectory is restricted in various windows placed in phase space, however the main difference is that BXD uses a square well restraint rather than a harmonic constraint, and BXD also provides simultaneous kinetics and thermodynamics without the need for compliacted unbiasing routines.

While the trajectory is in box m the time between successive collisions with the boundary ρ_{m-1} is recorded. These times are referred to as first passage times (FPTs). The set of FPTs for each boundary is used to calculate the rate constant for crossing the boundary:

$$k_{m,m-1} = \frac{1}{\langle \tau_{m,m-1} \rangle} \tag{2.6}$$

where $k_{m,m-1}$ is the rate constant for diffusion from box m into box $m-1$ and $\langle \tau_{m,m-1} \rangle$ is the mean of the first passage times (MFPT) on boundary ρ_{m-1}. Once the rate constant for diffusion in the opposite

direction is known then the equilibrium constant between the two boxes is calculated as

$$K_{m,m-1} = \frac{k_{m-1,m}}{k_{m,m-1}} \tag{2.7}$$

where $K_{m,m-1}$ is the equilibrium constant for diffusion between boxes m and $m-1$. The free energy difference between boxes m and $m-1$ is simply

$$\Delta G_{m,m-1} = -RT \ln K_{m,m-1}. \tag{2.8}$$

In this way it is possible to obtain the free energy along the whole reaction coordinate. It is also possible to calculate the probability of a trajectory populating any box m as a function of time. For a series of boxes the rate of change of the population of box m is

$$\frac{dn_m(t)}{dt} = k_{m-1,m} n_{m-1}(t) + k_{m,+1,m} n_{m+1} - n_m \left(k_{m,m-1} + k_{m,m+1}\right) \tag{2.9}$$

where $n_m(t)$ is the population in box m as a function of time. The right hand side of equation 2.9 is the flux entering the box minus the flux leaving. For a system of N boxes equation 2.9 can describe the population of every box in matrix form:

$$\frac{d\boldsymbol{n}(t)}{dt} = \boldsymbol{M}\boldsymbol{n}(t) \tag{2.10}$$

where $\boldsymbol{M}$ is an N by N matrix of rate constants and $\boldsymbol{n}(t)$ is a vector of length N of the populations of each box as a function of time. The solution to equation 2.10 can be expressed through the eigenvalues and eigenvectors of $\boldsymbol{M}$ as

$$\boldsymbol{n}(t) = \boldsymbol{U} e^{\boldsymbol{\lambda} t} \boldsymbol{U^{-1}} \boldsymbol{n}(0) \tag{2.11}$$

where $\boldsymbol{n}(0)$ is a vector containing the initial population of each box which is assigned from a Boltzmann distribution, $\boldsymbol{U}$ is the eigenvector

matrix obtained by diagonalising $\boldsymbol{M}$ and $\boldsymbol{\lambda}$ is the vector of corresponding eigenvalues. In general the eigenvalues are all negative and one will be separated from the others by several orders of magnitude. If an irreversible reaction is being studied then the flux out of the products box can be set to zero and the resulting lowest eigenvalue is taken to be the rate constant for the reaction.

BXD has much in common with other reactive flux sampling methods such as milestoning and forward flux sampling. All three methods place boundaries in phase space along a reaction coordinate, allowing the barriers to be crossed while calculating the free energy along the coordinate. However, BXD has the advantage of yielding kinetic and thermodynamic information simultaneously. BXD is also a very simple technique; a single trajectory can be left alone by the user to sample the reaction coordinate. In milestoning, multiple trajectories need to be initiated in different boxes which is more complicated and also requires prior knowledge of the conformational space. In forward flux sampling trajectories must be stored at various points and used to start a new batch of trial runs.

2.3 High Resolution BXD

The resolution of the free energy can be increased in the analysis of a BXD simulation. There are two methods of doing this: a WHAM approach or by box splitting.

2.3.1 WHAM for BXD

In our adaptation of the Weighted Histogram Analysis Method (WHAM) the phase space within each box is divided up into smaller bins, shown as dashed lines within each box in figure 2.5. Note that these bins are for analysis purposes only and are not present in the simulation. Once

the free energy has been calculated along the reaction coordinate the probability of the trajectory being in box m is given as

$$P(m) = \frac{1}{\sum_n e^{\frac{-\Delta G_n}{k_B T}}} e^{\frac{-\Delta G_m}{k_B T}} \tag{2.12}$$

where $-\Delta G_n$ is the free energy of box n and the summation is the total free energy of all the other boxes. The box to box free energies are normalised so that

$$\sum_n P(n) = 1. \tag{2.13}$$

Given that the trajectory is in box m, the probability $p(j)$ of being in the j^{th} bin within box m is the amount of time spent in bin j divided by the total time spent in box m. This probability is then multiplied by the probability of being in box m to give the normalised probability of finding the trajectory in bin j compared to the rest of the entire reaction coordinate:

$$P(j) = P(m)\, p(j) \tag{2.14}$$

In this way, the resolution of the free energy can be increased to an arbitrary level. This procedure is more simple and stable than the WHAM algorithm used to recover Boltzmann free energies from umbrella sampling.

2.3.2 Box Splitting

Box splitting is the other method of increasing the resolution of the free energy. In the analysis stage each box is split into smaller parts, shown in figure 2.6.

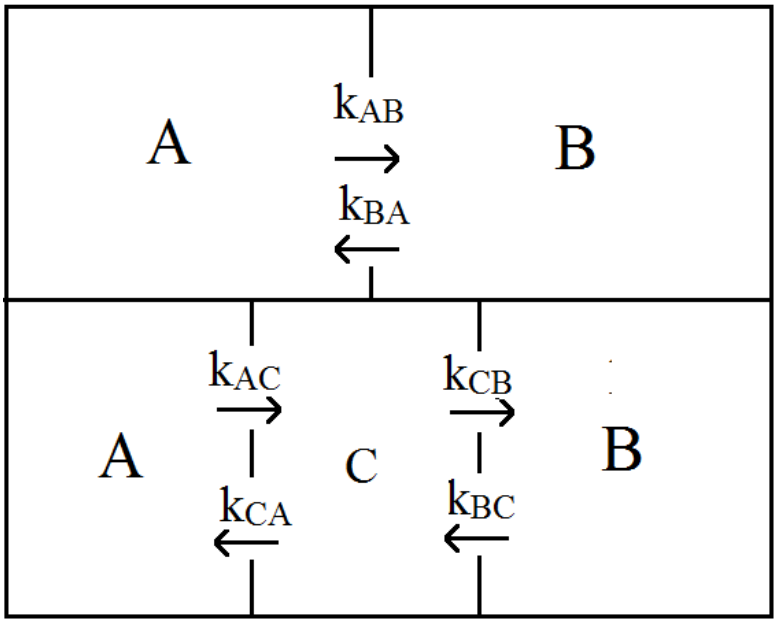

Figure 2.6: the resolution of the free energy can be increased by adding extra boxes in the analysis stage.

Note that the smaller box is only for analysis purposes and is not present in the simulation. The rate constant k_{AB} is calculated directly from the BXD simulation using equation 2.6. The lower half of figure 2.6 is similar to the set-up shown in figure 2.2 hence equation 2.1 applies, as k_{AB} can be written as the product of an accelerated rate constant and a correction factor:

$$k_{AB} = k_{CB} P_{CORR} \tag{2.15}$$

Because of equations 2.2 and 2.3 it is possible to rewrite equation 2.15 as

$$k_{AB} = k_{CB} \frac{1}{1 + K_{CA}} = k_{CB} \frac{1}{1 + e^{\frac{-\Delta G_{CA}}{k_B T}}} \tag{2.16}$$

where K_{CA} is the equilibrium constant between boxes C and A (equation 2.7) and $\frac{\Delta G_{CA}}{RT}$ is the free energy difference between C and A, calculated

as

$$\Delta G_{CA} = -k_B \ln \frac{P(C)}{P(A)} \tag{2.17}$$

where $\frac{P(C)}{P(A)}$ is the ratio of the probabilities of finding the trajectory in box C and box A. This ratio is equal to the ratio of the time spent in each box which is obtained from analysing the plot of reaction coordinate against time. By rearranging equation 2.16 and repeating for the rate constants between box C and B it is possible to double the resolution of both the free energy and the rate constants along the reaction coordinate. Figure 2.7 shows the effect of increasing the resolution of the free energy using the BXD implementation of WHAM.

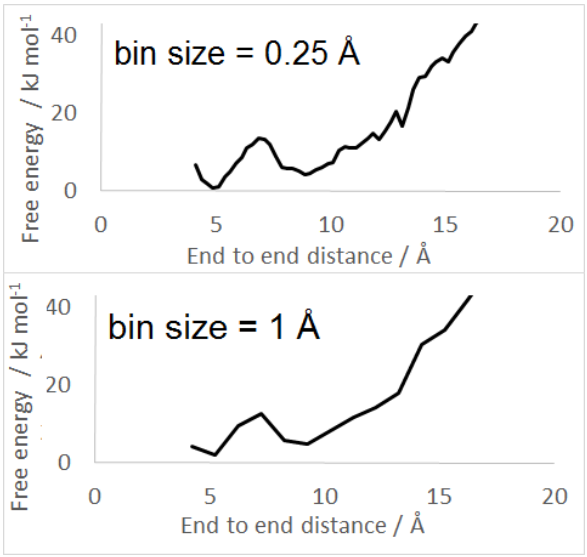

Figure 2.7: increasing the resolution of the free energy along end to end distance for peptide P1 (see Chapter 4). The resolution was increased from 1 Å to 0.25 Å using our WHAM approach.

Now that BXD has been introduced the assumptions on which it relies will be examined along with the conditions under which they are valid.

2.4 Decorrelation and Ergodicity

BXD relies on the assumption that the motion within a box is stochastic and that sequential hits and velocity inversions are uncorrelated, i.e. the time between hits must be longer than the correlation time. In general this is not the case; a trajectory reflected from a boundary can sometimes turn back quickly. These short-time-correlated events need to be accounted for. This is done by removing the FPTs which correspond to these fast events, a process known as decorrelation.[43] Figure 2.8 shows a set of FPTs from a boundary in a BXD simulation.

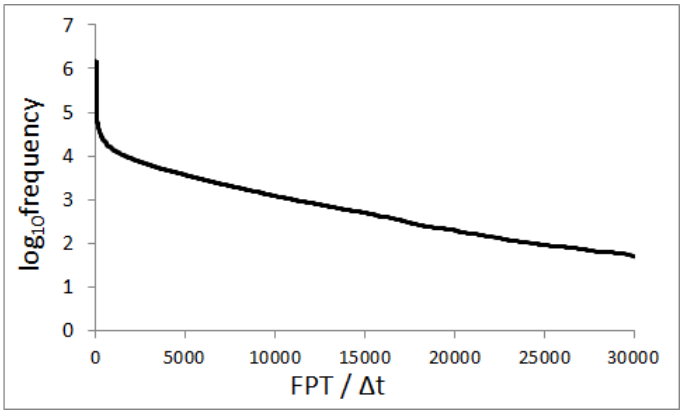

Figure 2.8: a typical distribution of first passage times for a boundary between boxes in a BXD simulation. The initial very steep section corresponds to fast correlated collisions which are removed from the set of FPTs used to calculate the rate constant. This ensures that the assumption of stochastic motion within each box is kept valid.

The log frequency distribution of FPTs shown in figure 2.8 is split into two parts. The initial steep section comes from fast correlated collisions against the boundary while the flatter section comes from ergodic collisions where the trajectory has had time to come to equilibrium in between hits. To decorrelate the statistics the FPTs from the steep section are removed. Then the rate constant in equation 2.6 will only be calculated with FPTs which are greater than the correlation time of the system τ_{cor}.

In practice this is done by defining a cutoff value τ_{cor} below which FPTs are ignored. The free energy is then calculated for different values of τ_{cor} which is increased until the free energy no longer changes. At this point the statistics are decorrelated and the value of τ_{cor} at which this happens is the characteristic correlation time for that system. An example of decorrelation from an actual BXD simulation of a cyclic peptide precursor S7 (see chapter 4) is shown in figure 2.9.

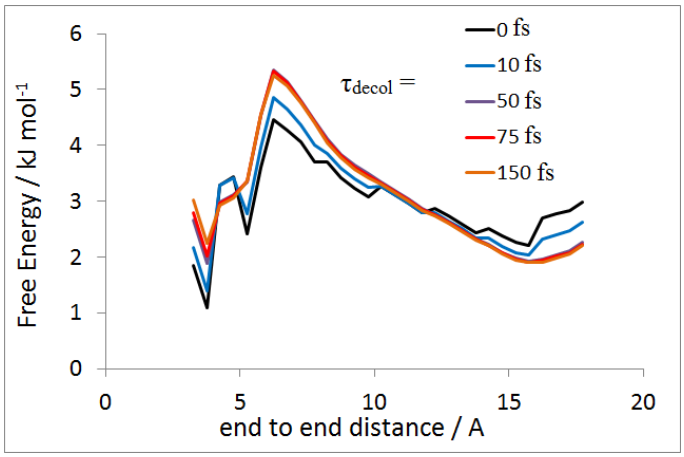

Figure 2.9: to remove FTPs corresponding to fast correlated motion a cutoff value τ_{cor} is defined. This is varied until the free energy no longer changes. In this example the free energy no longer changes when τ_{cor} is increased above 50 fs hence this was taken as the decorrelation time for this simulation of cyclic peptide precursor S7 (See Chapter 4).

Once the statistics have been decorrelated equation 2.1 is valid as dynamic recrossing no longer contribute to the MFPT.

The other condition that must be met is that the boxes must be in equilibrium with each other. This is because equation 2.2 assumes that the system is ergodic i.e. every point in phase space has an equal probability of being visited by a trajectory. This is possible only if the trajectory has time to forget its previous state after a collision. If the box is too small then this will not be possible as the time in between collisions with a boundary will be less than τ_{cor} so the trajectory will

never relax.

It is possible to check that the boxes are large enough by inspecting the plot of the reaction coordinate value against time. One such plot is shown in figure 2.10 for a box that is too small and a box that is sufficiently large. If the box is too small then the trajectory can be seen to hit a boundary very frequently whereas if the box is large enough then the trajectory has plenty of time to explore the box in between collisions with the boundary.

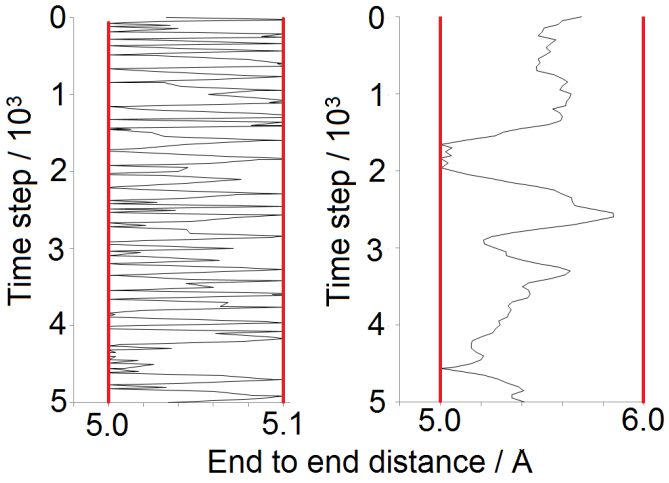

Figure 2.10: if a box is too small (left) then the trajectory hits the boundaries (red) too frequently and cannot equilibrate within the box. Decorrelation here is impossible as there are no FPTs longer than the characteristic decorrelation time. If the box is large enough (right) then decorrelation is possible as the trajectory can explore the box and come to equilibrium in between collisions with the boundary.

Another indication of the boxes being too small is if there is no value of τ_{decor} for which the free energy converges, as by definition if a box were smaller than the correlation length then decorrelation would not be possible.

For BXD to be accurate the rate constants must be converged. To check for convergence the data from a BXD simulation is split in half.

The free energy is calculated from half of the data as well as from the full dataset. If the two free energies are the same then the simulation has converged. See Appendix A for a more detailed description of how BXD is used and a worked example of obtaining a free energy profile.

BXD has recently been used to get converged rates and free energies for processes up to a timescale of seconds[44] including desorption of ions from monolayers, diamond etching and peptide cyclization. Now that BXD has been introduced along with an assessment of the conditions under which it is valid, the next chapters will introduce some recent applications and developments.

The aim of the first application reported here was to investigate whether BXD could be used to succesfully model AFM protein pulling: a very long time scale process with a parge protein molecule. By benchmarking against readily available experimental data it was possible to prove that BXD was capable of accuratly simulating a process which traditionally is very computationally demanding, and to push the limits of what could be achieved by using an enhanced sampling algorithm on modest hardware. The second application was undertaken to show that BXD has practical uses and can be used to help solve an urgent problem in medicinal science: the shortage of antibiotics.

Chapter 3

Atomic Force Microscopy Protein Pulling

3.1 Introduction

Proteins are an important component of all bimolecular machines where their mechanical properties are often crucial. Titin for example is a molecular spring which plays a part in human muscle function; defects on the part of titin have been linked to heart and lung failure.[45;46] Mechanical resistance is also important for many other proteins such as those that bind cells together in living tissue.[47] The mechanical properties of proteins have been studied by Atomic Force Microscopy (AFM), where an AFM probe is used to pull and unfold single protein domains or chains of domains. The aim of these experiments is to shed some light on protein folding and unfolding pathways and to discover why certain proteins are more mechanically robust than others.[45]

The first AFM studies of the mechanical properties of proteins was done by Rief, Gaub and Fernandez *et. al.* in 1995.[48;49] These early studies used an AFM probe to measure the interaction force between two strands of DNA. One strand of DNA was attached to a solid bead while another was attached to the AFM tip. The two strands were brought together for a certain amount of time and then separated, with the force

needed to separate the strands being measured by the AFM probe. The authors found that the force needed to separate the two strands varied depended on how long the two strands were left in contact before being pulled apart.

These studies were important because they took advantage of the then recent discovery that functional groups could be attached to AFM tips and used to pull apart large molecules, opening up a whole new field of experimental biochemistry. The dependence of the force of interaction between the DNA strands on the initial contact time suggested that some dynamic process such as conformational change was responsible for the interactions which were being probed.

This is highly relevent to computational studies because techniques such as MD can in principle be used to directly observe the relationship between the dynamics of a biomolecule and its resistance to force.

AFM is a technique whereby a sharp nano scale tip is attached to a thin cantilever. The position of the cantilever can be measured via laser reflection to a very high level of accuracy down to the nanometre scale. Once in contact with a material the deflection of the cantilever is known. Knowing this displacement as well as the force constant of the cantilever allows the force acting upon the tip to be calculated. In AFM protein pulling one end of a chain of protein domains, or concatemer, is attached to the cantilever tip while the other is anchored to a solid surface. The tip is then retracted causing the domains to unfold. This is shown below in figure 3.1.

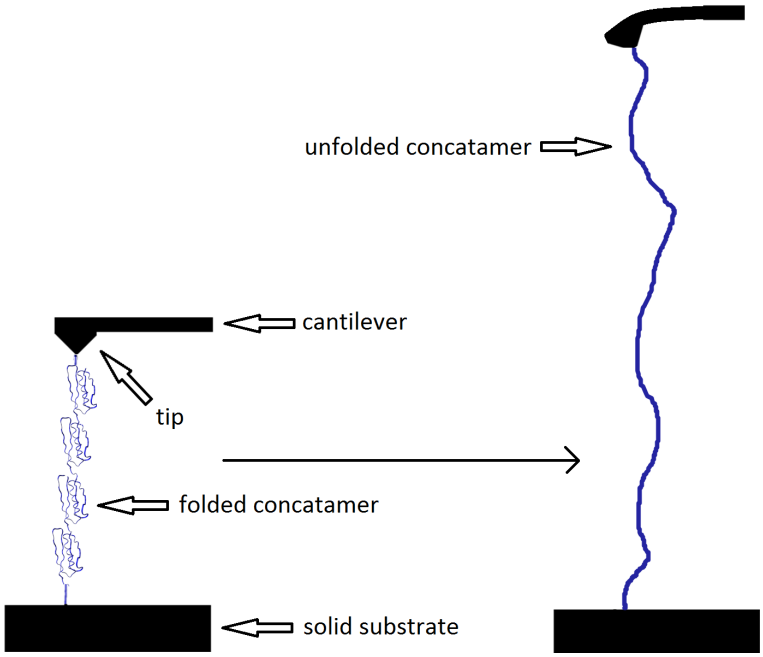

Figure 3.1: the AFM protein pulling experiment. An AFM tip pulls apart a chain of protein domains.

Two methods of AFM pulling are common: Force Clamp (FC) where the tip is retracted at a constant force, and Velocity Clamp (VC) where the tip is retracted at constant velocity. FC experiments[50–54] produce a plot of extension versus applied force or extension vs time at a range of forces, and VC experiments[55–59] record the force exerted by the cantilever against protein extension. Typical experimental traces from VC and FC methods are shown below in figure 3.2.

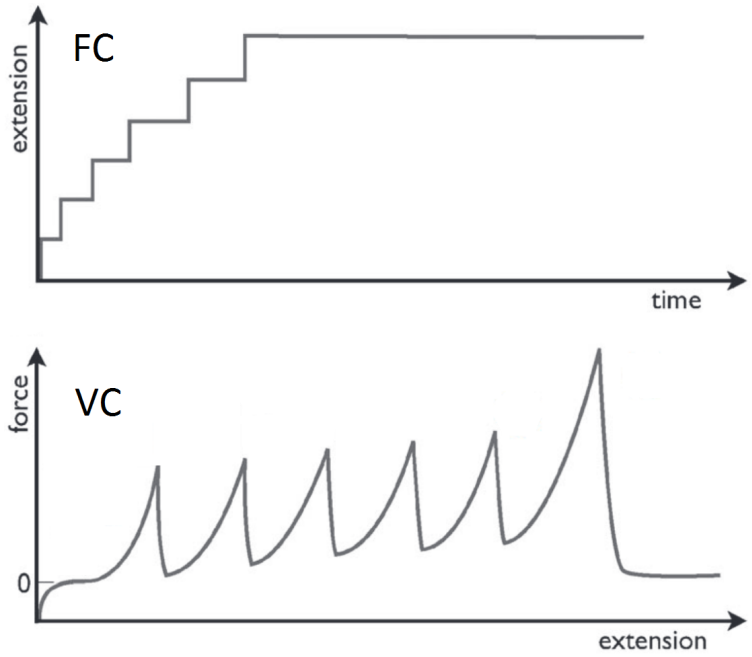

Figure 3.2: typical experimental traces from AFM pulling experiments. FC mode (top) keeps the force constant and records extension against time, with each step corresponging to a single domain unfolding. VC mode (bottom) extends the protein at a constant speed and records the force exerted against extension, with each peak in the trace corresponding to the unfolding of a domain. The increasing height of the peaks is due to the domains unfolding at higher forces due to the reduced elasticity in the chain as domains become straightened out. Reproduced from Ref.[60] with permission from The Royal Society of Chemistry.

Comparison of these experiments with MD simulations should be possible and will reveal otherwise inaccessible information about protein dynamics, which could lead to insights into the factors affecting the mechanical stability of proteins as well as the mechanism of their mechanical unfolding.

However the time scale of a typical AFM pulling experiment is much longer than anything that can be simulated by MD. VC experiments are performed on a timescale of microseconds to milliseconds while FC unfolding can take as long as seconds. The free energy of

AFM protein extension is very high meaning that it is a rare event not often sampled by conventional MD.

In addition to the standard long timescale problem the dynamics of AFM protein pulling varies with the pulling speed. High speed pulling experiments tend to involve dynamical pathways whereas low speed experiments involve a more stochastic process where more pathways and conformations are explored. Despite attempts to increase the timescale of simulation or reduce the timescale of pulling, simulation and experiment have not yet met in the middle.

In this chapter BXD is applied to the AFM protein pulling experiment to replicate both VC and FC modes.

Experimental data are reproduced and insights into the mechanical unfolding process are presented. AFM unfolding will be simulated for three protein domains: Protien L which binds antibodies, I27 which forms part of the muscle protein Titin and IM9 which plays a role in the immune system. The structures of these domains are shown below in figure 3.3.

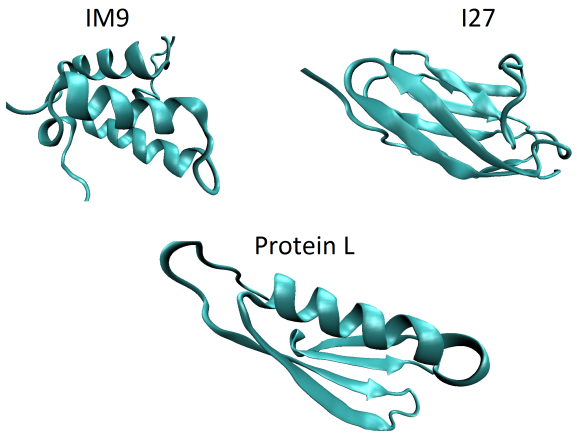

Figure 3.3: the structures of the protein domains that feature in this study. Structures are taken from the RCSB Protein Databank: 1IMQ (IM9), 1TIT (I27) and 1K53 (Protein L).

3.1.1 Existing Methods of Simulation

Most simulations of the AFM experiments involve the application of an artificial force to the system; a technique known as Steered Molecular Dynamics (SMD).[61] A virtual harmonic spring is attached to each end of the protein. Moving the springs apart at constant velocity applies a force which pulls the ends of the protein apart in a similar way to how the AFM experiment operates. The extension of the spring is recorded as the protein unfolds and this provides a plot of force versus extension which mimics the VC experiments. SMD can also mimic FC experiments by applying a constant force along the vector between the ends of the protein, pulling them apart with a constant force rather than at a constant speed.

SMD has been used to investigate the mechanical unfolding of a number of protein domains. The I27 domain of titin is an all beta sheet domain which is well studied by experiment and simulation. VC experiments have found that I27 has a high mechanical strength and consequently unfolds at a high pulling force of around 150 to 200 pN

at a range of pulling speeds between 1 and 5000 nm per second.[55;62] As the force builds up I27 extends by a very small amount until the peak unfolding force is reached, after which a rapid collapse of the structure leads to large extensions and a decrease of the force.[57;62] Theoretical studies[63–67] have used SMD to pull I27 and have found that the mechanical behaviour comes from the backbone hydrogen bonds between the terminal beta sheets.

Early SMD work by Schulten, Rief and Fernandez *et. al.*[55;64–67] reported the sequence of events undergone by I27 when a force is applied in simulation. The initial 10 Å of extention does not result in any structural change. After the extension increases to around around 14 Å the A' and G beta strands slide past each other as 6 hydrogen bonds between them rupture. Figure 3.4 shows these strands in I27 and the hydrogen bonds between them. After these strands separate the remaining strands fail one by one at lower pulling forces until the molecule is linear.

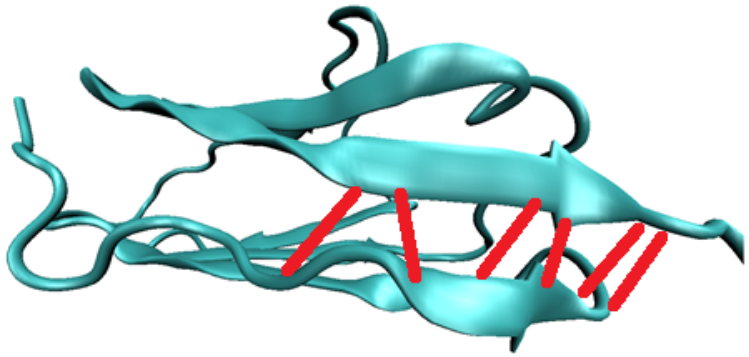

Figure 3.4: the structure of I27 and the hydrogen bonds between the A' and G strands which are responsible for the mechanical strength of the domain.

