Computer Simulations of Protein Folding

Haydn Wyn Williams, BSc.(Hons)

Thesis submitted to The University of Nottingham

for the degree of Doctor of Philosophy

September 2011

Abstract

Computer simulations of biological systems provide novel data while both supporting and challenging traditional experimental methods. However, continued innovation is required to ensure that these technologies are able to work with increasingly complex systems.

Coarse–grained approximations of protein structure have been studied using a lattice model designed to find low–energy conformations. A hydrogen–bonding term has been introduced. The ability to form β–sheet has been demonstrated, and the intricacies of reproducing the more complex α–helix on a lattice have been considered.

An alternative strategy, that of better utilising computing power through the technique of milestoning, has shown good agreement with previous experimental and computational work. The increased efficiency allows significantly less extreme simulation conditions to be applied than those used in alternative simulation methods, and allows more simulation repeats.

Finally, the principles of Least Action Dynamics have been employed to combine the two approaches described above. By splitting a simulation trajectory into a number of smaller components, and using the lattice model to optimise the path from a start structure to an end structure, it has been possible to efficiently generate dynamical information using an alternative method to traditional molecular dynamics.

Acknowledgements

Throughout my PhD studies I have been fortunate enough to work with many talented and knowledgable researchers, both staff members and other PhD students. Foremost amongst these are my supervisors, Prof. Hirst and Prof. Williams. I would like to sincerely thank them for the help and support they have provided over the past three years. My thanks also to Mark Oakley for discussion over the course of the project, and for developing the first version of the LaMP software.

In addition, other group members have made working in the laboratory both fun and scientifically exciting. Particular thanks are given to Craig Bruce for doing all things sysadmin, to Ben Bulheller for all manner of geeky conversation, and to Claire-Louise Evans for assistance with CHARMM.

On a more personal level, I thank my partner Rebecca for all the encouragement and support she has given me over the course of my research.

Finally, my thanks to the following groups for funding:

- Nottingham Nanoscience and Nanotechnology Interdisciplinary Doctoral Training Centre
- The School of Pharmacy, University of Nottingham
- The School of Chemistry, University of Nottingham

Contents

List of Abbreviations

Aβ – Amyloid-β.

ADLD – Action–Derived Lattice Dynamics.

ADMD – Action–Derived Molecular Dynamics.

AFM – Atomic force microscopy.

APP – Amyloid Precursor Protein.

BSE – Bovine Spongiform Encephalopathy.

CJD – Creutzfeldt-Jakob disease.

CF – Cystic fibrosis.

CFTR – Cystic fibrosis transmembrane conductance regulator.

DNA – Deoxyribonucleic acid.

EEF1 – Effective energy function.

GB – Generalised Born.

GBSA – Generalised Born Surface Area.

HP – Hydrophobic-Polar.

I27 – I27 domain of Titin.

Ig – Immunoglobulin.

MC – Monte Carlo.

MD – Molecular dynamics.

PDB – Protein Data Bank.

REx – Replica Exchange.

RMSD – Root Mean Square Deviation.

SMD – Steered molecular dynamics.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Protein Folding and Aggregation

1.1.1 Motivation for Understanding Protein Folding

Proteins are essential for life in all organisms, from the simplest examples of prion proteins and viruses, through to complex eukaryotic animals and plants. The pathway from genetic code to functional protein has been studied heavily for more than 50 years, but definitive answers about various aspects of the process remain elusive. One area which has received continual interest is that of how proteins fold into their unique three-dimensional shapes.

In recent decades a range of discoveries, from the elucidation of the structure of DNA through to developments in technologies such as NMR and x-ray crystallography, have greatly enhanced our understanding of each end of the protein folding process. The mechanism by which a protein chain is synthesised from a DNA sequence using amino acids is generally understood, and we are able to determine the final, folded three-dimensional structure of many proteins in their native states (and indeed in other, per-

turbed states). However, the middle stage in this process - that of the actual folding from an unstructured polypeptide chain to a functional three-dimensional structure - is still relatively unknown. A variety of methods have been proposed for this mechanism (see section 1.1.2), but it is important to understand why this particular problem has remained at the forefront of the scientific community's consciousness.

A Scientific Problem

The protein folding problem was brought to the fore by Levinthal in the 1960s (1). Prior to his work, it was assumed that proteins folded by randomly sampling conformations until the native state was discovered. Levinthal suggested that proteins used some more systematic approach to folding. His conservative estimates (discussed in more detail in section 1.1.2), suggested that a random search strategy would require sampling of an infeasible number of conformations. This simple observation resulted in a shift in thinking about the mechanisms of protein folding; researchers began to look for method in a process previously assumed to have none. The ability of proteins to fold is intriguing because it seems to take place in a directed fashion, but, in the majority of cases, without any external influence or guidance. Although chaperone proteins may assist in the folding or unfolding of other proteins, or prevent their aggregation into non–functional forms, they are by no means a prerequisite for these activities. This implies that all the information required to form a functional molecule is bound up in the sequence of each protein. That this folding process gives the same end result time after time is a remarkable feat, and one which has therefore understandably captured the imagination of many researchers over the years.

Treatment of Diseases

There are more practical reasons for developing a greater understanding of protein folding. The ability of a protein to fold into its native state is vital for it to be biologically active. Should a protein fold into an incorrect conformation, its ability to catalyse a reaction, remain mechanically strong, or transport another molecule may be diminished or entirely lost. Should such an error occur on a regular basis, disease symptoms can result (2). For example, mutations in the cystic fibrosis transmembrane conductance regulator (CFTR) gene can lead to the production of proteins which are unable to fold correctly. This in turn leads to a loss of function, and the disease symptoms of cystic fibrosis (CF). However, a mutated protein has a different sequence to the wild-type and so differences in the folding process may be expected.

Equally interesting are those proteins that cause disease by failing to fold correctly even though they are not mutated. Clinical symptoms of Alzheimer's disease are caused by aggregation of a mis-folded form of the amyloid-β protein, a protein which is normally present in the human body (3). Even when the sequence of this protein is not mutated, some (as yet unknown) event can cause spontaneous misfolding. Several misfolding events can be followed by aggregation and the formation of insoluble neuronal plaques. A number of diseases show similar forms of action, including Huntingdon's disease and Parkinson's disease.

Aggregation is also important in prion diseases, whereby correctly folded proteins aggregate in a similar fashion to that seen in Alzheimer's disease. There is obviously therefore a motivation to correct misfolding defects, either as a strategy to reduce symptoms or to try and prevent them altogether. Current research is having some success identifying potential means of preventing aggregation in vitro (4).

A firm understanding of processes upstream of aggregation would enable such diseases to be tackled at their root, rather than in a reactive manner.

1.1.2 Protein Folding Theories

The Sequence / Structure / Function Relationship

One of the most important realisations in the field of protein folding took place when Anfinsen showed that the three-dimensional structure of a protein was determined by the amino acid sequence (5). The dependence of a protein on its specific native state conformation for biological activity was already known at the time. However, Anfinsen's experiments involving the denaturation and renaturation of bovine pancreatic ribonuclease showed that when all tertiary structure is removed from a protein, the information necessary to re–form that structure (the native state) is still contained within the protein. He demonstrated this using ribonuclease which had been denatured using urea; when the denaturant was removed, the inactive ribonuclease regained its activity. Indeed, further work on a variety of enzymes showed that certain sections of a protein chain could even be cleaved without the protein losing its ability to fold.

As a result of this work, Anfinsen stated that "the native conformation is determined by the totality of interatomic interactions and hence by the amino acid sequence" (6). Another important conclusion of this work was that the native state of the protein was also the most energetically favourable. As each specific amino acid sequence would have only one global energy minimum, the protein would favour that conformation each time it folded. This became known as "the thermodynamic hypothesis", and thus the primary structure and native state of a protein were acknowledged to be intrinsically linked.

Anfinsen's work established where the information needed to fold a protein was stored, but thoughts subsequently turned to *how* proteins reach their native conformation.

Levinthal's Paradox

Around the same time that Anfinsen was finding the link between amino acid sequence and tertiary structure, Levinthal was also considering the protein folding problem. He pointed out that a protein could not fold by sampling every possible conformation until the global energy minimum was reached; the timescale for doing so would be infeasible (1). Such a process is described by a flat energy landscape, providing no guidance to the protein on taking a particular pathway to the native state (Figure 1.1). For example, in a 150 residue peptide with only three degrees of freedom in the backbone, there must be 3^{150} different conformations, or $3.7x10^{71}$. Assuming that 10^{12} conformations can be sampled every second, the protein would still require $7x10^{53}$ years to work through them all. This argument, countering the conclusions of Anfinsen, suggested that the native state of a protein was not the *global* minimum, but rather a local minimum. It is also worth noting that some metastable proteins can take on non-native structures with a lower energy than the native conformation (7). Levinthal's suggested process for folding was that of a 'pathway' to the minimum, rather than a random search. He also proposed that 'segments' of structure would form initially, and would make further condensation of the protein into recognisable structure elements more likely (8).

Anfinsen in fact supported this latter theory, by suggesting that proteins fold via "nucleation" events. He also discussed the idea of protein folding being a co-operative process, requiring the whole protein to come together in a coherent structure before becoming biologically functional. The results of experiments detailed in his 1973 paper indicated

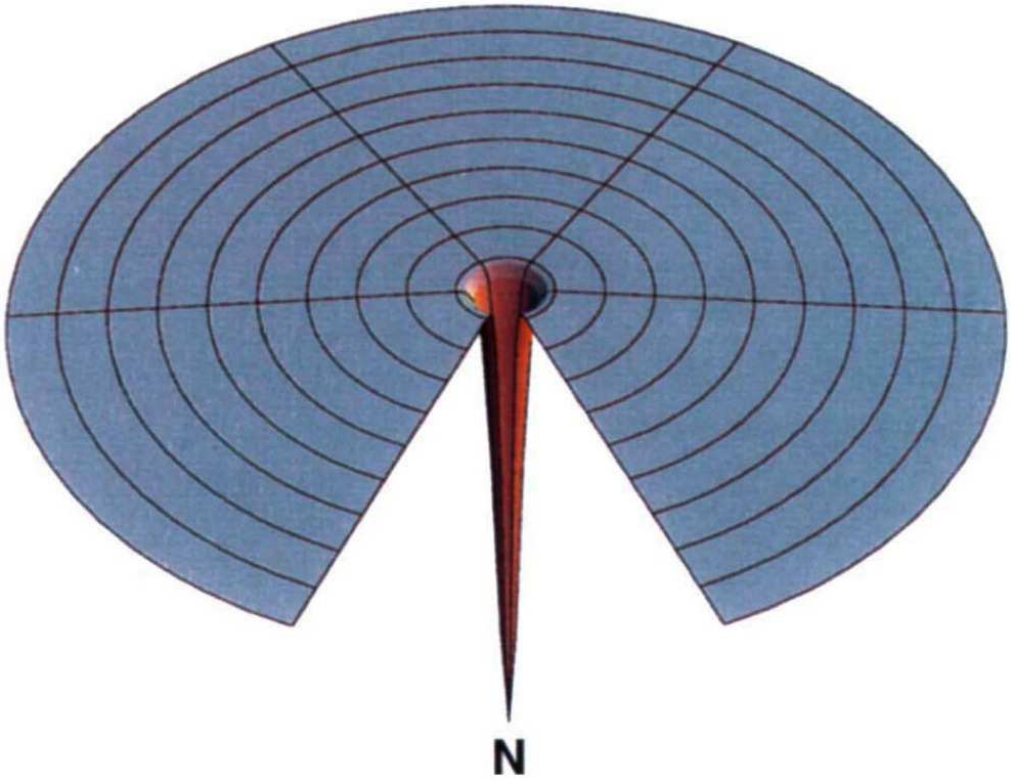

Figure 1.1: A flat energy landscape; the native state is point N. Reproduced from (9).

that folding could take place at any point with a complete protein, not only during its synthesis on the ribosome.

Levinthal's idea of pathways in folding can be visualised as an energy landscape which consists of a channel to guide the protein towards the native state (the local minimum on the landscape). An example of this is seen in Figure 1.2. By following this channel, a folding protein does not have to explore the entire energy landscape before finding the native state, thus overcoming Levinthal's numbers problem. The idea does suggest, however, that a folding protein will always follow the same path as it folds.

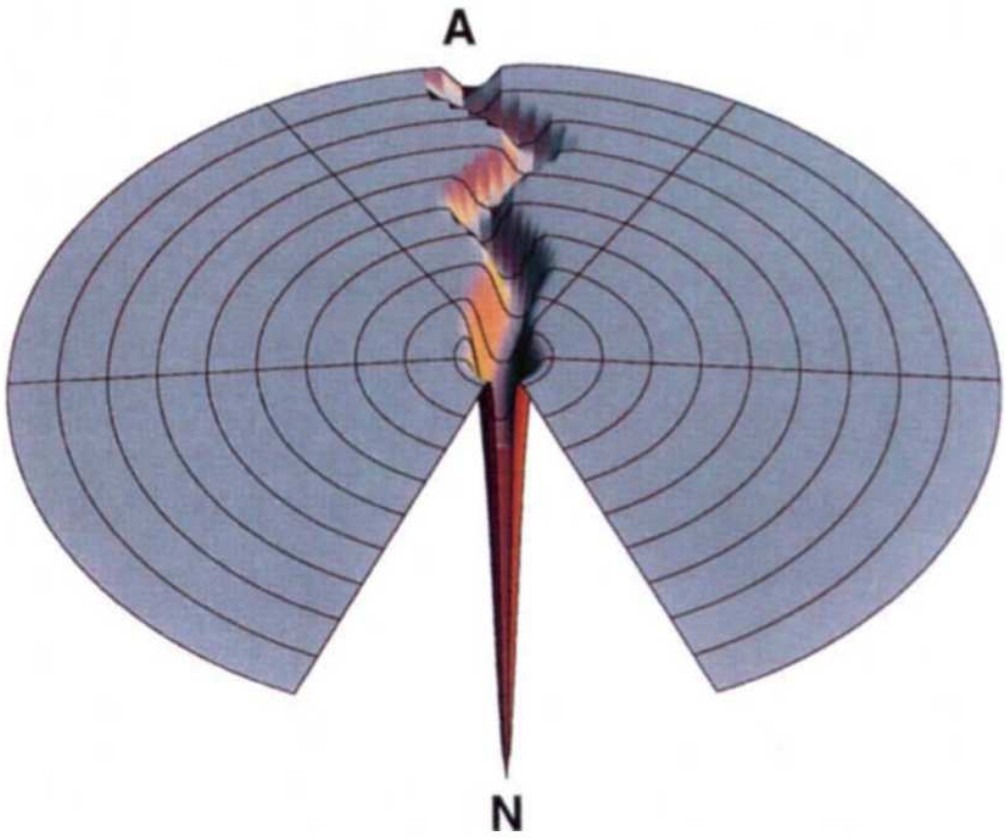

Figure 1.2: An energy landscape with a 'pathway' to guide the folding protein toward the native state N from the denatured state A. Reproduced from (9).

Folding Funnels

In the years following the development of the 'pathway' model, researchers realised that it was not sufficient to adequately explain all aspects of folding. The model was intended to represent a protein folding from a denatured conformation to the native state. If the native state is the lowest point on the landscape, then denatured states are represented by all other points. It is therefore possible to start from a denatured state which is not in the 'pathway', thus going against the principles of the model. Since denaturation experiments such as Anfinsen's had shown that proteins can re–fold to the native state from any number of different denatured conformations, the model is obviously deficient. A

logical expansion of the argument leads to the development of multiple pathways, which in turn can be developed into a generalised 'funnel' with a rugged landscape (Figure 1.3).

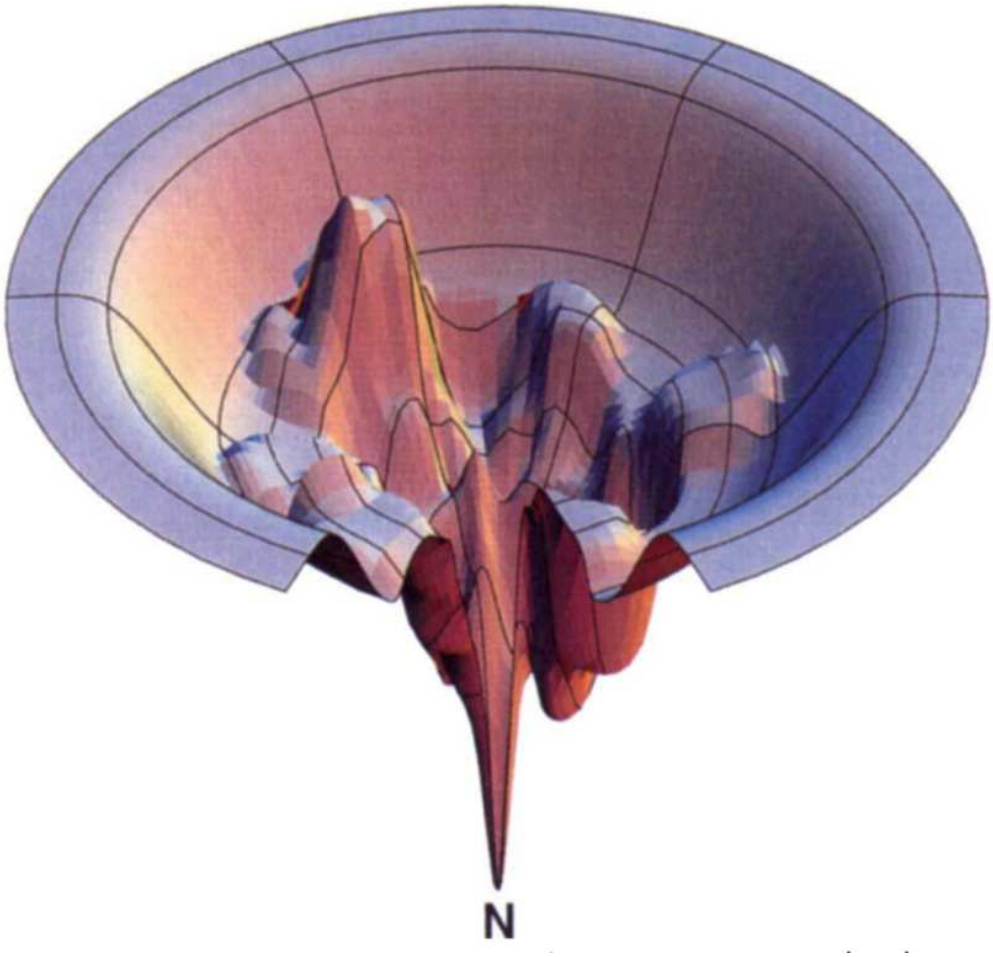

Figure 1.3: A true 'funnel' free energy landscape. Reproduced from (9).

This funnel model is sufficient to describe all the nuances of protein folding seen from Anfinsen's work through to current developments in the field. A protein can move from any point at the top of the funnel, *i.e.* any denatured structure, to the native state at the bottom of the funnel. The free energy reduces as the molecule moves from the top of the landscape to the bottom. While there is no single specific pathway to follow, the protein will be to some extent guided in its folding by the presence of 'hills' and 'valleys' in the

landscape (representing transition states and intermediates respectively).

While undoubtedly an improvement on previous concepts, the funnel model does not necessarily provide an accurate model of the whole protein folding picture. For example, it does not convey the contribution that conformational entropy plays in protein folding. Mirny *et al.* (10) point out that when moving from an intermediate to a transition state, the protein may actually increase in energy. At first glance, such an increase may seem counter-intuitive on a funnel diagram, but this is due to the fact that entropy is not represented explicitly. Prior to the transition state, folding is driven by a loss in entropy, counteracting any gain in energy. Once the transition state is passed, the reaction becomes energetically-driven and the concept of continually moving towards a lower energy state becomes valid once again. One way to consider this is for the width for the funnel to act as a general indicator of conformational entropy; as a protein progresses from the fringes of the funnel to the centre, entropy decreases even if energy increases. Despite such nuances, the funnel method has become widely used and is the most prevalent concept used to explain protein folding and energy landscapes at the current time.

The Mechanics of Folding

While protein folding funnels help us visualise how a particular protein may proceed from denatured to fully folded, they provide no detail about the three-dimensional structures adopted during the process. There have been several theories proposed regarding the actual physical mechanisms of protein folding. One of the first utilised the hierarchical nature of proteins' structural elements. Known as the 'framework model', it postulated that the secondary structure elements would form first, and that these would diffuse around before eventually interacting together to create the tertiary structure (11, 12).

This view has subsequently been replaced to some degree by the 'hydrophobic collapse' or 'nucleation–condensation' model (13). This entails the hydrophobic residues aggregating to escape the aqueous solvent as much as possible. This forms a hydrophobic core, with more hydrophilic elements on the exterior. True protein folding, however, is likely to be a mixture of the two methods, although there is still much research being undertaken in this area.