Protein L is a mixed alpha + beta domain which shows similar mechanical strength and unfolding behaviour to I27.[68] SMD was used to investigate the mechanical unfolding of protein L[61] which was found to

be similar to I27 in that a cluster of backbone hydrogen bonds between beta strands withstand a high force and then suddenly fail, leading to complete unfolding. SMD suggests[61] that when the force is initially applied the N terminal beta strand reorientates to align with the applied force. This minor conformational change yields a very small extension yet is responsible for the peak unfolding force. This rearrangement disrupts the hydrogen bonds between the N and C terminal beta strands, leading to a rupture of the contacts between them. The structure of Protein L and the hydrogen bonds responsible for the initial force resistance are shown below in figure 3.5. The main difference between Protein L and I27 was found to be that I27 populates a number of intermediate states along the unfolding pathway while Protein L unfolds via a simple two state system.[61]

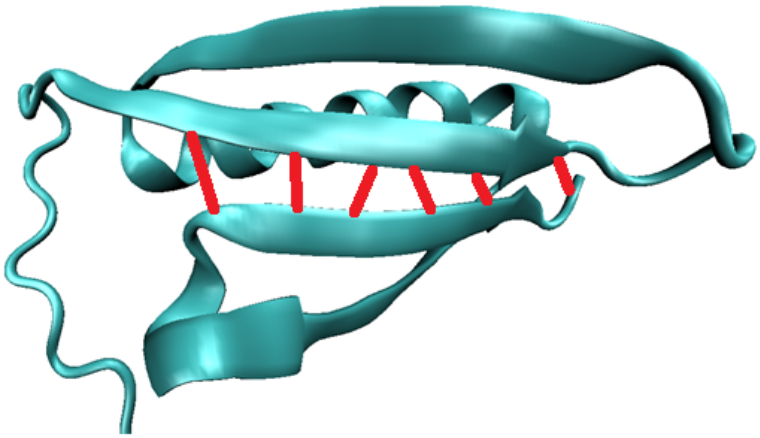

Figure 3.5: the structure of Protein L and the hydrogen bonds responsible for the initial resistance of the domain according to SMD studies[61].

All alpha domains such as IM9 unfold at much lower forces than all beta, alpha + beta or alpha/beta domains.[57;61] This is thought to be due to the absence of a cluster of backbone hydrogen bonds which are loaded all at once, and share the force as is found in beta sheets. Instead the backbone hydrogen bonds are loaded sequentially and fail one at a time, leading to a gradual unravelling at low force rather than the initial resistance followed by sudden failure displayed by I27 and Protein L.[61]

This view is confirmed by computational studies where SMD was used to unfold a number of all alpha domains.[69–71] In summary SMD has been used to show that the secondary structure of a protein or domain is a major factor in determining the resistance to pulling.[55] Another major factor is the direction in which the force is applied, with proteins being much stronger if their topography results in the force being applied along the long axis of betra strands which are hydrogen bonded to each other, due to the fact that the lead is then shared between multiple hydrogen bonds.[57;61], such as with I27 and the A'-G mechanical clamp (see figure 3.4).

SMD applies a pulling speed or force that is several orders of magnitude greater than that of the corresponding experiment. The effect of this on the validity of the simulations is under debate. Experiment has shown that for I27 the excessive pulling speeds and forces used have not negatively impacted the results of the simulations, however it is possible that it is a problem for systems that do not share the same mechanical characteristics as I27.[71] BXD differs in that no force or pulling speed is applied, instead the system diffuses along the reaction coordinate of end to end distance with no modification of the potential.

3.2 Method

Calculations were performed using the BXD subroutine implemented in CHARMM.[72] After every time step the BXD code recieves the atomic positions and velocities from the CHARMM integrator which are used to update the value of the reaction coordinate. If the value of the reaction coodinate has change such that a boundary has been crossed then the velocity of each atom is inverted with respect to the reaction coordinate (see figure 2.1), and the new inverted velocities are passed back to the integrator. The inversion is performed in the centre of mass frame

to converve linear and angular momentum.

The reaction coordinate was chosen as the distance between the two termini of the protein domain as this would correspond to the coordinate sampled by the AFM pulling experiments. This end to end distance coordinate was later translated to extension by subtracting the equilibrium value. The EEF1 (Effective Energy Function 1) implicit solvent model[73] and CHARMM 19 force field were used for the simulations along with the Langevin thermostat set to 303 K. To begin with the PDB structures for the three domains were equilibrated under the forcefield and solvent model for 500 ns before BXD simulations commenced.

For IM9 and Protein L boxes were placed at intervals of 0.75 Å from 11.25 Å to 320.25 Å, and from 21 Å to 320.25 Å respectively. For I27 the boxes were at intervals of 0.5 Å from 20 Å to 330 Å. Between 5000 and 2000 inversion events were required in each box before a boundary could be passed. Initially the reaction coordinate was sampled downwards from zero extension to the lowest boundary in order to fully explore the local minimum before the direction was reversed and the protein domain began to extend. Sampling was continued until a linear conformation had been achieved.

Simulations were also carried out with boxes at intervals of 0.5 Å, 0.25 Å and 0.125 Å in order to ascertain that the result was independent of box size and placing. The free energy along end to end distance for protein L, calculated at different box sizes, is shown in figure 3.6.

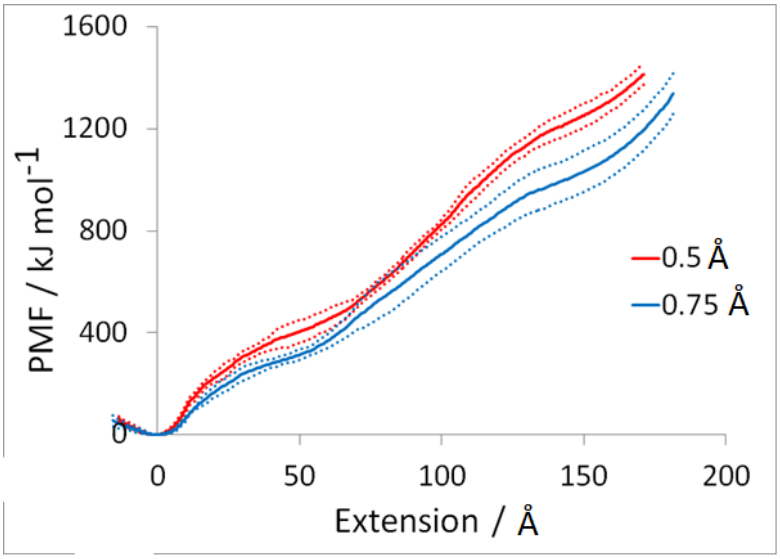

Figure 3.6: free energy along end to end distance for Protein L with box sizes of 0.5 Å (red) and 0.75 Å (blue). Solid lines represent the average free energy of 10 individual trajectories. Dotted lines show one standard deviation. The uncertainty resulting in the distribution of FPTs used to calculate the rate constants (see Chapter 2) is very small as the trajectory remains in each box for a long time. The more significant uncertainty shown here result from each trajectory sampling a slightly different pathway. The uncertainty at small extensions is very small as the unfolding pathways are all the same initially: The force builds up until the hydrogen bonds between the terminal beta strands rupture. At higher extensions the uncertainty increases as a wider range of pathways can be accessed as units of secondary structure break down.

It is clear from figure 3.6 that the free energy does change slightly with box size. However for Protein L, 0.5 Å was found to be the limit of how small the boxes could be before it became impossible to decorrelate the trajectories. Presumably this is because the correlation length for Protein L is slightly below 0.5 Å. With 0.75 Å boxes the decorrelation procedure worked well while with larger boxes the simulation took too long. Because of this it was decided to use a box size of 0.75 Å. For I27 the trajectories were well decorrelated with a box size of 0.5 Å while larger boxes did not allow the simulation to proceed, hence a box size of 0.5 Å was chosen. For each protein domain 30 unfolding BXD runs were

obtained. The total simulation time for each domain was of the order of tens of nanoseconds. The decorrelation time used in the procedure, which removes correlated short time events, was found to be 600 fs for each system. This value was obtained by calculating the free energy at a number of different decorrelation times and taking the smallest value for which it had converged (see figure 3.7).

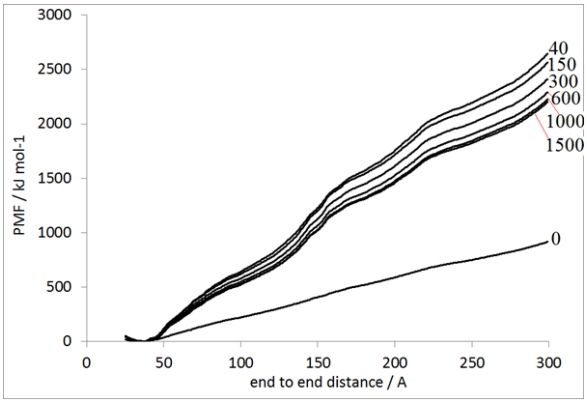

Figure 3.7: decorrelating the free energy for I27. Different decorrelation times, shown in femptoseconds by the number on the end of each line, were used to calculate the free energy. On going from 600 to 1000 fs the free energy no longer changed so 600 fs was taken as the correlation time for I27.

The set of box-to-box rate constants yields both the free energy and all the information necessary for a description of the kinetics of unfolding along the pulling reaction coordinate.

3.2.1 Treatment of Solvent Effects

Including water molecules in the unfolding simulations would be prohibitively expensive due to the very large box that would be needed to cover an extended protein. Therefore the solvent effects are treated implicitly with the EEF1 solvent model[73] which modifies the potential energy of the system to mimic the effects of water. The solvation free

energy of a molecule is modelled as

$$\Delta G_{slv}(total) = \sum_{i=1}^{N} \Delta G_i^{ref} - \sum_{i=1}^{N} \sum_{j \neq i}^{N} f_i(r_{ij}) V_j \tag{3.1}$$

where N is the number of solute atoms, ΔG_i^{ref} is the solvation free energy of atom i, and V_j is the volume of atom j. The function$f(r_{ij})$ is the solvation free energy density function

$$f(r_{ij}) = \frac{\alpha_j}{4\pi r_{ij}^2} e^{-\left(\frac{r_{ij}-R_j}{\lambda_j}\right)^2} \tag{3.2}$$

where r_{ij} is the distance between atoms i and j, R_j is the Van Der Waals radius of atom j, λ_j is the correlation length of atom j (3.5 Å for most atoms) and α_j is a coefficient of proportionality given by

$$\alpha_j = \frac{2\Delta G_j^{free}}{\lambda_j \sqrt{\pi}} \tag{3.3}$$

where G_j^{free} is the solvation free energy of the isolated atom j, which is close to ΔG_j^{ref}. The main point of equation 3.1 is that for each atom, the solvation free energy is taken to be that of itself fully immersed in water, minus a term to account for the atom being partially exposed. The solvation energy from equation 3.1 is added to the potential energy of the solute derived from the force field:

$$W_{EEF1} = U(r) + \Delta G_{slv} \tag{3.4}$$

where W_{EEF1} is the new effective energy under the EEF1 model, $U(r)$ is the potential energy from the force field and G_{slv} is the solvation free energy from equation 3.1. While the EEF1 model takes the solvent into account at the molecular scale, Langevin Dynamics is used to replicate the bulk properties of water. This is done by modifying the Newtonian equations of motion to take into account friction and random collisions with water molecules, which is added onto the force derived from the

gradient of $U(r)$, the potential energy derived from the force field. The Langevin equation is defined as

$$F_i(t) = m_i a + \gamma_i v_i m_i + \sqrt{2\gamma k_B T m_i} R_i(t) \tag{3.5}$$

where $F_i(t)$ is the force acting on atom i as a function of time, γ_i is the friction coefficient between water and atom i, k_B is the Boltzmann constant, T is the temperature of the simulation and $R_i(t)$ is a function introducing random collisions between atom i and water. Langevin Dynamics allows the temperature of the simulation to be controlled as T affects the amplitude of the thermal jostling experienced by the solute atoms. The friction coefficient γ_i allows the viscosity of bulk water to be replicated.

Mechanical unfolding has for I27 at least, been shown to be partly mediated by water as after the initial rupture of the A'-G hydrogen bonds (see figure 3.4) the hydrophobic core of the domain is destabilised by the water which acclerates the collapse of the protein structure.[55;64] It is also possible that the hydrogen bonds which are broken in the A'-G rupture are re-formed with water which would lower the barrier to unfolding. However the relaxation time of water is typically between 1 and 2 ns[74] meaning that, for a simulation using explicit water, unless a conformation persists for at least this time then the arrangment of the water molecules around it will not be realistic and the conformation will not be properly stabilised. Because accelerated sampling methods such as BXD and SMD often involve rapid conformational change, with SMD unfolding simulations often completly unfolding the protein in around a single nanosecond[63–67], it may be that implicit solvent models are the more appropriate choice.[61]

3.3 Results and Discussion

3.3.1 BXD Calculation of Unfolding Free Energies and Replication of VC Experiments

Free energies as a function of the end-to end distance and their gradients are shown in figures 3.8, 3.9 and 3.10 for I27, protein L and IM9 respectively. Points A to E indicate particularly important regions of the free energy, corresponding to minima and maxima of the force. The snapshots of the structures at those regions are also shown. The dashed line and dotted line represent one standard deviation from the calculated average. For the free energy profiles this was calculated by taking the standard deviation of the ensemble of free energies from each individual unfolding trajectory. Each free energy profile was converted into force by differentiation and then the average force and standard deviation were taken from the ensemble of individual force versus extension plots.

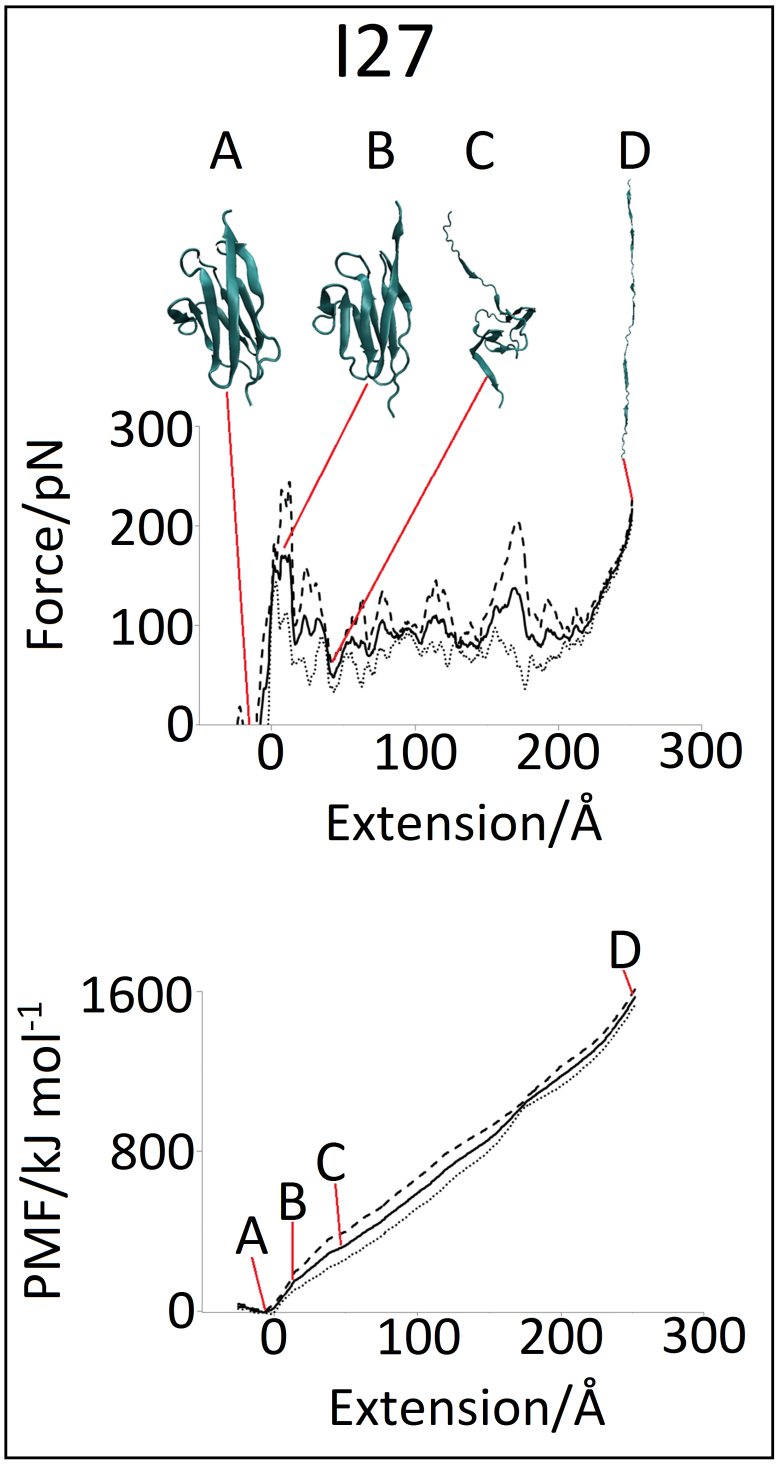

Figure 3.8: free energy and force along the pulling coordinate for I27.

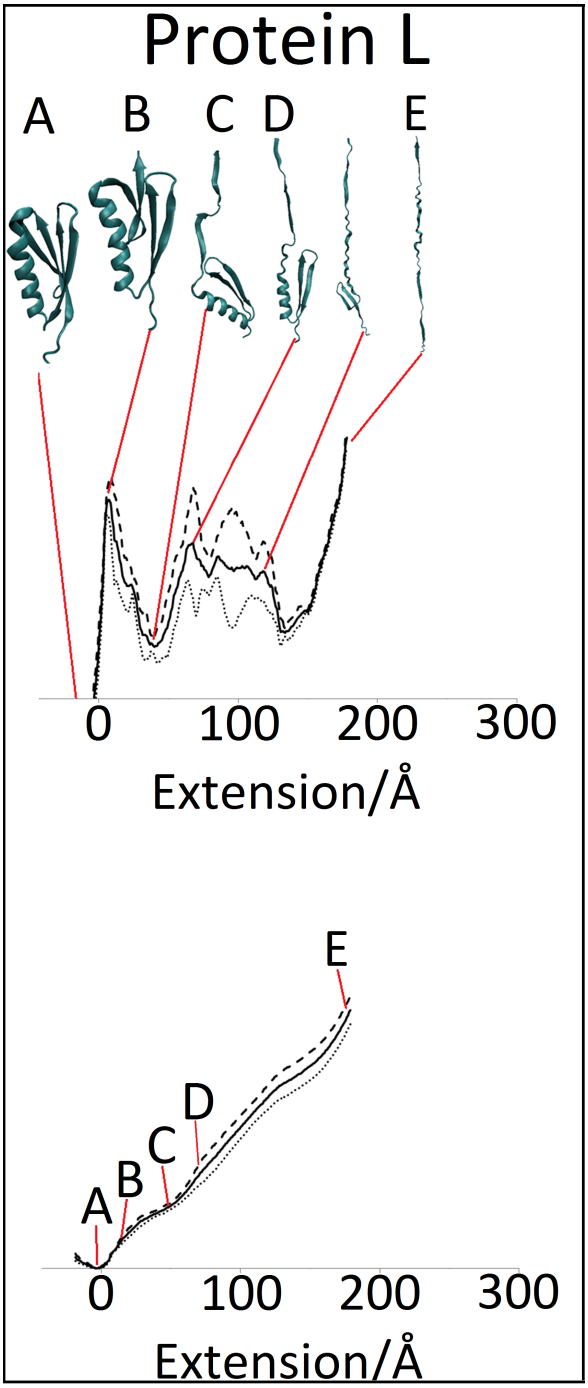

Figure 3.9: free energy and force along the pulling coordinate for Protein L.

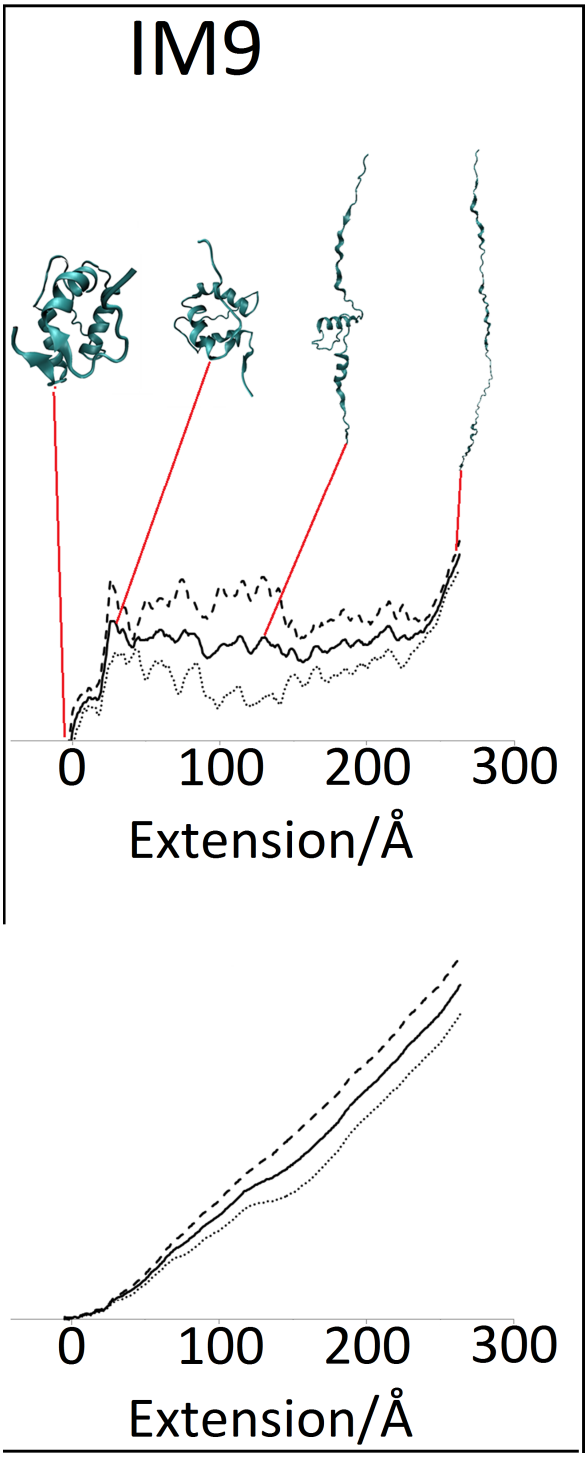

Figure 3.10: free energy and force along the pulling coordinate for IM9.

The main insights from the above figures are that the mechanical strength of a protein depends on the secondary structure of the region to which the force is initially applied, which can be determined by observing which region is the first to fail. For I27 the force increases rapidly, to a high value with little extension until hydrogen bonds bentween the A' and G beta strands (see figure 3.4) rupture, in agreement with SMD studies[63–67]. For IM9 the force is loaded onto alpha helix; the hydrogen bonds between them fail sequentially leading to gradual unfolding of the whole domain at low forces. Protein L has a mixture of alpha helixes and beta sheets and and a force-extension profile which is similar to that of I27. The initial force is loaded onto a series of hydrogen bonds between beta strands rather than onto an alhpa helix, supporting the suggestion that the resistance depends on the local structure of the region which is first to experience the force. The partially unfolded structures of Protein L and I27 show that beta sheets fail one at a time after the initial rupture which leads to more peaks at lower forces.

The structures along the unfolding pathway (figures 3.8 and 3.9) show that the initial rupture of I27 and Protein L exposes the hydrophobic core to the water which may result in issues with the implicit solvent model, as the hydrophobic effect and water stabilisation of ruptures hydrogen bonds will not be fully reproduced, however because of the fast time scale at which these conformations change in the simulation an implicit solvent model may still be better.[61]

It should be noted that mechanical unfolding is a completely different process to thermal unfolding *in vivo*.[75] While thermal unfolding free energies feature small barriers separating stable states, the AFM probe pulls the protein apart, forming an unnatural linear conformation with a very high free energy. This is supported by the fact that there is no correlation between the mechanical and thermal stability of a protein.[76] The difference between the two kinds of unfolding is illustrated in figure 3.11.

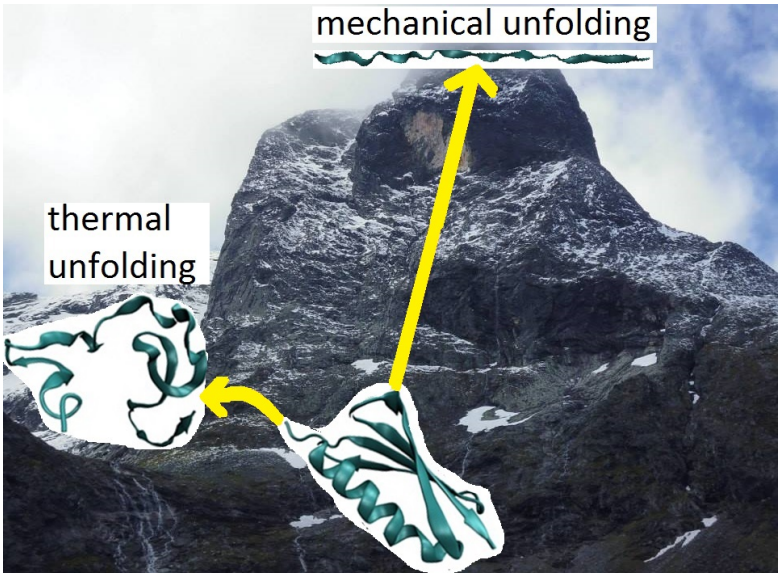

Figure 3.11: *in vivo* unfolding features stable states separated by small energy barriers. Unfolded states retain much of their secondary structure. With mechanical unfolding the AFM probe pulls the protein apart to higher and higher free energies along an unnatural pathway.

As the free energy represents thermodynamical free energy along the reaction coordinate x, by definition the change of the free energy is equivalent to mechanical work required for displacement along the reaction coordinate (protein extension in this case), and mechanical force is given by the gradient of the free energy,

$$F = \frac{dW}{dx} \approx \frac{dG}{dx} \tag{3.6}$$

where W is the work required, provided that the extension is slow enough for equilibrium thermodynamics to be valid. In the limit of higher speed the kinetics of protein pulling should be considered in more detail. As the kinetics of mechanical unfolding changes with pulling speeds a number of models have been developed to describe this dependence.[77–79] In principle BXD theory can be applied to a wide range of pulling speeds and the relationship of the pulling force and pulling speed can be obtained, but this study focuses on the limit of slow speed

where simple thermodynamical argument leads to equation 3.6.

Some features of the free energies are similar for all three protein domains. At low extensions near equilibrium there is a well, which is steep for the beta containing domains I27 and L, and relatively smooth and broad for the all alpha IM9 domain. This difference is due to the fact that, for I27 and protein L, the initial force is loaded onto beta sheets which hold out to high forces before suddenly rupturing, whereas with IM9 the initial force is loaded onto alpha helices which rupture gradually at low forces.[45;57;62] This well is situated around the global minimum along the extension coordinate and corresponds roughly to the equilibrium native structure of the domain under the conditions of the simulation.

The force curves generated by BXD have a similar shape to those produced by SMD, however the forces reported by SMD are at least an order of magnitude higher than those from BXD, and also from any experimental force curves. This discrepency in the SMD force curves is due to the artificial forces used in the simulation which are several orders of magnitude greater than those used in the AFM experiments.

For all three proteins the equilibrium well is followed by an inflection point before a region of less steep free energy is reached. This inflection point corresponds to breaking the native structure once a peak unfolding force is reached, corresonding to a transiton from point A to point B in figures 3.8 to 3.10. I27 and Protein L have a deeper equilibrium well and a more pronounced inflection point than that of the alpha domain IM9. This is due to the fact that beta sheets offer more mechanical resitance than alpha helixes.[45] On going from the bottom of the well to the inflection point does not result in significant change in the equilibrium structure, before a sudden rupture between beta sheets leads to a reduction in the force. This is due to the brittle nature of beta-sheets; initially the force is shared between multiple hydrogen

bonds which fail suddenly, allowing the beta sheet to unravel easily.[57;62] This is responsible for the sudden reduction in the force on going form point B to C in figures 3.8 to 3.10. The hydrogen bonds responsible for the initial resistance are shown in figure 3.12.

Next the other beta sheets are loaded and fail sequentially in a similar brittle manner, until they have all unfolded and the force again increases as a linear conformation is reached (point D). The shape of the free energy curve calculated from fully atomistic simulations is very much in line with the suggestions made to explain experimental results[80] and theoretical models.[81]

For the alpha-protein IM9 after the initial increase the force remains flatter and shows smaller peaks and troughs than those of the I27 protein. This is because the connections within alpha helices, and the helices themselves fail more gradually leading to lower forces. This is in agreement with the literature; Brockwell reports that the all alpha IM9 domain unfolds below the noise limit of the experiment[45] and SMD simulations[69–71] suggest that the mechanical weakness of alpha helices is due to the fact that only one backbone hydrogen bond is loaded at a time, leading to sequential failure and an unravelling of the helix.

Protein L is a combination of beta sheets and alpha helices. Initially the force profile and free energy is similar to that of I27 as the force is initially resisted by beta sheets in both proteins. After the first beta sheet fails (see figure 3.12) a strong intermediate structure remains. The resistance of a beta sheet in this intermediate is responsible for the large force peak at moderate extensions of Protein L (going from points C to D). After this beta sheet fails the force curve flattens out as an alpha helix unravels gradually before the remaining structure fails and a linear conformation is reached (point E).

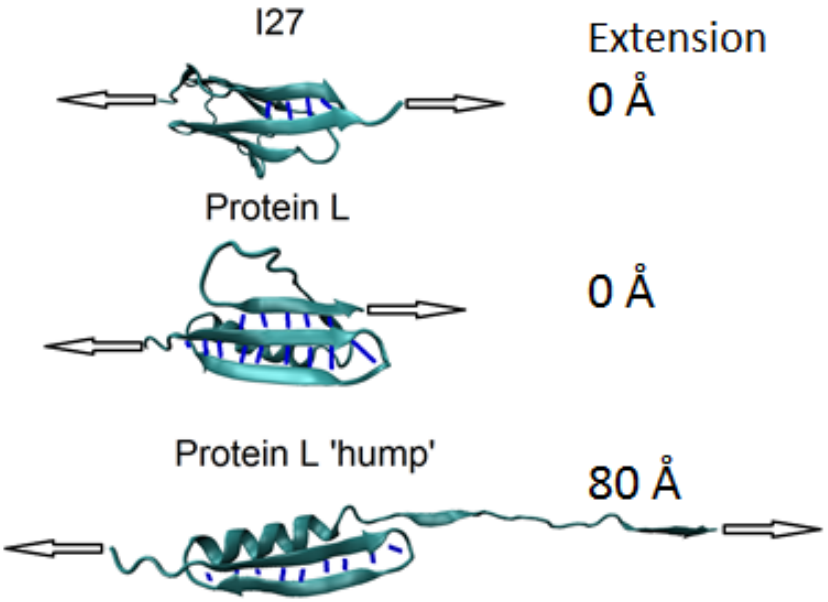

Figure 3.12: Structures responsible for force peaks in I27 and Protein L. Breaking the native structures (top two) requires the maximum pulling force as they represent the bottom of steep free energy wells around the native structure. The hydrogen bonds responsible for the initial resistance are shown as blue lines. These bonds rupture simultaneously causing a large structural change and rapid extension of the domain. In protein L there are two systems of hydrogen bonds, which unzip sequentially. The structure of the intermediate responsible for the hump in the force curve of protein L is shown at the bottom, corresponding to point D in figure 3.9. Arrows indicate the direction and points of application of the experimental pulling force.

In the experiment it is not a single protein domain but their sequence, a concatemer that is pulled. However, knowing the free energy of an individual domain and its gradient, and assuming that the domains extend independently and sequentially one by one allows reconstruction of the experimentally observed dependence of the force vs protein extension, shown by figure 3.13 alongside experimental force curves.

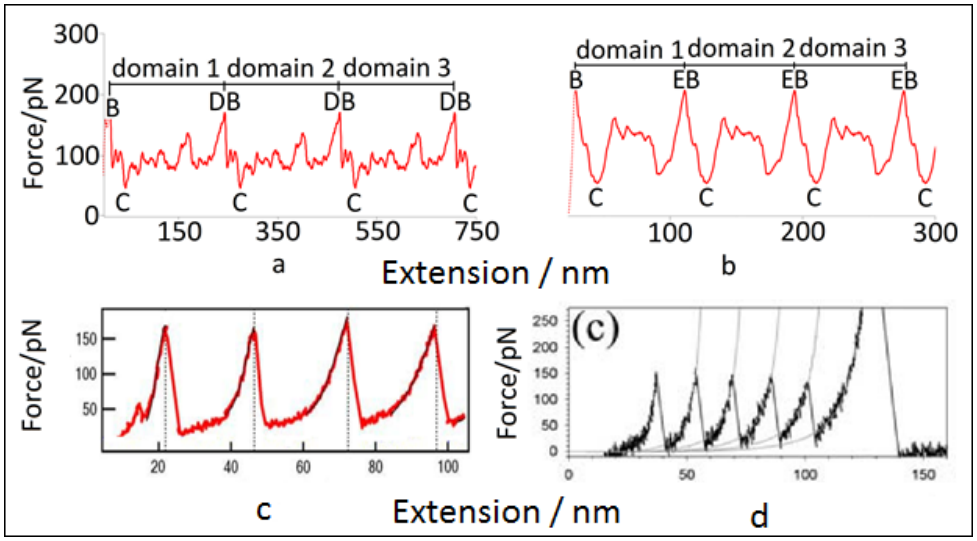

Figure 3.13: The teeth due to extension of a chain of the proteins for I27 (frame a) and Protein L (frame b). When one domain is extended to the point where the force increases to the value necessary to break out of the native well of the next domain, it's native structure quickly passes the inflection point and breaks down leading to the sharp fall of the pulling force. Then the process is repeated. Frames a and b correspond to I27 and Protein L respectively. The teeth are generated by repeating the section of the force curve beyond the initial peak force and the point where the force reaches this value again. This mimics the experimental force trace where a chain of domains unfolds one at a time. Experimental force curves for I27 (frame c) and Protein L (frame d) are shown for comparison. I27 experimental data is taken from Ref.[62] and Protein L data from Ref.[61] with permission from Elsevier.

As shown by figure 3.13 frame a, for I27 BXD qualitatively reproduces the overall structure of the experimentally observed teeth in the dependence of force on extension. The peak force and the length of the tooth are both well reproduced. Frame b of figure 3.13 is the same as frame a but for Protein L. The main peak (DB for I27, EB for Protein L) in both frames is reached when an expanding member of the concatemer reaches a linear extension and the force becomes sufficient to break the native structure of the next member of the chain. Then the process is repeated.

No multiple teeth can be obtained during the extension of a sequence of IM9 as there is no sharp decrease of the force at the point of inflection. The slope of the free energy for the alpha-protein IM9 in figure 3.10 is lower than that of protein L and I27 respectively. These observations are in agreement with experiment where alpha domains were found to be less robust and showed neither significant peaks nor teeth in the dependence of the force vs extension.

In a number of experiments the humps of the force were observed and also attributed to the structural changes in small proteins,[55;80] which are also present in our simulations. The experimental dependence of the force on the extension is typically less structured than those shown by figures 3.8 and 3.9 as BXD provides the free energy and force curves at a higher resolution that the AFM experiments. This has allowed the identification of an intermediate structure in the mechanical unfolding of Protein L (see figure 3.12.

The assumption that the unfolding events witnessed in the concatemer are equivalent to the unfolding in isolated domains is common.[50;82] However the extent to which it is valid is not fully resolved as cooperative motion of the concatemer and domain-domain interactions can affect the results of the experiment.[83] In certain systems domains have been shown to unfold together rather than independently[70] and in VC experiments the unfolding forces depend on the number of domains in the concatemer.[82;83]

In a recent study,[84] Uribe *et.al.* used umbrella sampling to calculate the free energy along end to end distance for a set of small peptides. The free energies obtained are very similar in shape to those shown in figures 3.8 to 3.10. In particular, the free enregies reported by Uribe show a distinct equilibrium well representing the native structure, followed by an inflection point leading to a region of less steep free energy.

These results are very much in line with the findings presented here and suggest that the free energies obtained by BXD are correct. Despite the free energies presented by Uribe being for small peptides of around 10 amino acids, the reaction coordinate used by Uribe is that of mechanical unfolding and the basic shape of the free energy profile is much the same as those generated by BXD for larger proteins, suggesting the mechanical unfolding free energies calculated by BXD are of the correct form.

3.3.2 BXD Kinetic Description of FC Experiments

Force Clamp is another mode of AFM pulling in which proteins are pulled with a constant force and the distribution of unfolding times is measured, giving the rate constant for unfolding at that particular force. Another quantity reported in the FC experiments is the protein extension as a function of the pulling force. Typical measured unfolding times (inverse rate constants) are in the order of seconds,[50–54] which nevertheless can be within the reach of BXD. Figures 3.14 and 3.15 summarize the results of modelling the FC experiments. It is clear from figures 3.8 to 3.10 that unfolded states would never be reached as the free energies are too high. It is thought that the application of force along the pulling coordinate tilts the free energy profile and allows mechanical unfolding to occur.[85;86] To test this theory the free energies obtained with BXD were adjusted to take a pulling force into account. If the force modified free energies allow FC experimental data to be reproduced then the theory of force tilted free energy profiles for mechanical unfolding is supported.

Frame A shows the force adjusted free energy obtained by modifying box-to-box rate constants with the factor $exp\left(\pm\frac{\Delta xF}{k_bT}\right)$ so that

$$k'_{n,n-1} = k_{n,n-1}exp\left(\frac{\Delta xF}{k_bT}\right), k'_{n-1,n} = k_{n-1,n}exp\left(-\frac{\Delta xF}{k_bT}\right) \qquad (3.7)$$

where Δx is the box size. For each free energy profile an end point was chosen to correspond to the maximum extension under that pulling force. The choice of end point depended on the shape of the free energy. For low forces ($<$ 30 pN) the free energy rises monotonically with no transition state or well present, which implies no or extremely slow unfolding. In this case the end point was defined as the inflection point in the free energy in the initial native well at which peak force is reached. For intermediate forces (30 to 50 pN) the free energy is tilted to show two wells, the native structure well at zero extension and a well at extensions of 50 to 100 Å. In this case the second well was chosen as the final state because it represents a transition from the native structure to a stable unfolded conformation. At higher forces ($>$ 60 pN) the force adjusted free energy shows three wells and the final well was chosen as a final state. The kinetic end points are shown by frame A in figures 3.14 and 3.15 by the arrows and the dependence of the endpoint on the external force, that is the extent to which the domain unfolds at each force, is shown by frame B in comparison with that of experiment.[52]

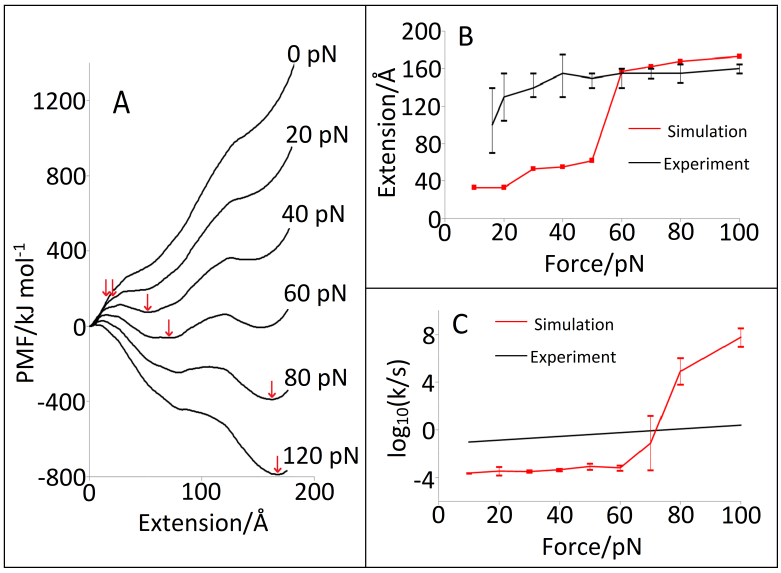

Figure 3.14: free energies for Protein L obtained with modified rate constants (equation 3.7) which include an additional factor taking into account the external force in the FC experiment (frame A). The assumed end points are shown by arrows and are also shown in frame B as a function of the applied force (red squares) and compared with the experimental data (black line). The estimated rate constant of unfolding is shown by the red line in frame C and compared with experiment shown by black line. For the force below 60 pN the theoretical rate constant remains flat, underestimating the experimental data by 2 orders of magnitude. The rate constent then grows fast at higher forces as the unfolding becomes barrierless. It must be noted that the experimental error bar is not known but may be significant.

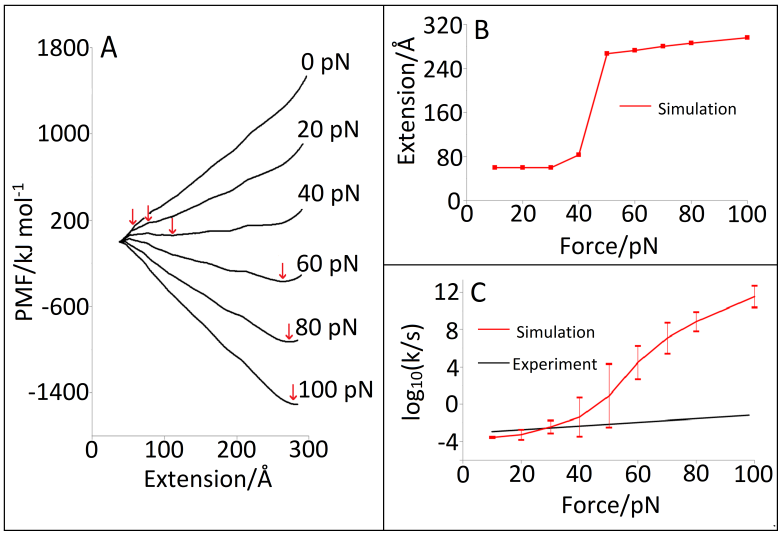

Figure 3.15: free energies for the protein domain I27 obtained with modified rate constants (equation 3.7) which include additional factor taking into account external force in the FC experiment (frame A). The assumed end points are shown by arrows and are also shown in frame B as a function of the applied force (red squares). Experimental data is absent here. The estimated rate constant of unfolding is shown in frame C by the red line and compared with experiment shown by the black line. For the force above 60 pN the theoretical rate constant starts growing fast as the unfolding becomes barrierless. It must be noted that the experimental error bar is not known but may be significant as the experiment is unable to detect very fast unfolding events.

There is significant difference between the calculation and experiment for the extension versus force data as well as with the unfolding rates versus applied force data (frames B and C). This is probabily because the rates of unfolding, as well as the amount to which a domain extends for a given force are highly dependent of the height of the free energy barrier for unfolding. The main barrier(s) to unfolding, for both I27 and Protein L, for forces above 20 pN are in a region of the PMF where the uncertainy is quite high (see figures3.8 to 3.10) and hence this slight lack of convergence in the important regions of the free energy could explain the large discrepency with experimental data. It should also be noted that quantitative agreement with all aspects of experimental data for such a long timescale process as AFM induced infolding

would represent an unprecedented achievement in atonistic simulation. Considering the modest resources used in the simulation and fact that BXD is a simple and fast algorithm, the results obtained with BXD, especially in the reproduction of the VC experiment (figure 3.13) are generally encouraging.

In order to estimate the unfolding times, the matrix of rate box-to-box rate constants $\boldsymbol{M}$ (see Chapter 2 equation 2.10) was truncated by removing all boxes beyond the box which represents maximum extension (frame B figures 3.14 and 3.15). The rate constant corresponding to motion from this last box to the box at a lower extension was then set to zero. This was done to mimic the FC experiment where the proteins are extended irreversibly. The eigenvalues of the kinetic matrix now determine the kinetic rate constants of unfolding at this particular force. Usually the lowest eigenvalue of a kinetic matrix $\boldsymbol{M}$ is separated from all other eigenvalues and determines the reaction characteristic rate constant for unfolding at that force. This was repeated for a range of experimental forces. Frame C in figures 3.14 and 3.15 compares the calculated rate constanta with experiment. For medium force the calculated characteristic times are within the range of 10^2 to 10^3 seconds which is approximately 2-3 orders of magnitude higher than the reported experimental result. At higher forces (c. 50 pN) the reaction becomes barrierless, unfolding becomes very fast and the rate constant increases dramatically.

Quantitative agreement with experiment has not been achieved here but it must be taken into account that BXD estimates the reaction rates in the order of minutes based on fully atomistic simulations without any modifications of the force field. However the qualitative agreement with the FC experiment suggests that the tilting of the free energy by the applied force does indeed allow mechanical unfolding to occur.

Comparison of BXD calculations with FC experiment allows modifications of the force field to be suggested which may improve the agreement with experiment. If the equilibrium well in the free energy is steeper and the region after the inflection point is flatter, then the agreement between experiment and theory in the frames on the right of figures 3.14 and 3.15 should improve. The former can be achieved by adjusting the parameters of hydrogen bonds responsible for the gradient of the free energy on the right hand side of the equilibrium well.

Also in these simulations an implicit solvent model was used to reduce the cost of the simulations, common practice in mechanical unfolding simulations.[61] There is evidence that the barrier to unfolding is lowered by the presence of water molecules. For example, with a protein such as I27 where unfolding is resisted by multiple hydrogen bonds in beta sheets it is thought that ruptured hydrogen bonds can reform with water rather than within the protein, lowering the energetic cost of rupture and hence reducing the height of the unfolding barrier.[55;64] Using explicit water instead of implicit solvent model also may improve the agreement with experiment reproduce this but this is computationally expensive.

Perhaps an approach similar to that outlined by Korotkin *et. al.* could be used, which treats explicitly only the water molecules which are in close proximity to the protein, and the rest of the water box by a hydrodynamic approach.[91]

Similarly to the velocity clamp experiments, force clamp experiments can be affected by the fact that AFM experiments pull a concatemer of many protein domains. It is assumed that each domain unfolds fully and independently of the others. Based on this assumption each of the steps in the extension vs time trace of FC pulling is assumed to be from a single domain and the data for each domain is averaged.

However if this independence of unfolding events is not wholly correct then the situation becomes complicated. Cooperative motion of the concatemer, interactions between domains and energy storage in the chain would make analysis of the experimental data much more difficult and inconclusive. The debate as to whether this central assumption is valid is ongoing[51] and AFM studies of the refolding of a concatemer show that cooperative motion drives the process.[92] Simulation of a concatemer and investigation of cooperative effects will be a future goal.

It is also possible that the measured force at which unfolding events occur does not exactly correspond to reality. Thermal fluctuations in the position of the AFM tip may cause protein domains to unfold prematurely at forces other than that registered by the experiment[53] and if the concatemer is not pulled perpendicular to the solid surface to which it is attached then the force along the pulling coordinate could be less than what is measured.[53]

A key difference between simulation and the AFM experiment is the dependence of unfolding force upon the direction in which the force is applied to the domain. Secondary structure alone does not fully explain the mechanical resistance of proteins,[87] as some proteins, such as I27 and TNfn3 with similar secondary structure in the region to which the force is apllied, show dissimilar unfolding forces. Brockwell *et. al.*[88] demonstrate that the dependence of the mechanical resistance upon the direction of applied force is due to the arrangement of the secondary structure with respect to the angle of the force. When the force is applied in parallel to beta strands the load is shared between multiple hydrogen bonds resulting in a strong resistance, and when the load is applied perpendicular to the beta strands the hydrogen bonds rupture one at a time, leading to less resistance.

Carrion-Vazquez *et.al.*[89] show that the choice of linkage between the protein domains in the concatemer affects the direction in which

force is applied to each domain and hence the force at which they unfold. As linkages are not included in the BXD simulations it is possible that this explains some of the discrepency between experimental data and the simulation. Zoldak and Rief[90] suggest that, as proteins unfold at different forces when pulled in different directions, end to end distance could be an unsuitable choice of reaction coordinate for mechanical unfolding, however for most systems the main features such as barrier height and position are captured sufficiently. BXD does not apply a force to the protein, instead the dynamics along the coordinate of end to end distance are simulated, with extension occuring along the vector between the N and C terminals. While this may be incorrect due to the discussed effects of pulling direction relative to the hydrogen bonds between beta strands, it is probable that BXD does not fail on this account because overall agreement with experimental unfolding forces are good (figure 3.13). Also the mechanical clamp between the A' and G beta strands is correctly identified as the key structure responsible for I27s mechanical strength (see figure 3.4), suggesting that the dynamics simulated with BXD are similar to those of the AFM experiment and the anisotropy of a domains mechanical strength has not had a significant effect on the results of the simulations.

After a domain unfolds the AFM tip recoils and the force is no longer applied to the concatemer for a certain amount of time.[93] Any unfolding events that take place in less time than this recoil time will be missed. In the experiments from which the data is taken for comparion to the BXD simulation this detection threshold is roughly 5 ms, corresponding to $log_{10}(k_{unfold}) \approx 2.3$. This effect would affect the experimental unfolding time vs force curve as fast unfolding events would be missed because the force is not applied for this period. Based on the above factors the results obtained by BXD are in good qualative agreement with the FC experiments.

Recently it has been discussed in the literature whether the kinetics of the FC experiment can be described by a single rate.[92] In principle BXD allows reproduction of complicated kinetics. For example it has been shown that given the initial conditions peptide cyclization kinetics can exhibit both simple single exponential and complicated power law behaviour.[94] Describing this phenomena in AFM experiments will require more work. However this study only shows that BXD is capable of recovering realistic rates and very long characteristic times observed in the FC experiments. In future, quantitative agreement with experiment will be attempted by addressing the issues discussed above.

3.4 Conclusions

The above discussion can be summarized as follows:

1)BXD reproduces realistic pulling forces (in the order hundreds of pN) observed in VC experiments.

2)Initially the force increases without affecting the equilibrium structure significantly. When a threshold force is reached a structural unit fails, followed by the sequential failure of the remaining structural units at lower forces until a linear conformation is reached.

3)Beta-sheets resist the force well before failing abruptly leading to pronounced peaks and troughs in the force vs extension curves, whereas alpha helices fail gradually at lower forces, leading to a flatter force curve.

4)The force vs extension curves reproduced by BXD reveal intermediate structures along the mechanical unfolding pathway. These structures can easily be visualized.

5)Qualitative agreement with the FC experiments has been achieved on unfolding processes up to a timescale of seconds. The assumption that

the application of force tilts the free energy landscape, thus allowing mechanical unfolding to proceed,[85;86] *has been confirmed.*

We believe that this is the first time that the mechanical unfolding of proteins has been simulated under realistic conditions without artifically high pulling forces, providing reliable confirmation of the factors affecting mechanical stability as well as the unfolding pathways and free energies

3.5 Further work

The following issues will be addressed later:

1) - The calculatons will be repeated with explicit solvent to achieve a better agreement with experiment.

2) - An attempt to go beyond the assumption of slow pulling will be made. This may allow investigation of the dependence of the maximum pulling force on the pulling speed observed in experiment. BXD is well suited for this as it provides kinetic information along the whole reaction coordinate.

3) - Simulating the extension of a concatomer may make it possible to assess the importance of cooperative effects and also answer the question whether the humps which correspond to the breaking of the structures within a single member of the concatomer are not detected. This is within the reach of BXD but will require more work.
Thus BXD is an efficient theoretical tool to study long time processes such as mechanical unfolding.

Chapter 4

In Silico Screening of Medicinal Cyclic Peptides

4.1 Introduction

This chapter presents another application of BXD: predicting the propensity of amino acid sequences to form cyclic peptides.

Traditional drug discovery has focused on designing small molecules which work by binding to a particular bimolecular target responsible for the disease or condition.[95;96] Binding to the target causes it to lose its biological activity, for example by blocking the active site of an enzyme. Despite these insights pharmaceutical research and development has been declining in efficiency for several decades.[97–99] The situation has become especially urgent in the field of antibiotics, where widespread resistance to conventional small molecule drugs is a major threat to world health.[100–102]

One factor in the reduced effectiveness of drug discovery is that, despite the recent increases in the understanding of the mechanisms of disease, the number of possible targets for a conventional drug or Active Pharmaceutical Ingredient (API) is highly limited. Indeed, traditional small molecule APIs are, in general, only effective against targets that

contain a discrete hydrophobic pocket or are integral to the outside or membrane of a cell. Together this accounts for only twenty percent of all known disease related targets.[98] Another problem facing drug development is the low solubility of many APIs[103;104] and their inability to be taken orally.[96;97;105]

Overcoming these problems will require a dramatic innovation in drug discovery along with the introduction of new classes of compounds.[97;98] Cyclic peptides are a promising candidate with much potential for providing new types of antibiotic and anti-infection drugs.[97;98;105;106;122] Consisting of a ring of between 5 and 30 amino acids, cyclic peptides have a number of features which allow them to target biomolecules which cannot be affected by traditional small molecule APIs.[106;107] For example, their large size and complexity allows more reliable and selective binding to large extended binding sites as well as the targeting of protein-protein interactions, both cases where conventional drugs struggle.[97;98;108–112]

Cyclic peptides provide a degree of pre-organisation which reduces the entropic cost of binding while retaining enough flexibility to interact well with dynamic protein targets which are very difficult for small molecule APIs.[106;107;113] There is also evidence to suggest that a cyclic conformation helps improve other areas of performance such as membrane permeability, metabolic stability and the overall interaction with the body.[106;107;114] as well as an increase in the overall stability of the molecule over a linear equivalent.[98;108–112]

Since the first cyclic peptide drug Gramicidin S (figure 4.1) was introduced as an antiseptic in 1944[115] there are now at least thirty cyclic peptides on the market with more in clinical trials[97] most of which are antibiotics, anti-infection or anti-cancer agents[97] which are in very high demand due to antibiotic resistance.[100–102] Recent advances in peptide synthesis have increased the ease of cyclic peptide production

and study[98] and hence the stage is set for cyclic peptides to become a valuable aspect of modern medicine.

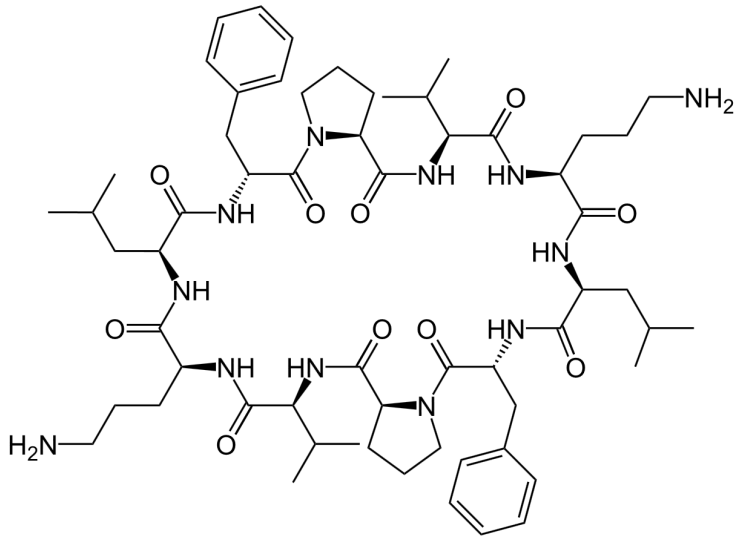

Figure 4.1: Gramicidin S, the first medicinal cyclic peptide, was discovered in 1944 and used to prevent infection in gunshot wounds.

The first step of cyclic peptide manufacture is to obtain the linear peptide precursor, by inserting a fragment of DNA that codes for the desired peptide into *E. coli* bacteria which then produce the peptide, or by chemical synthesis. Both methods are capable of producing low cost product on the scale of kilograms. Now that the linear precursor has been produced it must be cyclized - the most difficult part of the process.[116–120;141;143]

The three forms of cyclization are head to tail, where the C and N terminals come together to form the ring, side chain to side chain where two side chains join together (often via a disulphide bond) and head or tail to side chain. This work focuses on head to head cyclization as this is the approach taken by the experimental collaborators. There are two

approaches to cyclising the linear peptide: via conventional chemical synthesis or via enzymatic cyclization.