1.1.3 Protein Misfolding and Aggregation

Misfolding Diseases

A protein must be folded into a specific three-dimensional conformation to have any biological activity. A variety of factors may cause a protein to lose this specific shape, resulting in not only a loss of biological activity, but in some cases in toxicity as well. As Anfinsen discovered, the biological activity of a protein is directly linked to the amino acid sequence. Thus, any alteration in the sequence through genetic mutation can cause the protein to fold incorrectly or not at all. As mentioned previously, cystic fibrosis is caused by mutations in the CFTR gene (14). This gene encodes a chloride ion channel protein, which is important for the production of mucosal secretions. The ΔF508 mutation, for example, deletes a phenylalanine residue, and without it the protein produced from the mutated gene is unable to fold. Humans have two copies of the CFTR gene; if both are mutated in this way then no functional CFTR protein is produced, and CF results.

Another less well-known example affects lysosomes, the organelles responsible for catabolism of macromolecules within cells (15). When lysosomal proteins fail to fold, the organelles cannot operate effectively and a build-up of partially degraded material can occur within

them. The defective protein may be an internal component of the lysosome or part of the membrane. Regardless of the location, mutations in any one of a number of proteins can result in over 40 lysosomal diseases. In the examples above, and various others, the wild-type protein performs a specific job, and fails to do so only when the integrity of the genetic code for the protein is compromised.

Protein Aggregation / Prion Diseases

An alternative class of proteopathic diseases involves proteins in which the wild-type genes are unaffected, but the wild-type protein still suffers a loss of function through a failure to fold correctly. Such proteins may fold incorrectly immediately following synthesis on the ribosome, or as a result of spontaneous unfolding from the native state and re–folding (Figure 1.4). Regardless of the method by which the protein reaches the alternative conformation, the end result is aggregation of molecules of the protein into insoluble plaques. In Alzheimer's disease, the proteins involved are tau and amyloid-β (Aβ). Aβ is formed from Amyloid Precursor Protein (APP) by two successive cleavage events. β-Secretase splits a soluble extracellular fragment from the membrane-bound APP molecule. The main molecule is then cleaved again, in the transmembrane domain, byγ-Secretase, producing Aβ (16). The extracellular aggregations of Aβ protein are the bodies first seen by Alzheimer during his characterisation of the disease, and the plaques and their precursors are its main cytotoxic element (3). The cytotoxicity can be attributed both to the interactions of amyloidogenic proteins with other cell components, and to the sheer volume of aggregated protein (17).

Aggregation in this manner is seen in multiple other diseases, including Huntington's disease and Parkinson's disease (Table 1.1). A similar mechanism is also seen in the prion

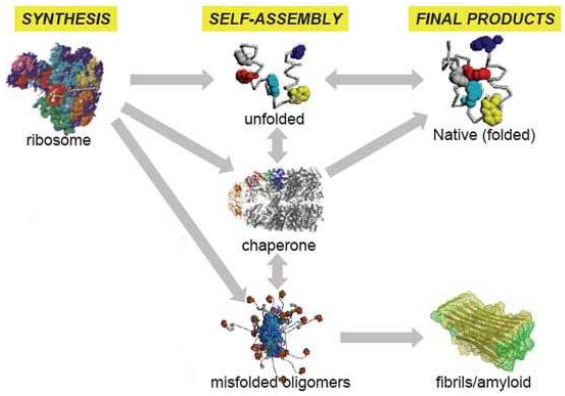

Figure 1.4: Possible pathways to protein misfolding of wild-type proteins (18).

diseases, which include Bovine Spongiform Encephalopathy (BSE) and Creutzfeldt-Jakob disease (CJD) (19). Prion diseases are caused by a protein folding from its normal form, PrP^{C}, to a protease-resistant conformer known as PrP^{Sc} (20). While the physiological function of PrP^{C} is still not fully understood, it is clear that PrP^{Sc} is the causative agent in prion diseases. Like the amyloid-β protein in Alzheimer's disease, mis–folded prion proteins aggregate in elongated fibres to form insoluble plaques. The mis–folding of one molecule of PrP^{C} can initiate a cascade effect, causing many other molecules to unfold and subsequently take on the PrP^{Sc} form (21). This effect leads to the phenomenon of the spongiform encephalopathies being transmissible diseases (22, 23, 24).

It is not only proteins involved in disease which form amyloid fibrils. A number of other proteins have been seen to form fibrils *in vitro* (Table 1.2), but not *in vivo* (17) .

Disease	Main Aggregate Component
Alzheimer's disease	Aβ peptides (plaques); tau protein (tangles)
Spongiform encephalopathies	Prion (whole or fragments)
Parkinson's disease	α–synuclein (wt or mutant)
Huntington's disease	Huntingtin
Primary systemic amyloidosis	Ig light chains (whole or fragments)
Secondary systemic amyloidosis	Serum amyloid A (whole or 76–residue fragment)
Senile systemic amyloidosis	Transthyretin (whole or fragments)
Injection–localised amyloidosis	Insulin

Table 1.1: Diseases of proteins involving amyloid fibril formation. Adapted from (17).

Amyloid Fibril Formation

Much like the protein folding problem, the motivation to understand the formation of amyloid aggregate formation is two–fold. Firstly, there is the obvious aim of developing treatments for the variety of diseases in which amyloid plaques are implicated. There is, however, an interesting phenomenon which is almost the opposite of that described in section 1.1.1. The observation that unique individual protein primary structures give rise to unique three-dimensional tertiary structures is not necessarily always true. Indeed, the diseases caused by amyloid formation show very similar, stable structures despite the disparity between the amino acid sequences of the different proteins involved (25, 26). This is, in effect, an interesting inversion of the traditional protein folding problem.

A large proportion of the research into the amyloidoses has revolved around Alzheimer's disease, due to its high-profile and incidence rate compared to prion diseases and other

Domain / Protein	Source
Endostatin	Human
Stefin B	Human
Amphoterin	Human
ADA2H	Human
Apolipoprotein CII	Human
Apomyoglobin	Equine
Fibronectin type III module	Murine
Phosphoglycerate kinase	Yeast
Fibroblast growth factor	*Notophthalmus viridescens*
Apocytochrome *c*	*Hydrogenobacter thermophilus*

Table 1.2: Proteins found to form fibrils *in vitro*. Adapted from (17).

diseases involving fibril formation. Work in the 1960s determined that the plaques seen in neuronal tissue of Alzheimer's disease sufferers are composed of aggregates of individual filaments (27, 28) of Aβ. Each of these has subsequently been found to have a diameter of approximately 7–10 nm and lengths of 100 nm upwards (29) (Figure 1.6). However, the inherent insolubility and non–crystalline nature of the highly stable filaments meant that further structural information could not be resolved using the methods employed at the time. These included x-ray crystallography and liquid-state nuclear magnetic resonance (NMR). Indeed, the protein responsible for forming the filaments was only isolated and sequenced in 1984 (30). Shortly after the discovery of the presence of fibrils in neuronal plaques, data from x-ray diffraction experiments showed that the fibrils consist predominantly of β–sheet. The distances of 4.75 Å and 9.8 Å obtained from the original experiments indicated β–strands spaced in parallel or anti–parallel configu-

ration, and several sheets lying alongside each other, respectively (31). This suggested a "cross-β" conformation, that is, the sheets run perpendicular to the fibril axis (32). More recent cryogenic electron microscopy and NMR work has supported this observation (33, 34). Each fibril consists of several protofilaments, in common with other amyloid–forming proteins (35, 36).

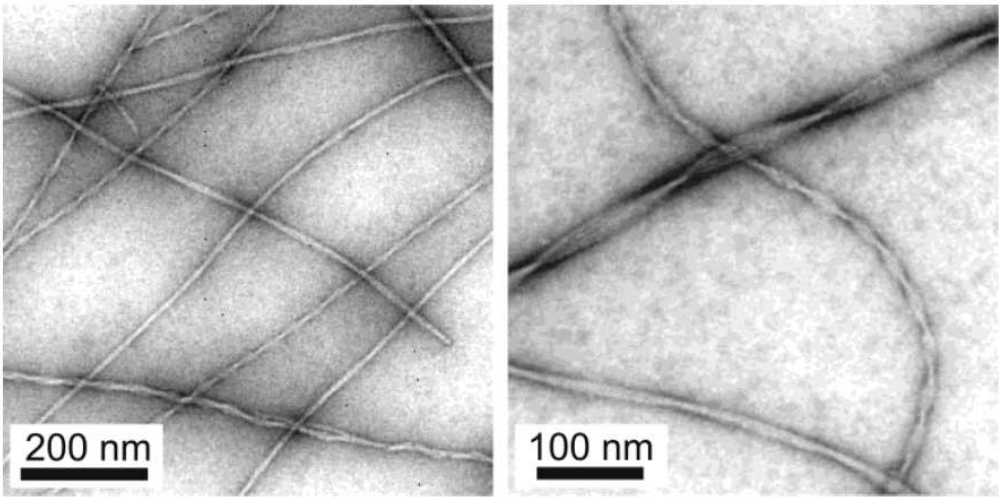

Figure 1.5: Transmission electron micrograph image of Aβ fibrils. Reproduced from (37).

In 2003, Tycko *et al.* brought together molecular constraint data from a variety of sources, including their own solid– and liquid–state NMR experiments, and developed a model for the structure of the Aβ protofilaments (37). The model indicated that they are constructed from two individual molecules twisted around each other. Each molecule contains two anti–parallel β–strands, in the cross–β formation. Stability is conferred by the confinement of hydrophobic residues to the interior of the protofilament, with polar and charged residues on the outside of the structure. Previous work on an Aβ fragment, containing only residues 34 to 42 (Aβ_{34-42}) also suggested that an anti–parallel conformation was most likely (38, 39). However, subsequent work on Aβ_{10-35} established that a parallel conformation may indeed be possible (40).

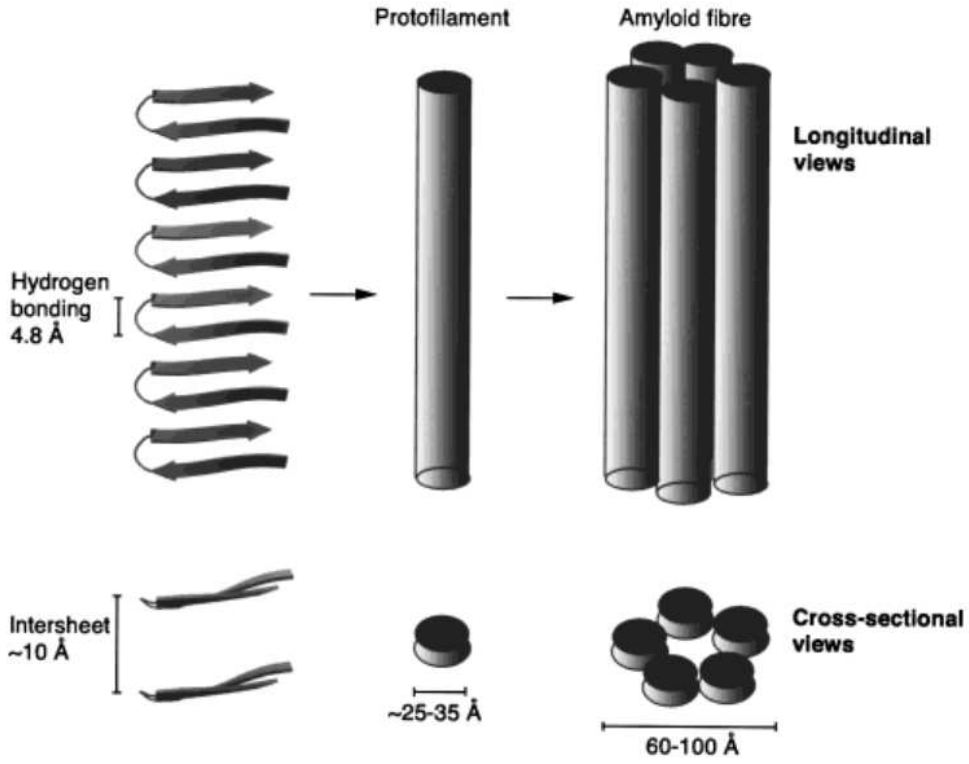

Figure 1.6: Diagrammatic representation of Alzheimer's amyloid–β fibril structure. Reproduced from (33).

Despite advances in the determination of the structure of the Aβ fibrils, relatively little is understood about the mechanism of protein aggregation. Contrary to initial assumptions that Aβ was entirely insoluble, it has been found to be soluble under certain conditions (41, 42). Suggested trigger factors for the switch from soluble α–helix to insoluble β–sheet have included localised changes in pH (43). It has recently become apparent that, whatever the mechanism, the species involved in the early stages formation of complete fibrils may be as important as the fibrils themselves. Cytotoxic effects have been caused by the pre–fibrillar aggregates of amyloid–forming proteins which are not seen in any natural diseases (19, 44). The methods of cytotoxicity vary, but include alterations in the level of intracellular Ca^{2+} ions, and in the exocytotic ability of cells.

Prion diseases have also been shown to consist of very similar structures to those seen in Alzheimer's disease; the yeast prion Sup35 has a cross-β structure (45) containing very little water (46), while the Ure2p prion has been shown to contain predominantly parallel β-sheet (47, 48). Molecular dynamics simulations of the Sup35 prion have indicated that aggregation may proceed from an anti–parallel starting structure towards a parallel conformation in more mature oligomers (49).

Computational models capable of determining low–energy structures may provide an insight into the native structures of functional proteins in their native state. However, they could also shed new light on diseases involving highly stable mis–folded molecules such as amyloid fibrils. The nature of these structures makes them hard study physically, and information derived from *in silico* work may help develop treatments for the numerous diseases involving fibril formation.

1.2 Computer Simulations

1.2.1 Background to Molecular Dynamics Simulations

Successes and Milestones

The field of molecular dynamics simulation has developed since the late 1950s, shortly after computers first became available to researchers (50). The first molecular dynamics programs used hard sphere models to simulate liquids (51). These were subsequently developed to model biological molecules such as proteins, initially in the absence of solvent (52), and then in simple lipid bilayer membranes (53). As available computing power has increased, the ability to simulate solvent, both implicitly and explicitly, has become commonplace.

One of the first major simulations of a biological molecule was that of bovine pancreatic trypsin inhibitor. This work by McCammon *et al.* was published in Nature in 1977, and represented a massive computational effort. McCammon used the x-ray structure of the inhibitor to perform dynamics calculations and investigate fluctuations about the average structure (52). The system consisted of 500 atoms in a vacuum, with a simulation duration of 9.2 ps. This work was fundamental in shifting the view of proteins from one of rigid structures to more dynamic systems with a certain level of flexibility and fluidity. Since that time, the field has expanded rapidly, and current simulations typically involve much larger systems, or longer timescales. The first simulations of DNA were completed in 1986 (54), and allowed the predicted structures seen in the model to be compared with NMR restraint data. Whilst the system size of eight base-pairs and approximately 1,200 water molecules is very modest by today's standards, this was an important development in modelling work.

In 1998 Duan and Kollman carried out the first 1 μs simulation of a protein in explicit water (55). They looked at the early stages of protein folding in the Villin headpiece, a system which also featured in the Folding@home project in 2006 (56). Around 500 μs of simulation in explicit solvent were generated by the massive parallel architecture of the latter project, demonstrating that an increasingly detailed picture of protein folding could be obtained from Monte Carlo as well as molecular dynamics simulations. Also in 2006, the entire Satellite Tobacco Mosaic virus (over one million atoms) was simulated by Freddolino and co-workers (57). The ability to study the entire viral system allowed them to establish that the protein capsid was significantly less stable in the absence of the RNA core. Their results also agreed with previous studies on the stability of the capsid and the core.

Principles of Molecular Dynamics

Molecular dynamics software packages such as CHARMM (58), Amber (59), NAMD (60) and GROMACS (61) all integrate Newton's equation of motion (Equation 1.2.1) to determine the positions of atoms at a future point in time.

$$\mathbf{F} = m\mathbf{a} \tag{1.2.1}$$

where $\mathbf{F}$ is force, m is mass and $\mathbf{a}$ is acceleration.

However, this is not a trivial operation, and a variety of algorithms are used in molecular dynamics simulations. It is assumed that the positions of the atoms at points forwards and backwards in time can be approximated by Taylor series expansions (Equation 1.2.2).

$$\begin{aligned} \mathbf{r}(t+\Delta t) &= \mathbf{r}(t) + \Delta t \cdot \mathbf{v}(t) + \frac{\Delta t^2 \cdot \mathbf{a}(t)}{2} + \frac{\Delta t^3 \cdot \mathbf{b}(t)}{6} \\ \mathbf{r}(t-\Delta t) &= \mathbf{r}(t) - \Delta t \cdot \mathbf{v}(t) + \frac{\Delta t^2 \cdot \mathbf{a}(t)}{2} - \frac{\Delta t^3 \cdot \mathbf{b}(t)}{6} \end{aligned} \tag{1.2.2}$$

where $\mathbf{r}$ is atomic position, t is time, $\mathbf{v}$ is velocity, $\mathbf{a}$ is acceleration and $\mathbf{b}$ is the rate of change of acceleration.

By adding the Taylor expansions, we can calculate the atomic position at $t + \Delta t$; this is the Verlet algorithm (Equation 1.2.3).

$$\mathbf{r}(t+\Delta t) = 2\mathbf{r}(t) - \mathbf{r}(t-\Delta t) + \Delta t^2 \mathbf{a}(t) \tag{1.2.3}$$

Note that the first– and third–order terms from the Taylor expansions have cancelled out, meaning that we need only calculate $\mathbf{a}(t)$, the acceleration. This is just the force divided by the mass (Equation 1.2.1). The force can be calculated as the derivative of the energy

with respect to the change in position of the atom (Equation 1.2.4), and the mass of each atom is known. In this manner, the Verlet algorithm can be solved.

$$\mathbf{F} = -\frac{dE}{d\mathbf{r}} \tag{1.2.4}$$

where $\mathbf{F}$ is force, E is energy and $\mathbf{r}$ is atomic position.

One disadvantage this method has is that it does not explicitly calculate velocity values for the atoms in the system. While they can be estimated, there are some related algorithms which are often used in preference to the basic Verlet algorithm.
The "leapfrog" algorithm calculates velocities at the half–step (Equation 1.2.5), and these therefore "leapfrog" the coordinates, which are calculated at the full step (Equation 1.2.6).

$$\mathbf{v}\left(t + \frac{\Delta t}{2}\right) = \mathbf{v}\left(t - \frac{\Delta t}{2}\right) + \left[\Delta t \cdot \mathbf{a}\left(t\right)\right] \tag{1.2.5}$$

$$\mathbf{r}\left(t + \Delta t\right) = \mathbf{r}\left(t\right) + \left[\Delta t \cdot \mathbf{v}\left(t + \frac{\Delta t}{2}\right)\right] \tag{1.2.6}$$

The lack of $\Delta t^2 \cdot \mathbf{a}\left(t\right)$ term means that this algorithm is more accurate than the standard Verlet algorithm. It also requires only one set of atomic coordinates to be stored in memory at any given time, rather than the three sets used with standard Verlet.

Another popular algorithm used is the "Velocity Verlet" algorithm. This has the advantage of calculating both the atomic positions and coordinates at the full step, *i.e.* in synchrony. Firstly the positions at $t + \Delta t$ are calculated (Equation 1.2.7).

$$\mathbf{r}(t+\Delta t) = \mathbf{r}(t) + [\mathbf{v}(t) \cdot \Delta t] + \frac{\mathbf{a}(t) \cdot \Delta t^2}{2} \tag{1.2.7}$$

The velocities at the half–step are then calculated (Equation 1.2.8), and from this the positions at the full step (Equation 1.2.9).

$$\mathbf{v}\left(t + \frac{\Delta t}{2}\right) = \mathbf{v}(t) + \frac{[\Delta t \cdot \mathbf{a}(t)]}{2} \tag{1.2.8}$$

$$\mathbf{r}(t+\Delta t) = \mathbf{r}(t) + \left[\mathbf{v}\left(t + \frac{\Delta t}{2}\right) \cdot \Delta t\right] \tag{1.2.9}$$

The velocities at the same point can then be calculated (Equation 1.2.10).

$$\mathbf{v}(t+\Delta t) = \mathbf{v}(t) + \frac{[\mathbf{a}(t) + \mathbf{a}(t+\Delta t)] \cdot \Delta t}{2} \tag{1.2.10}$$

The molecular dynamics force field describes how each of the different forces on an atom contributes to its overall energy. The force field is thus made up of several different components, which can be calculated differently depending on the software package being used. As a general principle, the force field includes two types of interaction, bonded and non- bonded (Equation 1.2.11).

$$V_{total} = V_{bonded} + V_{non-bonded} \tag{1.2.11}$$

where V is potential energy.

Each of these is in turn composed of several different terms, which combine to give the total energy of a system at a given point in time.

The bonded terms collectively refer to any term which describes the impact of a property of one or more covalent bonds on a system's energy. It is made up of four main terms (Equation 1.2.12).

$$V_{bonded} = V_{length} + V_{angle} + V_{dihedral} + V_{improper} \tag{1.2.12}$$

Bond length, or stretching, relates to the length of a covalent bond between two atoms, and its deviation from the equilibrium length. The greater the deviation, the more energy is required to maintain the conformation. While a Morse potential gives a better representation of the term, many systems use a simpler harmonic potential to reduce calculation time. The "bond angle" term is effectively a measure of the deviation from the equilibrium angle formed by three atoms which are bonded together in a linear fashion. A harmonic potential is also most often used to model this interaction.

The dihedral bond angle, or torsion angle, applies where two pairs of covalently bonded atoms are themselves joined by another covalent bond (Figure 1.7). It is the angle between the planes formed by the two atom pairs.

The rotatable nature of the central bond means that a harmonic energy function is not appropriate, and so a function of periodic form is used instead. In practice, the energy is given by the sum of the term over all dihedral angles in the molecule (Equation 1.2.13). This term is sometimes referred to as the proper dihedral angle.

$$v_{dihedral} = \sum_{dihedrals} \frac{1}{2} V_n [1 + \cos(n\phi - \delta)] \tag{1.2.13}$$

where v is energy, V_n is the force constant, n is periodicity and δ is phase.

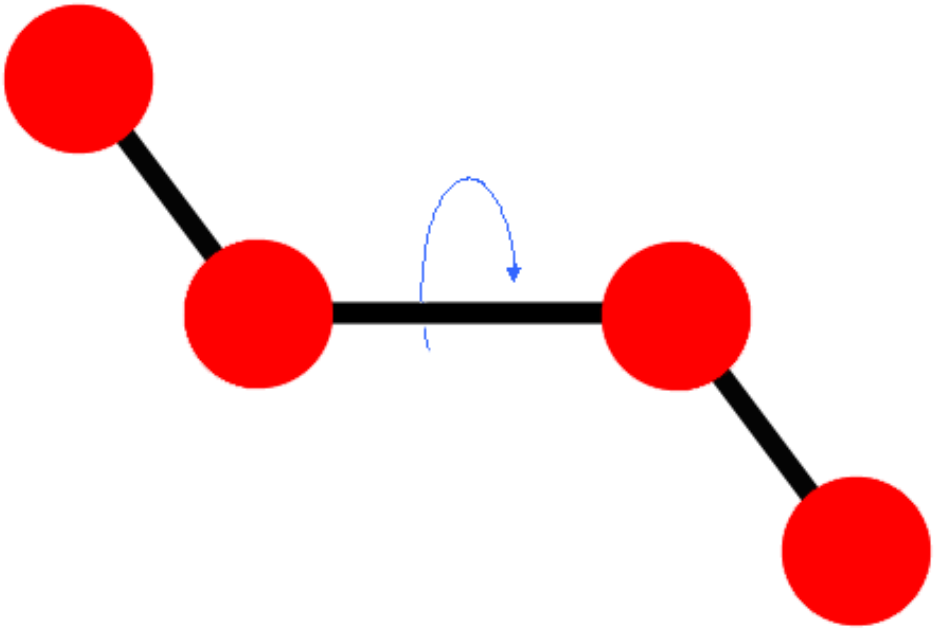

Figure 1.7: A dihedral angle.

n is the periodicity of the angle, while δ is the phase. v_n is the force constant. By varying the periodicity and force constant, it is possible to model the number and relative size of the minima present in the function.

The improper dihedral or torsion angle is used to correct for the out-of-plane motions of an atom which is bonded to three others in a non-linear fashion (Figure 1.8).

It is defined as the angle between the planes formed by atoms *i*,*j*,*k* and *j*,*k*,*l*. Some algorithms use a different parameter, that is, the distance between the central atom, *j*, and the plane formed by atoms *i*,*k*,*l*.

The non–bonded terms are the more complex terms making up the force–field. They model the forces on atoms caused by electrostatic interactions and van der Waal's forces. Van der Waal's forces are the attractive and repulsive forces caused by the electrons clouds of adjacent atoms. When atoms are very close together, they are subjected to a strong repulsive force as their electron clouds overlap. As the distance increases, con-

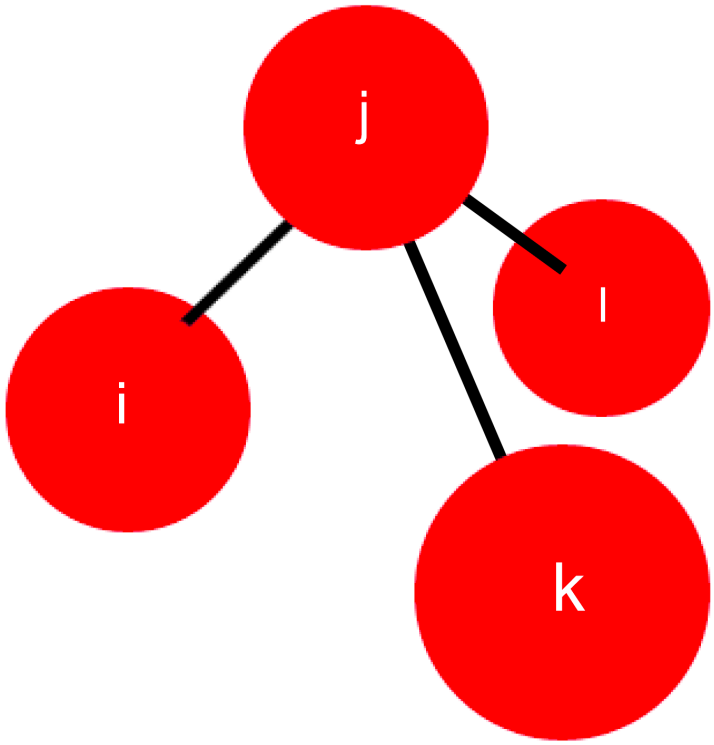

Figure 1.8: An improper dihedral (torsion) angle.

tinual fluctuations in the distributions of electrons cause instantaneous dipoles to form. The presence of a dipole on one atom can cause another dipole to form in a different nearby atom. The attractive force between dipoles becomes dominant at this distance, and then diminishes as the distance between the atoms increases. The van der Waal's interactions are commonly represented by the Lennard-Jones potential (Figure 1.9) (62).

Whereas van der Waal's forces are caused by fleeting interactions between electrons with constantly shifting electron distributions, electrostatic interactions are a result of partial or whole charges on atoms of a molecule with an overall neutral charge. As with van der Waal's interactions, electrostatic interactions occur between each pair of atoms in a system.

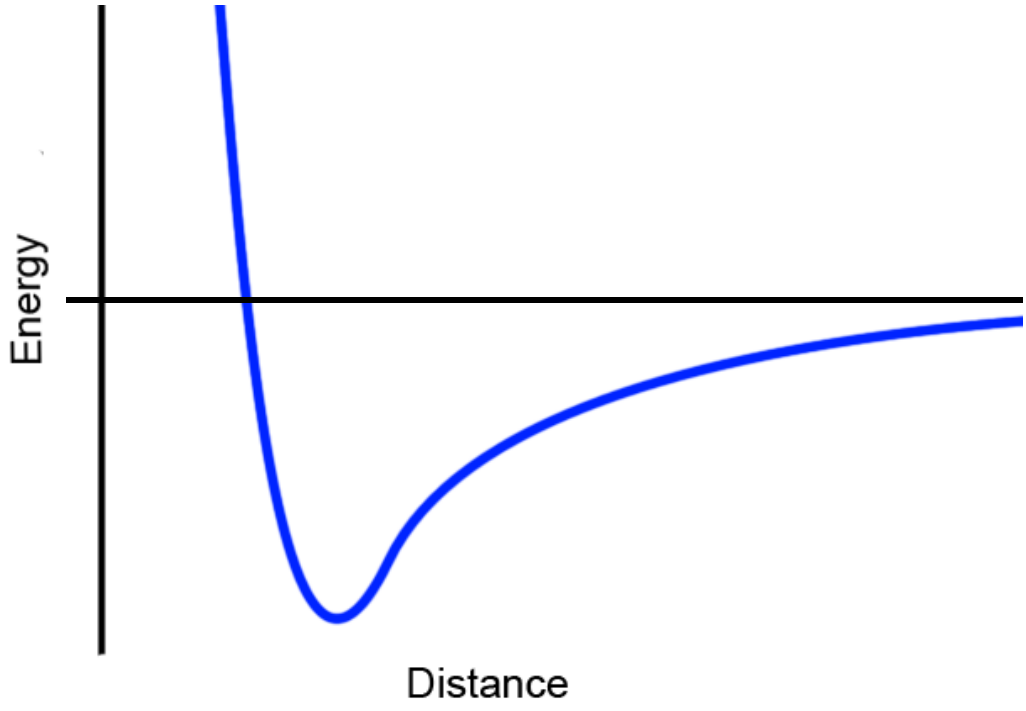

Figure 1.9: The Lennard-Jones potential.

To avoid the high computational cost of calculating long–range non–bonded interactions which normally have little impact on the system, a cut–off system may be implemented. For example, the 6-12 Lennard-Jones potential used to represent van der Waal's interactions tends towards zero as the distance between two atoms increases. A cut–off function can specify that there is no interaction beyond a certain distance, that distance being less than the range over which interactions are normally seen (typically around 9 Å). At the relevant distance the energy becomes zero, and remains so for all atom pairs further apart than the cut–off. The algorithm does not need to calculate any van der Waal's forces beyond the threshold distance, and so computational time is saved.

As can be seen from Figure 1.10, use of a cut–off distance alone can cause a discontinuity in the energy at the cut–off. As the calculation of force is based on the derivative of the energy, a discontinuity can prove problematic. A switching function can be used to

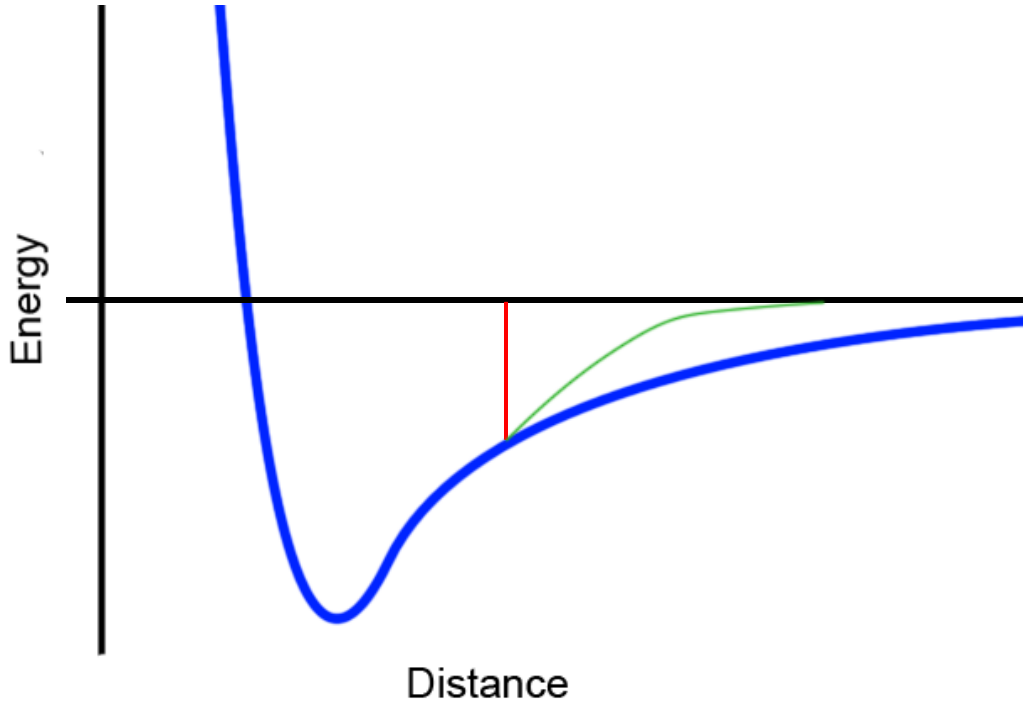

Figure 1.10: Cut–off (red) and switching (green) functions.

modify the potential in a manner which ensures that it truncates smoothly at the cut–off distance. Other methods designed to save time in the calculation of these forces and long–range electrostatics are Particle Mesh Ewald (PME) (63) and Particle-Particle Particle Mesh (P3M) (64), both of which entail fitting atomic charges onto a regular grid to facilitate faster computation.

There are a variety of ways in which the solvent surrounding macromolecules in molecular dynamics simulations can be treated. In early models, solvent was neglected altogether in vacuum simulations (52). The next major step was to represent the solvent in an implicit manner, a method which is still in widespread use. While atoms in the solute are represented explicitly, the solvent does not contain any explicit atoms. Instead, the solvent is represented by a bulk continuum around the solute, with a higher dielectric

constant. The aim is to determine the solvation free energy of the solute, that is, the free energy change when the molecule is moved from a vacuum to the medium. The Poisson-Boltzmann equation can be solved to determine the solvation energy, but doing so is very computationally expensive. While some programs do exist to calculate solutions numerically (65), they run relatively slowly.

The solvation free energy can be considered as the sum of a cavity term, a van der Waal's term for solute-solvent interactions, and a solute-solvent term to polarise the solvent around the solute (Equation 1.2.14).

$$G_{sol} = G_{cav} + G_{vdW} + G_{pol} \tag{1.2.14}$$

where G_{sol} is solvation free energy, G_{cav} is the cavity solvation energy term, G_{vdW} is solvation energy of solute-solvent interactions and G_{pol} is the polarisation component.

The first two terms in Equation 1.2.14 can be related to the sum of solvent accessible surface areas, as seen in Equation 1.2.15 (66).

$$\Delta G_{sol} \approx \sum_i \sigma_i ASA_i \tag{1.2.15}$$

where σ_i is an empirical atomic solvation parameter and ASA_i is solvent–accessible surface area.

This leaves only the electrostatic polarisation term, G_{pol}, to calculate.

The Generalised Born equation states that:

$$G_{pol} = -166\left(1 - \frac{1}{\epsilon}\right)\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{q_i q_j}{f_{GB}} \tag{1.2.16}$$

where q_i and q_j are the partial charges of atoms i and j, ϵ is the dielectric constant and r_{ij} is the distance between the two atoms.

where:

$$f_{GB} = r_{ij}^2 + b_i b_j \exp\left(-\frac{r_{ij}^2}{4 b_i b_j}\right) \tag{1.2.17}$$

It is important that b_i and b_j be calculated correctly; these are the Born radii of the atoms. The atomic radius is not appropriate, because the atom is not fully solvent–exposed. The effective Born radius is a parameter which reflects the degree by which the atom is "buried" within the solute molecule. By reducing the Born radius, we reduce the effect of the atom on the solvation energy. The aim is therefore to determine a Born radius which accurately reflects the contribution of the solvent–exposed area of the atom to the polarisation energy.

Some of these methods may also be supplemented with details of non-polar contributions to the solvation energy, calculated from the solvent–accessible surface area. These combined methods are often known as GBSA, or Generalised Born Surface Area, methods.

The effective energy function (EEF1) is a modification to the CHARMM19 polar energy term (67). It allows more accurate calculation of the solvation energy of the solute. This is achieved through solvent–exclusion principles, whereby the solvation energy of an atom is reduced in proportion to the amount by which other solute atoms shield it from

the solvent. The sum of these individual atomistic contributions gives the solvation energy of the solute.

Explicit solvent models represent each atom in the solvent in the same manner as those in the solute. The major drawback of this methodology is the requirement to calculate trajectories for all the solvent atoms; this requires a large amount of computing power. In biological systems the solute is usually water, and a variety of different models have been developed to simulate the interactions between the solvent and solute. The TIP3P model (68) uses a rigid representation of the three atoms making up each water molecule. Each atom is assigned a point charge, and this relatively simple model results in fast calculations. Other models may use different geometries, for example SPC (69), which uses an ideal tetrahedral geometry rather than the experimentally-derived geometry for water. In some cases a Lennard-Jones potential may also be applied to certain atoms (typically the oxygen) in addition to partial charges. A popular derivative of TIP3P is the four–site TIP4P model. In addition to the three atoms found in TIP3P, the model also includes a "dummy" atom with a partial charge designed to represent the oxygen lone pairs. Other more sophisticated models may include charges for each individual lone pair (70), and six–point models which combine the extra sites found in both TIP4P and TIP5P (71). While the various models have been shown to reproduce phenomena such as melting more accurately than other models, their increased complexity results in a greater computational requirement.

When using explicit solvent, the size of the solvent box surrounding the solute must be taken into account. The box must be large enough that its edges do not affect the solute, but not so large that valuable computational time is spent calculating trajectories for water molecules which have no discernible affect on the solute. Periodic boundary conditions are often implemented to overcome these problems. The main simulation cell

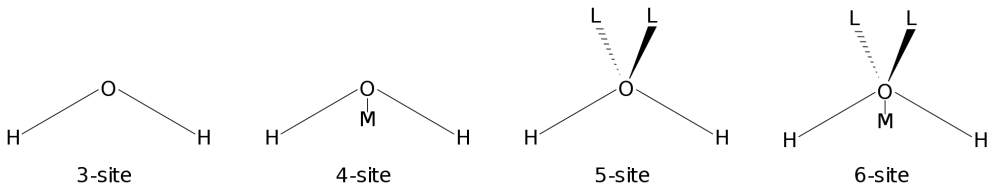

Figure 1.11: Schematic representation of 3- to 6-site water models. Exact geometries vary, depending on the particular model. *M* represents a "dummy" atom, *L* represents a lone pair.

is surrounded by a number of identical cells in all dimensions, the movements of which also mirror those of the main cell. If a solvent molecule moves out of one side of the main cell, an identical one therefore moves in to the cell from the opposite side. In this way, the total number of solvent molecules remains constant, and there are no edge or surface effects. It is important to ensure that there is sufficient distance between the solute and the edge of the cell to prevent solute molecules in neighbouring cells interacting with one another. While a cubic cell shape is common, a truncated octohedral cell can reduce the number of water molecules which need to be simulated, and thus speed up the calculation.

Another consideration is the treatment of temperature and pressure during simulations. Temperature is derived from the mean kinetic energy of the atoms in the simulation (Equation 1.2.18), while pressure can be thought of as the force applied to the boundary "walls" of the simulation box by the atoms within.

$$\bar{E}_k = \frac{3}{2}NkT \tag{1.2.18}$$

where $\bar{E}_k$ is average kinetic energy, N is the number of degrees of freedeom, k is the Bolztmann constant, and T is temperature.

The simplest ensemble type is where the number of atoms (N), the volume of the simulation (V) and the energy (E) remain constant throughout the simulation. This is known as the microcanonical ensemble. The pressure and temperature of the simulation cell may alter; this allows time–averaged values to be calculated, as all microstates in the ensemble are equally likely to be sampled.

In a canonical ensemble, the number of atoms (N), the volume of the simulation (V), and the temperature (T) all remain constant. This is also known as NVT dynamics. NVT simulations use a thermostat to ensure that the temperature of the system does not vary. The simplest method for maintaining temperature is the direct scaling of atomic velocities, but small temperature fluctuations as would be expected in biological systems are inhibited by this method. The Berendsen thermostat rescales atomic velocities in a similar manner, but the rate of temperature change is kept proportional to the temperature difference (69). This results in the system tending towards the specified temperature, rather than always being reset to the same value. A more tightly-regulated version of NVT can be implemented with the Nosé-Hoover algorithm, which models heat exchange between the system and a heat bath (72, 73). This entails the addition of an extra, dimensionless, degree of freedom to the system, which is then used to apply friction to the system and therefore reduce atomic velocities and temperature. Other heat bath systems have been proposed, such as that developed by Andersen (74) whereby the velocity of a randomly selected atom is replaced by one from a Maxwell-Boltzmann distribution.

An isothermic-isobaric ensemble involves maintaining the system at a constant temperature but also a constant pressure (also known as NPT). For this reason, a thermostat similar to those used in NVT simulations is required, along with a barostat. A number of barostat models have been proposed to work in similar ways to thermostats. Berendsen's method (69) involves scaling the size of the simulation cell and the atomic coordinates

to maintain a constant pressure (as opposed to scaling velocities to maintain temperature). Other systems have been developed to maintain pressure in a manner more akin to the Nosé-Hoover thermostat, that is, the addition of an extra degree of freedom to effect changes in pressure. This type of system can be conceptualised as the addition of a piston to the simulation, albeit one which adjusts the system volume in an isotropic or anisotropic manner rather than a traditional cylindrical piston.

Monte Carlo Simulations

As discussed previously, the native state of a protein is a low–energy conformation, as are structures seen in amyloid plaques. For this reason it is often desirable to derive low–energy structures computationally. While this can be done by studying how a system evolves over time, there are alternative ways to sample a protein's configurational space. One commonly used method in such work is that of Monte Carlo simulation.

A starting structure is passed to the algorithm, and the energy calculated. A change is made to the structure, and the energy is re–calculated. If the new structure has a lower energy than the original, it is carried forward and the process repeated. If the new structure is less stable, there is still the possibility of accepting it, according to a probability known as the Metropolis criterion (Equation 1.2.19).

$$p = e^{-\Delta E/RT} \tag{1.2.19}$$

where p is the probability of accepting a move, E is potential energy, R is the ideal gas constant, and T is temperature.