4.1.1 Conventional Synthesis

The linear peptide is anchored to an insoluble polystyrene bead. This prevents the peptides reacting with each other as the insoluble bead immobilises the peptide,[116] preventing them from reacting with each other rather than cyclising. This technique allows the necessary dilute reaction conditions to be mimicked without compromising on yield. The cyclization reaction step is carried out with a Palladium catalyst before the product is cleaved from the support. In order to prevent a head or tail to side chain cyclization a number of protecting groups (small chemical motifs which attach to a group to stop it taking part in the reaction) are added to the peptide and must be removed later.

The advantage of this method is that overall yield is high, as intermolecular reactions are prevented by the beads, and the protecting groups ensure that only head to tail cyclisaion takes place. Purification of the product is easy as the large size of the insoluble beads means impurities can be rinsed away. This is a significant advantage of this process as product purity is critical; impurities such as leftover solvent and catalyst are major safety concerns which are heavily regulated.[121]

This method is limited by the ability of the linear precursor to preorganise into a conformation where the N and C terminus are close to each other so that the reaction can occur. This conformation is thought of as the transition state for cyclization and its adoption by the linear precursor is the limiting factor of conventional cyclic peptide synthesis.[116–120;141;143] For a peptide to change conformation there must be rotation around the bonds in the backbone. Rotation around the peptide bond (the bond between amino acid residues) is limited due to it being a partial double bond caused by the resonance shown in figure

4.2. The other two bonds can rotate but the accessible torsion angles are limited due to interactions between the side chains.

All naturally occurring amino acids are in L stereochemistry hence the peptide bond is in the *trans* conformation. For an all L peptide to adopt a cyclic conformation large amounts of steric strain build up in the backbone, because the *trans* conformations force the other two backbone bonds to adopt unfavourable torsion angles far from their equilibrium values.[116] Many organisms which produce cyclic peptides overcome this by including D amino acids into the sequence, inducing a turn in the linear peptide which makes a cyclic conformation less disfavoured. This is because the backbone torsion angles do not need to move as far from their equilibrium values to complete the ring. D amino acids are introduced in nature by changing the stereochemistry of L amino acids with the enzyme L Amino Acid Oxidase. Unfortunately D amino acids are expensive so some other method is needed for synthesis.[123]

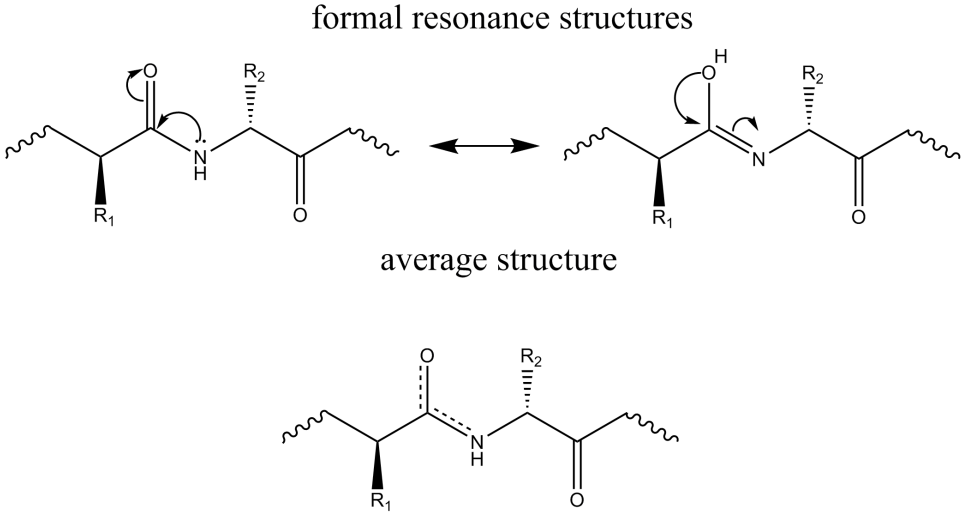

Figure 4.2: resonance forms of the peptide bond lead to a bond order greater than one which hinders rotation.

Another way to introduce *cis* peptide bonds into the precursor is

to include the artificial amino acid Pseudoproline in the precursor. This induces a turn in the peptide backbone as shown in figure 4.3. After cyclization the pseudoproline is converted into Threonine. The peptide termini can also be brought closer together by using a metal ion to chelate to the peptide backbone, lowering the energy of the cyclization transition state as the interactions with the metal holds the near cyclic conformation in place, shown in figure 4.4.

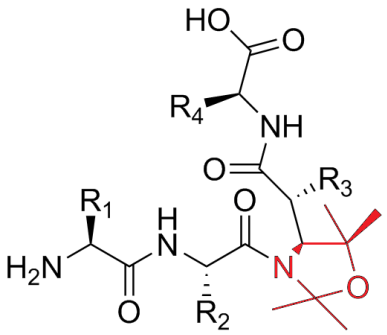

Figure 4.3: addition of pseudoproline (red) in the linear precursor reduces the steric strain of the cyclization transition state due to the turn in the peptide backbone.

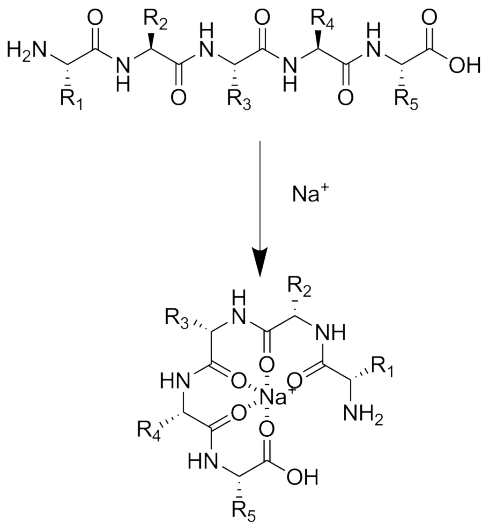

Figure 4.4: chelation to a metal ion can stabilise the cyclization transition state.

A wide range of organic catalysts can temporarily react with specific side chains to hold the peptide in a horseshoe conformation, reducing the distance between the termini. See reference[116] for details. Despite all these methods of reducing the energy barrier to cyclization the efficiency of conventional synthesis varies depending on the peptide to be synthesised, there being some sequences for which no method is successful.[116;124]

4.1.2 Enzymatic synthesis

Many cyclic peptides are found in nature[125–128] where they are synthesised without employing any of the methods described above. Instead the linear peptides are produced by the ribosome and are then cyclized by enzymes. One well studied enzyme is PatG, found in the cyanobacteria Prochloron. Prochloron lives in a symbiosis with the marine organism *Lissoclinum patella* A.K.A. the Indo-Pacific Sea Slime (figure 4.5). Prochloron bacteria produce a variety of cyclic peptides which have shown anti cancer activity,[125] the most famous being Patellamide A (figure 4.6) which has anti cancer properties and was discovered in 1982.[129]

Figure 4.5: *Lissoclinum patella* or Indo-Pacific Sea Slime. The green colouring comes from symbiotic cyanobacteria which produce anti cancer cyclic peptides using the PatG enzyme. Author: Nick Hobgood. Reproduced from Wikipedia Commons under the Creative Commons Attribution-Share Alike 3.0 Unported licence. See htt ps://creativecommons.org/licenses/by-sa/3.0/deed.en for licence certificate.

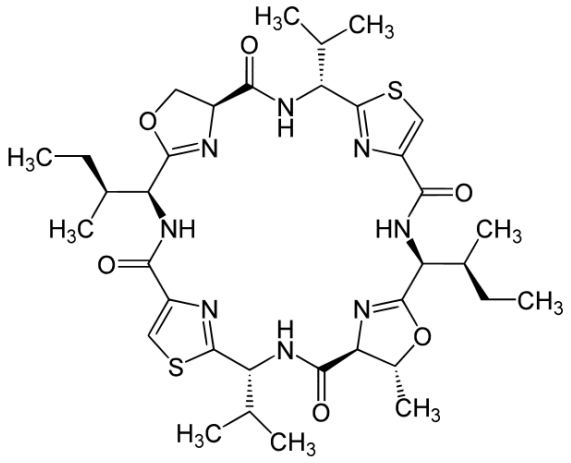

Figure 4.6: Patellamide A, the most well known cyclic peptide produced by Prochloron cyanobacteria.

PatG works by taking a linear peptide and catalysing the cyclization reaction. Before cyclization begins the linear precursor is sometimes modified by other enzymes, a number of amino acids can be converted into other heterocyclic groups. The linear substrate of n amino acids (represented by X) can be written as $X_n...X_2X_1AYDG$ where X_1 must be Proline. This is because the peptide bond between X_1 and X_2 must be in the *cis* conformation for the enzyme to work, and Proline is the only natural amino acid which makes the isomerism possible at room temperature.

The $AYDG$ section is a recognition tag which binds to the active site of PatG, the *cis* geometry of the $X_2 - Pro$ bond allows the peptide chain to stick out away from the body of the enzyme. If the X_1X_2 bond were *trans* then a large steric clash between enzyme and substrate would prevent binding. The necessary presence of a Proline or some other modified amino acid which increases the probability of a *cis* peptide bond occurring, is probably common to most enzymatic cyclization reactions as Proline is found in most short cyclic peptides.[130] The proposed mechanism of PatG cyclization is shown below in figure

4.7.

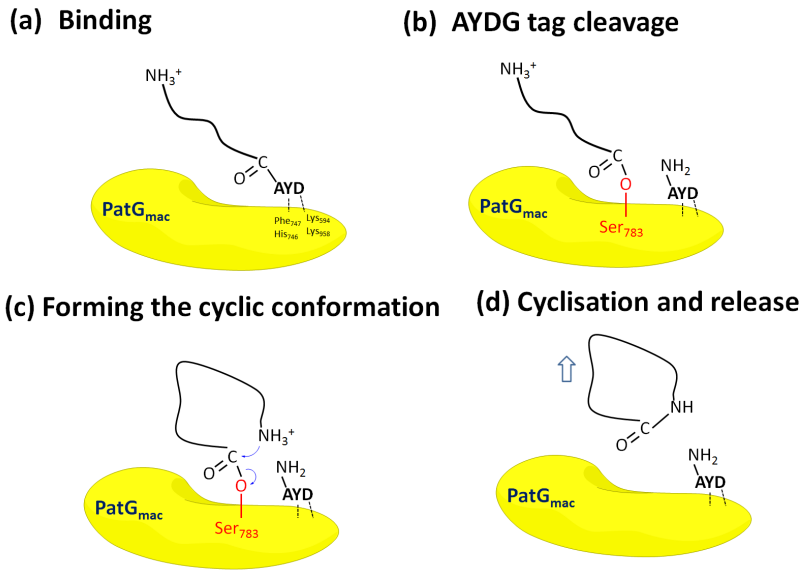

Figure 4.7: simplified scheme of PatG cyclization.

In more detail, the steps of the reaction are as follows:

1) - The AYDG tag binds to the active site of PatG (figure 4.7 frame a).

2) - Cleavage of the AYDG tag and binding of the Proline terminal oxygen to the Ser_{783} group on PatG (figure 4.7 frame b).

3) - The bound peptide adopts a conformation where the free N terminus is close to the anchored Proline (figure 4.7 frame c). This is referred to as the Pre Cyclization Conformation (PCC).

4) - Head to tail cyclization reaction.

5) - Release of cyclic product (figure 4.7 frame d).

If the peptide does not cyclize within a certain time after binding then the AYDG tag will be cleaved via a spontaneous hydrolysis reaction

and the peptide will be released in the linear form. This time limit is not known and is currently under investigation. The structure of PatG is shown in figure 4.8.

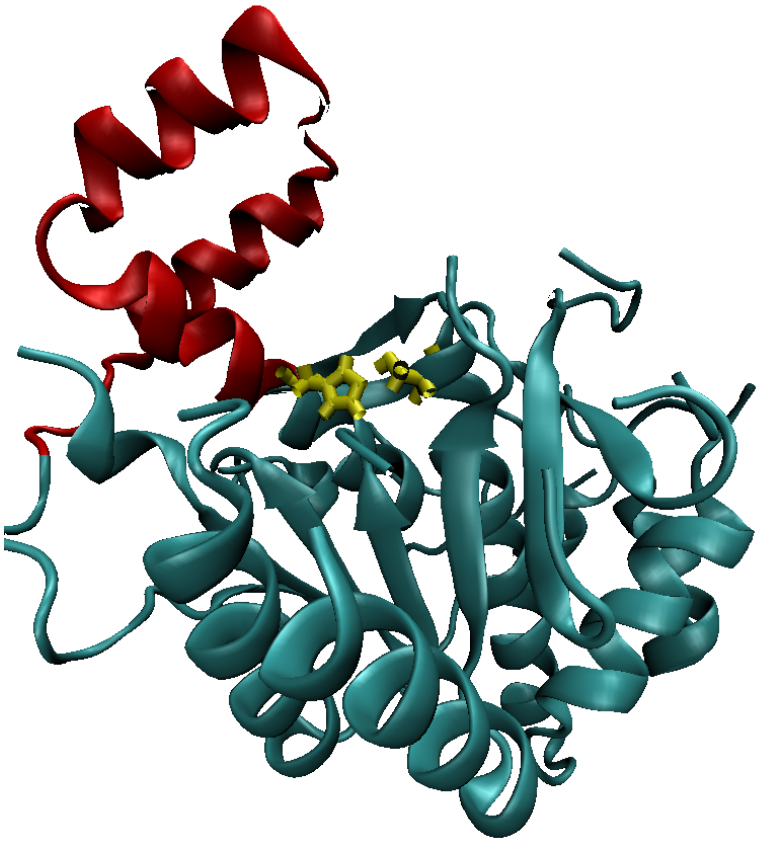

Figure 4.8: the PatG enzyme. The groups which form the active site and bind to the AYDG tag are shown in yellow. The red structure has been proposed as a plug which, after peptide binding, moves in to prevent water from interfering with the head to tail cyclization reaction. The gaps in the structure are due to sections of flexible linker which are not resolved by X-ray crystallography. Structure taken from RCSP data bank (4AKS).

The mechanism shown in figure 4.7 represents the best current knowledge of PatG cyclization,[126] however the point at which the linear substrate adopts the PCC has not been confirmed. It could occur before binding (frame a figure 4.7), between binding and AYDG tag cleavage (frame b figure 4.7) or after tag cleavage (as proposed in frame c figure

4.7).

It is also not known to what extent the enzyme assists the substrate in adopting the PCC. It is possible that the enzyme mechanically folds the substrate into the PCC after binding, or this step could be diffusional with the enzyme waiting for the peptide to find a cyclic conformation. Also, it has not been confirmed whether the conformational search to adopt the PCC, assisted or otherwise, is rate limiting or whether the rate of cyclization depends on other steps such as binding, tag cleavage or the final cyclization reaction. Another unresolved aspect of the cyclization mechanism is the role of the proposed hydrophobic plug, shown in red on figure 4.8. It is thought that this structure could prevent water from accessing the active site once the peptide is bound,[126] as water molecules would interfere with the head to tail cyclization reaction.

4.1.3 Objectives

The aims of this study are to use BXD to address the following:

1) - To Develop a method of predicting whether or not a sequence of amino acid residues can be cyclized by the PatG enzyme.

2) - Make a high throughput *in silico* screening tool for designing medicinal cyclic peptides.

3) - Investigate the mechanism of PatG cyclization. Is the adoption of the PCC diffusional or enzyme assisted, and is it the rate limiting step?

4) - Test the model on other enzymes to determine if the mechanism of PatG cyclization is common to other enzymes.

5) - Find some general rules for what makes a peptide cyclisable.

4.2 Existing Methods of Predicting Peptide Cyclization

A major limitation of any method of cyclic peptide synthesis is the lack of understanding of what factors affect the propensity of a linear precursor to adopt a cyclic conformation, necessary for the reaction to occur.[141] Most computational studies on cyclic peptides focus on the prediction of the structure of the cyclic product[108–112] rather than the probability that the linear precursor will cyclize. However some theoretical studies have been made and are discussed in the following.

In 1992 Cavelier-Frontin *et. al.* investigated the effect of the different ways of arranging the amino acids in the linear precursor.[141] The cyclic tetrapeptide Chlamydocin consists of four amino acids $Aib-Phe-Pro-Ala$, where Aib is aminoisobutyric acid which induces a peptide to form alpha-helixes.[142] Chlamydocin can be synthesised from four linear peptides: $Aib-Phe-Pro-Ala$ (A) $Phe-Pro-Ala-Aib$ (B), $Pro-Ala-Aib-Phe$ (C) or $Ala-Aib-Phe-Pro$ (D). Precursor C is the only one that cyclizes.

The authors first investigated whether steric effects in the different precursors hinder ring closure to different extents. This was done by taking the experimental structure of each amino acid in the cyclic product and using them to reconstruct the precursors A,B,C and D. Single point energy calculations were then carried out for each of the linear structures at close end to end distances to calculate the potential energy of the cyclization transition state for each peptide. No significant difference in these energies was found so the authors ruled out this factor.

Cavelier-Frontin *et. al.* then used an energy minimisation on each of the four transition states to find the relaxed structure for each linear

peptide for which the potential energy was lowest. Single point calculations were carried out on these structures and the difference in energy between them and the transition state was calculated and taken to be the activation energy for cyclization. It was found that linear peptide C, the only one which can cyclize, had the lowest activation energy. After this successful trial the same method was extended to four other four membered cyclic peptides and the correct precursor was identified in each case.

Despite its success this method is limited due to the fact that it does not consider the conformational space of the peptide. Only single, static structures are taken for the transition and extended states of the peptide and the cyclization reaction is assumed to be a function of the difference in potential energy between these structures. This is a static view of the system which is not suited to the dynamical nature of biomolecules. Also the structures considered were generated by simple energy minimisations which do not explore the complicated conformational space of the peptide; there is no way of knowing whether the structure chosen as the extended conformation is in fact the global free energy minimum or one of many local minima which exist in equilibrium.

While these limitations may not be critical for very short peptides it is unlikely that this method could be extended to large systems where the conformational space is larger and more complicated. This was the case when Besser *et. al.*[143] applied the above method to other four membered cyclic peptides and found it to be unsuccessful. Despite these shortcomings credit should be given for what was achieved at a time when a more rigorous calculation or simulation would have been very expensive.

In their work Besser *et. al.* assumed that the ability of a linear sequence to cyclize depends on the probability of finding the system in a cyclization transition state structure, that is with the termini close

together. This was done by first generating an ensemble of possible conformations of the linear peptide. For this a Monte Carlo Multiple Minimisation (MCMM)[131] method was used. A starting structure is chosen and rotations of the peptide backbone bonds are introduced by changing the angle of torsion of randomly selected bonds. An energy minimisation is performed on the new structure which is then compared to the previous step. Only the backbone bond angles were varied in order to focus on the 'signal' of peptide conformational change rather than the 'noise' of side chain fluctuations. If it is sufficiently different it is accepted and the angles are perturbed again and the process is repeated, the new structures being compared to all the previous accepted ones. Every time a structure is accepted it is added to an overall set of potential energy minima. Besser *et. al.* used this process until they had a large number of structures from which they selected the 300 that had the lowest potential energy.

Each of these was then used as a seed in a separate MCMM calculation until a very large number of conformations had been gathered. MCMM generates an ensemble corresponding to a particular temperature which determines the magnitude of the torsion angle change, higher temperatures lead to larger perturbations. The next step was to apply MCMM method to starting structures where the ends of the peptide were close together, representing a conformation from which the cyclization reaction could occur.

During the MMCM procedure some structures were constrained so the ends of the peptide remained close together, resulting in an ensemble of transition state structures. In order to estimate the probability of a transition state occurring each of the structures in the unrestricted ensemble were compared to the set of transition state structures. The total number of structures from the ensemble which were found to be close to a transition state structure was divided by the total number of structures in the ensemble to give the probability of finding the peptide

at the transition state. This is illustrated in figure 4.9.

This approximation of transition state probability relies on the system being sampled ergodically so that a fair Boltzmann weighted ensemble is produced. Because of this the MCMM run used to generate an ensemble was continued for around 300000 steps.

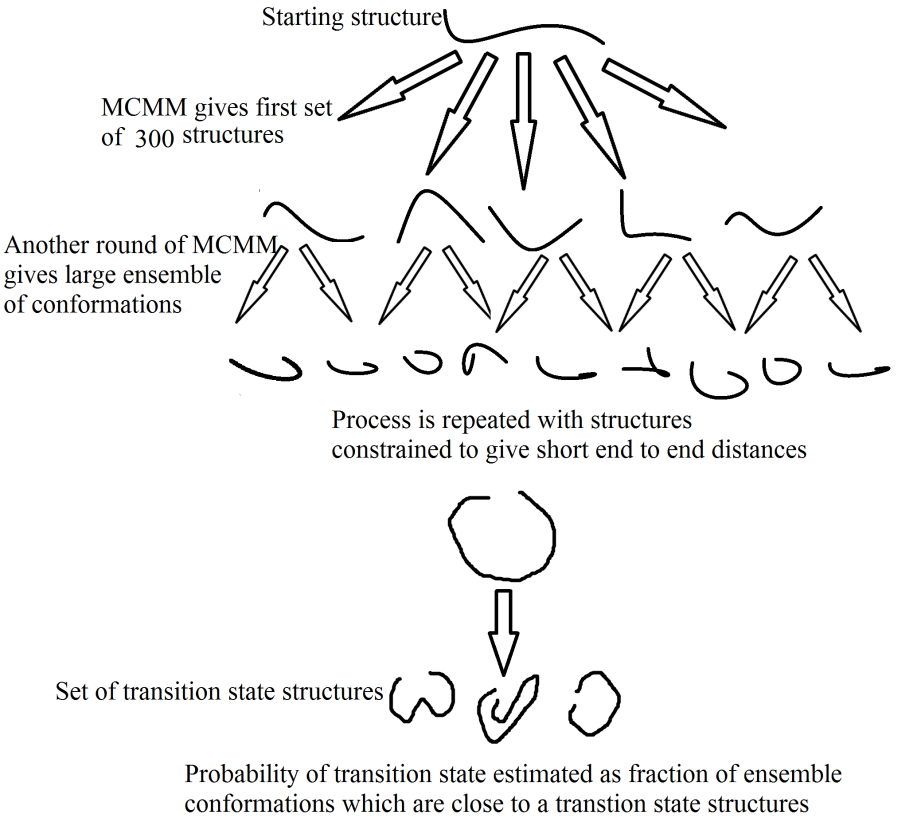

Figure 4.9: the Monte Carlo Multiple Minimisation procedure used by Bresser *et. al.*[143] to generate an ensemble of conformations of a peptide.

The cyclization probability of five peptides calculated in this way were compared to experimental data. At 300 K the coefficient of correlation between theoretical and experimental data was only 0.59 implying a poor performance of the model.[143] By changing the temperature in the MCMM procedure and generating a new ensemble of structures at

1000 K the coefficient of correlation was improved to 0.92 and after the temperature was again increased to 5000 K the coefficient fell to 0.78. The authors attribute this effect to discrepancies in the force field used. At all temperatures the most and least likely peptides to cyclize were ranked correctly, the ordering of the others being incorrect. Despite this seemingly low success rate the peptides which were furthest apart in cyclization probability correspond to a factor of 2 in the experimental rate of cyclization. This limit on the accuracy of the method is quite good considering that the rate constant for cyclization of peptides can vary by a factor of 500.[132] The authors also report that each conformational search took around 4 hours[143] on a cluster of 8 400 MHz processors, which is a very short simulation time especially considering the very modest processors used.

In a recent study of peptide cyclization[133] Yongye *et. al.* used conventional unbiased MD to investigate the cyclization ability of three small peptides. After running 20 ns trajectories of each peptide they collected the most common conformations that were sampled. To predict which sequences would cyclize they compared the number of extended conformations where the ends are far apart to the number of transition state structures. This method is conceptually similar to that of Besser *et. al.* the difference being MD was used to sample the conformational space rather than MCMM. Two out of the three peptides are known from experiment to cyclize readily.

The peptide which cyclizes best in experiment was found to have well populated conformations where the ends of the peptide are held close together by backbone hydrogen bonds, presumably making the cyclization reaction more likely. The peptide which does not cyclize was found to have well populated extended conformations where the termini are far apart and few with the ends close together, making cyclization less likely.

The authors reported that the third peptide provided inconclusive results as the backbone was stiff,[133] meaning that conformational sampling was poor during the simulation due to the large energy barrier of rotation of backbone bonds. This prevented meaningful results from being obtained as the conformational space must be well sampled for the relative numbers of extended and transition state structures to match the true Boltzmann ensemble.

BXD has previously been used to investigate peptide cyclization by Shalashilin et. al. in 2012.[94] The aim of this study was to reproduce the findings of Volk and Hochstrasser[134–136] who investigated the kinetics of peptide cyclization. A cyclic peptide was produced with the ends joined together by a sulphur bridge. Flash photolysis is used to break this bond leaving a sulphur radical on each end of the peptide, bond cleavage taking place on the sub picosecond timescale. The peptide is then free to diffuse until the ends come together again and the two radicals recombine to reform the sulphur bridge and cyclize the peptide again. The presence of the radicals leads to a characteristic absorption which is monitored over time and the rate at which the strength of the absorption decreases is related to the rate of the cyclization of the peptide. The procedure is shown in figure 4.10.

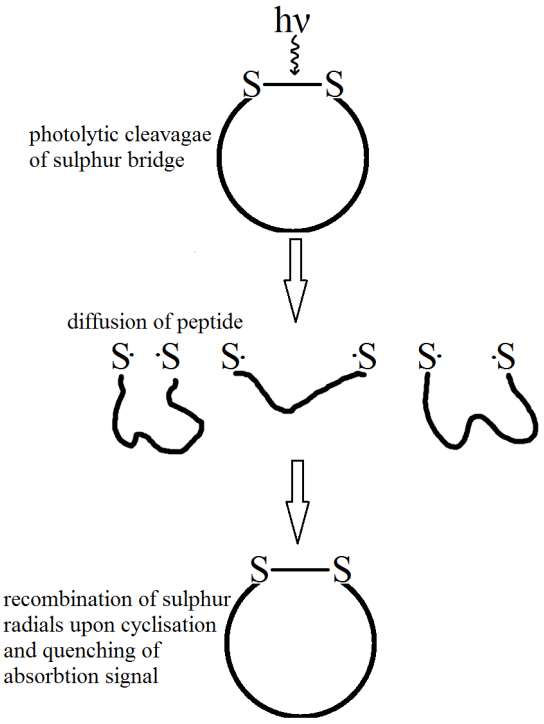

Figure 4.10: monitoring the rate at which the absorption from the sulphur radicals is quenched allows the rate constants for cyclization to be calculated.

Shalashilin *et. al.* used BXD to calculate the rate constants of cyclization for the peptides studied by Volk et. al.[134–136] The reaction coordinate was defined as the distance between the two ends of the peptide and boundaries were placed along it at distances of between 3 and 40 Å. The phase space along the coordinate was sampled as the trajectory diffused through the boxes (see figure 2.3) allowing box to box rate constants to be determined (equation 2.6) which were then used to calculate the free energy along the reaction coordinate (equations 2.6 to 2.8). The resolution of the rate constants was increased by using the procedure described in reference[94] (see equations 2.15 to 2.17) and equations 2.9 to 2.11 were used to calculate the rate of cyclization.

In the experiments[134–136] the cyclization of the peptide is irreversible as it involves the recombination of the sulphur radicals. To

take this into account in the analysis of the BXD rate constants a transition state for cyclization is set at 5 Å and the rate constant for crossing the boundary placed at 5 Å to higher extensions is set as zero, mimicking the irreversible nature of the radical recombination. The rate constants for cyclization were estimated by calculating the population of the box at 5 Å with respect to time and were in excellent agreement with the experiments of Volk *et. al.*,[134–136] matching the experimental data over several orders of magnitude. Each peptide took a few CPU days to converge.

No other literature on the prediction of peptide cyclization could be found, and no studies could be found where cyclization was predicted for the enzymatic route rather than for conventional synthesis. Given that BXD has already successfully been applied to peptide cyclization and is a proven way of overcoming the long timescale problem[44] that affected the work of Besser *et. al.*[143] it was decided that BXD would be used to undertake the blind test.