This methodology leads to a situation whereby more stable conformations are always accepted, *i.e.* movement towards the bottom of the energy landscape. However, the pos-

sibility of moving to higher–energy structures is always present, and this prevents the simulation becoming stuck in local minima. By altering the temperature, the probability of accepting less stable structures can be amended, and thus the tendency of the model to "jump" around the energy landscape can be influenced. Such methods have been in use since the 1970s, when they were used to probe secondary–structure elements (75), along with basic models of protein folding (76, 77). More recently, alternative methods to explore landscapes have been developed. For example, "model–hopping" (78) involves swapping the energy functions used for the various simulations in an ensemble. One simulation may utilise several different functions over its duration, enabling variations in the fidelity of its "walk" around the energy landscape. Again, this reduces the probability of the simulation becoming held in a local minimum.

1.2.2 The Problems Faced

There are several problems to be overcome when carrying out computer simulations of biological systems. The main restriction on the development of *in silico* models has always been the availability of sufficient computing power to deal with larger systems. However, there are other factors which can prevent a system being simulated sufficiently accurately to produce a worthwhile model.

Accuracy of the Models

Careful consideration of all the parameters involved in a simulation is necessary if the data obtained are to be of any value. There are a variety of factors which can cause a simulation to fail; some are avoidable, others are assumptions inherent in the theories and models being used.

Of critical importance is the selection of the time step used in the simulation. Obviously a large time step is desirable as it minimises the number of calculations required and maximises the timescale of the simulation. However, the time step must remain shorter than the fastest vibrational motion in the system. In biological systems this is most often the covalent bonds to hydrogen. As a result, the time step must be kept to 1 fs. When using implicit solvent in the absence of hydrogen atoms, a 2 fs time step may be used. If explicit solvent is being used, it is possible to fix the covalent bond and prevent its vibration (for example, the SHAKE algorithm (79)). This also allows a 2 fs time step. Algorithms have been developed to use multiple different time steps in a single simulation and thereby increase the speed of the calculations, but these are not in widespread use (80, 81).

The accuracy of the force field being used is one area in which a variety of approximations are introduced. These are often a bid to increase computational efficiency, with the resultant increase in simulation speed offset against a minimal decrease in the accuracy of the model. For example, the Lennard-Jones potential was originally described as a generic term, $1/r^n$ - $1/r^m$, with the subsequent use of $1/r^{12}$ and $1/r^6$ in molecular dynamics being based on the computational efficiency of squaring the latter to obtain the former. Switching and shifting functions allow researchers to neglect long–range non–bonded interactions, to increase the speed of simulations. Indeed, all the approximations indicate a balance between accuracy and speed of simulation, a point noted by van Gunsteren in 1990 (82) but just as valid today.

The methods reviewed here are all designed to increase the accuracy of an MD force field, but there are some properties which remain difficult to model. The fixed point charges commonly placed on a solvent molecule in MD are not an accurate reflection of the true charge distribution. The distribution is dynamic in nature, moving around the

molecule rather than remaining static, and solvents are therefore polarisable. Although many models do not include the added complexity of polarisability in their consideration of system energy, a number of polarisable force fields for water do exist (83, 84, 85). More generally, they cover a range of solvents (86), and may focus on specific systems such as protein–ligand interactions (87). However, specialised force fields have also been developed for application to macromolcular biological systems (88, 89, 90, 91).

Regardless of the accuracy of the force field, a simulation must sample the conformational space sufficiently. A failure to do so could result in the model not finding the "best" structure, pathway or event in a given scenario. The algorithm chosen must allow barriers in the energy landscape to be overcome, but it is also important that sufficient temporal sampling also takes place. While the simulation of processes such as protein folding will obviously require a long time to model, it is also important to ensure that, for example, Monte Carlo simulations are given sufficient time to explore the conformational space thoroughly.

Computational Power

One of the main factors preventing continued development of molecular dynamics methods is the nature of the pair-wise interactions, the properties of which must be calculated at each step. Consequently as the system size increases, the number of interactions, and therefore calculations, also increases. As the development of computing power seems to be matching Moore's law, the capabilities of current machines are much greater than those of even a few years ago. For example, in 1990 van Gunsteren wrote that a "simulations of liquids typically involve 10^2 or 10^3 atoms" (82), whereas is it not unusual for current simulations to contain over 10^4 atoms. Even with increasing computing power, the time taken to run simulations of events like protein folding, especially given the

probabilistic nature of such events, can still make them impractical.

Recently, research has been focusing on ways of leveraging cluster CPU power in a more parallel fashion. There are a wide range of possible cluster setups, ranging from local systems with high–speed interconnects to "grid" systems with nodes at a number of geographical locations (for example, the Folding@Home project (18)). There are also various ways to split the data between the computing nodes. However, increases in computing power are not the only solutions being investigated, with the continued development of coarse-grain models (see section 1.2.3) to maximise the use of existing facilities.

1.2.3 Coarse-Grain Models

Given the discrepancy in time between *in vivo* and *in silico* protein folding events, there has always been a drive to increase the efficiency of the computational models. One way to achieve this is to move from atomistic representations of proteins into simpler models. This may be a simplification of the conformational space, or in the representation of the residues themselves.

Lattice Models

Any algorithm must be capable of efficiently searching the conformational space of the molecule being studied. In the case of atomistic models, that space is usually so large that it takes a long time to sample it sufficiently. One way to reduce the size of the search space, without reducing the conformational space too severely, is to implement a lattice framework. This entails restricting all atoms, or residues, to points on a regular lattice. The protein structure then takes the form of a self-avoiding walk. The coordination number of the lattice is important (Figure 1.12); a higher coordination number provides

a better fit to biological structures, but the more complex lattice also results in longer computation times.

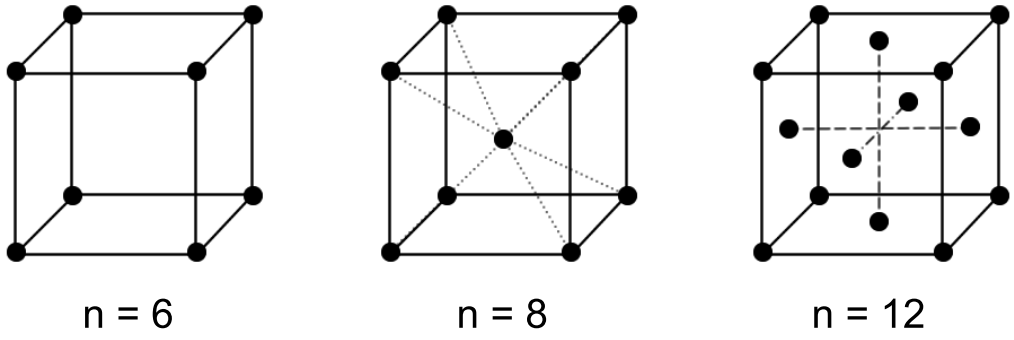

Figure 1.12: Cubic, body-centred cubic and face-centred cubic lattices, and their coordination numbers (92).

Some of the simplest lattice models were first developed to try and explain the fundamental principles behind protein folding. By using small peptides on simple lattices, several groups were able to reveal factors which influence protein folding. One model often used on a lattice, although not restricted to coarse–grain applications, is that developed by Gō *et al.* (93). It specifies that native contacts have a favourable energy, while other contacts have zero energy. This bias ensures that a search for the lowest energy conformation will be guided towards the native state. In this way, an extended protein conformation can be provided to the algorithm, and the development from that extended state to the native state can be seen. Similar work with lattice models has revealed the role of funnel energy landscapes (94) and the importance of hydrophobic interactions (95). Li *et al.* (96) used both 2D (square) and 3D (cubic) lattices to look at what they termed the "designability" of proteins. They found evidence for the enhanced stability of regular structures, such as secondary structure elements. The phase diagram of a simple protein, showing transitions between molten globule, random coil and "frozen" native state was derived by Dinner *et al.* using a 2D lattice (97).

Other models have subsequently looked at the folding of specific proteins, that is, the manner in which individual proteins fold on lattices rather than the general principles behind all folding. They primarily involve the development of potentials to simulate folding of proteins, and are based on statistical data such as that found in the Protein Data Bank (PDB). The models include work by Miyazawa and Jernigan (98, 99), who developed potential data for each of the pairwise interactions between the 20 amino acids. Skolnick and Kolinski (100) used an advanced potential to fold a globular protein on a 210 lattice, resulting in a structure similar to the native state. Their potential combined a slight bias towards the native state with hydrophobicity data from the Miyazawa–Jernigan work.

Model Resolution

While a lattice can, in theory, be employed on any resolution model, they are most commonly used in models employing a reduced resolution. Rather than including each atom in the molecule, these systems group heavy atoms together into larger particles. The method was first introduced by Levitt and Warshel in 1975 (101), and has remained in constant use since. Their simulations of pancreatic trypsin inhibitor used two centres for each residue, one on the C_α and one on the C_β to represent the side–chain. Other variations since have used single centroids for each residue, thus further reducing the number of particles in the system (102). The benefit of these changes is a reduction in the total number of particles in the system, and therefore a reduction in the number of pairwise interactions to be calculated at each time step. This in turns means that less time is taken to carry out the simulation.

Hydrophobic-Polar Models

The Hydrophobic-Polar (HP) model was developed as a way to simplify simulations by reducing the complexity of interactions between particles in a system. It was proposed in 1985 by Dill *et al.* (103), and is based on the observation that interactions between hydrophobic residues are important for protein folding ("hydrophobic collapse"). These residues are attracted to each other and favour a position in the centre of the protein, away from the polar solvent. By classifying each of the 20 amino acids as either hydrophobic or polar, and assigning a favourable energy to interactions between hydrophobic residues, this phenomenon can be reproduced. The HP model has been used extensively in lattice simulations (95, 96), on both 2D and 3D lattices. Modifications have been made to the model. Blackburne and Hirst used a 3D diamond lattice to investigate the 'designability' of proteins (104). Their model utilised the 'shifted-HP' variant of the HP technique, which includes a term to prevent very compact structures being formed. Decatur developed the standard HP potential to include varying hydrophobicity values for each of the amino acids (105).

Multiscale Models

Coarse–grain models offer a significant time saving over atomistic models. However, it is also possible to combine both methods. One option is to use the two techniques independently; de Mori *et al.* used a coarse–grain Monte Carlo search algorithm to generate compact structures from an extended conformation of the villin headpiece (106). These candidate structures were then fed into an atomistic model for more detailed simulations. Aspartic proteases have been studied using conventional molecular mechanics around the active site region, but single particles centred on C_α atoms around the rest of

the protein (107). Using both representations in the system allowed the authors to maintain a more detailed model in the complex active site, while retaining the influence of the rest of the molecule. This is essentially a similar principle to that of QM/MM, where quantum mechanics (QM) are paired with molecular mechanics (MM). Other multiscale methods entail moving molecule representations between different resolution models or force fields. The "ResEx" model developed by Lyman (108) allows simulations to swap between coarse-grain and atomistic levels of detail. This permits greater sampling, thus avoiding techniques such as increased temperature, but retains detailed simulations as well. "Model Hopping" (78) modifies this concept by alternating between different force fields rather than resolutions. Moving models between force fields helps them overcome energy barriers, for example, by moving to a potential where the limiting factor does not influence the energy as much.

1.3 Forced Protein Unfolding

Anfinsen's work showed that all the information necessary to fold a protein was contained within the primary sequence, and that denatured proteins could re–fold under favourable conditions (6). His work used chemical denaturants to unfold a simple protein, but more recently unfolding experiments have been able to focus on mechanical proteins. These are molecules which are subject to force as part of their native function, and are therefore resistant to unfolding. Some may even repeatedly fold and unfold naturally as part of their biological activity. The development of methods which apply forces to these proteins (both *in vitro* and *in silico*) has allowed the investigation of not only mechanical strength, but also folding/unfolding energy landscapes, folding barriers, intermediates, and natural re–folding processes.

1.3.1 Experimental Techniques

Atomic Force Microscopy

The atomic force microscope (AFM) is a type of scanning probe microscope (109), developed from the scanning tunnelling microscope in 1986 (110). It utilises a cantilever with a probe tip mounted at one end. The cantilever is positioned so that the tip is brought almost into contact with the surface of the sample. Over the course of the experiment, the tip is moved across the surface. Non–bonded interactions between the tip and the surface cause the tip to move up and down in parallel with the surface topography, deflecting the cantilever by varying amounts in the process. This deflection can be recorded and used to reconstruct a representation of the surface. In tapping mode, variations in the oscillation frequency of the tip are used to determine surface features. A small tip is crucial to AFM, with a typical radius of curvature of less than 50 nm. Depending on the experimental setup, AFM can be capable of a vertical resolution of 0.01 nm.

While AFM is often used as a technique for probing non–biological surfaces to determine their topography, it has also been used more widely. For example, conformational changes in lysozyme were observed from a change in the height of the molecule (112). Florin *et al.* (113) measured the force required to separate a variety of ligand–receptor pairs, by coating a surface with receptor and the AFM tip with ligand. By first bringing the AFM tip onto the surface to allow binding, and then pulling it away again, the force required to separate the two molecules could be measured. A similar procedure was carried out with streptavidin and biotin by Allen *et al.* (114). Measurements on intra– and inter–molecular bonding in DNA were subsequently carried out by Lee *et al.* (115). By 1997, AFM was being used to determine a number of properties of biological molecules, including elasticity, variations in sample surface composition and various length mea-

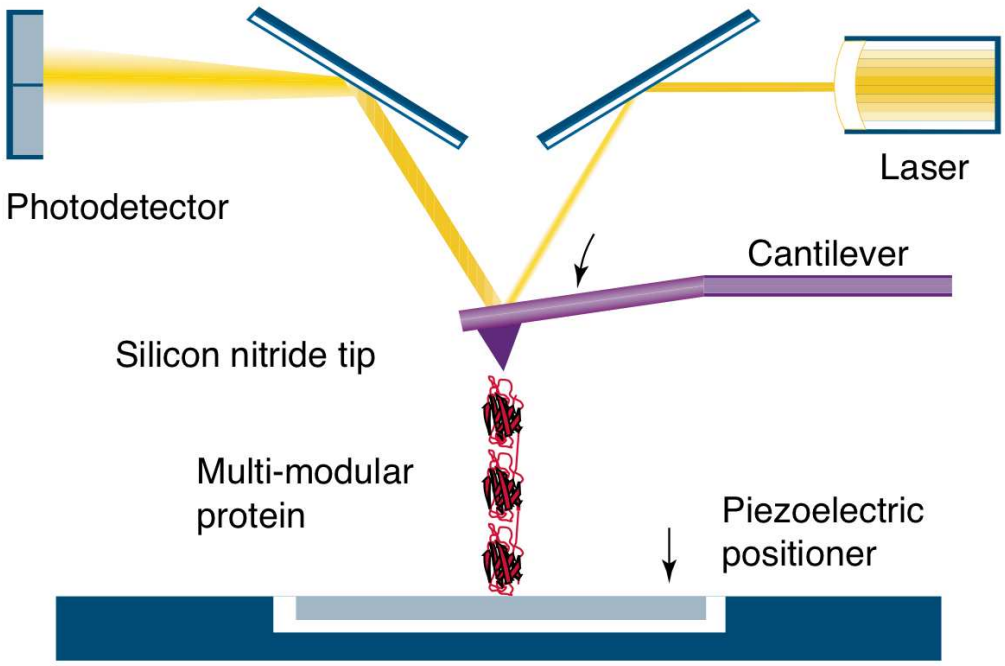

Figure 1.13: Typical AFM apparatus for forced unfolding of a multi–domain protein. Reproduced from (111).

surements of cells and DNA (116). Fluorescence and circular dichroism had previously suggested that the muscle protein titin unfolded in a reversible manner (117, 118). The protein proved a suitable target for studies investigating the force required to simulate natural unfolding via the breaking of intramolecular bonds. Rief *et al.* (119) carried out the first of a series of experiments, and saw the now well–recognised sawtooth pattern in the force curves (Figure 1.14). This indicated that individual domains within the molecule unfolded sequentially rather than cooperatively. These were the first of many experiments on titin and its component domains (discussed in further detail in section 3.1). The presence of an intermediate in the stepwise unfolding of spectrin has also been demonstrated (120), the modular nature of the extra-cellular matrix protein tenascin confirmed (121), and the effect of force on energy landscapes has been investigated (122).

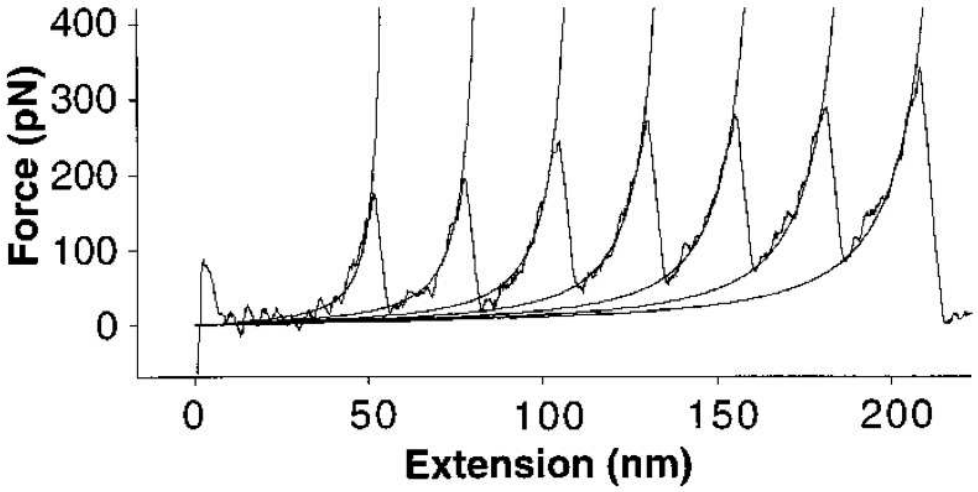

Figure 1.14: Typical sawtooth force extension curve for multimeric titin I27. Each peak represents a single I27 domain unfolding under force (119).

1.3.2 Molecular Dynamics Simulations of Protein Unfolding

Developments in the application and accuracy of AFM have allowed researchers to probe new aspects of unfolding processes and energy landscapes. The ability to simulate these computationally has developed at the same time, and provides complementary data to those obtained from experimental work. However, the length of time taken to carry out atomistic simulations has meant that several strategies have been applied to facilitate unfolding within a feasible timescale.

Elevated Temperature

In some cases, the temperature of the simulations can be increased in order to promote unfolding. The first simulations of protein unfolding, which looked at non–mechanical proteins, utilised temperatures of 498 K to unfold an 83–residue protein in explicit solvent over 2.2 ns (123). While these results, and others (124) suggest that elevated temperatures may be a useful way of reducing computational overhead, their application

should be carefully considered. Paci and Karplus compared molecular dynamics simulations at elevated temperatures (300 K to 500 K) to unfolding by force (125). They found that the two methods of unfolding gave different results, with the force–induced unfolding revealing intermediates which were not present in the temperature–induced unfolding pathways. The high temperature runs showed a much smoother unfolding profile than the steered molecular dynamics (SMD) runs, with the N and C termini not moving apart until late in the simulation.

Constant Velocity vs. Constant Force

SMD simulations by Lu and Schulten (126) revealed, in detail, the breaking of crucial hydrogen bonds between β–strands in the structure of the I27 domain of titin. Their simulations were carried out in both constant force and constant velocity modes. The former entailed the application of a number of different forces to the N–terminus of the protein, while the position of the C–terminus remained fixed. Despite the fact that AFM work had previously shown only 150 pN of force were required to induce unfolding over a period of milliseconds (119, 121), Lu and Schulten had to apply forces between 750 pN and 1200 pN to witness unfolding over 1.25 μs of simulation time. In this case, excessive force had to be applied in order to cause the protein to unfold within a computationally accessible timescale. Similarly, constant force simulations resulted in a high peak force of 2100 pN. Similar forces had been required in previous SMD simulations by the group (127).

The speed of constant velocity simulations was also increased to ensure that the molecule unfolded. AFM work had pulled I27 apart at a speed of 1 μm/s, whereas simulations required speeds of between 0.1 and 0.5 Å/ps. The latter equates to 5 x 10^7 μm/s, or a difference of more than seven orders of magnitude. Lu and Schulten suggest that their

results are qualitatively similar to those obtained from AFM, and that they can be scaled to give agreement with the experiments. However, any improvements to molecular dynamics protocols which bring them closer to physiological conditions would be useful; this fact is also acknowledged by Gao *et al.* (128). As researchers are provided with ever more computing power, so the fidelity of simulations can be increased and the severity of the simulation conditions reduced. This in turn should allow simulations to be carried out using, for example, temperatures and forces much more akin to those found physiologically than have previously been required. Rather than relying solely on increased computing power, it is important to optimise use of the existing resource through the development of new methods. Only by combining these tactics will *in silico* work be able to inform and complement *in vitro* studies.

1.4 Aims and Objectives

The aim of this work is to investigate methods which may increase the efficiency and scope of computer simulations of protein folding and unfolding. The ability to better simulate physiological conditions is also considered.

In Chapter 2, a lattice model of protein folding is developed. The model uses a coarse–grain representation of a polypeptide, with each residue bound to a lattice point. This reduces both the number of pairwise interactions to be calculated, and the search space, resulting in a program which can quickly generate structures and determine those with the lowest energy.

The technique of milestoning is discussed in Chapter 3. Unlike the lattice model, this method retains all–atom representations of proteins. In order to increase the simulation efficiency, a trajectory is split into a number of sections. This allows each section to be

run in parallel, but also has a number of other benefits that increase the probability of witnessing rare events and allow the application of lower forces than has previously been possible.

Chapter 4 utilises the efficiency of the lattice model to generate a dynamic trajectory using an alternative method to the integration of Newton's laws of motion. By considering the principle of Least Action, the most probable path from a start structure to an end structure is derived, providing a number of structures between to make up a full trajectory.

Chapter 2

LaMP - A Lattice Model of Proteins

2.1 Origins of the Model and Minor Developments

The long-term aim of the LaMP project is to develop a coarse–grain lattice model to study protein folding and aggregation. This builds upon existing work which has utilised one or both of the main elements of this model. The first coarse–grain model to be successfully applied to protein folding was developed by Levitt and Warshal in 1975 (101). They acknowledged that the number of degrees of freedom in an atomistic simulation was too great to model effectively, so sought a method to reduce this. Each amino acid was represented by two simulation particles. As their simulations progressed and suggested stable folded structures, they moved towards increasingly less coarse–grain models, eventually creating final candidate structures from fully atomistic simulations. Their conclusions also touched on the general subject of how folding takes place, an issue which was still being debated fiercely at the time. Miyazawa and Jernigen defined a simple residue:residue potential in 1985 (98), allowing amino acids to be treated as single particles in simulations rather than aggregations of multiple atoms or points. While

originally discussed by Levitt and Warshel in 1975, the development of methods to re–construct all–atom models from α–carbon backbones took many years but has proved complementary (129, 130). While many coarse–grain and lattice methods have simulated protein folding, no definitive model has emerged. The combination of these features, in an efficient model capable of accurately creating tertiary protein structure, has yet to be achieved. Current coarse–grain models have progressed beyond the HP model which has persisted for many years, with the off–lattice BLN model proteins forming a target for research (131). These group residues into the traditional hydrophobic (H), but also hydrophilic (L) and neutral (N). These have been used to model generic tertiary structure elements such as the β–hairpin and β–barrel (132, 133). Nevertheless, the HP model remains an important tool which has been extended in a similar manner to create the generic HPNX model and derivatives (134). Search strategies are also growing to encompass alternative methods such as genetic algorithms (135, 136). Lattice models, meanwhile, have also grown in complexity, not least in the lattice types used. The first models used the simple cubic lattice (137) moving on to the more complex cubo–octohedral (138) as available computing power increased, and finally on to face–centred cubic (139). Some models have also experimented with the imposition of conditions on the ability of atoms to occupy particular points on a conventional Bravais lattice, resulting in the 210 lattice, for example (100, 140). Consideration has been given to the initial transfer of the protein to a lattice (141), the move sets used (137), and to the representation of water both implicitly and explicitly (142, 143). The simulation of hydrogen bonding and the subsequent formation of secondary structure has proved a more difficult challenge. Both water models have been employed in attempts to create energy functions which accurately promote the formation of α–helix (144, 145) and β–sheet (139, 146). While models have been reported as capable of forming both helix and

sheet (102), no preferred method has yet emerged which displays sufficient homology with experimental results to be applied to larger systems without prior knowledge of the folded structure.

LaMP functionality was, in common with most protein folding models (147), to be accomplished by presenting the program with a starting all–atom, off–lattice structure and having it generate the lowest–energy permutation of that structure (according to the force field parameters). As native and aggregated states represent energy minima, the resultant structures could prove useful in a number of research areas. The LaMP package consists of two main programs; *PDB2Lattice* converts off–lattice structures into a lattice format, and *LatticeSearch* uses the resultant lattice structures to search for the most stable conformation.

While the first version of the model was capable of finding very low energy structures very quickly, this was subject to the limitations of the force field. As a result, the structures generated were not biologically relevant, and the package had the potential to become more user–friendly. While the development of the force field was the primary objective, discussed in section 2.2.2, a number of smaller improvements to both the usability and accuracy of the model are discussed below.

2.1.1 PDB2Lattice

The first step in running a lattice simulation is the conversion of a standard Protein Data Bank (PDB) file from an all-atom, off-lattice structure to a coarse-grain lattice representation. The *PDB2Lattice* program is used to carry out this conversion. The program extracts the coordinates of each Cα molecule from the PDB file. Coordinates for all other atoms are discarded. Since version 1.0 of the program, compatibility with the PDB file format (148) has been significantly increased. Each of the Cα atoms is used to represent the cen-

tre of the entire residue in which it was originally incorporated. A lattice grid is overlaid onto the off–lattice structure. Each residue is moved to the nearest lattice point. Finally, a number of Monte Carlo search steps (discussed in section 2.1.2) are carried out to minimise the RMSD of the lattice structure compared to the off-lattice structure.

The actual fitting process has also been improved in the latest version of the program. Originally, a search of 10,000 Monte Carlo fitting steps was carried out regardless of the size of the protein. Whilst sufficient with smaller polypeptides, this gave poorer results with large proteins. The relatively small number of steps did not sufficiently search conformational space. The number of search steps is now therefore proportional to the size of the molecule being fitted, with a minimum value of 10,000 steps. This resulted in the model being able to find a lower energy structure on approximately 20% more occasions than with the default number of steps.

Another problem with the first version of the program was a tendency to create overlapping residues, that is, for two residues to be occupying the same lattice coordinate, and therefore the same point in space. Whilst these overlaps were detected, and the RMSD of the fit weighted to reduce them, the situation was not ideal. The current version has minimised this problem by detecting such clashes, and moving one of the residues to the next-nearest unoccupied lattice point. In a test set of 109 proteins with a total of 9062 residues, this reduced the total number of overlapping residues from 29 residues in 22 proteins to zero. The mean RMSD to off-lattice structure was 1.4077 Å prior to the modification, and 0.86 Å after. The model is therefore closer to guaranteeing a true 'self–avoiding walk' for lattice structures.

2.1.2 Lattice Search

Metropolis Monte Carlo Search

Monte Carlo methods provide a way of simulating systems using random sampling. While there are a number of related methods, the general principle involves defining a set of possible inputs for the system, and then repeating three steps. Inputs are generated at random and a deterministic calculation is performed on the input; these two steps are repeated multiple times, before the results are aggregated. Monte Carlo methods are often used when there is no simple deterministic solution to the problem being investigated, and they are therefore particularly suitable for molecular modelling applications. The methods were first described in 1949 by Metropolis *et al.* (149), and have been used extensively since the 1970s in computational work relating to protein folding, solvent models, forced unfolding of proteins, and protein aggregation (68, 76, 150, 151).

In LaMP, the Metropolis Monte Carlo method (152) is used to search for the lowest energy structures. A starting structure is provided to the algorithm, following which a single alteration is made to the structure. The energy of the molecule is calculated; the individual terms used are discussed in more detail in Section 2.2.1. If the new structure is more stable than the original, it becomes the new starting structure. If the structure is less stable, it is accepted according to a probability represented by equation 2.1.1.

$$p = e^{-\Delta E/RT} \tag{2.1.1}$$

where p is the probability of accepting a move, E is potential energy, R is the ideal gas constant, and T is temperature.

The new structure has an alteration made to it, and the cycle is repeated a number of times. This allows progression to lower energies, while still allowing the model to escape from local energy minima, which may otherwise act as traps. This behaviour can be modified by adjusting the parameter T, the temperature of the system. Since both the temperature and the energy of the model are arbitrary, the ideal gas constant, R, is set to one and thus R and T can be treated as unity. Using a low temperature gives a more thorough search of a small area of the energy landscape, but may result in searches becoming trapped in local energy minima. A higher temperature allows the search to move out of such minima and across a larger portion of the landscape. However, this high temperature can result in the global potential energy minimum also no longer being the global free energy minimum.

The treatment of temperature is therefore an important factor when running Monte Carlo simulations, and a number of optimisations have arisen as a result. A plain Monte Carlo search utilises a constant temperature, T, but choosing this temperature can be problematic. Ideally, the goal is to combine the thorough searching at low temperatures with the wider landscape searching of higher temperatures. Simulated Annealing (SA) (153) involves beginning the simulation at a high temperature, and reducing it over the duration of the calculation. In LaMP, the temperature is reduced in a linear fashion over the course of the simulation. Replica Exchange (REx) (154) entails setting up multiple simulations (replicas) at regular intervals between an upper and a lower temperature. At regular time intervals during the search, structures from two of the replicas are exchanged. Thus any one search can run at both high and low temperatures over the course of a search, providing an improved combination of search breadth and fidelity.

LaMP provides plain, SA and REx searches, the parameters for which are set in the search input file. One potential optimisation of the model would be the utilisation of multiple processors in parallel during REx searches.

Lattice Types

Using a lattice reduces the conformational space being searched, and therefore increases search speed. Different lattice types have differing degrees of accuracy (101, 155). While many folding models use square and cubic lattices (143, 145, 156), LaMP uses the face–centred cubic (FCC) by default.

There are a number of factors to be considered when choosing a lattice. The atomic packing factor for the cubic lattice is 0.524, whereas the FCC lattice has packing factor of 0.740. This indicates that the latter fills more space than the former; the "dead" space not filled by the model is conformational space which is not accessible by the search. Hence, FCC is preferable for an accurate model. The hexagonal close-packed (HCP) lattice has a similarly high atomic packing factor, but it has been suggested that protein crystal structures are more akin to FCC (157, 158). Although more accurate than simple lattices such as cubic, the FCC lattice is more complex due to a higher coordination number. The coordination number of 12 is the highest possible bulk coordination number.

Altering the lattice type can provide more accurate results as a product of an increased atomic packing factor, but the downside of this method is one familiar to almost all areas of computational chemistry; increased calculation complexity and therefore increased run time for simulations. For example, a cubic lattice provides n^6 possible conformations for a peptide of n residues in length. Moving to an FCC lattice gives n^{12} possibilities, and therefore requires a much longer search to enumerate all structures (or more realistically, a large proportion of them).

To increase the scope of the model, the ability to read in custom lattice formats has been incorporated. This functionality allows the type of lattice to be defined. The lattice must be described using an XML file (Figure 2.1) which contains information about the geometry of the lattice, including both the distances between residues (not shown), and the moves available on the lattice. Custom lattice types can be used in both the *PDB2Lattice* and *Lattice Search* programs.

```
<coordination>8</coordination>
<moveset>
      <x moves="1,1,0,0,-1,-1,0,0" />
      <y moves="0,0,1,1,0,0,-1,-1" />
      <z moves="1,-1,1,-1,1,-1,1,-1" />
</moveset>
```

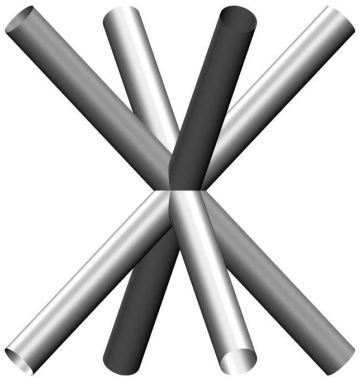

Figure 2.1: Excerpt from a custom lattice file for body-centred cubic, along with the moves represented.

Move Sets

During a lattice search, a single alteration is made to the structure on the lattice between each Monte Carlo energy evaluation. This may be the movement of a single residue, or entire sections of the protein. It is important to ensure that the set of possible moves addresses both local and non-local movements, that is, both the movement of individual residues and a more generalised movement of larger sections of the molecule. To this end, a variety of possible moves are available within the model. Each of the possible moves has an equal chance of being selected. The available moves are described below.

The 'moveBy2Bonds' move involves the transposition of a single residue from its current position to a new, randomly-selected position (Figure 2.2). If the new position is already occupied, or results in the chain being broken, the move is abandoned.

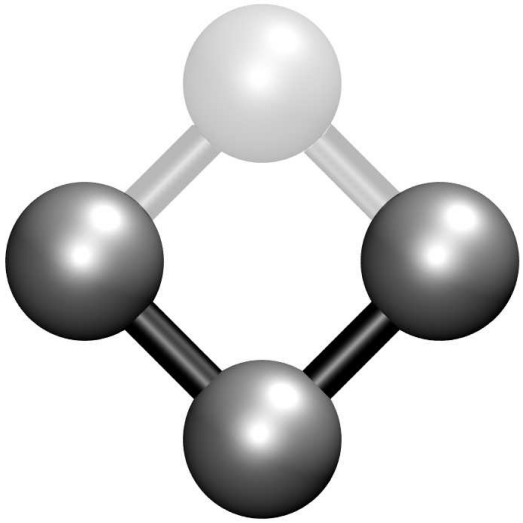

Figure 2.2: The 'move by two bonds' move - a single residue moves from the top position to the bottom, causing two bonds to also move.

This move was optimised from the initial version of the software, as some moves are not possible from certain starting conformations. For example, when three residues lie in a straight line it is not possible to move the middle residue and retain the correct lattice distances between all residues. If this move is attempted on such a conformation, it will be abandoned immediately rather than taking time to fail all possible moves.

'moveByReptation' moves the first residue in the chain to a randomly-selected new position, and moves each subsequent residue to the position previously occupied by its neighbour. This method has the effect of 'pulling' the chain along the lattice, and was first developed to describe the movement of a polymer in a gel (159).

The 'invert' move reverses the sequences of residues in the current structure, thus causing the C-terminus to move to the current position of the N-terminus and vice versa. All

other residues also move to their respective inverted positions. The final move in the set is a more localised alteration, 'moveByTerminalSwing', in which either the C or N terminus is chosen at random. The selected terminus is moved to a free lattice point nearby, the selection of which is again random.

2.1.3 Development

Initial development of the LaMP package was by Oakley *et al.* (160). The package was written using version 1.5 of the Java programming language (Sun Microsystems). Documentation for the entire project may be generated using the JavaDoc feature. All development was carried out using the Eclipse IDE, with JUnit unit testing procedures and continuous integration support from CruiseControl. The package was tested on a combination of openSUSE Linux, Apple Mac and Microsoft Windows operating systems. Individual calculations used a variety of Linux and Apple Mac computers, through a Condor computing pool.

2.2 Force Field Development

While it is computationally feasible for a model to find the lowest energy structure according to a particular function, it is the quality of that energy function which determines the quality of the resultant structures. A variation in the function will result in a subsequent variation in the tendency of the model to create conformations with particular features. Version 1.0 of the software utilised a limited function, causing it to create structures that, while the lowest energy for that particular version, were not so biologically relevant (Figure 2.3).

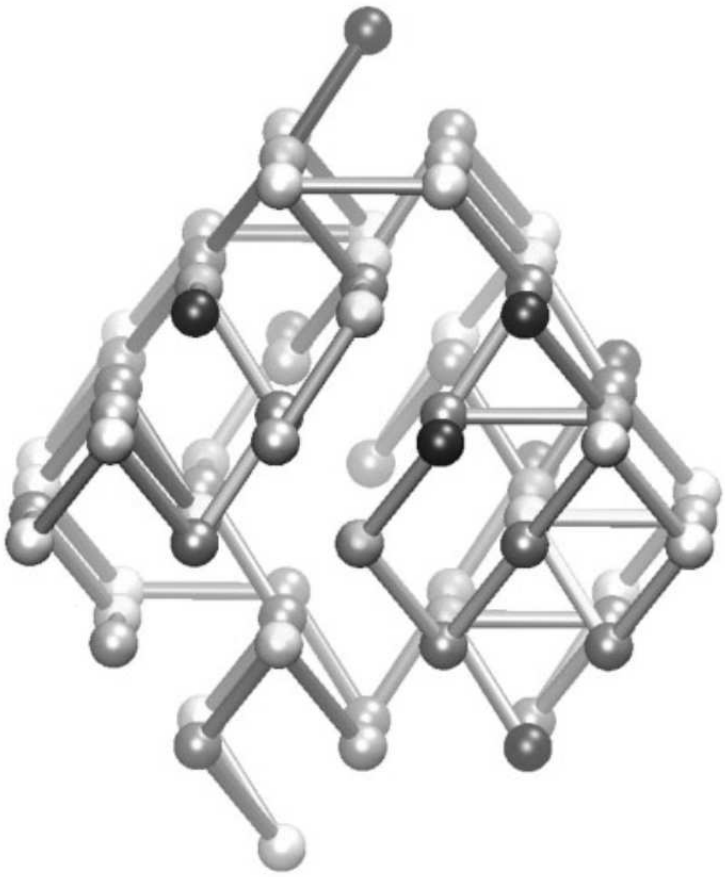

Figure 2.3: LMVGGVVIA octamer structure generated by LaMP version 1.0 (160). Note the compact conformation and lack of secondary structure.

Important aspects such as hydrogen bonding were missing, resulting in the model being unable to generate, for example, secondary structure elements. It is therefore important that the force field be defined in a manner which either incorporated these aspects, or which generates final structures with similar features to those found in nature (*i.e.* secondary structure elements).

2.2.1 Existing Terms

The force field used in initial work by Oakley *et al.* (160) consisted of three terms, discussed below (Equation 2.2.1). The energy values discussed in each of the following

sections are contained as default values within the LaMP programs, and relate to the default FCC lattice. However, they can all be overridden using a custom force–field file at runtime if required.

$$E = \Sigma(E_p^C + E_p^r) + \Sigma E_{backbone} + \Sigma E_{hydrogen-bonding} \tag{2.2.1}$$

where E is potential Energy, E_p^C is the attractive energy, E_p^r is the repulsive energy, $E_{backbone}$ is the backbone angle energy and $E_{hydrogen-bonding}$ is the hydrogen bonding energy.

The major term deals with the interactions between pairs of residues which are next to each other on the lattice, and is based on work by Miyazawa and Jernigan (99). They studied the interactions between each of the possible pairs of the twenty amino acids. This was based on a sample of over 1,100 interactions in proteins, and built on their earlier work which had used a smaller sample set (98). By analysing the frequency with which each pair of residues was found in contact, they generated an attractive term to reflect the propensity for two residues to be close to one another (Figure 2.4). The LaMP model does not contain explicit water molecules, but the work by Miyazawa and Jernigen implicitly includes solvent effects. Their 1996 paper states that this includes hydrophobic effects, although subsequent work has questioned the accuracy of the interactions between hydrophobic residues and has attempted to correct for this (161).

	Cys	Met	Phe	Ile	Leu	Val	Trp	Tyr	Ala	Gly	Thr	Ser	Asn	Gln	Asp	Glu	His	Arg	Lys	Pro	
Cys	–5.44	–4.99	–5.80	–5.50	–5.83	–4.96	–4.95	–4.16	–3.57	–3.16	–3.11	–2.86	–2.59	–2.85	–2.41	–2.27	–3.60	–2.57	–1.95	–3.07	Cys
Met		–5.46	–6.56	–6.02	–6.41	–5.32	–5.55	–4.91	–3.94	–3.39	–3.51	–3.03	–2.95	–3.30	–2.57	–2.89	–3.98	–3.12	–2.48	–3.45	Met
Phe			–7.26	–6.84	–7.28	–6.29	–6.16	–5.66	–4.81	–4.13	–4.28	–4.02	–3.75	–4.10	–3.48	–3.56	–4.77	–3.98	–3.36	–4.25	Phe
Ile				–6.54	–7.04	–6.05	–5.78	–5.25	–4.58	–3.78	–4.03	–3.52	–3.24	–3.67	–3.17	–3.27	–4.14	–3.63	–3.01	–3.76	Ile
Leu					–7.37	–6.48	–6.14	–5.67	–4.91	–4.16	–4.34	–3.92	–3.74	–4.04	–3.40	–3.59	–4.54	–4.03	–3.37	–4.20	Leu
Val						–5.52	–5.18	–4.62	–4.04	–3.38	–3.46	–3.05	–2.83	–3.07	–2.48	–2.67	–3.58	–3.07	–2.49	–3.32	Val
Trp							–5.06	–4.66	–3.82	–3.42	–3.22	–2.99	–3.07	–3.11	–2.84	–2.99	–3.98	–3.41	–2.69	–3.73	Trp
Tyr								–4.17	–3.36	–3.01	–3.01	–2.78	–2.76	–2.97	–2.76	–2.79	–3.52	–3.16	–2.60	–3.19	Tyr
Ala									–2.72	–2.31	–2.32	–2.01	–1.84	–1.89	–1.70	–1.51	–2.41	–1.83	–1.31	–2.03	Ala
Gly										–2.24	–2.08	–1.82	–1.74	–1.66	–1.59	–1.22	–2.15	–1.72	–1.15	–1.87	Gly
Thr											–2.12	–1.96	–1.88	–1.90	–1.80	–1.74	–2.42	–1.90	–1.31	–1.90	Thr
Ser												–1.67	–1.58	–1.49	–1.63	–1.48	–2.11	–1.62	–1.05	–1.57	Ser
Asn													–1.68	–1.71	–1.68	–1.51	–2.08	–1.64	–1.21	–1.53	Asn
Gln														–1.54	–1.46	–1.42	–1.98	–1.80	–1.29	–1.73	Gln
Asp															–1.21	–1.02	–2.32	–2.29	–1.68	–1.33	Asp
Glu																–0.91	–2.15	–2.27	–1.80	–1.26	Glu
His																	–3.05	–2.16	–1.35	–2.25	His
Arg																		–1.55	–0.59	–1.70	Arg
Lys																			–0.12	–0.97	Lys
Pro																				–1.75	Pro

Figure 2.4: Amino acid contact energies derived by Miyazawa and Jernigan (99).