4.3 Method

4.3.1 The Blind Test

The Jaspars group at Aberdeen University are developing a method of cyclic peptide synthesis which uses the PatG enzyme to cyclize linear peptides into designer cyclic products, avoiding many of the limitations of traditional cyclic peptide synthesis. The aim is to use a single enzyme that can cyclize a general linear peptide which will lead to easier production than conventional synthetic methods which have variable yields depending on the sequence.[116;124] Enzymatic cyclization also has the advantage that no harmful catalysts or other reagents must be added leading to a safer and cheaper production of APIs as purification and quality control would be easier.

Seeking to improve the efficiency of their enzymatic production the Jaspars group presented a challenge: given a number of amino acid sequences could we predict which ones would cyclize with their enzymes? Two sets of sequences were given: one from PatG and one from an enzyme from another species of cyanobacteria, AcyG. These peptides that do cyclize were mixed in with other sequences which are known from experiment not to cyclize. The experimental results of cylisation were withheld until a prediction had been made. Thus the model and all the various parameters would have to be developed in the dark - a blind test. If the model were successful then this would aid the Jaspars group design medicinal cyclic peptides as less time would be spent trying to produce products which cannot be cyclized.

4.3.2 Predicting Cyclization with BXD

BXD was used to calculate the free energy of cyclization for the peptide sequences provided by the Jaspars group. The reaction coordinate chosen was the distance between the two ends of the peptide as it is these termini which must come together to cyclize the substrate thus this coordinate differentiates between cyclized and linear states, and the free energy along it should correspond to the free energy of cyclization. This is shown below in figure 4.11.

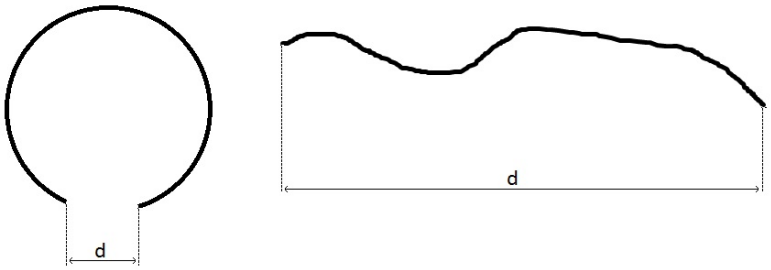

Figure 4.11: end to end distance is a reaction coordinate which describes peptide cyclization.

Boundaries were placed at intervals of between 1 and 2 Å along the reaction coordinate and the number of collisions required to pass into the next box was set at 1000. The boxes started at 3 Å and continue upwards to a length at which the peptide is judged to be mainly linear. This upper limit was produced empirically and was usually around 2.5 Å multiplied by the number of amino acids in the chain. This limit reflects a compromise between effective sampling and simulation efficiency. If the peptide does not extend to high enough free energies then the number of pathways sampled on the way down the reaction coordinate will be limited, however if too much time is spent at high extensions then the simulations will be slow.

For each peptide 20 trajectories were simulated with BXD, starting from the linear conformation with a different initial distribution of velocities. This ensures that the phase space is well sampled as there are 20 sets of initial conditions. Simulations were continued until the free energy had converged for each peptide. Convergence was checked by taking the data from each of the 20 trajectories and splitting it in half. Each half dataset was used to calculate the free energy and this was compared to the free energy calculated with all the data. If the two free energies were similar then that trajectory was said to have converged, and if not then more simulations were carried out until each peptide had 20 converged trajectories. Usually this occurred when a trajectory had completed around 20 to 50 cycles in both directions (similar to those shown in figure 2.4) through the reaction coordinate. The free energy was then calculated for each of the 20 trajectories and then averaged. The EEF1 implicit solvent model was used, as in Chapter 3 for the AFM unfolding study. While the EEF1 model is not the most accurate solvent model it is suitable for this application as it is very fast, which is a requirement as 20 trajectories were simulated for each member of a large set of peptides. The EEF1 model is very basic however this may be more appropriate for this application as it is not known to what extent the bound substrate is exposed to water, as the role of the hydrophobic

plug (red structure in figure 4.8) has not been resolved hence a more accurate solvent model may not be appropriate.

The set of sequences provided by the Jaspars group for AcyG and PatG are shown below:

PatG
8 residue
WAPWVWLP (8a)
EDWYFDHP (8b)
MDCWINYP (8c)
VIQHYLFP (8d)

10 residue
YSNKPSDFSP (10a)
QENHVFIQFP (10b)
TSQIWGSPVP (10c)
PTGIPDHCEP (10d)

12 residue
ILGEGEGWNYNP (12a)
NEFMQTGSYSGP (12b)
YWRNNTPKPMYP (12c)
LTPGQWHMKWVP (12d)

15 residue
HAFIGYDQDPTGKYP (15a)
VPYMPKADKFCMSCP (15b)
KHLRHHQLQVHSHEP (15c)
TLGCMNGTERCLGLP (15d)

20 residue
TYFAVTLTSRIWCLWFYYEP (20a)

WGNGTGLDWKLLTGGISASP (20b)

miscellaneous
VALKLALKLALPRGPRP (S15)
VGAGIGFP (S10)
VPAPIPFP (S7)

The miscellaneous sequences were provided some weeks later as more experimental data became available.

The following sequences were provided for PatG. Modified residues are denoted by X and Y. Italic letters indicate D stereochemistry rather than the usual L. The groupings of the sequences are arbitary and arranged for convenience of plotting the free energies.
set A
VGAGIGFX (B1) X=Pesudoproline
VGAGIGFX (B2) X=Aib
VGAGIGFX (B3) X=Piperidine

set B
VGAGIGX (B4) X=Piperidine
CITXC (B5) X=Propyn-Alanine
GSKLQIDP (B6)

set C
EDWYFDHP (B7)
QENHVFIQFP (B8)
NEFMQTGSYSGP (B9)
LTPGQWHMKWVP (B10)

set D
RTVXMTVX (B11) X=ThH
VTMXVTRX (B12) X=ThH

VTRXVTMX (B13) X=ThH
MTVXTRVX (B14) X=ThH

set E
XSKLQIDP (B15) X=Z-Fmoc
XSKLQIDP (B16) X=Z-TFA
XSKLQIDP (B17) X=Z-Ac
XSKLQIDP (B18) X=Z

set F
XSKLQIDP (B19) X=Z-Fmoc
XSKLQIDP (B20) X=Z-TFA
XSKLQIDP (B21) X=Z-Ac
XSKLQIDP (B22) X=Z

set G
XSKLQIDP (B23/n28) X=Ser-Ac
XSKLQIDP (B24/n29) X=T-Ac
DXYSKLQP (B25) X=Piperidine Y=Cbz
TDXYSKLQP (B26) X=Piperidine Y=Cbz

set H
DCSPAKCSLLCSNP (B27)
VALKLALKLALPRGPRP (B28)
VCGETCVGGTCNTPGCTCSWPVCTRNGLP (B29)

The modified residues are shown below in figure 4.12

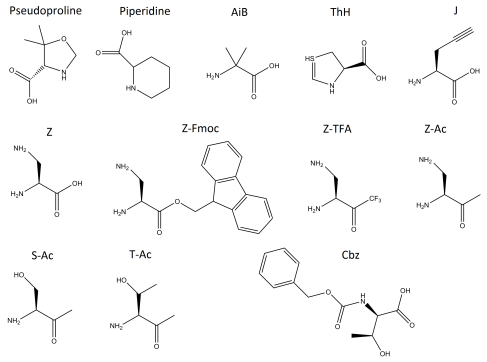

Figure 4.12: modified residues present in the PatG sequences.

The decorrelation procedure was used on the rate constants calculated by BXD was done as described in section 2.1.1 by varying the cuttoff τ_{decor} until the free energy no longer changed. For all peptides τ_{decor} was found to be 120 fs. An example of the procedure is shown in figure 4.13.

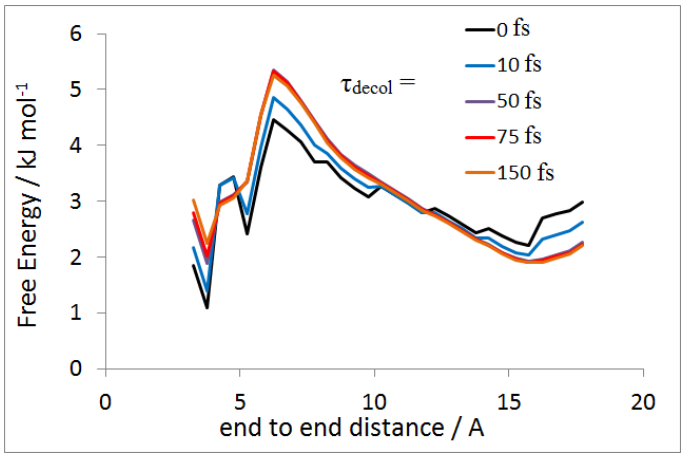

Figure 4.13: decorrelation of the free energy along end to end distance for the peptide S7. A value of 50 fs was chosen as the decorrelation time as the free energy no longer changed when it was increased any further.

4.4 Theory

The following assumptions are made about enzymatic cyclization:

1) - The rate limiting step is the formation of the PCC by the substrate (frame c figure 4.7).

2) - The formation of the PCC is diffusional: the peptide finds a cyclic conformation through its own dynamics without assistance from the enzyme.

3) - The sequence of residues in the substrate is the only factor affecting the rate at which the PCC is adopted.

These assumptions are made to simplify the model. Without assumption 1 the chemical reactions of binding and cyclization would have to be included which would necessitate the simulation of quantum dynamics and electronic structure. This would be prohibitively expensive for any high throughput screening tool. Assumption 2 simplifies the theory further by allowing the peptide to be simulated without the enzyme. This makes the simulations much cheaper. Assumption 3 allows the model to be used for different systems as only the identity of the substrate needs to be considered.

Based on the above assumptions the rate constant of enzymatic cyclization is

$$k_c = P^{\neq} K \tag{4.1}$$

where k_c is the overall rate constant of cyclization, $P^{\neq}$ is the probability of the peptide adopting the PCC and the factor K, which is assumed to be the same for all peptides, accounts for all the stages in figure 4.7 apart from formation of the PCC. The PCC is defined as when the termini are 4 Å away from each other. Because of the assumptions made in the model this probability is the only factor in equation 4.1 which changes from peptide to peptide. $P^{\neq}$ is calculated from the equation

below

$$P^{\neq} = \frac{e^{\frac{-\Delta G_{\neq}}{RT}}}{\sum_{i=1}^{Nbox} e^{\frac{-\Delta G_i}{RT}}} \tag{4.2}$$

where the numerator is the Boltzmann factor for the box covering the transition state at 4 Å and the denominator is the partition function for all the boxes along the reaction coordinate. The assumption that only $P^{\neq}$ varies between peptides is made to increase computational efficiency as it means the peptides can be simulated in isolation. Fortunately this simple treatment increases the usefulness of the study; if the model is accurate and the free energy along the cyclization coordinate is enough to predict enzymatic cyclization then the PatG enzyme only holds one of the peptide and cyclization is diffusional. If the model fails to make accurate predictions then the enzyme must do more to catalyse the cyclization. The model will not only help design cyclisable sequences but the it will illuminate the mechanism of PatG cyclization.

4.4.1 Asessing Model Performance

The model ranks the peptides in order of how likely they are to cyclize, whereas the experimental data only states whether or not that sequence can be cyclized with a binary yes or no. To compute the accuracy of the model it is necessary to compare the same data; the ranking of sequences must be converted into a yes/no prediction. To do this the Jaspars group provided us with the relative numbers of peptides that do and do not cyclize. Once this number is known the sequences are ranked according to their value of $P^{\neq}$ calculated by BXD. If we know the number N of peptide which do cyclize then the top N peptides in the ranking are said to cyclize and the others are said not to. A direct comparison between model and experiment is then possible.

4.5 Results

In general the set of 20 free energies for each peptide are well converged with a standard deviation from the mean of around 1 $kJmol^{-1}$. An example of an average free energy with one standard deviation is shown in figure 4.14.

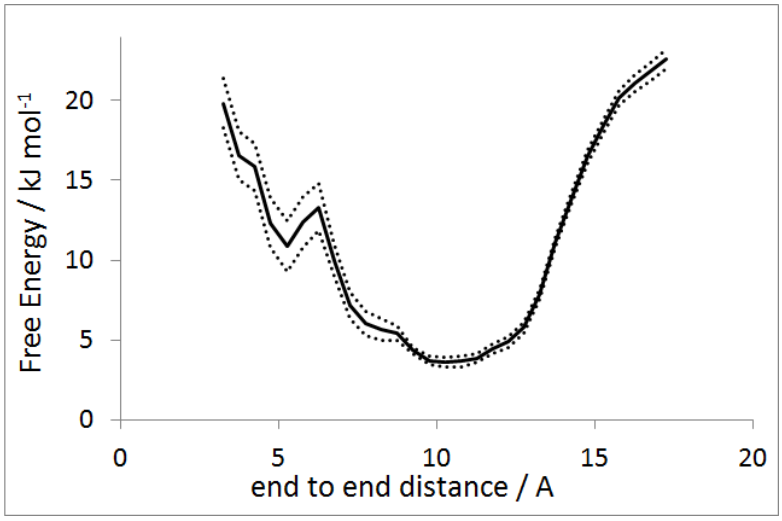

Figure 4.14: example of average free energy showing average of 20 trajectories plus or minus a standard deviation. Taken from the peptide S10 with the EEF1 solvent model.

4.5.1 AcyG

For the AcyG dataset the free energies along the cyclization coordinate are shown below in figures 4.15 to 4.20. The transition state at 4 Å is shown by the grey line.

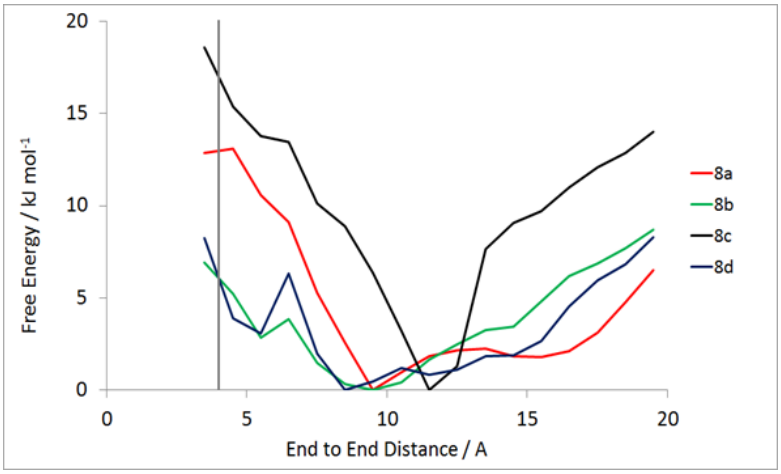

Figure 4.15: Free energy along end to end distance for 8 membered peptides.

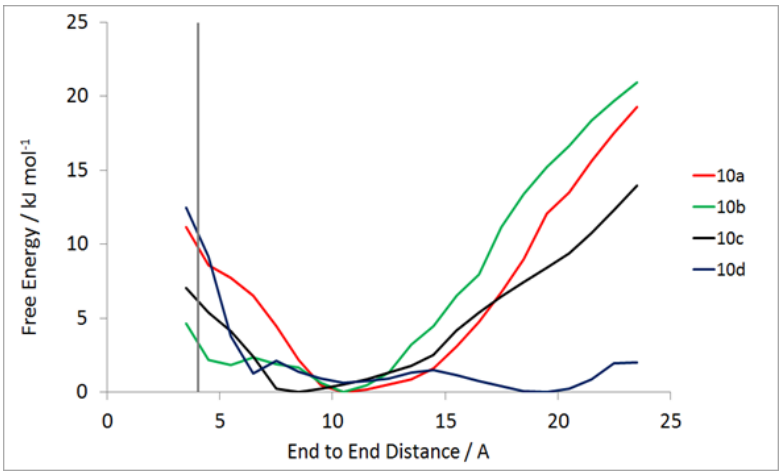

Figure 4.16: Free energy along end to end distance for 10 membered peptides.

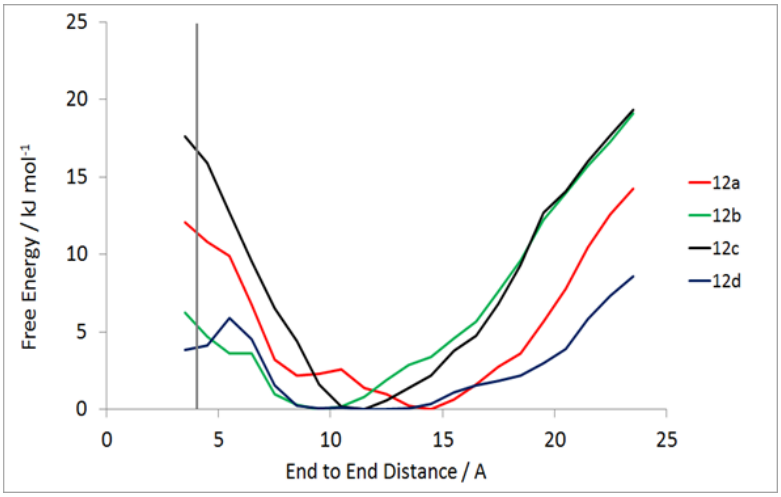

Figure 4.17: Free energy along end to end distance for 12 membered peptides.

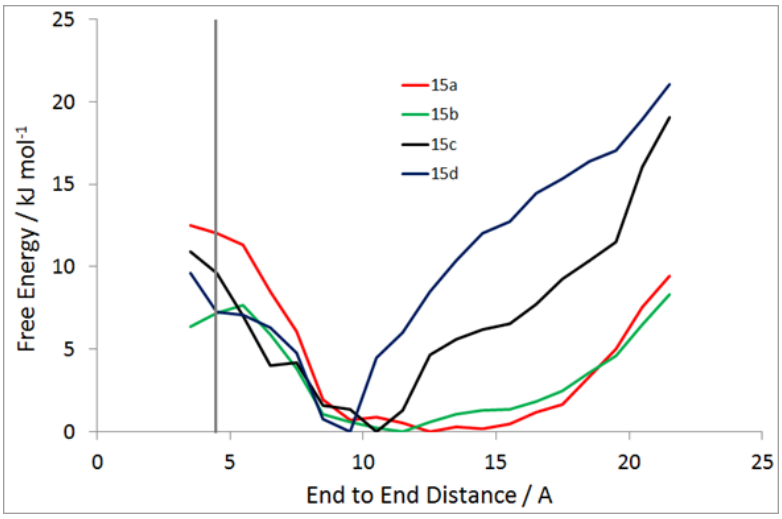

Figure 4.18: Free energy along end to end distance for 15 membered peptides.

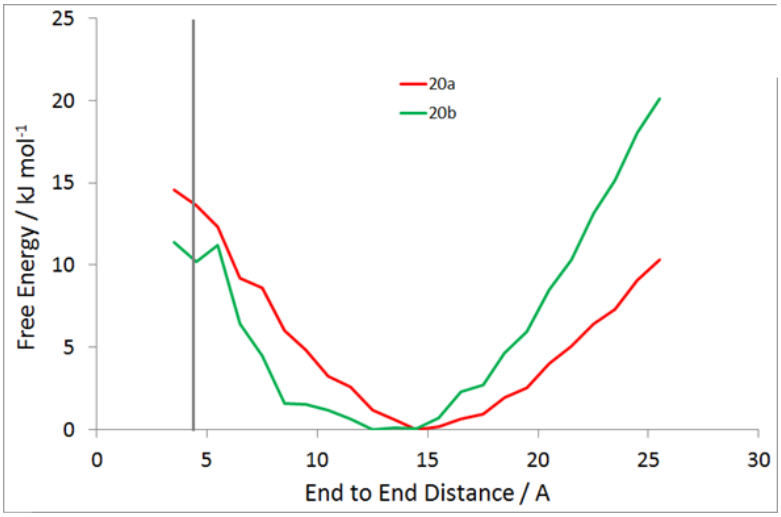

Figure 4.19: Free energy along end to end distance for 20 membered peptides.

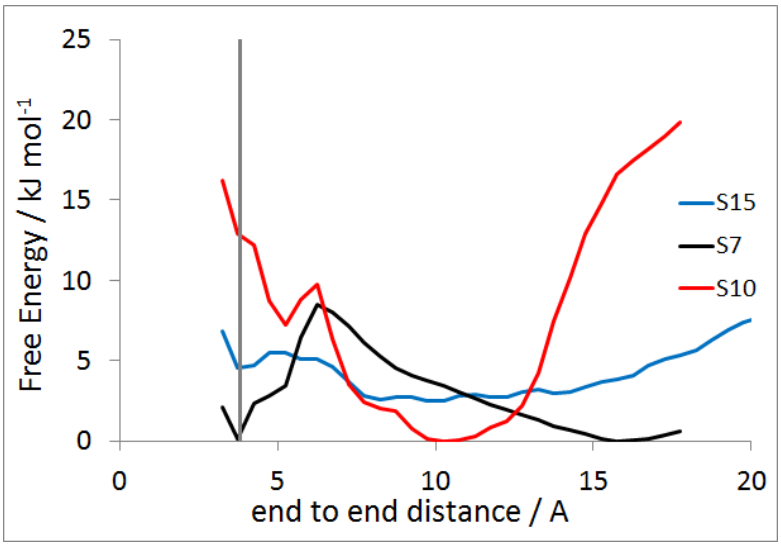

Figure 4.20: Free energy along end to end distance for miscellaneous peptides.

4.5.2 PatG

For the PatG dataset the calculations were carried out both with and without the AYDG tag which is shown in figure 4.7. This was done to

investigate whether the substrate adopts the PCC before or after the AYDG tag is cle aved. Most of the PatG sequences are not yet converged so the results presented here are preliminary. When the ADYG tag was included the model performed very poorly and the results were random. All the cyclization probabilities were ver y low, of the order of 0.00001. With the AYDG tag removed the performance was better, the early results are shown in table 4.2. The free energies along the cyclization coordinate, with the AYDG tag removed, are shown below in figures 4.21 to 4.28.

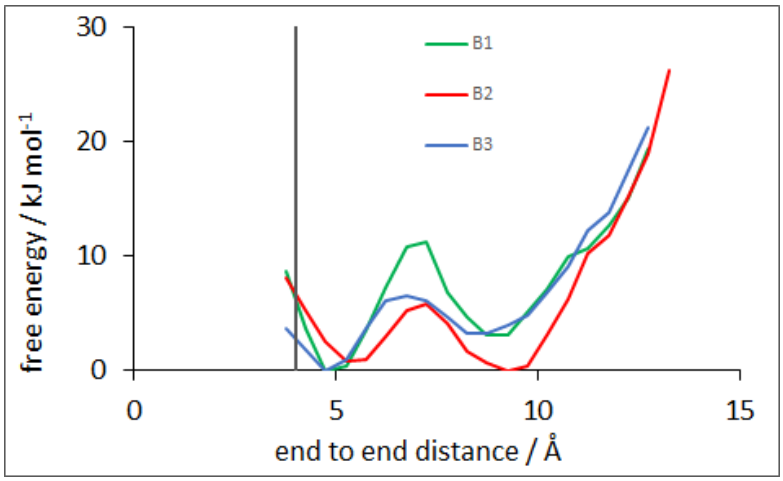

Figure 4.21: Free energy along end to end distance for set A.

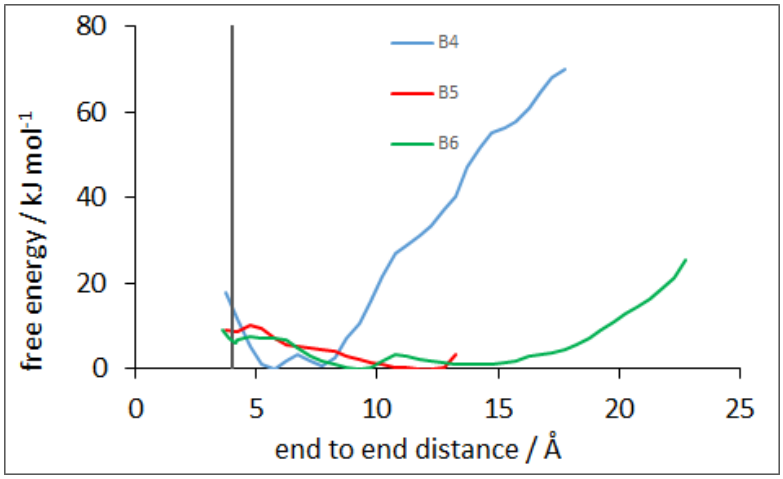

Figure 4.22: Free energy along end to end distance for set B.

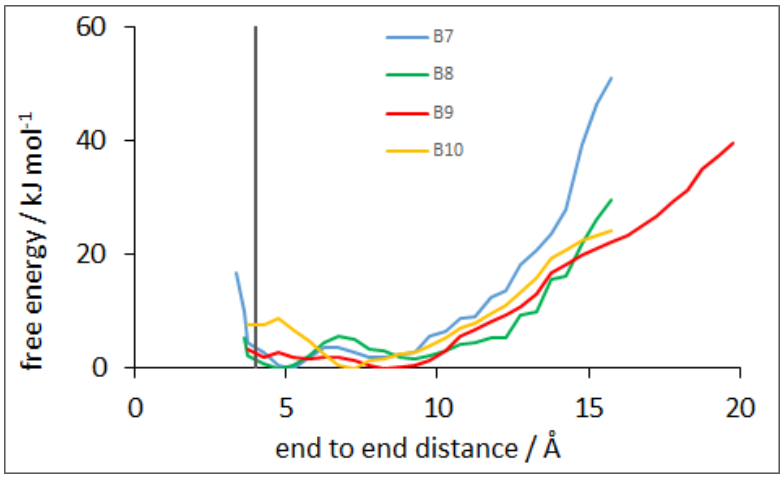

Figure 4.23: Free energy along end to end distance for set C.

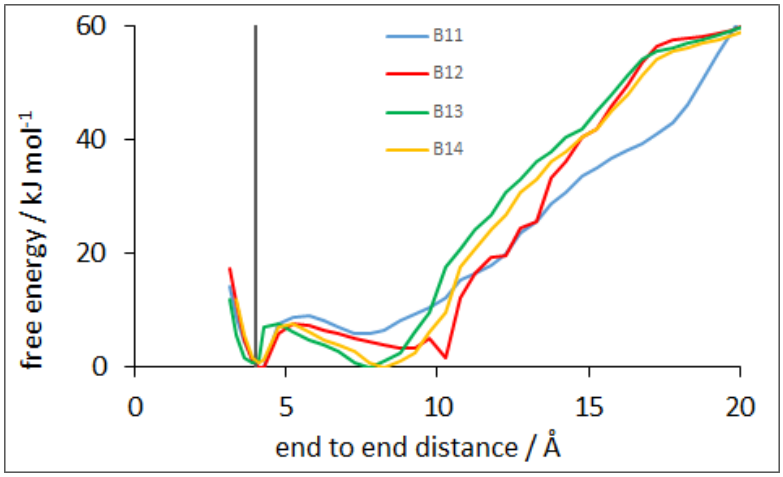

Figure 4.24: Free energy along end to end distance for set D.

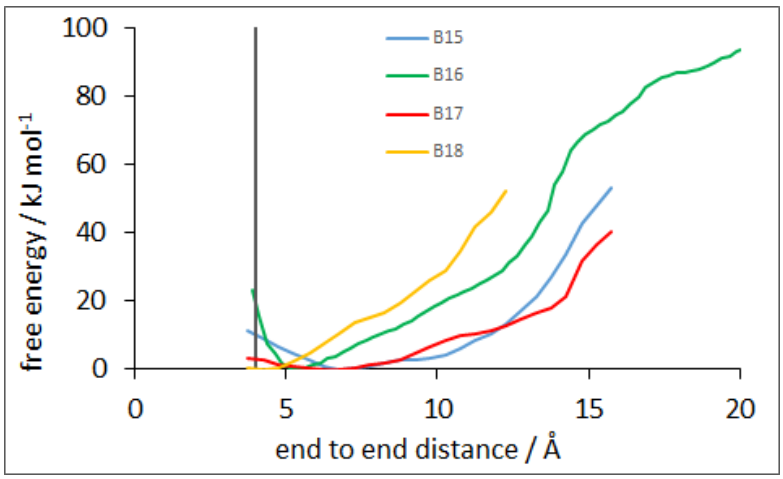

Figure 4.25: Free energy along end to end distance for set E.