There are no distance–dependent relationships on the lattice, only a binary distinction of 'in contact' or 'not in contact' depending on whether two residues are on adjacent sites. For this reason, all distance–dependent terms in the original Miyazawa–Jernigan potential have been discarded.

The work by Miyazawa and Jernigan also included another term. A purely attractive term, as described above, would result in tight, compact structures with minimal distance between residues. To prevent such conformations, and allow the formation of more extended features, such as α–helix and β–sheet, a repulsive term is included. This stipulates that any residue with a number of contacts above a certain threshold (specified per amino acid) should have a positive, or unfavourable, energy associated with it. This increases the overall energy of the conformation, and thus forces the model to move away from dense, closely–packed structures.

The nature of a lattice model means that only certain angles can be made between the residues forming the Cα backbone of the protein. In the cubic lattice, all angles are at 90°, while on the face-centred cubic, 60°, 90°, 120° and 180° are all available. The backbone angles of naturally–occurring proteins tend to, for example, avoid tight turns. Some backbone angles are, therefore, more prevalent than others. Version 1.0 of the model included weightings added to a conformation's energy, based on the frequency of each of the available angles. On a cubic lattice this concept is redundant as all bonds are at 90°, but default weightings for the FCC lattice were included (Table 2.1).

Backbone Angle	Energy
60°	1.032
90°	0.353
120°	0.146
180°	0.000

Table 2.1: Default backbone energy values for the FCC lattice.

2.2.2 Hydrogen Bonding

While hydrogen bonding plays an important part in many areas of protein biology, in this instance we have attempted only to simulate that which is involved in the formation of secondary structure. This, for the reasons outlined above, is the most important aspect of hydrogen bonding in the lattice model, and therefore the first to be addressed.

Hydrogen bonds within lattice structures are treated as a stabilising influence on a conformation. Hydrogen bonds in each type of secondary structure identified within the model (parallel and anti–parallel β–sheet and α–helix) have a negative energy associated with them. The united atom nature of the model means that bonds are created between residues rather than between specific functional groups. The restrictions of the lattice also ensure that hydrogen bonds must only be created between adjacent residues. This leads to some disparity between the length of actual hydrogen bonds and those created by the model.

β–sheet is a form of secondary structure commonly found in proteins. Two areas of protein backbone are held together by hydrogen bonding between the NH groups of one peptide group, and the C=O groups of the other peptide group. Multiple bonds, between two areas of the same chain or between different molecules, hold the strands

together. Multiple strands may thus align to form a sheet, which exhibits a twist over its gross structure. Two strands lying alongside one another may run in the same direction, forming a parallel β sheet, or in opposing directions, forming an anti–parallel sheet. The latter is marginally more stable than the former due to the more linear nature, and shorter length, of the hydrogen bonds (Figure 2.5). The anti-parallel configuration is also entropically more favourable.

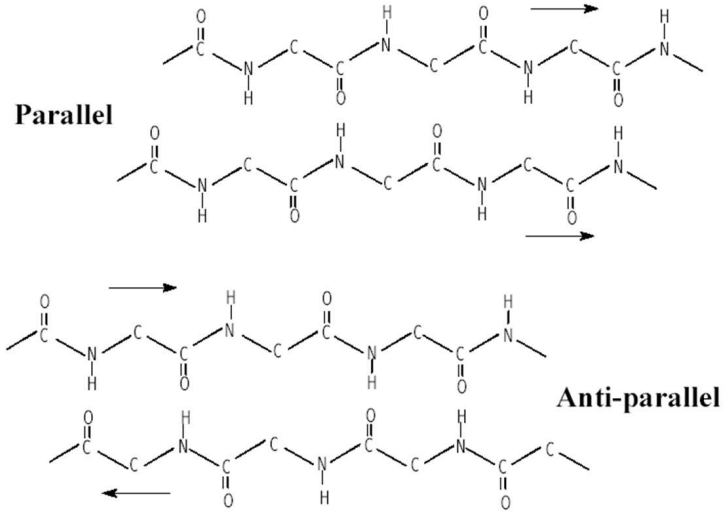

Figure 2.5: Hydrogen bonding in parallel (top) and anti–parallel (bottom) β–sheet.

The ability to model these interactions is of particular importance due to their inherent stabilising property within proteins, their ability to promote the formation of tertiary structures via intra–molecular bonding, and the prevalence of β–sheet in diseases such as amyloidoses.

The LaMP model calculates the number of parallel and anti–parallel sheet hydrogen bonds present in a structure, and then adds a proportional negative energy to the overall energy of the conformation. The simplest example of a hydrogen bond in a sheet conformation involves two residues forming a bond. While not technically a sheet at this stage,

this provides a starting point from which a sheet may later extend. Two residues lying next to each other, *i.e.* in contact on the lattice, is not a sufficient criterion; we must also ensure that the two strands are lying parallel to one another. From Figure 2.6, it can be seen that for an anti–parallel sheet, this requires not only residues *i* and *j* to be in contact, but also for *i+1* and *j-1* to be paired, along with *i-1* and *j+1*.

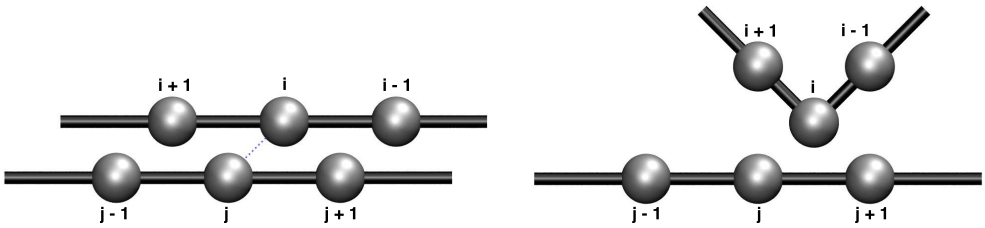

Figure 2.6: A single hydrogen bond between two anti–parallel strands (left). In the right–hand conformation, no bond is formed as residues *i+1* and *j-1* are not in contact, nor are *i-1* and *j+1*.

This prevents bonds forming between two strands which, for example, run orthogonal to each other on the lattice. A similar term exists for the initiation of parallel sheets. This strategy can be extended down the polypeptide chain, with each residue being assessed in turn for proximity to another chain, followed by an assessment of the proximity of neighbouring residues. If the relevant requirements are not met, the 'ladder' of hydrogen bonds is broken.

The chemical nature of amino acids mean that each residue (excluding the R group) does not make more than two hydrogen bonds. This is reflected in the model; once two bonds have been formed, the residue cannot participate in any further hydrogen bonding. The technique outlined above, of working along a chain and adding hydrogen bonds as soon as they are found, gives a bonding pattern that includes 'offset' bonds, whereby residue *i* bonds not only with residue *j*, but also with either *j+1* or *j-1*. This pattern works well with two β–strands, but was less effective for larger sheets consisting of three strands

or more. In these cases, the offset bonds take up the second hydrogen–bond available to each residue, preventing it from making any more. Residues in the second chain are therefore unable to bond with those in the third, as they each already have the maximum of two bonds. For multiple chains, this resulted in a network of 'fingers' (Figure 2.7), bound only by residues at the ends of chains, which did not suffer the 'offset bonds' problem.

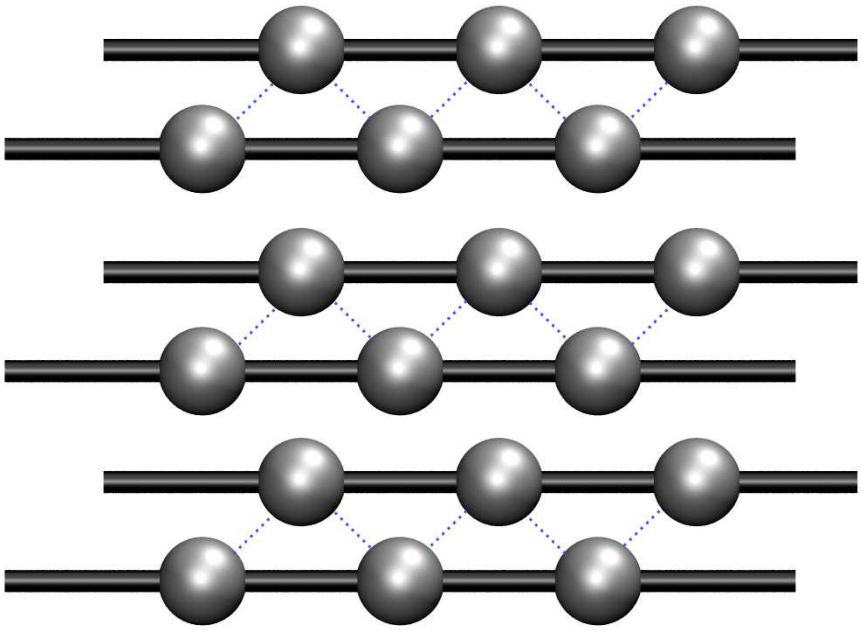

Figure 2.7: Incorrect pattern of hydrogen bonding with multiple strands. Note that pairs of strands are separate from one another.

The model was altered to remove the offset hydrogen bond from each residue between the first and second chains, and replace it with a direct one between the second and third chains. This resulted in a pattern of bonds which was much more consistent with those found in nature. The definition successfully recreated the bonding pattern along β–sheets composed of multiple chains (Figure 2.8).

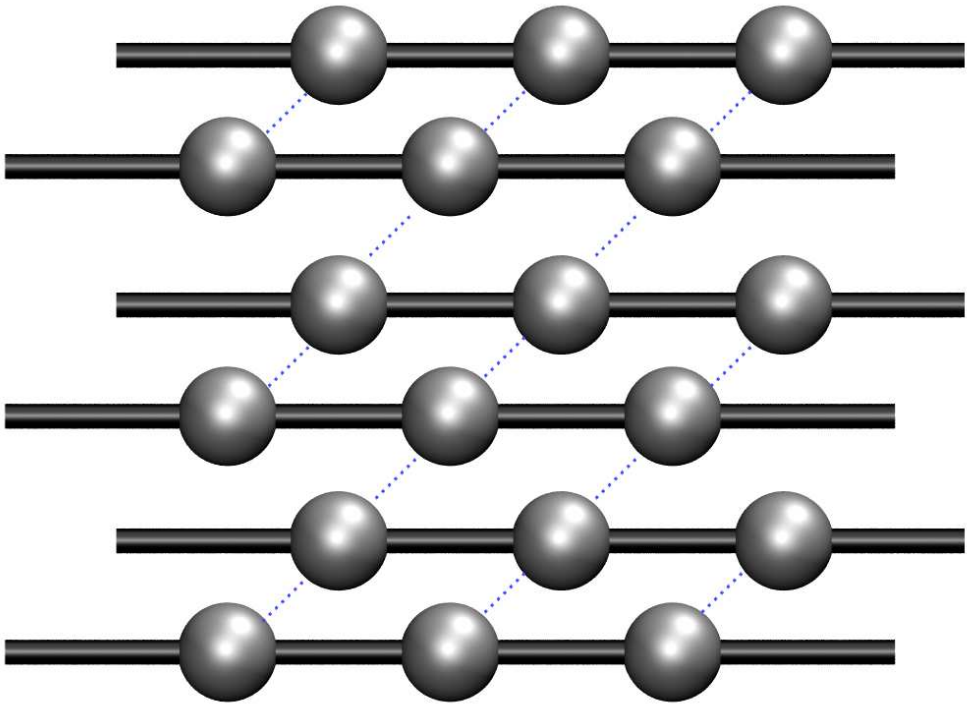

Figure 2.8: Corrected bonding patterns in a β–sheet composed of multiple chains.

As previously, by considering both possible orientations of the chains, this definition of bonding can be used for both parallel and anti–parallel conformations.

To this point, the bonding definition has been considered only in two dimensions, rather than the three available on the lattice. This two–dimensional testing assumes that multiple chains all lie in the same plane; all test structures used prior to this point followed this rule. While bonding between strands in the same plane is a possibility on the lattice, it is far more likely that chains would lie at different angles, and thus the accuracy of the definition must be confirmed on the lattice in three dimensions. The strength of the bonds being created to stabilise the various conformations should also be considered. The stronger the interactions used in stabilising a particular feature, the more likely that feature is to form, and to dominate other structural features, which may also be desirable. For this reason, the calculations carried out on test structures used a variety of

energies. When a highly negative energy is associated with hydrogen bonds, fewer such bonds are required to confer a highly stable (more negative) energy on a conformation. In effect, the magnitude of the hydrogen bonding energy is so great that it overrides any positive energy which may result from other terms such as the backbone or repulsive terms. Given that several hydrogen bonds should exist in a starting structure which already displays secondary structure, even if the model only retains a certain proportion of these, they should reduce the energy of the structure significantly, and "lock" it into its starting conformation or similar. When the model is run with a less negative energy for the hydrogen bonding term, the influence of the term is diminished, and the secondary structure elements are less likely to form or be retained. It is important to ensure that a sufficiently large conformational search space is used to test the model. This gives the greatest possibility of finding more stable search structures which prevent the original conformation becoming "locked" in to place, thus testing the bonding definition most thoroughly.

The main test structure used to assess the effectiveness of the β–sheet hydrogen bonding definition was the I27 domain of Titin (PDB reference 1TIT). This structure was chosen not only because other work within this Thesis uses it, but also because it includes one parallel and one anti-parallel β–sheet. The domain is 89 residues in length. This is larger than any structure applied to version 1.0 of LaMP, the largest of which was an octamer of a 9-mer polypeptide.

The native structure of I27 was provided to the model, and run with hydrogen bonding forces of -10 and -20 units. Control runs were carried out with the bonding term active, but a hydrogen bond energy of zero. In this model, no α–helical hydrogen bonding term was present. All searches ran for 10^7 steps using Replica Exchange Monte Carlo

searches. Five replicas were used with temperatures ranging from T = 1.0 to T = 10.0. The structures resulting from the runs can be seen in Figure 2.9.

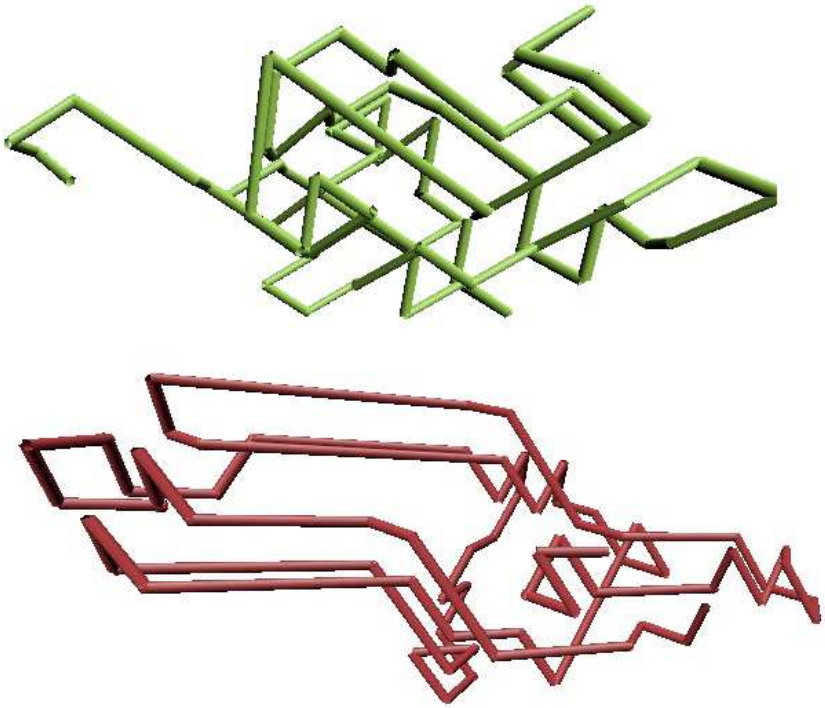

Figure 2.9: Top - Control run with no beta–sheet hydrogen bonding. Bottom - Linear strands in contact when energy = -10.

As expected, the run without any favourable energy term for sheet hydrogen bonds resulted in a structure with no discernible secondary structure. The structure is quite compact, and there are no sections where strands run alongside each other for more than one lattice unit in length. In contrast, the structure obtained with an interaction energy of -10 units per hydrogen bond shows more promise. In a similar fashion to the native state starting structure, the molecule shows a more compact and less structured side at one end, while the other half of the molecule appears to contain anti–parallel beta sheet. The strands are stacked, rather than forming individual sheets of only one strand thickness. The hydrogen bond definition does not restrict this in any fashion, and this could be an interesting target for further refinement of the model.

If the hydrogen bonding force is increased, the structure begins to again resemble the native conformation. This is most likely a result of LaMP identifying certain hydrogen bonds almost immediately the calculation begins. These bonds are given a strong favourable energy, and the structure becomes locked; any move to an alternative structure results in too great a gain in energy. With a weaker energy of -10, this problem is less severe, and the model can explore the conformational space more freely.

Helices are the most common secondary structure element, with α–helix being present in biological structures much more frequently than 3–10 and π–helices. The α–helix consists of a right–handed helix with a width of 5.4 Å. Each residue forms a 100 $^\circ$ section of the helix, giving 3.6 residues per full helix turn. This arrangement permits hydrogen bonding between the carbonyl group of one residue and the amine group of the residue four positions further along the peptide chain (Figure 2.10). This hydrogen bonding is responsible for the stability of the helix, and the reproduction of such an effect is required for any model to represent natural protein structures accurately.

Recreating any secondary structure motif is inherently difficult when restricted to a lattice framework. The primary obstacle with fitting an α–helix is the directionality imposed by the lattice. The axis of a natural helix can run in any direction in three–dimensional space. Even if a good approximation of a helix can be established in one orientation on a particular type of lattice, subsequent rotation of the helix axis can cause the representation to break down if there is no rotational lattice symmetry. In the case of the FCC lattice, the default in LaMP, there is six–fold rotational symmetry, allowing helices to take on any one of six directions, *i.e.* three different orientations.

The method utilised in LaMP involves identifying when a structure resembles a helix, and giving that structure a favourable (*i.e.* more negative) energy. The defining characteristic of an α–helix is the hydrogen bonding between residues i and i+4. This bonding

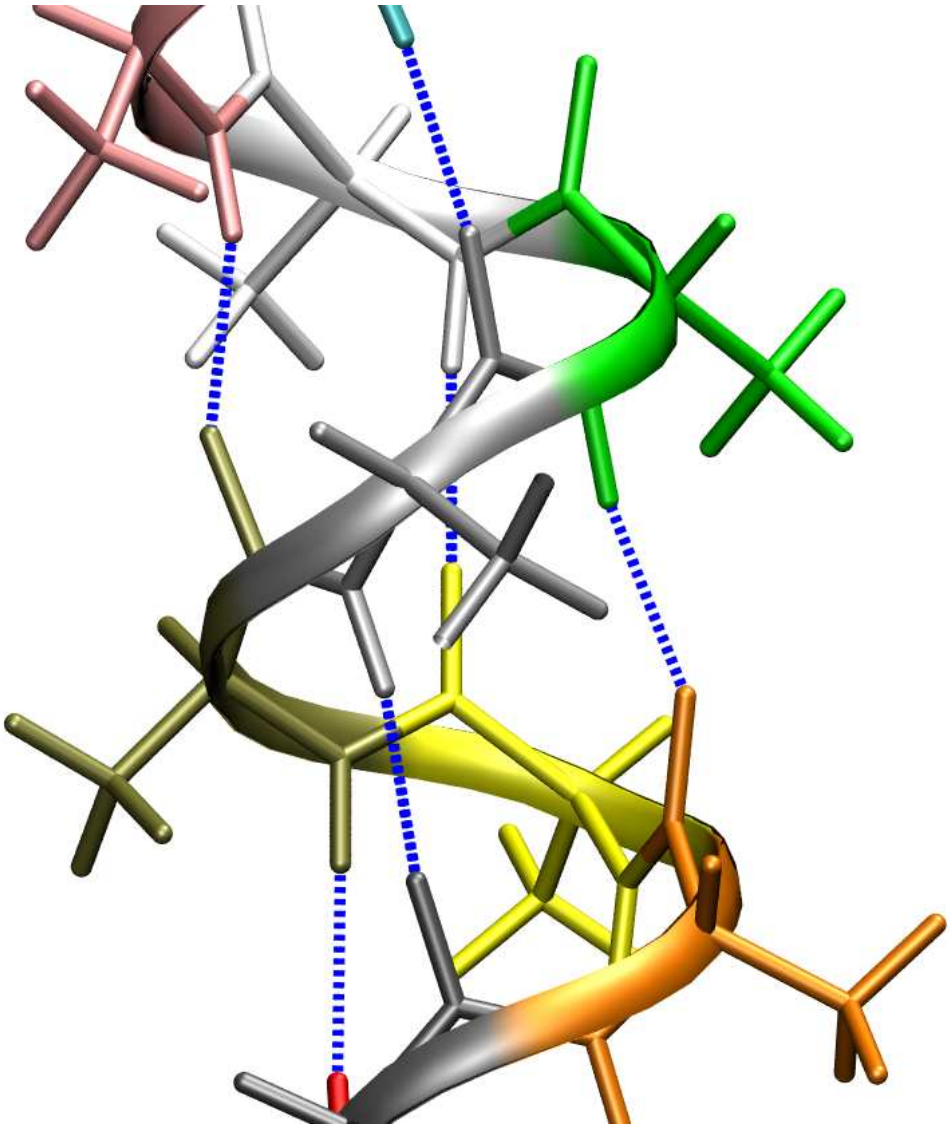

Figure 2.10: *i:i+4* hydrogen bonding in the alpha–helix.

pattern repeats down the length of a helix, and is distinct from the type of bonding found in β–sheet. As this is an inherent part of the helix definition, the same definition was initially used to identify a helix on the lattice. The model stipulated that residues i and $i+4$ should be in contact for a hydrogen bond to result. However, results from trial runs using this definition gave uncoordinated structures with no definition (Figure 2.11). The starting structure was a geometrically perfect right–handed helix, created in CHARMM (58) using phi and psi angles of 58° and 47° respectively. This was converted to a lat-

tice structure with the LaMP *PDB2Lattice* package on the FCC lattice. 10^9 steps of plain Monte Carlo search were carried out, at $T = 1.5$.