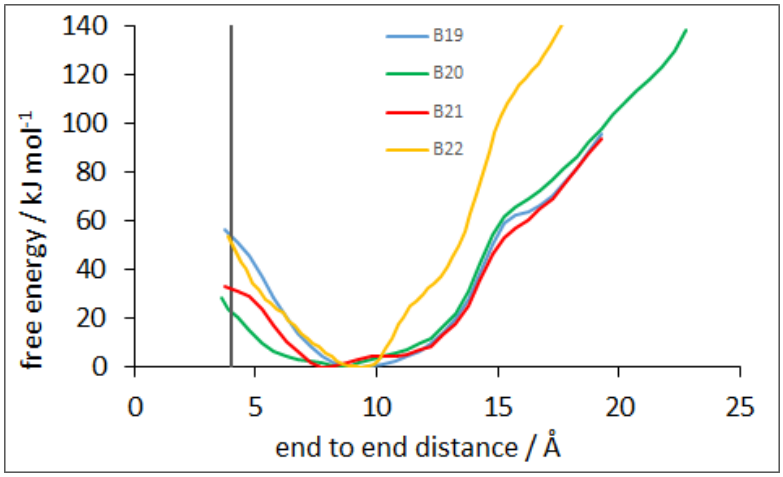

Figure 4.26: Free energy along end to end distance for set F.

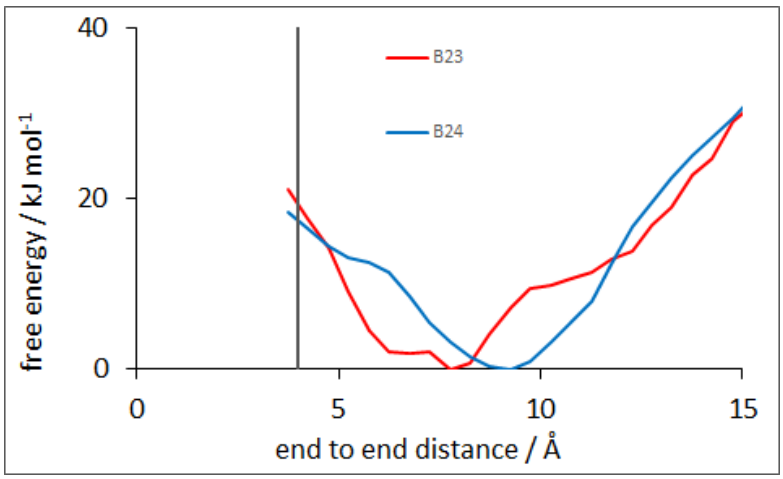

Figure 4.27: Free energy along end to end distance for set G.

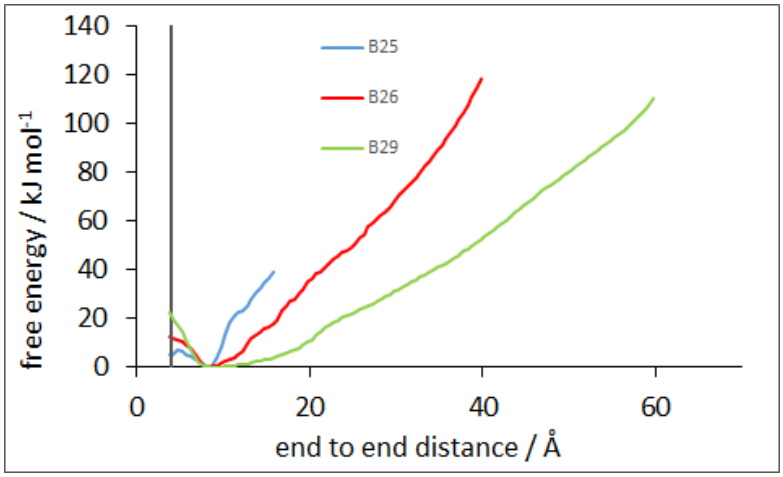

Figure 4.28: Free energy along end to end distance for set H.

For the Acy dataset most free energies have a minimum at aroun 15 Å. This is probably due to the fact that at a medium end to end distance the peptide backbone angles are more favourable and result in less torsion strain than at longer or shorter extensions. For the PatG dataset the free energies of the sequences that do cyclize show a well at much shorter extensions, this could be because of the inclusion of artificial and heterocyclic amino acids such as ThH, which induce a turn in the peptide backbone and reduce the energetic cost of bringing the termini close together. An example of this is the AcyG sequence S7, which unlike mose AcyG sequences, contains several proline residues.

The results of applying equation 4.2 to the free energies shown above are shown below in tables 4.1 and 4.2 for the AcyG and PatG datasets respectivly. The sequences are ranked according to their values of $P^{\neq}$. If N peptides are known to cyclize then the top N in the ranking are predicted as 'yes' by BXD. This prediction is compared to the experimental result to assess if the model was correct.

Table 4.1: results of the Blind Test for AcyG. The number of sequences which cyclize was known to be 12. The accuracy here is 15 out of 21 or 71 %. The experimental result column shows a yes/no answer to whether the data provided by the Jaspars group shows the peptide to cyclize.

Peptide	$P^{\neq}$	Rank	Prediction	Experimental result	BXD correct?
10b	0.067	1	yes	yes	yes
S7	0.042	2	yes	yes	yes
8d	0.033	3	yes	yes	yes
12b	0.024	4	yes	yes	yes
15d	0.023	5	yes	yes	yes
8b	0.022	6	yes	yes	yes
12d	0.018	7	yes	no	no
10c	0.017	8	yes	yes	yes
15b	0.007	9	yes	no	no
15c	0.006	10	yes	no	no
10a	0.005	11	yes	yes	yes
20b	0.002	12	yes	yes	yes
10d	0.002	13	no	no	yes
12a	0.002	14	no	yes	no
15a	0.001	15	no	yes	no
S15	0.001	16	no	no	yes
8a	0.001	17	no	yes	no
8c	0.001	18	no	no	yes
S10	0.001	19	no	no	yes
20a	0.001	20	no	no	yes
12c	0.000	21	no	no	yes

Table 4.2: results of the Blind Test for PatG. The number of sequences which cyclize was known to be 11. The accuracy here is 23 out of 29 or 79 %.

Peptide	$P^{\neq}$	Rank	Prediction	Experimental result	BXD correct?
B13	0.849	1	yes	yes	yes
B11	0.838	2	yes	yes	yes
B12	0.583	3	yes	yes	yes
B18	0.554	4	yes	yes	yes
B14	0.338	5	yes	yes	yes
B8	0.191	6	yes	yes	yes
B4	0.126	7	yes	yes	yes
B7	0.092	8	no	yes	no
B9	0.090	9	yes	yes	yes
B17	0.083	10	no	yes	no
B1	0.074	11	yes	yes	yes
B27	0.024	12	no	no	no
B3	0.019	13	yes	no	no
B10	0.019	14	no	no	yes
B15	0.013	15	no	no	yes
B5	0.008	16	yes	no	no
B6	0.005	17	no	no	yes
B2	0.002	18	no	no	yes
B28	0.001	19	no	no	yes
B16	0.000	20	no	no	yes
B25	0.000	21	no	no	yes
B26	0.000	22	yes	no	no
B20	0.000	23	no	no	yes
B23	0.000	24	no	no	yes
B29	0.000	25	no	no	yes
B24	0.000	26	no	no	yes
B21	0.000	27	no	no	yes
B22	0.000	28	no	no	yes
B19	0.000	29	no	no	yes

4.6 Discussion

Across both datasets BXD correctly predicted the cyclization of 38 out of 50 peptides, achieving an accuracy of 76 percent. If the model was

random and guessed yes or no then the chance of correctly guessing 38 out of 50 would be 0.0002 %. There is a strong correlation between the free energy along the cyclization coordinate and the ability of the enzyme to cyclize the peptide which suggests that the rate limiting step of cyclization is the peptide adopting the PCC. This suggests that the enzyme does not actively bring the ends of the peptide together, rather it waits for cyclization to happen diffusively.

This result supports the assumptions made in equation 4.1 that only the rate at which the peptide adopts the PCC varied from sequence to sequence, the binding, cleavage and release steps[143] happening at approximately the same rate between sequences. When the AYDG tag was added to the simulation the model performed very poorly. Performance improved dramatically when the tag was removed. This strongly implies that the sequence of events proposed in figure 4.7 is correct, i.e. the AYDG tag is cleaved before the peptide adopts a cyclic conformation.

It should be noted that the cyclization probabilities for AcyG are all low. Besser *et. al.* report cyclization probabilities of between 0.1 and 0.6 for several peptides. This difference is probably due to the fact that the peptides studied here are all L sequences with no turn inducing residues in the backbone. For the PatG dataset the probabilities of cyclization are much higher, in the range given by Besser *et. al.*. The PatG sequences contain many Proline residues and heterocycles which are not present in the AcyG dataset. Hence the higher cyclization probabilities of the PatG sequences is concurrent with the theory that heterocycles and Proline residues increase the rate of cyclization by inducing a turn in the backbone.[117–120]

4.6.1 Factors Affecting Cyclization

What factors affect the ability of a cyclic peptide to cyclize? Although all the discussion in the literature focuses on what makes a peptide cyclize in traditional non enzymatic synthesis the conclusions might still be valid for the case of enzymatic cyclization, as the blind test above concluded that cyclization probability correlates with the properties of the independent precursor peptide. In this section the results of the BXD simulations will be checked against the conclusions drawn in the literature. This analysis will only be performed with the structures in the AcyG dataset as the literature studies focus on peptides comprised of the naturally occuring amino acids..

Some studies[116;137] report that chain length is a factor; longer sequences are easier to cyclize because the total flexibility is higher and the termini can come close together without creating as much strain in the backbone. To see whether or not this is the case for the peptides studies here a plot of cyclization probability against chain length is shown in figure 4.29.

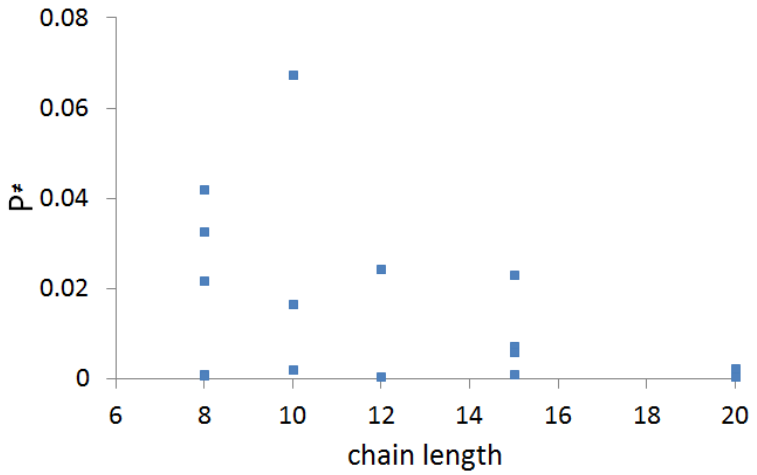

Figure 4.29: transition state probability against chain length for successfully predicted peptides.

It is clear from figure 4.29 that there is no correlation between chain length and cyclization probability. While longer peptides are indeed more flexible this might not have as large an effect as expected as a longer peptide will have more conformational space to search before it arrives at the cyclization transition state. This is equivalent to an entropy penalty (transition states are less likely to be found as overall phase space volume is higher) offsetting the enthalpy bonus of the torsion angles in the backbone being more favourable in the transition state.

Yongye *et. al.* found that hydrogen bonding between the peptide termini in the transition state lowered its energy and made cyclization more likely.[133] The occupancy of each hydrogen bond was calculated, that is the fraction of trajectory frames for which the hydrogen bond was present. This occupancy was found to correlate with the experimental rate of cyclization. To investigate whether or not this is the case here the trajectories from the BXD simulations were analysed. The structures of the cyclization transition states for peptides that were correctly predicted are shown below in figure 4.30 for those that do cyclize, and by figure 4.31 for those that do not. Structures were calculated by taking all the trajectory frames for which the termini were 4 Å apart and averaging them. Side chains were not included in the calculation of the average and are not shown in the diagrams because their motion is fast compared to backbone movement which is more important for conformational change. Clustering analysis was performed to check the multiplicity of conformations at the transition state. For all the trajectories one transition state conformation was present much more frequently than any other, this conformation was used to calculate the average. Hydrogen bonds are shown by purple lines.

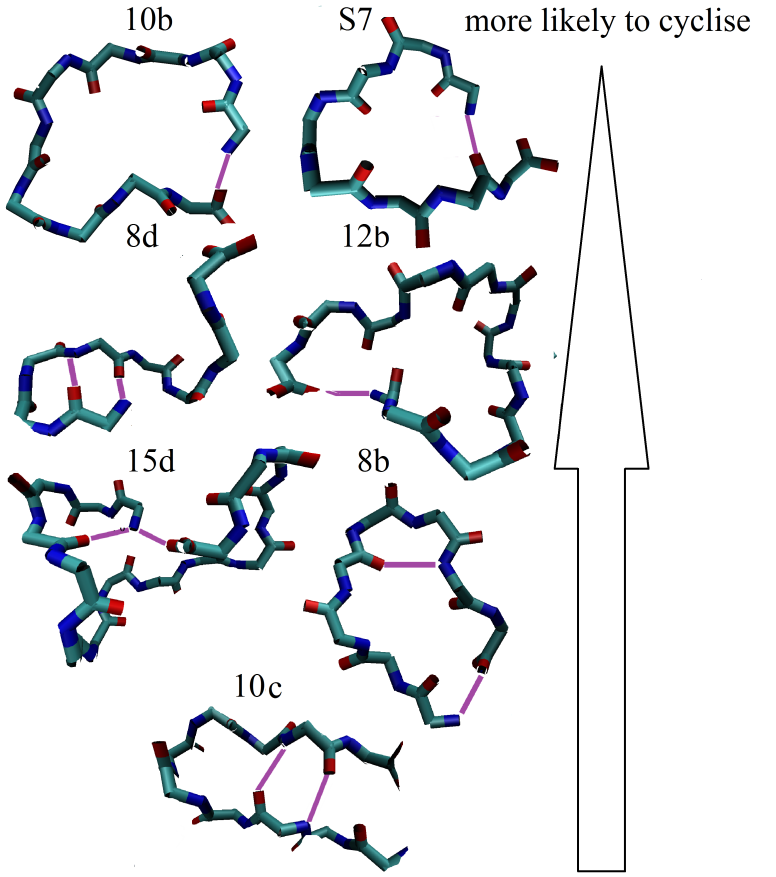

Figure 4.30: transition state structures of peptides that were correctly predicted to cyclize.

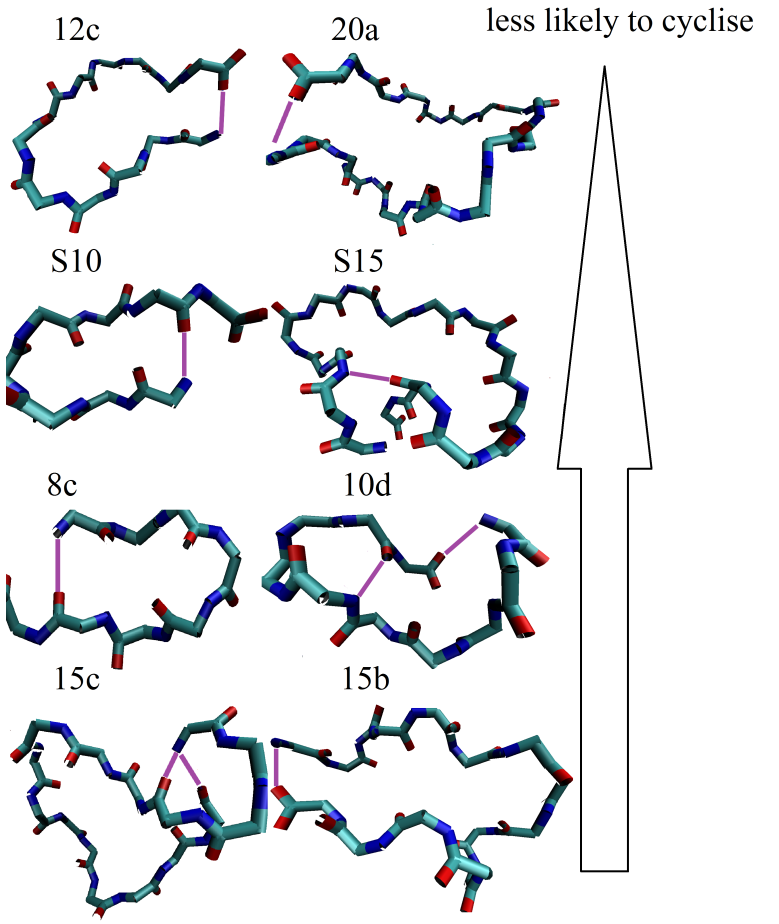

Figure 4.31: transition state structures of peptides that were correctly predicted to cyclize.

Hydrogen bonds were defined as existing when the donor and acceptor atoms are within 3 Å of each other, and form an angle of less than 20 degrees with the proton, which is consistent with the accepted definition in the literature.[138] The occupancy of each hydrogen bond was calculated as the fraction of frames corresponding to a transition state in which the bond was present, rather than over the whole trajectory as was done by Yongye *et. al.* This was done because Yongye *et. al.* used unconstrained MD simulations where the ensemble of conformations reflected the true Boltzmann weighted ensemble, with a BXD simulation the trajectory spends more time in high energy regions so

hydrogen bond occupancy averaged over the whole trajectory is not meaningful. Figure 4.32 shows the calculated cyclization probability against the occupancy of the inter terminal hydrogen bond.

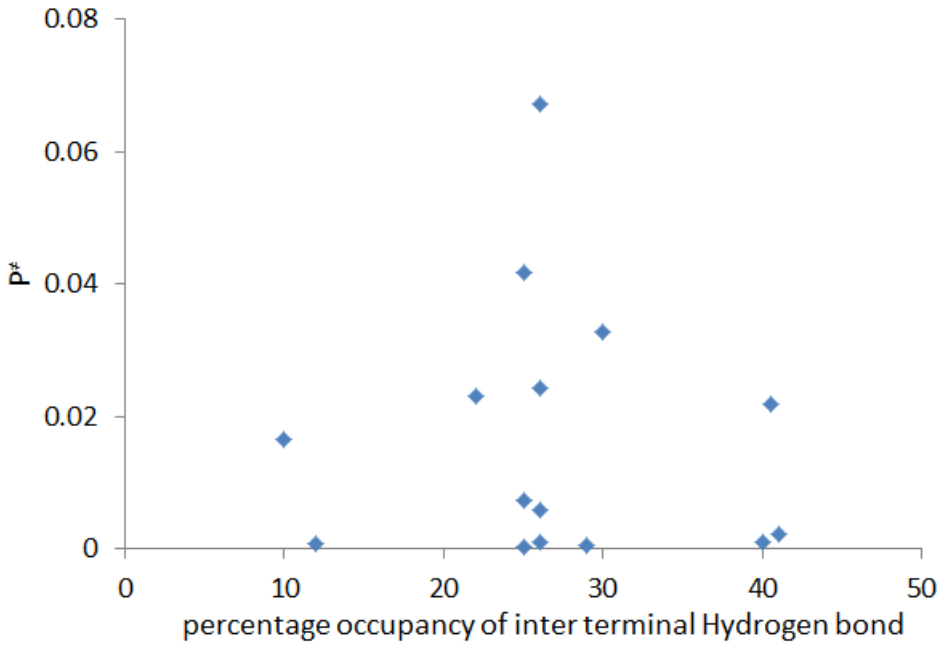

Figure 4.32: cyclization probability against occupancy of inter terminal hydrogen bond.

It is clear from figure 4.32 that there is no correlation between the strength of the inter terminal hydrogen bond and the cyclization probability. In their study of the cyclization rates of a large library of linear peptides[137] Thakkar *et. al.* reported that peptides with large side chains on their C termini cyclize less readily. Figure 4.33 shows cyclization probability against the molar volume of the side chain on the C terminus.

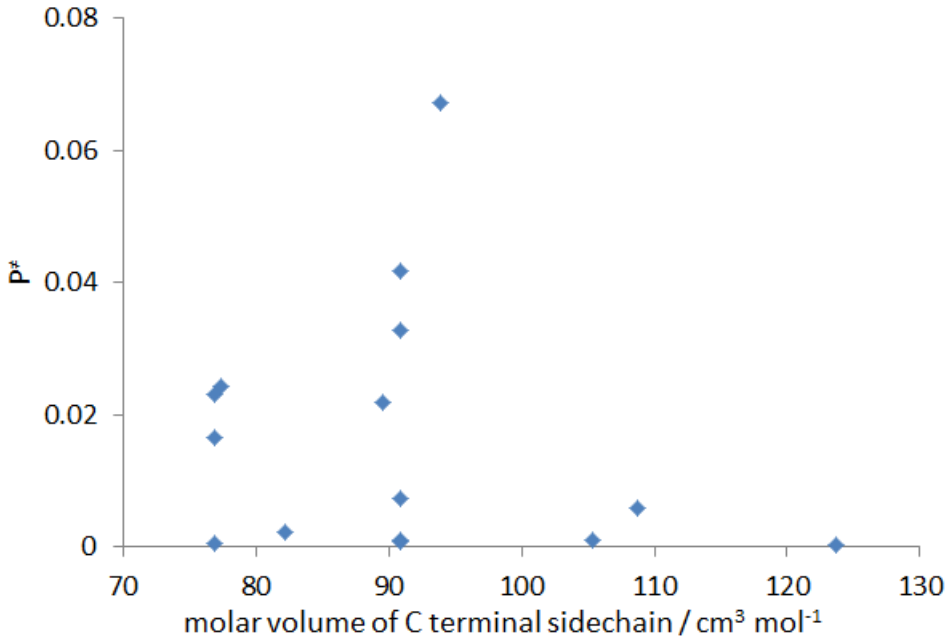

Figure 4.33: cyclization probability against molar volume of side chain on C terminus. Volumetric data taken from reference [139].

Again there is no real correlation between the molar volume of the C terminal side chain and the cyclization probability. It might be said with caution that amino acids with the largest side chains should not be placed on the C terminus of the linear precursor but there are only three data points for these larger side chains. However it is plausible that the steric bulk of these side chains prevents the terminal carbon and nitrogen atoms from coming within reactive range of each other. Thakkar *et. al.* also report that many of the sequences that do not cyclize are rich in Lys, Arg and Thr or contain the motifs ArgArg, LysLys, ArgLys, LysArg, ThrThr, ThrLys or LysThr. This correlation is not present in the sequences studied here; the sequences alone give no hint of whether or not they will cyclize.

Cavelier-Frontin *et. al.* examined the values of the backbone angles in the transition states of the peptides they studied. The distribution

of the backbone angles Ψ and Φ (see figure 4.34) is often plotted in a Ramachandran plot (see figure 4.35).

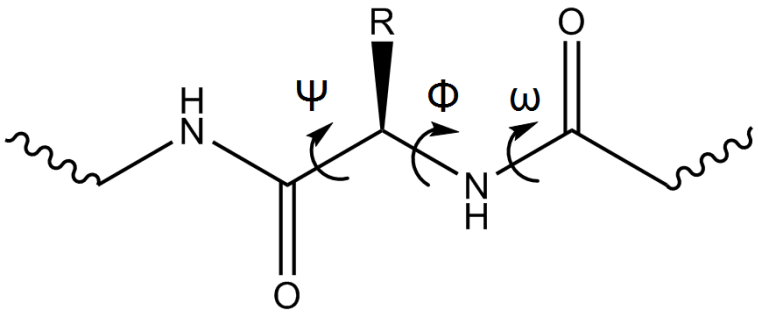

Figure 4.34: the angles used to create Ramachandran space. The third angle ω is not included as it is usually close to 180 degrees due to its partial double bond character.

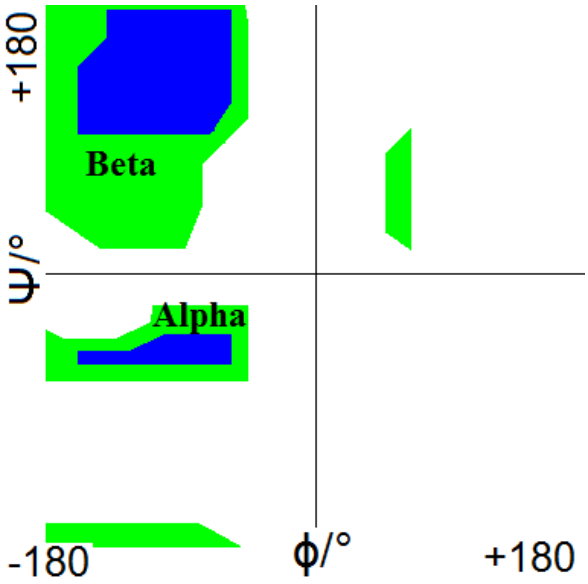

Figure 4.35: the allowed regions of Ramachandran space. Blue areas are highly favoured and green areas moderately so. Certain regions correspond to alpha helices or beta sheets.

Each amino acid residue contributes one point to the Ramachandran plot. Certain areas of Ramachandran space correspond to beta sheets or alpha helices. The Ramachandran plots of the transition states

of the peptides which were correctly predicted to cyclize are shown in figure 4.36 and in figure 4.37 for those that were correctly predicted not to.

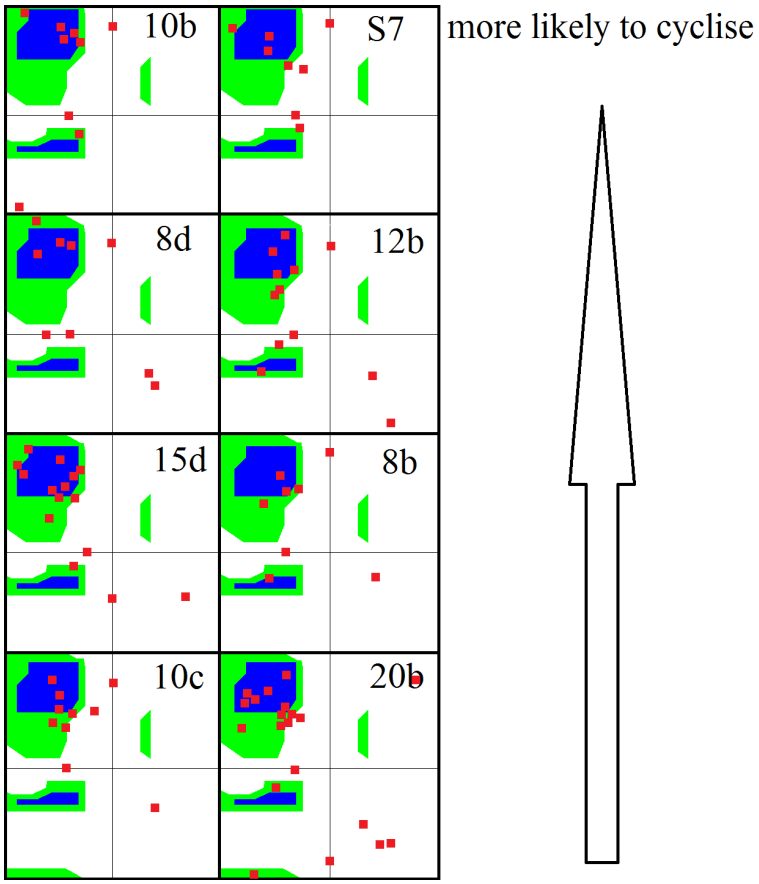

Figure 4.36: Ramachandran plots for peptides correctly predicted to cyclize.

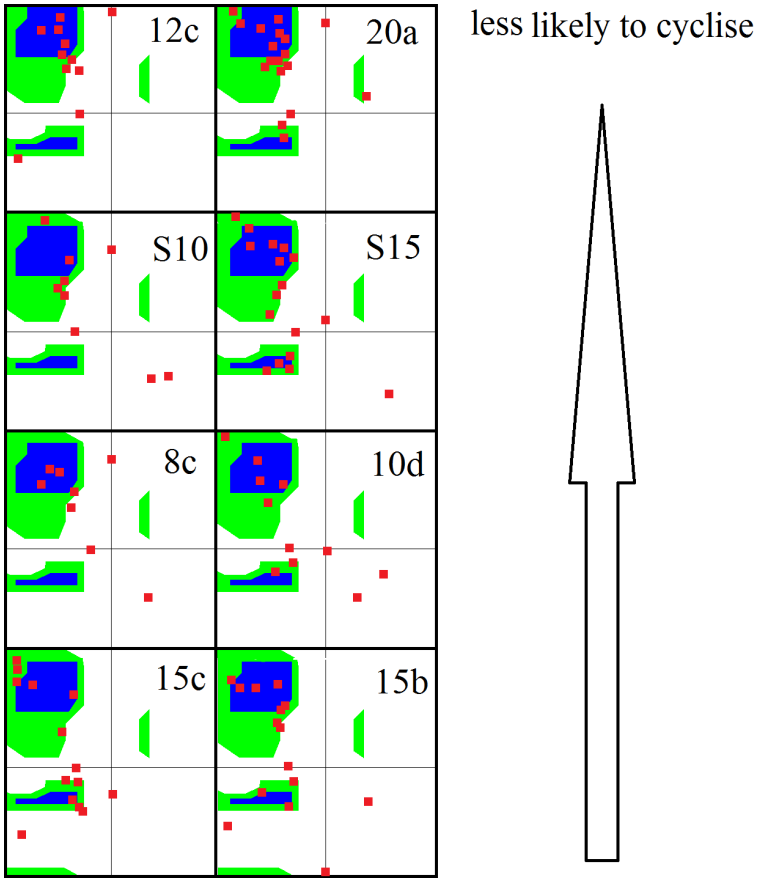

Figure 4.37: Ramachandran plots for peptides correctly predicted not to cyclize.

Cavelier-Frontin *et. al.* noticed that the peptide that cyclized most readily had a transition state which featured backbone angles that occupied the centre of the Ramachandran plot. Inspection of figures 4.36 and 4.37 shows that this is not the case in this study; there appears to be no link between the cyclization probability of a peptide and the Ramachandran plot of its cyclization transition state; the transition state Ramachandran plots of sequences that do and do not cyclize are very similar.

In a 2013 study[140] Diadone *et. al.* showed that the cyclization ability of peptides depended on the structure of the extended conformation.

Peptides that cyclize more readily were found to have Hydrogen bonds that stabilised a turn in the lowest free energy open conformation, making cyclization easier. This conclusion was drawn for larger peptides of around 20 or more amino acids which consisted of repeating GlySer units. To see if this conclusion can be carried over to shorter peptides containing different amino acids the lowest free energy, or extended, structures of the peptides are shown in figures 4.38 and 4.39.

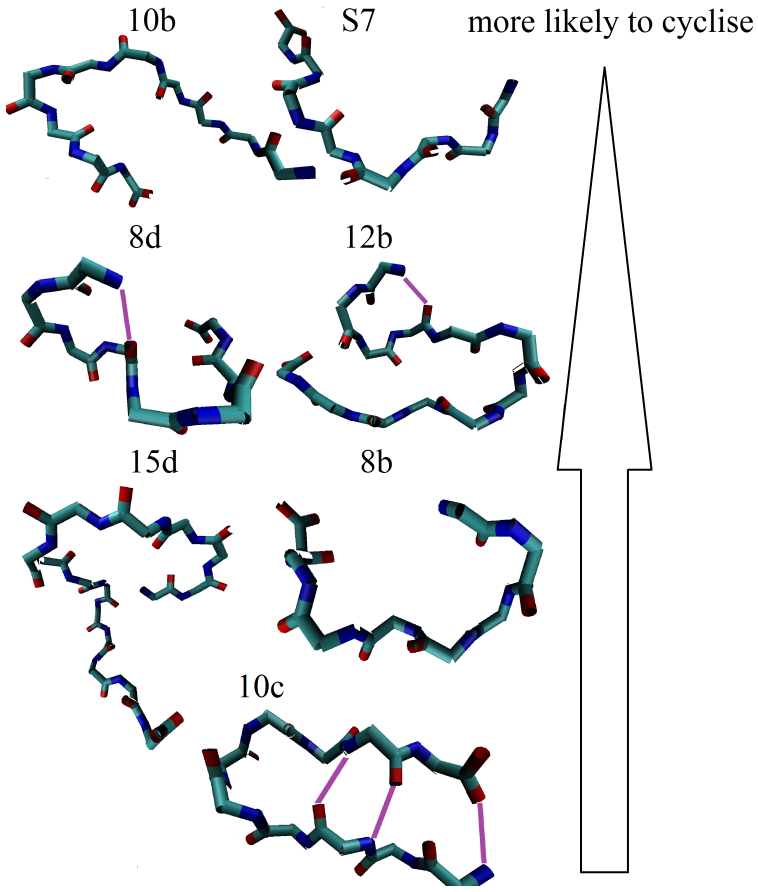

Figure 4.38: extended structures of peptides that were correctly predicted to cyclize.

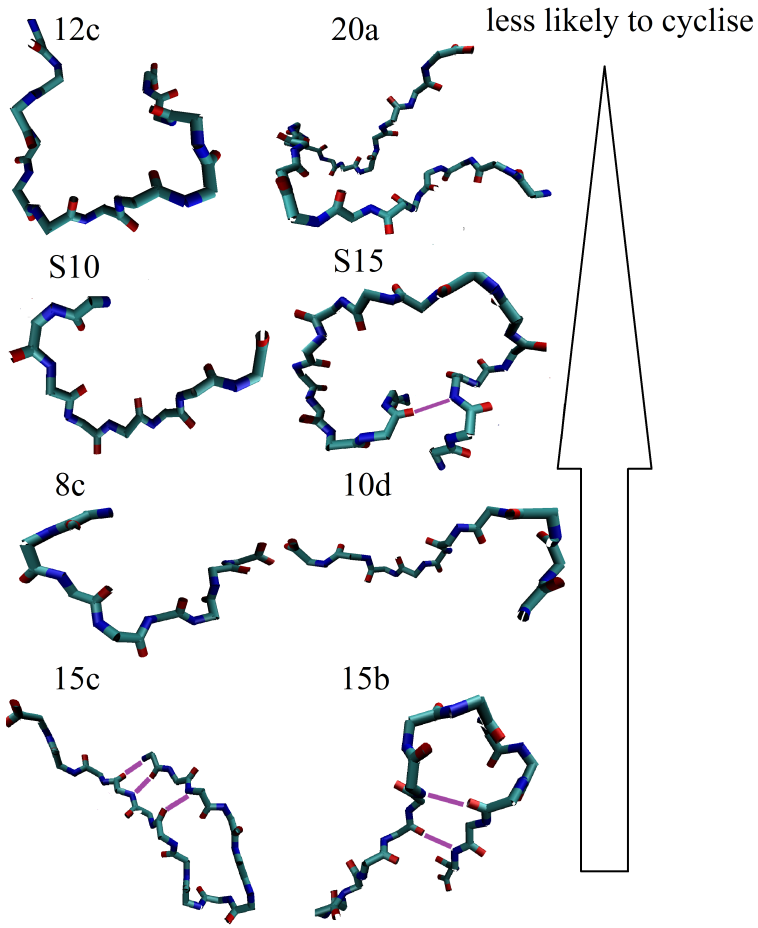

Figure 4.39: extended structures of peptides that were correctly predicted not to cyclize.

Figures 4.38 and 4.39 show that there is no correlation between the probability of cyclization and hydrogen bond stabilised turns in the extended conformation. Peptides 10c, 15b and 15c show this characteristic but they are poor to middling in terms of cyclization.

After having examined the results of this study in light of the conclusions drawn in the literature it seems that we have not yet been able to find any rule of thumb to explain the different rates of cyclization. Rather than inspecting the transition states and extended structures it is necessary to extensively sample the conformational space of a peptide

to predict its cyclization probability. No simple dependence on sequence could be made out from this dataset however this is not a problem as the calculations are cheap (roughly one CPU day per peptide per trajectory) hence high volume *in silico* screening is possible.

The performance of the model was good given its simplicity in modelling only the peptide diffusion rather than including any of the enzyme of the chemical reactions. Given the models simple implementation with a cheap solvent model and the presence of the substrate only, an accuracy of over 75 percent can be seen as a success. The incorrect predictions could be due to limitations in the assumption that the only factor affecting cyclization probability is the ability of the linear precursor to adopt a cyclic conformation and that the enzymes play no active role in assisting this. Despite this it seems that these assumptions are mostly correct.

The inaccuracy in the model could also come from the parameter chosen for the distance between the ends of the peptide at the transition state. The 4 Å transition state distance is close to that used in the literature[141;143] of around 3 Å, however it is much harder to sample the conformational space at a separation of 3 Å than at 4 Å due to the much increased free energy at closer distances. The extra Angstrom of separation should not make much difference as the transition state found by a classical forcefield cannot accurately reflect that of a real chemical reaction where bonds are breaking and being formed.

4.7 Conclusion

The work presented in this chapter can be summarised as follows:

1) - BXD can accurately predict whether or not a given sequence can be cyclized by the Acy or PatG enzyme. The model achieved an accuracy of 76 %.

2) - The rate limiting step of enzymatic cyclization with PatG and AcyG is the substrate peptides conformational search to adopt a cyclic conformation (PCC).

3) - The peptide is not assisted by the enzyme while it adopts the PCC, it is a diffusional process.

4) - Cleavage of the AYDG tag occurs after the peptide is bound to the enzyme and before the PCC is adopted.

5) - The rates of binding, AYDG cleavage and product release do not vary significantly between substrates.

6) - Based on analysis of the AcyG dataset there is no simple rule of thumb has been found to predict whether or not a sequence can be cyclized.

7) - Predictions were made quickly and cheaply, each trajectory converged in between 1 and 5 CPU days.

4.8 Further Work

Further work will focus on making predictions for more sequences and for different enzymes in order to increase the scope of the model and

test the validity of its assumptions on more systems. The insights about the mechanism of cyclization are not obtained directly and are somewhat speculative. More sophisticated BXD simulations, including the enzyme or other stages of cyclization, could be undertaken to carefully check the conclusions drawn here.

We are also working with experimentalists who are now trying to measure the rates of PatG cyclization rather than providing a simple yes/no answer. This will allow direct comparison of the BXD results with the experimental rankings as the experimental data will consist of a list of peptides ranked in order of their rates of cyclization.

Chapter 5

Conclusion and Further Work

BXD has been shown to be a useful tool for efficient simulation of long timescale processes. Insights have been into the mechanical unfolding of proteins, using what we believe to be the first all atom simulation of AFM protein pulling without the use of high artificial pulling forces. BXD has been shown to be a very powerful tool, as experimental data was accurately reproduced for the VC experiments and insights were gained into the mechanical unfolding of three protein domains. All this was done with trajectories that ran on a single CPU core for between one and two weeks. Accurately simulating a biomolecular process which occurs over a time scale of miliseconds to seconds would usually require specially built supercomputers, convoluted algorithms or very large additional forces.

BXD has also been used to build a fast and accurate *in silico* screening tool for the production of medicinal cyclic peptides, as well as to shed light on the mechanism of enzyatic peptide cyclisation. This shows that BXD can be applied to areas of real world importance such as the search for novel antibiotics. Because of the high speed at which calculations converge, roughly two days on a single core, and the fact that the high accuracy was maintained accross a dataset of 50 sequences, the BXD screening tool presented here is applicable to the pharmaceutical industry where very large libraries of candidates must be checked.

BXD is a novel technique in that, for these applications, both kinetic and thermodynamic information is provided quickly and on modest hardware, without the need for powerful computers or convoluted biasing routines or modification of the potential energy landscape. Further work will focus on the following three areas:

1) - AFM protein pulling will be simulated with concatamers rather than individual domains, and explicit water will be used.

2) - Simulation of peptide cyclisation will continue with more datasets from different enzymes.

3) - BXD will be added to other MD packages such as GROMACS in order to benafit from GPU acceleration.

Appendix A

Worked Example of Free Energy Calculation

This chapter gives a detailed account of how to calculate free energy using BXD. The example chosen is the small peptide B11 (see Chapter 4) with end to end distance as a reaction coordinate. For the input files and other analysis scripts used in BXD contact the author at jjbooth1989@yahoo.com. Many of the processes outlined here are close to being automated so for future studies BXD will be much faster and easier to carry out.

A.1 Box Placing

Placing the boundaries along the reaction coordinate (see Chapter 2) is a balance between computational efficiency and the quality of the sampling. If the boxes are too large then the simulation will be slow as the trajectory will still have to travel to higher regions of free energy without assistance from the ratched effect of the boundaries. If the boxes are too small then the kinetics and thermodynamics obtained will not be valid as the trajectory will not relax between collisions and the decorrelation proceedure, which ensure that the mathematical foundation of BXD is valid, will not work. In general it is best to make the boxes as

large as possible without sacrificing the speedup provided by BXD.

To determine the best arrangement of boundaries, a quick BXD run is carried out where the boxes are spaced widely apart. The number of collisions required to pass into the next box is set to a very low number such as 5 in order to quickly determine the optimimum boundary locations. If the trajectory does not quickly cycle through the reaction coordinate then the boxes are too large, as anything other than a fast cycle time will increase to a much slower time when the numbner of collisions requiured is increased from 5 to a much larger number for the production run.

In this example, initial boundaries were placed at intervals of 1 Å, from 4 Å to 26 Å, with 5 collisions required to pass into the next box. These boundaries were too widely spaced as a trace of the reaction coordinate against simulation time (figure A.1) shows that the trajectory stalls and does not quickly complete a cycle of the reaction coordinate.

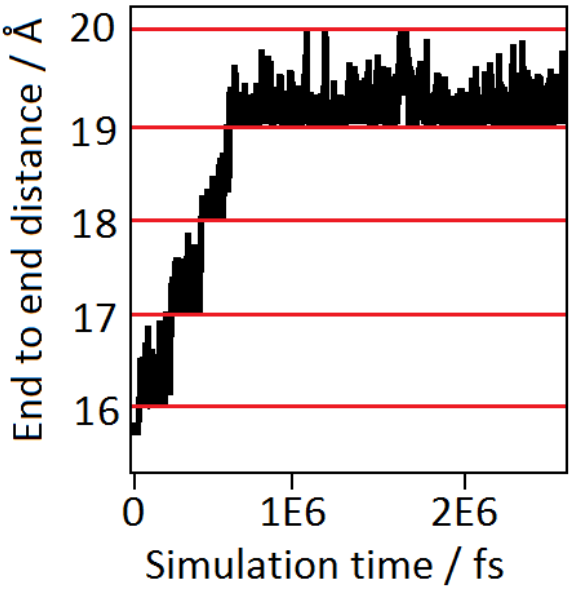

Figure A.1: In this box placing BXD run the boundaries (red lines) are too far apart resulting in a stalled simulation, as the trajectory (black line) does not leave the upper box.

The box sizes are reduced in areas where the trajectory stalls and the box placing runs are repeated until the trajectory can cycle through the reaction coordinate quickly. In this case the final boundries went in intervals of 0.5 Å from 3.5 Å to 6 Å, and from 22 Å to 26 Å, with the region in between 6 Å and 22 Å having boundaries at intervals of 1 Å. In general the regions of denser boundaries represent areas of steeper free energy where the trajectory requires more assistance. The trace of reaction coordinate against time for the final satisfactory arrangement of boundaries is shown below in figure A.2.

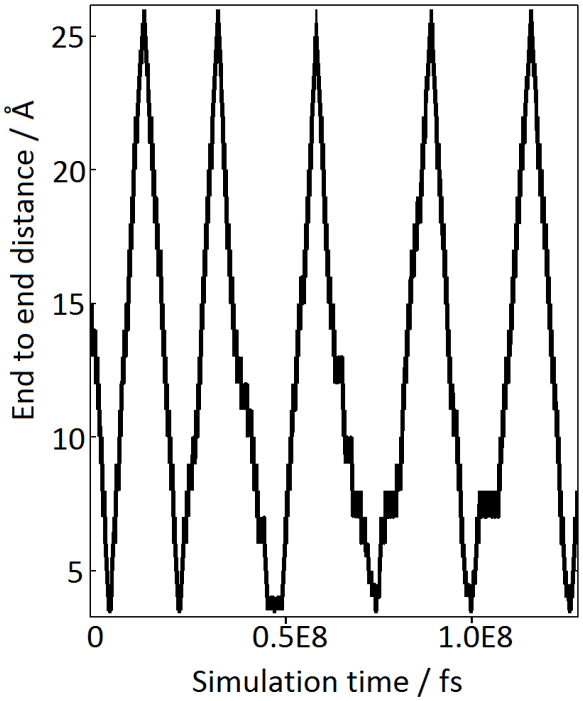

Figure A.2: a good distribution of boundaries allows the reaction coordinate to be sampled efficiently.

Care must be taken not to make the boxes too small, which would prevent the decorelation proceedure (see Chapter 2) from working and undermine the results provided by BXD. To check if this is the case the trajectory can be inspected to see how frequently it collides with the boundaries of each box. Figure A.3 shows an example where the boxes

are far enough apart for decorelation to be successful and where they are so close together that the trajectory rattles between the boundaries without any ergodic sampling of the phase space within the box.

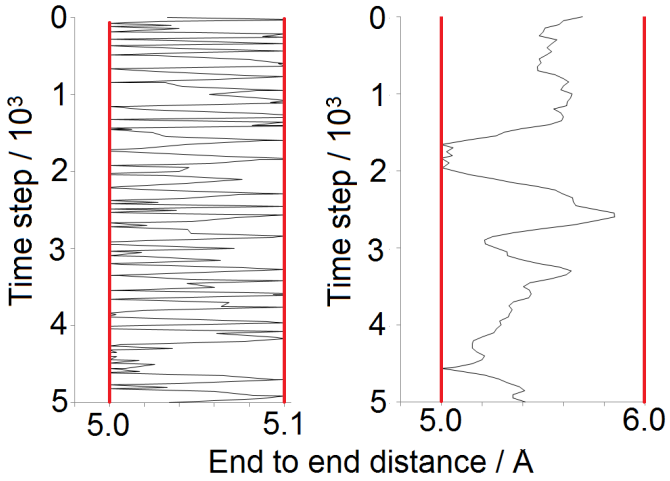

Figure A.3: if a box is too small (left) then the trajectory hits the boundaries (red) too frequently and cannot equilibrate within the box. Decorrelation here is impossible as there are no FPTs longer than the characteristic decorrelation time. If the box is large enough (right) then decorrelation is possible as the trajectory can explore the box and come to equilibrium in between collisions with the boundary.

Once an efficient distribution of boundaries has been found the number of collisions should be increased from 5 to around 1000 for the production run. As of 2016 this proceedure has been fully automated by the Glowacki group in Bristol and boundaries are adjusted on the fly to optimise the efficiency of the simulations.

A.2 Collision Threshold

The number of collions required to pass from one box to the next must be high enough for the trajectory to equillibriate and sample each box well, but not so high that the trajectory spends too long in each box and does not cycle through the reaction coordinate many times, which would limit the range of different pathways that the trajectory samples.

The ideal value for this parameter has been empirically determined to be around 1000.

A.2.1 Decorrelation

Once the boundaries have been placed and a production run is underway the decorrelation time for the system needs to be found. After the trajectory has completed around 20 cycles through the reaction coordinate the free energy is calculated with a decorelation time of 0 fs. The free energy is calculated again with an increased decorelation time until it no longer changes. The decorelation time τ at which the free energy has converged is taken as the decorelation time of the system and is used to calculate the final free energy of the production run. This process is illustrated below in figure A.4 for the peptide used in this example. For a theoretical description of decorelation see Chapter 2.

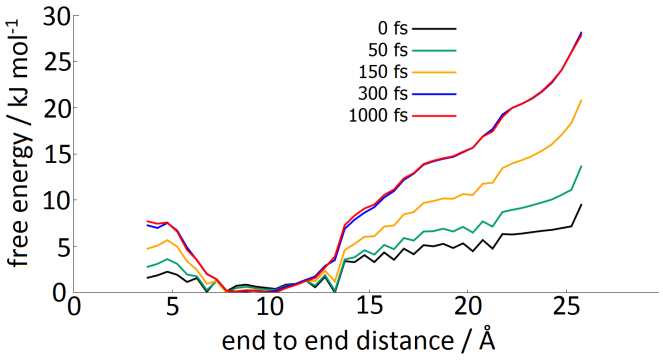

Figure A.4: the free energy is calculated at different decorrelation times until it no longer changes. The time τ at which the free energy no longer changes is taken to be the decorrelation time of the system, in this case 300 fs.

A.3 Convergence

To check whether a simulation has converged, while a production run is still progressing the free energy is calculated after a certain time has passed, usually one day, and again after double that period. If there is no significant change between the two free energies then that trajectory is said to be converged. If the free energy has not yet converged then the test is repeated at longer time intervals until convergence is reached. Figure A.5 shown an example of a free energy which has converged and one which has not.

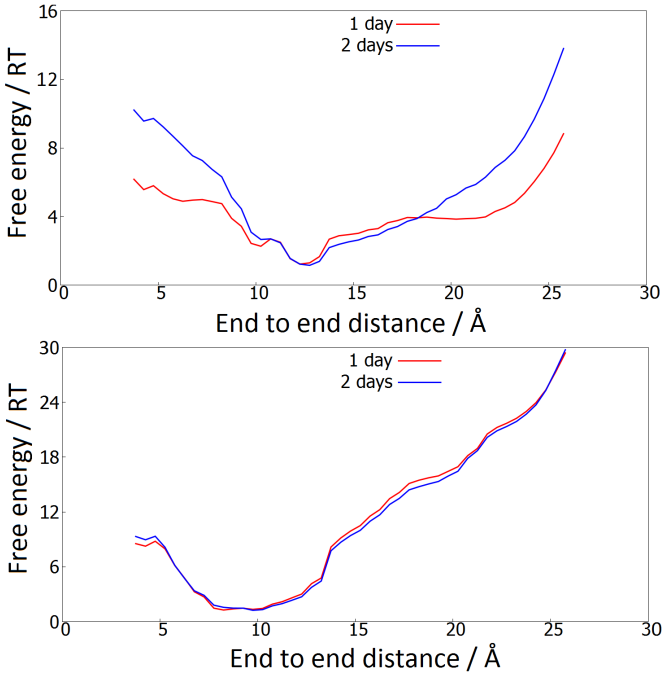

Figure A.5: to test for convergence the free energy is calculated after 1 day and again after 2 days. If the free energy changes significantly (top frame) then it has not converged and the calculations continue, the convergence check being repeated at intervals of 2 days and 4 days.. If there is no significant chnge (bottom) then the free energy has converged.

Note that the procedure described above is for a single trajectory. To better sample the phase space multiple independent trajectories should be generated and their individual free energies calculated in

the above way, until they have all converged and are then averaged.

Bibliography

[1] O'Connor, C.; Adams, J., 'Essentials of Cell Biology', Cambridge, MA: NPG Education, **2010**. 1

[2] Damodaren, S., 'Beyond the hydrophobic effect: Critical function of water at biological phase boundaries A hypothesis' *Adv. Colloid Interface Sci.*, **2015**, *221*, 22-33. 2

[3] Oldfield, C.; Dunker, K. 'Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions' *Ann. Rev. Biochem.*, **2014**, *83*, 553-584. 6

[4] Adcock, S. A.; McCammon, J. A., 'Molecular dynamics: Survey of methods for simulating the activity of proteins' *Chem. Rev.*, **2006**, *106*, 1589-1615. 6, 7

[5] Brooks, B. R.; Brooks, C. L.; Makerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang. W.; York, D. M.; Karplus, M., J., 'CHARMM: The Biomolecular Simulation Program' *J.Comp. Chem.*, **2009**, *30*, 1545-1613.

[6] Karplus, M.; McCammon, J. A., 'Molecular dynamics simulations of biomolecules' *Nature Struct. Biol.*, **2002**, *9*, 646-652. 6, 7

[7] Alder, B. J.; Pople, J. A., '3rd Virial Coefficient for Intermolecular Potentials with Hard Sphere Cores' *J. Chem. Phys.*, **1957**, *26*, 325. 7

[8] McCammon, J., 'Molecular Dynamics study of the bovine pancreatic trypsin inhibitor in models for Protein Dynamics, CECAM: Orsay, France, **1976**, 137. 7

[9] Yang, W.; Gao, Y. Q.; Cui, Q.; Ma, J.; Karplus, M., 'The missing link between thermodynamics and structure in F1-ATPase', *Proc. Natl. Acad. Sci. U.S.A.,*, **2003**, *100*, 874-879. 7

[10] Daggett, V., 'Protein folding-simulation' *Chem. Rev.*, **2006**, *106*, 1898-1916. 7

[11] Day, R., 'All-atom simulations of protein folding and unfolding' *Adv. Prot. Chem.*, **2003**, *66*, 373-403.

[12] Daggett, V., 'Molecular dynamics simulations of the protein unfolding/folding reaction' *Acc. Chem. Res.*, **2002**, *35*, 422-429.

[13] Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. A.; Salmon, J. K.; Shan. Y.; Wriggers, W., 'Atomic-Level Characterization of the Structural Dynamics of Proteins' *Science*, **2010** , *330*, 341-346.

[14] Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E., 'How Fast-Folding Proteins Fold' *Science*, **2011**, *334*, 517-520. 7

[15] Roux, B., Schulten, K., 'Computational studies of membrane channels' *Structure*, **2004**, *12*, 1343-1351. 7

[16] Beckstein, O.; Tai, K.; Sampson, M., 'Not ions alone: Barriers to ion permeation in nanopores and channels' *J. Am. Chem. Soc.*, **2004**, *126*, 14694-14695. 7

[17] Deng, Y., Roux, B., 'Molecular dynamics calculation of absolute binding free energy: Aromatic ligands bind to a nonpolar cavity of T4 lysozyme.' *Abstr. Pap. Am. Chem. Soc.*, **2004**, *228*, 254. 7

[18] Hockney, R.; Eastwood, J., 'Computer Simulation Using Particles', McGraw-Hill, New York **1981**. 8

[19] Dill, K. A.; MacCallum, J. L., 'The Protein-Folding Problem, 50 Years On' *Science*, **2012** , *108*, 1042-1046. 8

[20] Pianna, S.; Lindorff-Larsen, K.; Shaw, D., 'Atomistic Description of the Folding of a Dimeric Protein' *J. Phys. Chem. B.*, **2013**, *117*, 12935-12942. 8

[21] Pianna, S.; Lindorff-Larsen, K.; Shaw, D., 'Atomic-level description of ubiquitin folding' *Proc. Natl. Acad. Sci. U.S.A.*, **2013**, *110*, 59155920.