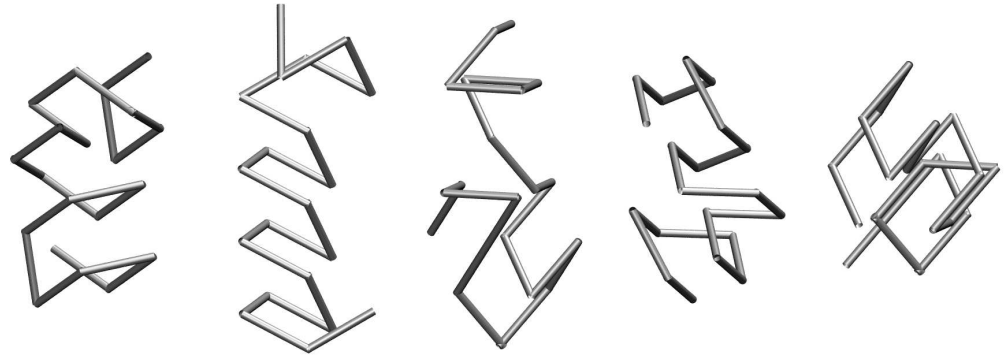

Figure 2.11: Example structures resulting from *i:i+4* contact hydrogen bond definition.

Inspection of these structures showed that the model was correctly identifying and assigning a negative energy to *i:i+4* pairings. However, this definition imposed no restrictions on residues *i+1*, *i+2* and *i+3*. These residues were free to move to any legal lattice location, and thus generate structures which did not necessarily represent helix. The discrete nature of the lattice means that a geometrically perfect helix will never fit directly onto the model (Figure 2.12).

Therefore, to obtain the best possible representation of a helix using the *i:i+4* definition, it is necessary to exert some influence over the locations of the intermediate residues. An analysis of the lattice moves made in a number of naturally-occurring helices was carried out, to determine which ones occurred most frequently in helix configurations. A dataset consisting of 226 individual helices was obtained from a set of 177 PDB structures. These structures were chosen on the basis that the only secondary structure elements they contained were α–helices. The relative frequencies of the 12 available lattice moves on the FCC lattice are shown in Table 2.2.

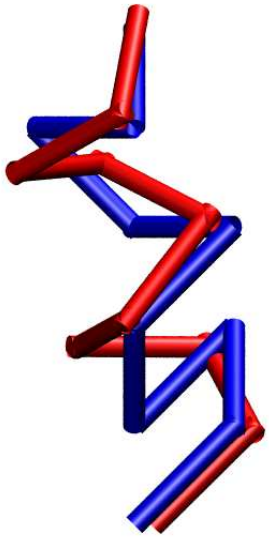

Figure 2.12: A geometrically-perfect α-helix (red), and the best–fit FCC lattice equivalent (blue).

This work showed that there is no major distinction between the relative frequencies of the FCC lattice moves. This is most likely due to the differing orientations of the helices relative to the lattice. While a number of helices oriented in the same direction could be expected to show a more distinctive pattern of moves, this would require the model to restrict the direction of helix formation, and would also be dependent on the type of lattice. Such a definition is not conducive to the construction of an accurate model.

In light of this discovery, it became apparent that there was no way of accurately restricting the intermediate *i+1*, *i+2* and *i+3* residues to locations which would facilitate the formation of helix between *i* and *i+4*. A number of alternative metrics were therefore investigated to determine whether any could be used to identify the presence of a helix. The same dataset of 226 individual helices was used, along with another set of structures representing β–sheet. These were obtained from a search in the PDB for structures containing 60 - 100% β–sheet, with no greater than 90% homology between selected structures.

Move	Relative Frequency
0 -1 1	0.0893
0 -1 -1	0.0876
0 1 1	0.0828
0 1 -1	0.0822
1 0 -1	0.0896
1 0 1	0.0815
1 1 0	0.0799
1 -1 0	0.0842
-1 1 0	0.0787
-1 -1 0	0.0828
-1 0 -1	0.0821
-1 0 1	0.0794

Table 2.2: Relative frequencies of the twelve possible moves on an FCC lattice, in a set of 226 helical fragments.

As a common feature along the length of the helix, the distance between each pair of residues should remain constant. In contrast, the distance between any such pair in a β–sheet cannot be accurately predicted without knowledge of the length of the sheet. On this basis, it should be possible to observe a difference in the *i:i+4* distances in the two different secondary structure type, and there are two reasonably distinct groups (Figure 2.13).

The more compact helix structures exhibit shorter separation lengths, with over half being around 2.5 lattice units, and more than 75% of helix *i:i+4* interactions happening over a distance of 2.5 lattice units or less. In contrast, most β–sheet *i:i+4* distances are 3.5 lat-

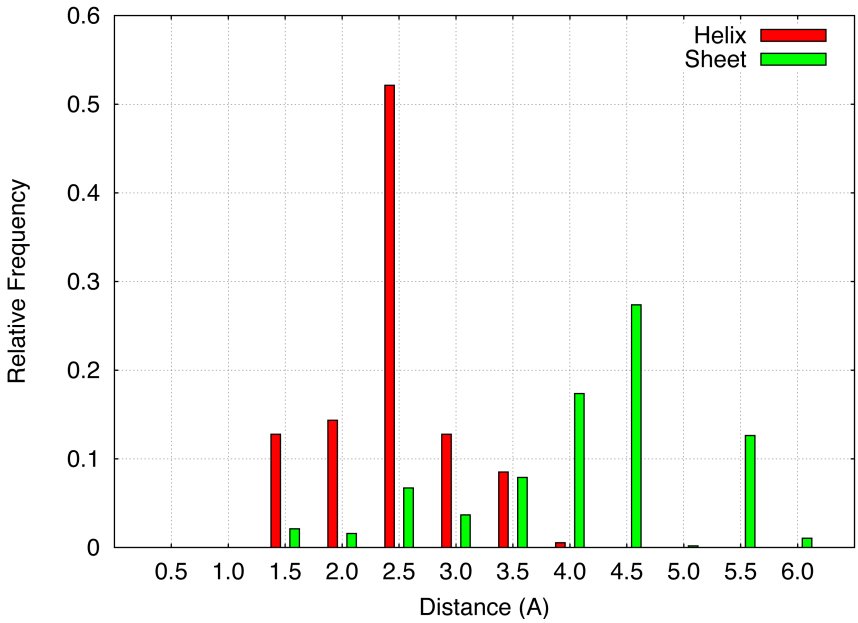

Figure 2.13: *i:i+4* distance in α–helices (red) and β–sheets (green).

tice units or more. There is, however, some overlap between the two sets of data, and this parameter is not sufficient to be used in isolation as a definition of helix bonding. Backbone angles on the lattice are already employed to ensure structures do not tend towards biologically unfavourable structures, for example those involving high frequencies of 60° bends. Analysis of all backbone angles along the entire length of each protein fragment in the test data sets showed that an angle of 180° is unlikely in helices (Table 2.3). These tend to take on tighter 60° turns to form the helix. Any tendency to take on 60° turns for this reason should be made energetically more favourable than the penalty imposed by the backbone term. Again, while a general distinction can be seen between the two data sets, there is insufficient evidence to support the creation of a hydrogen bonding helix rule on these data alone.

Between residues *i* and *i+4* there are three backbone angles. In a helix it is necessary to continue a tight spiral formation, while a sheet is more likely to run in a linear fashion.

Backbone Angle	Relative Frequency	
	α–helix	β–sheet
60°	0.609	0.058
90°	0.183	0.135
120°	0.183	0.640
180°	0.000	0.167

Table 2.3: Relative backbone angle frequencies in α–helix and β–sheet.

On this basis, taking the mean of these angles for each secondary structure type should give a more acute angle for helices than for sheets. While this is generally true (Figure 2.14), there is again an overlap which prevents the parameter from defining an accurate distinction between the two structure types.

A similar result was obtained when taking vector angles between pairs of residues. The first vector was always assigned to *i:i+1*, and angles were measured between this and a number of other vectors (Table 2.4).

Vector One	Vector Two
$\overrightarrow{i : i+1}$	$\overrightarrow{i+1 : i+2}$
$\overrightarrow{i : i+1}$	$\overrightarrow{i+2 : i+3}$
$\overrightarrow{i : i+1}$	$\overrightarrow{i+3 : i+4}$

Table 2.4: Backbone vector angle definitions.

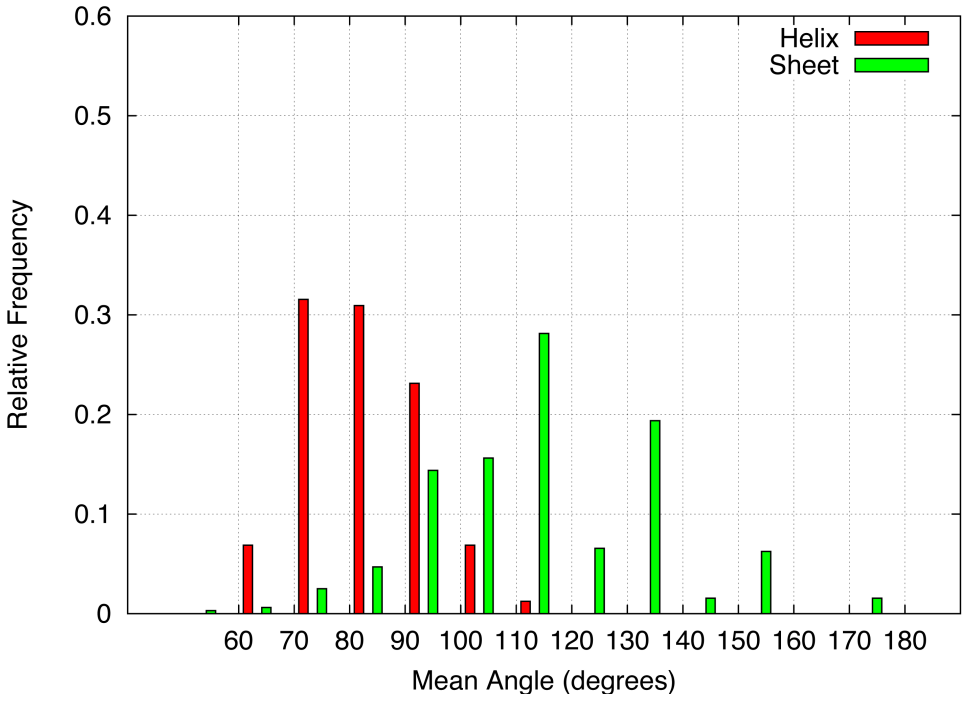

Figure 2.14: Mean backbone angle between residues *i* and *i+4*.

Taking the angle between the vectors $\overrightarrow{i : i+1}$ and $\overrightarrow{i+2 : i+3}$ gives the backbone torsion angle. Due to the constraints of the lattice this angle, like the backbone angle, can only take on certain discrete values. There is again an overlap between the α–helix and β–sheet distributions, although nearly 80% of helices have torsion angles of 120° or 180° (Table 2.5).

None of these parameters gave a definitive separation between helical and sheet structures, so a combination of factors was considered as a means of identifying relevant motifs. The LaMP model was amended to apply a helix hydrogen bond between residues *i* and *i+4* if the distance between the two residues was less than or equal to 2.5 lattice units and if the backbone torsion angle was greater than or equal to 120°. When applied to a set of test molecules, the definition gave encouraging results. The number of hydrogen bonds discovered by the model in each structure was compared to the number

Angle	Relative Frequency	
	α–helix	β–sheet
0°	0.001	0.261
60°	0.058	0.445
90°	0.136	0.134
120°	0.645	0.134
180°	0.136	0.026

Table 2.5: Relative frequencies of backbone torsion angle.

expected. In structures composed entirely of parallel or anti–parallel β–sheet, the model correctly identified that no *i:i+4* hydrogen bonds were present. A set of 72 helical structures was then tested, in which the model correctly identified the *i:i+4* bonds in all but one structure. The exception to this success was a helix in which several tight 60° turns were present.

Unfortunately, this accuracy did not transfer to biological molecules. When a short (nine residue) α–helix was presented to the model, and 10^7 steps of Replica Exchange Monte Carlo simulation carried out, the model was unable to hold the structure in place using predominantly hydrogen bonds (Figure 2.15).

Longer helices also produced similar results after 10^9 steps; simulations on the poly-alanine 20-mer used earlier in this work produced structures which were much more compact than the extended starting helix (Figure 2.16). The structures generated by LaMP had a mean radius of gyration of 4.497 Å, over 15 repeats, versus 8.879 Å for the starting structure. While a regular repetitive pattern can be seen in some of the structures, this takes the form of a left–handed helix, and formed in only two of the fifteen

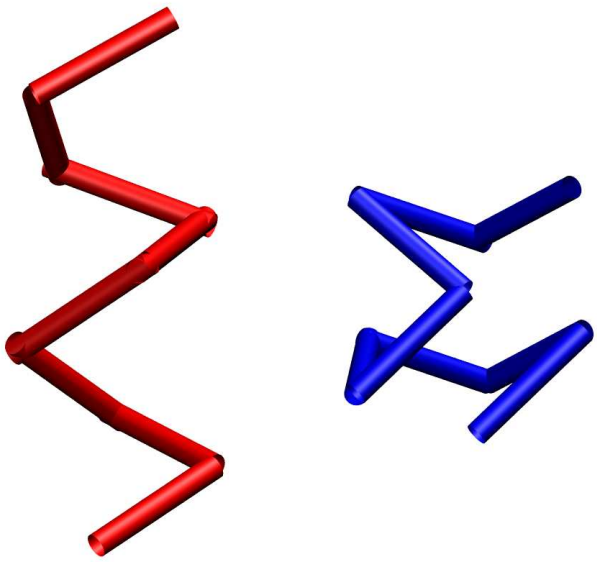

Figure 2.15: A sample test helix (helix extracted from PDB code 1HTA) and the resulting structure after 10^7 steps with the final helix definition.

repeats. The modal number of hydrogen bonds formed was 15, with a minimum of 13 and a maximum of 16. This represents, in all cases, an increase from the nine seen in the starting conformation. This is likely to be due to the more compact nature of the structures, allowing more residues to be in contact and form hydrogen bonds. However, the presence of these bonds appears to have prevented the structures from becoming as globular as in previous versions of the model, and forced them to take on a more elongated structure. While a useful first step towards the formation of helix, it is clear that the definition requires further development before α–helices will be reliably reproduced.

Future Considerations

In addition to developing the α–helix term, future development of the model will require work to combine the two hydrogen–bonding terms. Work so far has been carried out with only one term activated at any one time. Calculations involving both terms at

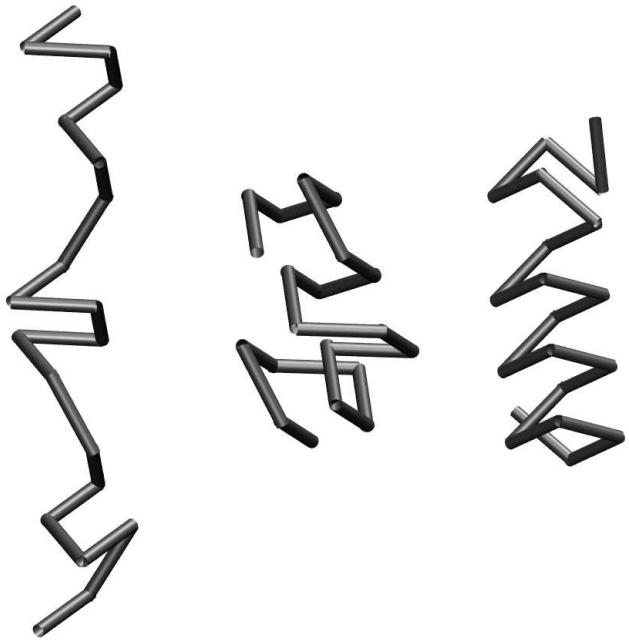

Figure 2.16: Two structures (left, middle) resulting from 10^7 calculation steps. The starting structure was the poly-Ala helix 20-mer fitted to the lattice (right).

once are affected by the restriction on the number of hydrogen bonds each residue can form. If, for example, α–helix bonds are assessed first and result in the formation of two hydrogen bonds on a residue, then no β–sheet bonds can then be made on that residue. Thus, the prevalence of a particular type of secondary structure can be influenced by the order in which the different bond types are assessed. Another important factor is the refinement of the individual hydrogen bonding term energies. The helix testing described in this Thesis was carried out with a highly negative energy, to ensure that any desired features which were identified remained in place; a more suitable value would need to be determined through further testing of the model. While some models have a fixed value for the strength of hydrogen bonds (162), it is not unusual for it to be presented as a parameter to be adjusted. For example, Kirov *et al.* describe the ability to alter the strength of hydrogen bonding (163) in an off–latice coarse–grain model. The more complex model developed by Fujitsuka *et al.* (164) features more advanced modification of

the bonding strength (based on buriedness of the bond, for example), but still contains a term for the relative strength of the bonds in relation to other forces. In LaMP, while the two energies can be assigned independently in a custom specification file, the relative strengths may need to be adjusted to provide a suitable mix of the two secondary structure types. If necessary, this may include adaptation of the term on a residue–by–residue basis to take account of the different propensities of amino acids to form each of the two structures.

2.3 Forced Protein Unfolding

The model, with a functional β–sheet energy term, was applied to a problem involving a protein containing a high proportion of sheet. The mechanical stability of the 27^{th} immunoglobulin domain of the protein Titin has been studied extensively, and forms the focus of the second chapter of this Thesis. Previous research (126, 165, 166) has indicated that when the termini of the I27 domain are pulled in opposite directions, the important unfolding events occur within the first 15 Å of extension. The LaMP model has been used to elucidate stable structures at 1 Å intervals along this trajectory, allowing comparison between these and other conformations previously proposed as being important in the stability of the domain.

2.3.1 Methods

A set of fifteen structures generated by Toofanny *et al.* (167) was used to represent regular 1 Å intervals across the first 15 Å of forced extension. These were converted from all–atom to lattice structures using the LaMP *PDB2Lattice* package, with a mean RMSD of 0.873 lattice units. The N– and C–terminal residues were held in place using the new

RESTRAIN function. The search input file contains a list of residues to be restrained, which are then prevented from moving through the simulation. This allowed the length of each starting structure to remain constant throughout the calculations. Each conformation was then run for 10^9 steps of Replica Exchange simulation, with ten replicas and temperatures between 1.000 and 5.000. The default FCC lattice was used, along with β–sheet energy of -10 for hydrogen bonds in both parallel and anti–parallel conformations. This value is the same as that used in the successful trials which demonstrated the formation of sheet previously. Similarly, the α–helix term was deactivated; as I27 contains β–sheet exclusively, this should not have adversely affected the results. All runs took place on Intel processor–based systems through a Condor compute pool.

2.3.2 Results

The unfolding of Titin I27 under force, which is discussed in depth in the next Chapter, is thought to consist of two main unfolding events. The A and B strands separate, after which the A′ and G strands are pulled apart. This latter step removes all structural stability from the protein, which causes it to rapidly unfold to its contour length. The lattice structures obtained from the LaMP calculations are distributed over a series of N–to–C terminus distances, and may therefore be considered to offer a coarse trajectory representing the unfolding process (Figure 2.17).