[22] Pianna, S.; Lindorff-Larsen, K.; Shaw, D., 'Protein folding kinetics and thermodynamics from atomistic simulation' *Proc. Natl. Acad. Sci. U.S.A.*, **2012**, *109*, 1784517850. 8

[23] Pianna, S.; Klepeis, J.; Shaw, D., 'Assessing the Accuracy of Physical Models Used in Protein-Folding Simulations: Quantitative Evidence from Long Molecular Dynamics Simulations' *curr. opin. struct. biol.*, **2014**, *24*, 98105. 9

[24] Torrie, G. M.; Valleau, J. P., N-PHYSICAL SAMPLING DISTRIBUTIONS IN MONTE-CARLO FREE-ENERGY ESTIMATION - UMBRELLA SAMPLING' 'Non-Physical Saimpling Distributions in Monto-Carlo Free-Energy Estimation - Umbrella Sampling' *J. Comp. Phys.*, **1977**, *23*, 187-199. 11

[25] Mills, M.; Andricioaei, I., 'An experimentally guided umbrella sampling protocol for biomolecules' *J. Chem. Phys.*, **2008**, *129*, 114101. 11

[26] Kumar, S.; Bouzida, D.; Swendsen, H.; Kollman, P. A.; Rosenburg, J. M., 'The Weighted Histogram Analysis Method for Free ENergy Calculation on Biomolecules .1. the Method'. *J. Comp. Chem.*, **1992**, *13*, 1011-1021. 11

[27] Mezei, M., 'Adaptive Umbrella Sampling - Self-Consistent Determination of the non-Boltzmann Bias' *J. Comp. Phys.*, **1987**, *1*, 237-248. 12

[28] Laio, A.; Parrinello, M., 'Escaping free-energy minima' *Proc. Natl. Acad. Sci. U.S.A.*, **2002**, *99*, 12562-12566. 12

[29] Faradjian, A.; Elber, R., 'Computing time scales from reaction coordinates by milestoning' *J. Chem. Phys.*, **2004**, *120*, 10880-10889. 14

[30] Vanden-Eijnden, E.; Venturoli, M.; Ciccotti, G.; Elber, R., J., 'On the assumptions underlying milestoning' *J. Chem. Phys.*, **2008**, *129*, 174102. 14

[31] Vanden-Eijnden, E.; Venturoli, M., 'Markovian milestoning with Voronoi tessellations' *J. Chem. Phys.*, **2009**, *130*, 194101. 16

[32] Martinez-Nunez, E.; Shalashilin, D., 'Acceleration of classical mechanics by phase space constraints' *J. Chem. Theory Comput.*, **2006**, *2*, 912-919. 13, 21

[33] Voter, A., 'A method for accelerating the molecular dynamics simulation of infrequent events' *J. Chem. Phys.*, **1997**, *106*, 4665. 13, 25

[43] Glowacki, D. R.; Paci, E.; Shalashilin, D. V., 'Boxed Molecular Dynamics: Decorrelation Time Scales and the Kinetic Master Equation' *J. Chem. Theory Comput.*, **2011**, *7*, 1244-1252. 23, 33

[35] Allen, R.; Valeriani, C.; Rein ten Wolde, P., 'Forward flux sampling for rare event simulations' *J. Phys.: Condens. Matter*, **2009**, *21*, 463102. 16

[36] Dellago, C.; Bolhuis, P.; Csajka, F.; Chandler, D., 'Transition path sampling and the calculation of rate constants' *J. Chem. Phys.*, **1998**, *108*, 1964-1977. 17

[37] Van Erp, T.; Moroni, D.; Bolhuis, P., 'A novel path sampling method for the calculation of rate constants' *J. Chem. Phys.*, **2003**, *118*, 7762. 17

[38] Huber, G.; Kim, S., 'Weighted-ensemble Brownian dynamics simulations for protein association reactions' *Biophys. J.*, **1996**, *70*, 97-110. 17

[39] Ren, W.; Vanden-Eijnden, E., 'Transition pathways in complex systems: Application of the finite-temperature string method to the alanine dipeptide' *J. Phys. Chem. B*, **2005**, *109*, 6688. 17

[40] Pan, A.; Weinreich, T.; Shan, Y.; Scarpazza, D.; Shaw, D., 'Assessing the Accuracy of Two Enhanced Sampling Methods Using EGFR Kinase Transition Pathways: The Influence of Collective Variable Choice' *J. Chem. Theory Comput.*, **2014**, *10*, 2860-2865. 18

[41] Ray, J.; Moody, M., 'Calculation of Elastic-Constants Using Isothermal Molecular-Dynamics' *Phys. Rev. B*, **1986**, *33*, 895-899. 18

[42] Sugita, Y.; Okamoto, Y., 'Replica-exchange molecular dynamics method for protein folding' *Chem. Phys. Lett.*, **1999**, *314*, 141-151. 19

[43] Glowacki, D. R.; Paci, E.; Shalashilin, D. V., 'Replica-exchange molecular dynamics method for protein folding' *J. Chem. Theory Comput.*, **2011**, *7*, 1244-1252. 23, 33

[44] Booth, J.; Vazquez, S.; Martinez-Nunez, E,; Marks, A.; Rodgers, J.; Glowacki, D.; Shalashilin, D., 'Recent applications of boxed

molecular dynamics: a simple multiscale technique for atomistic simulations' *Phil. Trans. R. Soc. A.*, **2014**, *372*, 20130384. 36, 90

[45] Hedberg, C.; Toledo, G.; Gustafsson, C.; Larson, G.; Oldfors, A.; Macao, B., 'Misfolding of fibronectin III 119 subdomain in titin results in hereditary myopathy with early respiratory failure' *Nuromuscul. Disord.*, **2014**, *24*, 832-832. 37, 57, 58

[46] Charton, K.; Sarparanta, J.; Vihola, A.; Suel, L.; Daniele, N.; Hackman, P.; Udd. B.; Richard, I., 'The relationship of calpain 3 and titin in the M-band' *Nuromuscul. Disord.*, **2014**, *24*, 885-885. 37

[47] Hur, J.; Darve, E., *Centre for Turbulence Research Annual Research Briefs*, **2003**, , 425-438. 37

[48] Moy, V.; Florin, E.; Rief, M.; Lehmann, H.; Ludwig, M.; Gaub, H.; Dornmair, K., 'Probing the Forces Between Complementary Strands of DNA with the Atomic Force Microscope' *P. Soc. Photo-Opt. Inst.*, **1995**, *2384*, 2-12. 37

[49] Florin, E.; Rief, M.; Lehmann, H.; Ludwig, M.; Dornmair, K.; Moy, V., 'Sensing Specific Molecular Interactions with the Atomic Force Microscope' *Biosens. Bioelectron.*, **1995**, *10*, 895-901 37

[50] Brujic, J.; hermans, R.; Walther, K.; Fernandez, J., 'Single-molecule force spectroscopy reveals signatures of glassy dynamics in the energy landscape of ubiquitin' *Nat. Phys.*, **2006**, *2*, 282-286. 39, 61, 62

[51] Garcia-Manyes, S.; Brujic, J.; Badilla, C.; Fernandez, J., 'Force-clamp spectroscopy of single-protein monomers reveals the individual unfolding and folding pathways of I27 and ubiquitin' *Biophys. J.*, **2007**, *93*, 2436-2446. 68

[52] Liu, R.; Garcia-Manyes, S.; Sarkar, A.; Badilla, C.; Fernandez, J., 'Mechanical Characterization of Protein L in the Low-Force Regime by Electromagnetic Tweezers/Evanescent Nanometry' *Biophys. J.*, **2009**, *96*, 3810-3821. 63

[53] Oberhauser, A.; Hansma, P.; Carrion-Vazquez, M.; Fernandez, J., 'Stepwise unfolding of titin under force-clamp atomic force microscopy' *Proc. Natl. Acad. Sci. U.S.A.*, **2001**, *98*, 468-472. 68

[54] Schlierf, M.; Li, H.; Fernandez, J., 'The unfolding kinetics of ubiquitin captured with single-molecule force-clamp techniques' *Proc. Natl. Acad. Sci. U.S.A.*, **2004**, *101*, 7299-7304. 39, 62

[55] Marszalek, P.; Lu, H.; Li, H.; Carrion-Vazquez, M.; Oberhauser, A.; Schulten, K.; Fernandez, J., 'Mechanical unfolding intermediates in titin modules' *Nature*, **1999**, *402*, 100-103. 39, 43, 45, 50, 61, 67

[56] Brockwell, D., 'Probing the mechanical stability of proteins using the atomic force microscope' *Biochem. Soc. Trans.*, **2007**, *35*, 1564-1568.

[57] Crampton, N.; Brockwell, D., 'Unravelling the design principles for single protein mechanical strength' *Curr. Opin. Struct. Biol.*, **2010**, *20*, 508-517. 43, 44, 45, 57, 58

[58] Hann, E.; Kirkpatrick, N.; Kleanthous, C.; Smith, D.; Radford, S.; Brockwell, D., 'The effect of protein complexation on the mechanical stability of Im9' *Biophys. J.*, **2007**, *92*, L79-L81.

[59] Sadler, D.; Petrik, E.; Taniguchi, Y.; Pullen, J.; Kawakami, M.; Radford, S.; Brockwell, D., 'Identification of a Mechanical Rheostat in the Hydrophobic Core of Protein L' *J. Mol. Biol.*, **2009**, *393*, 237-248. 39

[60] Hoffmann, T.; Dougan, L., 'Single molecule force spectroscopy using polyproteins' *Chem. Soc. Rev.*, **2013**, *41*, 47814796. xiii, 40

[61] Brockwell, D.; Beddard, G.; Paci, E.; West, D.; Olmsted, P.; Smith, D.; Redford, S., 'Mechanically unfolding the small, topologically simple protein L' *Biophys. J.*, **89**, *2005*, 506-519. xiii, xv, 42, 43, 44, 45, 50, 55, 60, 67

[62] Kawakami, M.; Byrne, K.; Brockwell, D.; Radford, S.; Smith, A., 'Viscoelastic study of the mechanical unfolding of a protein by AFM' *Biophys. J: Biohpys. Lett.*, **2006**, *91*, L16-L18. xv, 43, 57, 58, 60

[63] Fowler, S.; Best, R.; Herrera, J.; Rutherford, T.; Stewart, A.; Paci, E.; Karplus, M.; Clarke, J., 'Mechanical unfolding of a titin Ig domain: Structure of unfolding intermediate revealed by combining AFM, molecular dynamics simulations, NMR and protein engineering' *J. Mol. Biol.*, **2002**, *322*, 841-849. 43, 50, 55

[64] Lu, H.; Isralewitz, B.; Krammer, A.; Vogel, V.; Schulten, K., 'Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation' *Biophys. J.*, **1998**, *75*, 662-771. 43, 50, 67

[65] Lu, H.; Schulten, K, 'Steered molecular dynamics simulations of force-induced protein domain unfolding' *Proteins*, **1999**, *35*, 453-463.

[66] Lu, H.; Schulten, K, 'The key event in force-induced unfolding of titin's immunoglobulin domains' *Biophys. J.*, **2000**, *79*, 51-65.

[67] Gao, M.; Lu, H.; Schulten, K., 'Unfolding of titin domains studied by molecular dynamics simulations' *J. Muscle Res. Cell. Motil.*, **2002**, *23*, 513-521. 43, 50, 55

[68] Crampton, N.; Alzahrani, K.; Beddard, S.; Connel, S.; Brockwell, D., 'Mechanically Unfolding Protein L Using a Laser-Feedback-Controlled Cantilever' *Biophys. J.*, **2011**, *100*, 1800-1809. 43

[69] Chng, C.; Kitao, A., 'Mechanical unfolding of bacterial flagellar filament protein by molecular dynamics simulation' *J. Mol. Graph. Model.*, **2010**, *28*, 548-554. 45, 58

[70] Law, R.; Carl, P.; Harper, S.; Dalhaimer, P.; Speicher, D.; Discher, D., 'Cooperativity in forced unfolding of tandem spectrin repeats' *Biophys. J.*, **2003**, *84*, 533-544. 61

[71] Lee, E.; Hsin, J.; Sotomayor, M.; Comealles, G.; Schulten, K., 'Discovery Through the Computational Microscope' *Structure*, **2009**, *17*, 1295-1306. 45, 58

[72] Brooks, B.; Brooks III, C.; Mackerell, A.; Nilsson, L.; Petrella, R.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Din ner, A.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R.; Post, C.; Pu, J.; Schaefer, M .; Tidor, B.; Venable, R.; Woodcock, H.; Wu, X.; Yang, W.; York, D.; Karplus, M., 'CHARMM: The Biomolecular Simulation Program' *J. Comp. Chem.*, **2009**, *30*, 1545-1615. 45

[73] Laziardis, T.; Karplus, M., 'Effective energy function for proteins in solution' *Proteins*, **1999**, *35*, 133-152. 46, 48

[74] Mashimo, S.; Kuwabara, S.; Yagihara, S.; Higasi, K., 'Dielectric-Relaxation Time and Structure of Bound Water in Biological-Materials' *J. Phys. Chem.*, **1987**, *91*, 6337-6338. 50

[75] Yew, Z.; Olmsted, P.; Paci, E., 'Using the folding landscapes of proteins to understand protein function' *Single-Molecule Biophysics: Experiment and Theory*, **2012**, *146*, 395-417. 55

[76] Irback, A.; Mitternacht, S., 'Thermal versus mechanical unfolding of ubiquitin' *Proteins*, **2006**, *15*, 759-766. 55

[77] Dudko, O,; Hummer, G.; Szabo, A., 'Intrinsic rates and activation free energies from single-molecule pulling experiments' *Phys. Rev. Lett.*, **2006**, *96*, 108101. 56

[78] Hummer, G.; Szabo, A., 'Free energy surfaces from single-molecule force spectroscopy' *Acc. Chem. Res.*, **2005**, *38*, 504-513.

[79] Hummer, G.; Szabo, A., 'Kinetics from nonequilibrium single-molecule pulling experiments' *Biophys. J.*, **2003**, *85*, 5-13. 56

[80] Rico, F.; Gonzalez, L.; Casuso, I.; Puig-Vidal, M.; Scheuring, S., 'High-Speed Force Spectroscopy Unfolds Titin at the Velocity of Molecular Dynamics Simulations' *Science*, **2013**, *342*, 741-743. 58, 61

[81] Friddle, R.; Noy, A.; De Yoreo, J., 'Interpreting the widespread nonlinear force spectra of intermolecular bonds' *Proc. Natl. Acad. Sci. U.S.A.*, **2012**, *109*, 13573-13578. 58

[82] Zinober, R.; Brockwell, D.; Beddard, G.; Blake, A., Olmsted, P.; Radford, S.; Smith, D., 'Mechanically unfolding proteins: The effect of unfolding history and the supramolecular scaffold' *Protein Sci.*, **2002**, *11*, 2759-2765. 61

[83] Brujic, J.; Hermans, R.; Garcia-Manyes, S.; Walther, K.; Fernandez, J., 'Dwell-time distribution analysis of polyprotein unfolding using force-clamp spectroscopy' *Biophys. J.*, **2007**, *92*, 2896-2903. 61

[84] Uribe, L.; Gauss, J.; Diezemann, G., 'Comparative Study of the Mechanical Unfolding Pathways of alpha- and beta-Peptides' *J. Phys. Chem. B*, **2015**, *119*, 8313-8320. 61

[85] Mitternacht, S.; Luccioli, S.; Torcini, A.; Imparato, A.; Irback, A., 'Changing the Mechanical Unfolding Pathway of FnIII(10) by Tuning the Pulling Strength' *Biophys. J.*, **2009**, *92*, 429-441. 62, 71

[86] Kirmizialtin, S.; Huang, L.; Makarov, D., 'Topography of the free-energy landscape probed via mechanical unfolding of proteins' *J. Chem. Phys.*, **2005**, *122*, 234915. 62, 71

[87] Hughes, M.; Dougan, L., *Rep. Prog. Phys.* **2016**, Accepted. 68

[88] Brockwell, D.; Paci, E.; Zinober, R.; Beddard, G.; Olmsted, D.; Perham, R.; Radford, S., 'Pulling geometry defines the mechanical resistance of a beta-sheet protein' *Nature Struct. Biol.*, **2003**, *10*, 731-737. 68

[89] Carrion-Vazquez, M; Li, H.; Lu, H.; Marszalek, P.; Oberhauser, A.; Fernandez, J., 'The mechanical stability of ubiquitin is linkage dependent' *Nature Struct. Biol.*, **2003**, *10*, 738-743. 68

[90] Zoldak, G.; Rief, M., 'Force as a single molecule probe of multi-dimensional protein energy landscapes' *Curr. Opin. Struct. Biol.*, **2013**, *23*, 48-57. 69

[91] Korotkin, I.; Karabasov, S.; Nerukh, D.; Markesteijn, A.; Scukins, A.; Farafonov, V.; Pavlov, E., 'A hybrid molecular dynamics/fluctuating hydrodynamics method for modelling liquids at multiple scales in space and time' *J. Chem. Phys.*, **2015**, *143*, 014110. 67

[92] Lannon, H.; Vanden-Eijnden, E.; Brujic, J., 'Force-Clamp Analysis Techniques Give Highest Rank to Stretched Exponential Unfolding Kinetics in Ubiquitin' *Biophys. J.*, **2012**, *103*, 2215-2222. 68, 70

[93] Kaestner, J., 'Umbrella sampling' *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, , *1*, 932-942. 69

[94] Shalashilin, V.; Beddard, G.; Paci, E.; Glowacki, D., 'Peptide kinetics from picoseconds to microseconds using boxed molecular dynamics: Power law rate coefficients in cyclisation reactions' *J. Chem. Phys.*, **2012**, *137*, 165102. 70, 88, 89

[95] Surade, S.; Blundell, T., 'Structural Biology and Drug Discovery of Difficult Targets: The Limits of Ligandability' *Chem. and Biol.*, **2012**, *19*, 42-50. 72

[96] Lipinski, C.; Lombardo, F.; Dominy, B.; Feeney, P., 'Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings' *Adv. Drug Delivery Rev.*, **1997**, *23*, 3-25. 72, 73

[97] Giordanetto, F.; Kihlberg, J., 'Macrocyclic Drugs and Clinical Candidates: What Can Medicinal Chemists Learn from Their Properties?' *J. Med. Chem.*, **2014**, *57*, 278-295. 72, 73

[98] Kotz, J., 'Bringing Macrocycles Full Circle' *SciBX*, **5**, *45*, doi:10.1038/scibx.2012.1176. 73, 74

[99] Scannell, J. W.; Blanckley, A.; Boldon, H.; Warrington, B., 'Diagnosing the decline in pharmaceutical R and D efficiency' *Nat. Rev. Drug Discovery*, **2012**, *11*, 191-200. 72

[100] Professor Dame Sally Davies (U.K. Chief Medical Officer 2011 to present) *2013 Chief Medical Officer annual report volume 2:Infections and the rise of antimicrobial resistance*, 72, 73

[101] World Health Organisation *The evolving threat of antimicrobial resistance: options for action.* **2012**,

[102] Centers for Disease Control and Prevention, *Antibiotic resistance threats in the United States*, **2013**. 72, 73

[103] Williams, H.; Trevaskis, N.; Charman, S.; Shanker, R.; Charman, W.; Pouton, C.; Porter, C., 'Strategies to Address Low Drug Solubility in Discovery and Development' *Pharmacol. Rev.*, **2013**, *65*, 315-499. 73

[104] Booth, J.; Shimizu, S.; Abbott, S., 'Mechanism of Hydrophobic Drug Solubilization by Small Molecule Hydrotropes' *J. Phys. Chem. B*, **2012**, *116*, 14915-14921. 73

[105] Cain, C., 'Excited about cycling', *BioCentury* **2012**, *20*, A7-A13. 73

[106] Mallinson, J.; Collins, I., 'Macrocycles in new drug discovery' *Future. Med. Chem.*, **2012**, *4*, 1409-1438. 73

[107] Driggers, E.; Hale, S.; Lee, J., 'The exploration of macrocycles for drug discovery - an underexploited structural class' *Nat. Rev. Drug Discovery*, **2008**, *7*, 608-624. 73

[108] Riemann, R.; Zacharias, M., 'Reversible scaling of dihedral angle barriers during molecular dynamics to improve structure prediction of cyclic peptides' *J. Peptide. Res.*, **2004**, *63*, 354-364. 73, 83

[109] Keasar, C.; Rosenfeld, R., 'Empirical modifications to the Amber/OPLS potential for predicting the solution conformations of cyclic peptides by vacuum calculations' *Folding and Design*, **1998**, *3*, 379-388.

[110] Rayan, A.; Senderowitz, H.; Goldblum, A., 'Exploring the conformational space of cyclic peptides by a stochastic search method' *J. Mol. Graph. Mod.*, **2004**, *22*, 319-333.

[111] Beaufays, J.; Lins, L.; Thomas, A.; Brasseur, R., 'The CCK(-like) receptor in the animal kingdom: Functions, evolution and structures' *J. Pept. Sci.*, **2011**, *18*, 17-24

[112] Yu, H.; Lin, Y., 'Toward structure prediction of cyclic peptides' *Phys. Chem. Chem. Phys.*, **2015**, *17*, 4210-4219. 73, 83

[113] DeLorbe, J.; Clements, J.; Whiddon, B.; Martin, S., 'Thermodynamic and Structural Effects of Macrocyclic Constraints in Protein-Ligand Interactions' *A.C.S. Med. Chem. Lett.*, **2010**, *1*, 448-452. 73

[114] Gause, G.; Bogdan, A.; Davies, N.; James, K., 'Comparison of diffusion coefficients for matched pairs of macrocyclic and linear molecules over a drug-like molecular weight range' *Org. Biomol. Chem.*, **2011**, *9*, 7727-7733. 73

[115] Brazhnikova, M., 'Gramicidin Sand its use in the treatment of infected wounds' *Nature*, **1944**, *154*, 703. 73

[116] White, C.; Yudin, A., 'Contemporary strategies for peptide macrocyclization' *Nature Chemistry*, **2011**, *3*, 509-524. 74, 75, 76, 78, 90, 111

[117] Yu, Z.; Yu, X.; Chu, Y, 'MALDI-MS determination of cyclic peptidomimetic sequences on single beads directed toward the generation of libraries' emphTetrahedron. Lett., **1998**, *39*, 1-4. 110

[118] Tang, Y.; Xie, H.; Tian, G.; Ye, Y., 'Synthesis of cyclopentapeptides and cycloheptapeptides by DEPBT and the influence of some factors on cyclization' *J. Pept. Res.*, **2002**, *60*, 95-103

[119] Brady, S.; Varga, S.; Freidinger, R.; Schwenk, D., 'Practical Synthesis of Cyclic-Peptides, with an Example of Dependence of Cyclisation Yield Upon Linear Sequence' *J. Org. Chem.*, **1979**, *44*,

[120] Caumes, C.; Fernandes, C.; Roy, O.; Hjelmgaard, T.; Wenger, E.; Didierjean, C.; Taillefumier, C.; Faure, S., 'Cyclic alpha,beta-Tetrapeptoids: Sequence-Dependent Cyclization and Conformational Preference' *Org. Lett.*, **2013**, *14*, 3626-3629. 74, 75, 110

[121] 'Guideline on the specification limits for residues of metal catalysts', European Medicines Agency **2007** Print. 75

[122] Stewart, M.; Fire, E.; Keating, A.; Walensky, L., 'The MCL-1 BH3 helix is an exclusive MCL-1 inhibitor and apoptosis sensitizer' *Nat. Chem. Biol.*, **2010**, *6*, 595-601. 73

[123] Fialho, A.; Chakrabarty, A., 'Emerging Cancer Therapy: Microbial Approaches and Biotechnological Tools', Wiley, **2010**, Print. 76

[124] Chiou, A.; Ong, G.; Wang, K.; Chiou, S.; Wu, S., 'Conformational study of two linear hexapeptides by two-dimensional NMR and computer-simulated modeling: Implication for peptide cyclization in solution' *Biochem. Biophys. Res. Comm.*, **1996**, *219*, 572-579. 78, 90

[125] Jaspars, M.; Houssen, W., 'Azole-Based Cyclic Peptides from the Sea Squirt Lissoclinum Patella: Old Scaffolds, New Avenues' *Chembiochem.*, **2010**, *11* 1803-1815. 78

[126] Koehnke, J.; Bent, A.; Houssen, W.; Zollman, D.; Morawitz, F.; Shirran, S.; Vendome, J.; Nneoyiegbe, A.; Trembleau, L.; Botting, C.; Smith, C.; Jaspars, M.; Naismith, J., 'The mechanism of patellamide macrocyclization revealed by the characterization of the PatG macrocyclase domain' *Nat. Struct. Mol. Bio.*, **2012**, *19* 767-772. 81, 82

[127] Koehnke, J.; Bent, A.; Zollman, D.; Smith, K.; Houssen, W.; Zhu, X.; Mann, G.; Lebl, T.; Scharff, R.; Shirran, S.; Botting, C.; Jaspars, M.; Schwarz-Linek, U.; Naismith, J., 'The Cyanobactin Heterocyclase Enzyme: A Processive Adenylase That Operates with a Defined Order of Reaction' *Angew. Chem. Int. Ed.*, **2013**, *52*, 13991-6.

[128] Koehnke, J.; Morawitz, F.; Bent, A.; Houssen, W.; Shirran, S.; Fuszard, M.; Smellie, I.; Botting, C.; Smith, C.; Jaspars, M.; Naismith, J., 'An Enzymatic Route to Selenazolines' *Chembiochem.*, **2013**, *14*, 564-567. 78

[129] Ireland, C.; Durso, A.; Newman, R.; Hacker, M., 'Anto-Neoplastic Cyclic-Peptides from the Marine Tunicate Lissoclinum-Patella', *J. Org. Chem.*, **1982**, *47*, 1807-1811. 78

[130] Schmidt, U.; Langner, J., 'Cyclotetrapeptides and cyclopentapeptides: Occurrence and synthesis' *J. Pept. Res.*, **1996**, *49*, 67-73. 79

[131] Chang, G.; Guida, W.; Still, W., 'An Internal Coordinate Monte-Carlo Method for Searching Conformational Space', *J. Am. Chem. Soc.*, **1989**, *111*, 4379-4386. 85

[132] Mutter, M., 'Macrocyclization Equillibria of Polypeptides' *J. Am. Chem. Soc.*, **1977**, *7*, 8307-8314. 87

[133] Yongye, A.; Li, Y.; Giulianotti, M., 'Modeling of peptides containing D-amino acids: implications on cyclization' *J. Comput. Aided Mol. Des.*, **2009**, *23*, 677-689. 87, 88, 112

[134] Volk, M.; Kholodenko, Y.; Lu, H.; Gooding, E.; DeGrado, W.; Hochstrasser, R., 'Peptide conformational dynamics and vibrational stark effects following photoinitiated disulfide cleavage' *J. Phys. Chem. B*, **1997**, *101*, 8607-8616. 88, 89, 90

[135] Metzler, R.; Klafter, J.; Jortner, J.; Volk, M., 'Multiple time scales for dispersive kinetics in early events of peptide folding' *Chem. Phys. Lett.*, **1998**, *293*, 477-484.

[136] Milanesi, L.; Waltho, J.; Hunter, C.; Shaw, D.; Beddard, G.; Reid, D.; Dev, S., Volk, M., 'Measurement of energy landscape roughness of folded and unfolded proteins' *Proc. Natl. Acad. Sci. U.S.A.* , **2012**, *109*, 19563-19568. 88, 89, 90

[137] Thakkar, A.; Trinh, T.; Pei, D., 'Global Analysis of Peptide Cyclization Efficiency' *A.C.S. Comb. Sci.*, **2013**, *15*, 120-129. 111, 115

[138] Sheu, S.; Yang, D.; Selzle, H.; Schlang, E., 'Energetics of hydrogen bonds in peptides' *Proc. Natl. Acad. Sci. U.S.A.*, **2003**, *100*, 12683-12687. 114

[139] Sirimulla, S.; Lerma, M.; Herndon, W., 'Prediction of Partial Molar Volumes of Amino Acids and Small Peptides: Counting Atoms versus Topological Indices' *J. Chem. Inf. Model.*, **2010**, *50*, 194-204. xix, 116

[140] Daidone, I.; Neuweiler, H.; Doose, S.; Sauer, M.; Smith, J., 'Hydrogen-Bond Driven Loop-Closure Kinetics in Unfolded Polypeptide Chains' *PLoS Comp. Biol.*, **2010**, *6*, 1000645. 119

[141] Cavelier-Frontin, F.; Pepe, G.; Verducci, J.; Siri, D.; Jacquier, R., 'Prediction of the best Linear Precursos in the Synthesis of Cyclotetrapeptides by Molecular Mechanic Calculations', *J. Am. Chem. Soc.*, **1992**, *114*, 8885-8890. 74, 75, 83, 122

[142] Toniolo, C.; Bonora, G., 'Preferred Conformations of Peptides Containing Alpha, Alpha-Disubstituted Al[ha-Amino-Acids' *Biopolymers*, **1983**, *22*, 205-215. 83

[143] Besser, D.; Olender, R.; Rosenfeld, R.; Arrad, O.; Reissmann, S., 'Study on the cyclization tendency of backbone cyclic tetrapeptides' *J. Pept. Res.*, **2000**, *56*, 337-345. xvii, 74, 75, 84, 86, 87, 90, 110, 122