Unlike earlier iterations of the model, the software produced structures which were not compact in nature. Although the N– and C–terminal residues were restrained, the main bulk of each structure remained stable throughout the calculations. Each conformation also showed good retention of β–sheet structure. Explicit solvent molecular dynamics studies have suggested that the loss of the majority of sheet structure only occurs after

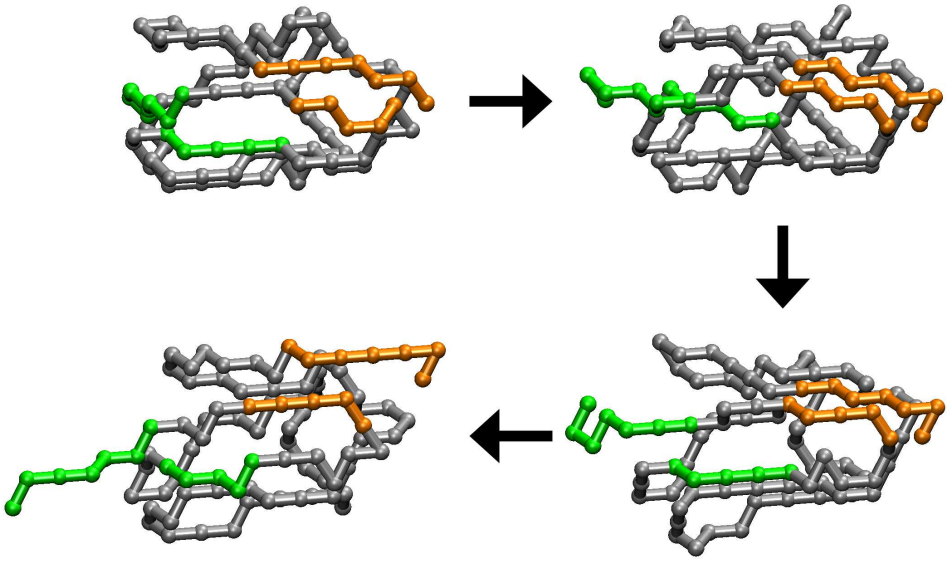

Figure 2.17: Unfolding of I27 on an FCC lattice, via milestoning structures. A and B strands are coloured green, A′ and G strands are orange.

the A′ and G strands have separated, and as the protein unfolds rapidly thereafter. It can be seen that the lattice structures follow a similar unfolding pattern, with the A and B strands becoming separated first, followed by the A′ and G strands (Figure 2.17). While the extended nature of the structures is in part due to the restraints placed on the termini, there is a marked improvement over the performance of previous versions of the model which did not have the β–sheet term (Figure 2.18).

These results do not provide any new mechanistic insights into the forced unfolding of the I27 domain. However, they act as a useful real–world validation of the model's β–sheet definition, especially when compared with the previous version. Both the continued presence of secondary structure elements, and the similar sequence of unfolding events, further suggest that the bonding definition is capable of reproducing or maintaining sheet secondary structure in an accurate manner.

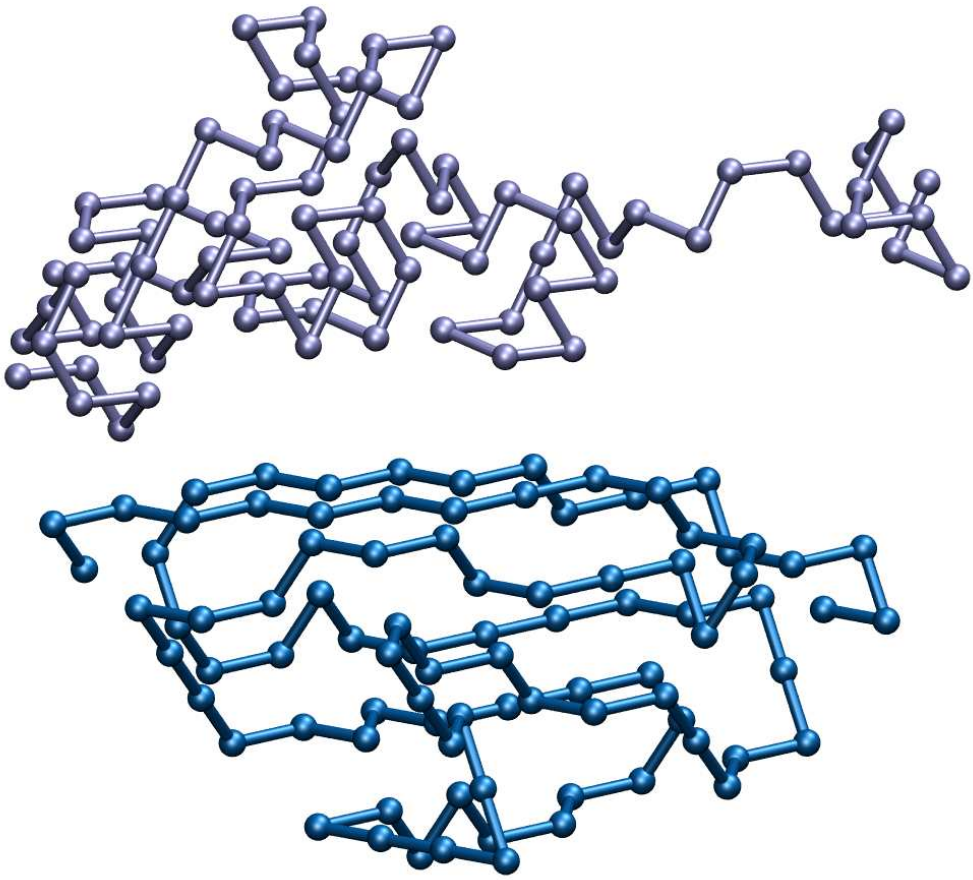

Figure 2.18: Example structures from lattice unfolding of I27 on an FCC lattice, in the absence (top) and presence (bottom) of hydrogen bonding energies.

2.3.3 Conclusions

This work has demonstrated the development and implementation of hydrogen bonding terms in a lattice model of proteins. The system has been updated to take these bonds into account when determining the lowest energy structure for a molecule. The β–sheet term has been shown to align strands in both parallel and anti–parallel orientations, and to maintain the structure of molecules already containing sheet secondary structure. In the absence of this term, the model produced compact, globular structures. The development of an α–helix term proved more complex, and relies on a number of parameters to establish the presence of helices. While the combination of these factors has resulted in

less globular structures, the relative complexity of the helix, combined with the difficulty of fitting it to the coarse–grain lattice, have meant that this definition has been less successful. Nevertheless, the introduction of these hydrogen bonding terms and numerous other improvements to the model have produced a significant improvement in the biological relevance of the low–energy structures produced. The model has been proven to elucidate these conformations in an efficient manner, and therefore continues to display potential for further development in the future.

Chapter 3

Forced Protein Unfolding

Protein folding work has long been complemented by studies looking at the forced unfolding of proteins. Historically, proteins with high mechanical stability have proved good targets for computational studies. As the computing power available to researchers increases, these simulations are able to probe ever more delicate processes. Nevertheless, it is important to recognise that there are alternative strategies to just increasing the computational resource available to simulations. The use of techniques to lever existing facilities, as well as to increase the chances of the event in question occurring, are also important. These factors combined should allow researchers to access more biologically relevant timescales, probe processes in finer detail than before, or to alleviate the use of extreme simulation parameters to promote events of interest.

3.1 Titin I27

The giant filamentous protein Titin, also known as connectin, is found in vertebrate cardiac and striated muscle, and is the third most plentiful protein within such tissue (after

myosin and actin). Despite its presence in large quantities, Titin was not discovered until relatively recently due to its size (168). Conventional electrophoresis gels of striated muscle showed a large undefined band at the top, which for years was not investigated. However, unconventional 2% gels showed that the band in fact represented a massive protein. This protein was subsequently named Titin, and it has since been the subject of much research due to its varied roles. Titin has a molecular weight of over 3 MDa, making it the largest protein in the human body (169). Each filament is approximately 1 μ m in length with a diameter of 4 nm (169), and consists of approximately 30,000 amino acids. The carboxy terminus of the protein is bound to myosin filament in the M-line, and the molecules reaches from here to the Z-disk, thus spanning half a sarcomere across both the I-band and A-band regions (Figure 3.1).

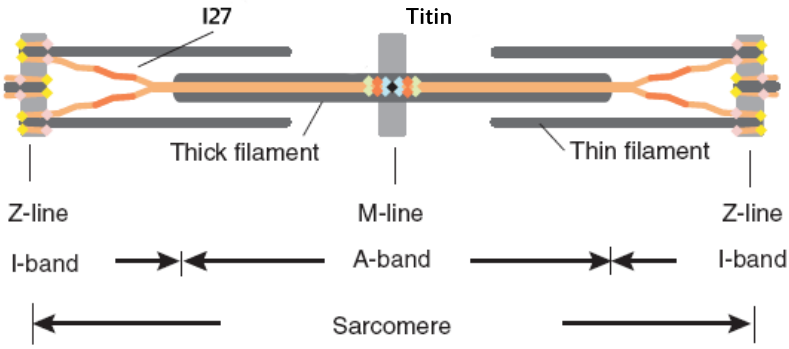

Figure 3.1: Titin's place in the sarcomere. Adapted from (169).

The protein consists of tandem arrays of immunoglobulin and fibronectin domains, interspersed with PEVK motifs, unique sequence insertions and other less structured sequences. The immunoglobulin domains feature primarily in the I-band region, along with the PEVK repeats (areas of sequence where the protein is composed of up to 70% Proline, Glutamic acid, Valine and Lysine). The arrangement of these domains and their modular structure is determined by differential gene expression in different types of

muscle tissue. The carboxy terminus has a serine/threonine kinase domain close to it, thought to be responsible for the reduction of transcription of serum response transcription factor (SRF) as a result of mechanical stress (170). A binding point for the muscle-specific calpain protease p94 has also been found within the molecule (169, 171). p94 has the ability to cleave Titin at two points, although one of these can be absent as a result of differential splicing. While the purpose of these cleavage sites is still unclear, suggested possibilities include a role in sequestering p94, or a mechanism for the control of Titin stability (171).

Titin has several roles within the body, including acting as a biomechanical sensor and playing a role in signalling mechanisms controlling expression of muscle-related genes (169, 171, 172). These functions also mean that mutations within Titin can result in a disease state. Indeed, Titin has been implicated in hereditary myopathy with early respiratory failure (HREF) through mutations to the kinase domain, and also in increased muscle stiffness when muscles are over-worked (172). In addition, autoantibodies specific for Titin have been found in sufferers of the condition myasthenia gravis. While the primary auto–immune response in these patients is directed towards the acetylcholine receptor, the secondary response focuses on Titin. All patients have one binding site, but approximately 40% also show a second site further along the molecule, raising the possibility of reactivity moving along the length of the Titin molecule as the disease progresses (173).

Titin's primary role, however, involves muscle contraction and passive elasticity (174). The molecule acts as a form of stabiliser, to help ensure that the sarcomere maintains a regular arrangement during muscle expansion and contraction (175). The passive elasticity operates by two different mechanisms. Under normal (low) forces, an entire Titin

filament can act as an entropic spring (176), including extension of the PEVK domains. However, when force is increased, elasticity is achieved through reversible unfolding of immunoglobulin-like (Ig) domains along the structure (168).

Such unfolding was first studied in entire Titin molecules, with their various component domains. Rief *et al.* (119) carried out forced unfolding using the atomic force microscope; when they plotted the extension of the molecule versus the force applied, they saw the now well-known saw-tooth pattern (Figure 3.2). Each peak or tooth corresponds to a single domain unfolding. The domains can therefore be seen to unfold sequentially on an individual basis, rather than concurrently.

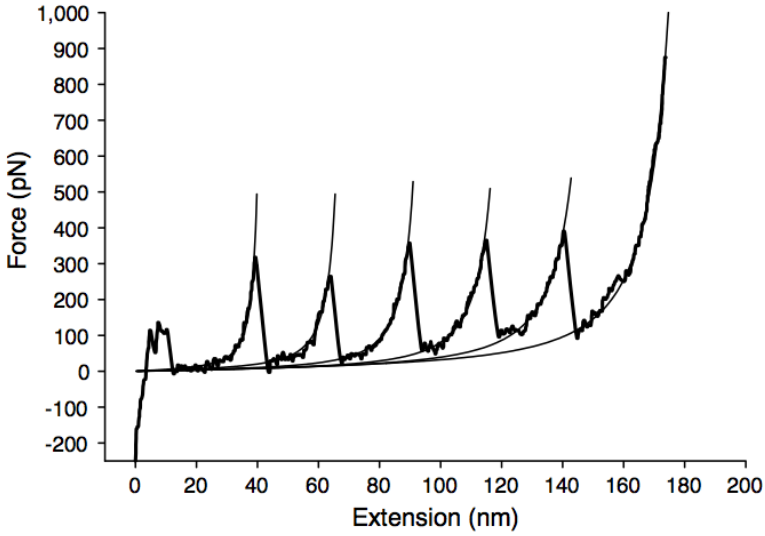

Figure 3.2: The distinctive 'sawtooth' extension plot of I27 under force (166).

The variation in the domains present in Titin makes them potentially difficult to study. It has become common, therefore, to synthesise proteins consisting of multiple repeats of a single domain (166). This ensures that only the effects of the domain being studied are measured.

One domain which has been studied in such a way is the 27th immunoglobulin domain of Titin. Known as I27, it is situated in the I-band. At 89 residues, it is relatively small and therefore provides a good model for molecular dynamics simulations as calculations are not too complex. As predicted by Politou *et al.* in their 1994 paper, the sections of Titin containing Ig domains consist mostly of β-sheet secondary structure (177). The I27 domain is composed of seven β-strands, arranged in two anti-parallel β-sheets (Figure 3.3).

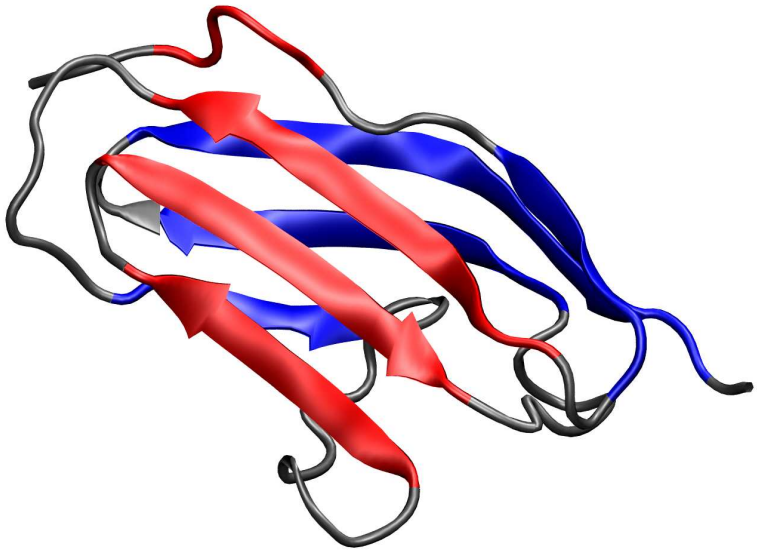

Figure 3.3: The I27 domain of Titin, showing the two β-sheets (front and back).

Lu *et al.* carried out some of the first work looking at the forced unfolding of I27 in 1998 (165), utilising the newly-developed technique of steered molecular dynamics (SMD). SMD involves the application of force to molecules within a simulation, in order to influence the behaviour of the system. The technique was first developed in the mid-1990s, and provided a convenient way to model I27 unfolding on a shorter timescale than would otherwise be possible. Their SMD simulations fixed the Cα of the N-terminal leucine residue, and applied a force to the Cα of the Glu88 residue. This force pulled the

atom at a constant velocity (0.5 Å/ps and 1.0 Å/ps) to extend the molecule. The work was amongst the first to show single unfolding peaks on a force versus extension curve in computer simulations (165).

Subsequent AFM work by Marszalek *et al.* showed the presence of an unfolding intermediate, approximately 7 Å along the unfolding coordinate. The group also carried out SMD simulations, the results of which showed good agreement with the AFM data (166). The unfolding to the intermediate state seen by Marszalek *et al.* in their subsequent AFM and simulation work was reversible, leading them to hypothesise that such activity may be important in conferring passive elasticity on Titin.

More SMD studies by Lu and Schulten led to the potential discovery of another intermediate at around 14 Å along the unfolding coordinate (126). Using the same software and similar methods, they extended their protocol to include both constant velocity and constant force simulations over 1 ns in explicit solvent. The 14 Å intermediate observed seemed to provide a more significant barrier to unfolding than the first 7 Å intermediate structure.

Both of the unfolding events are thought to occur as a result of the breaking of hydrogen bonds between different β-sheet strands of the I27 domain. The first intermediate is observed when H-bonds between the A and B strands are broken. This leads to the two strands becoming detached from one another. Continued application of force leads to the breaking of H-bonds between the A' and G strands, and subsequently the unfolding of the molecule. This second event is referred to by Lu and Schulten as the 'key event' in I27 unfolding, as it is the main barrier preventing the entire protein from unfolding. In their 2000 paper, the authors suggest that while the H-bonds between the A and B strands break relatively easily, those holding the A' and G strands together are only likely to be broken under the influence of water molecules in the solvent surrounding the protein.

The solvent molecules effectively out–compete for the H–bonds; the interaction of these molecules with backbone atoms weakens the residue–to–residue bonds, and the force applied by the AFM is sufficient to then break them (126).

In 2002, Fowler *et al.* produced I27 mutants in which the A strand had been altered to destabilise it, or removed entirely (178). A combination of AFM, NMR and molecular dynamics simulations indicated that neither modification had a major effect on the forces required to unfold the molecule. This would indicate that the A strand exhibits very little resistance to mechanical force. The force required to promote unfolding is therefore likely being used to unfold the intermediate seen by Marszalek *et al.*, rather than to unfold the native state conformation.

3.2 Milestoning

3.2.1 Milestoning Theory

The principle of milestoning was first introduced by Faradjian and Elber in 2004 (179). The technique represents a method by which molecular dynamics simulation timescales can be greatly increased, and is therefore useful for the investigation of long-timescale events such as protein folding and unfolding. Rather than running a single, lengthy simulation, milestoning relies on the simulation trajectory being split up into a number of discrete parts (Figure 3.4). These can be run as separate processes, and the results compiled into a single ensemble for analysis.

A key element of milestoning simulations is the determination of suitable milestones, that is, the hyperplanes which split the overall reaction coordinate into several smaller sections. A natural parameter choice when studying force–induced protein unfolding

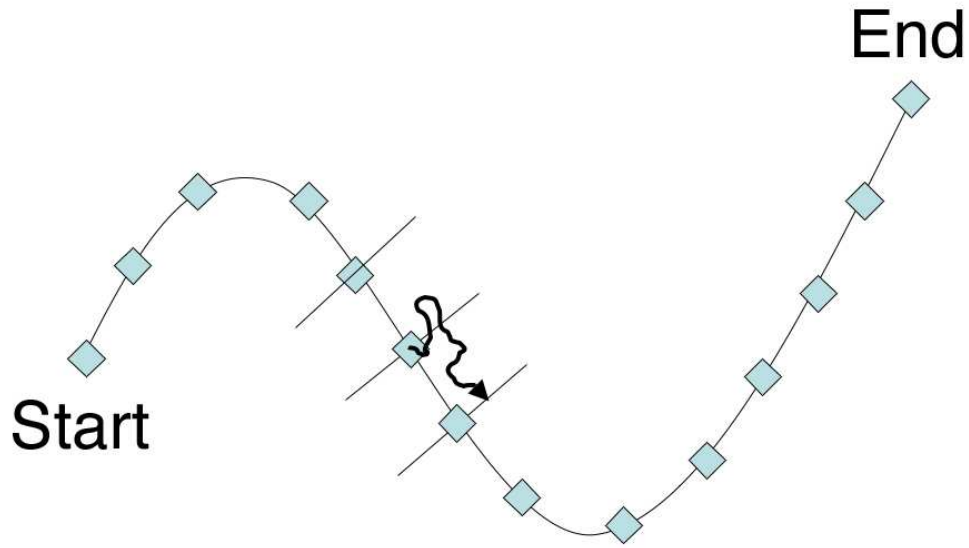

Figure 3.4: Milestoning involves splitting a simulation into several components and running them as individual simulations. The trajectory shown could also move to the previous milestone.

is the distance between the N and C termini of the protein. This provides a convenient measure in this scenario, but many other easily measurable parameters could be used, such as energy or radius of gyration. Independent simulations are started at each of the milestones, and ended when they reach either the previous or next milestone. Each of these calculations is therefore much shorter than one running for the entire length of the overall trajectory.

There are several benefits to milestoning (180): the most obvious is that each simulation can be run as a separate process. This gives the advantage of parallel computing without the inter-process overhead and software customisation normally associated with parallelisation. However, milestoning can also help increase the chances of events such as unfolding actually taking place during the simulations. Barriers on the energy landscape will prevent certain events happening regularly, and the chance of a simulation

overcoming these barriers is proportional to the size of the barrier. By creating smaller trajectories with lower individual barriers, an exponential decrease in the number of trajectories required to overcome the barrier is achieved.

These benefits are substantial, but the method is subject to certain assumptions which must remain true for it to be effective. In particular, there is a requirement for each simulation to be independent of others, *i.e.* there must be "loss of memory" between milestones. This Markovian nature, where the important factor is the time between milestones rather than time across the entire ensemble, is only maintained if milestones are a certain distance apart. The average time between each milestone must be significantly greater than the time taken for equilibration at any point. West *et al.* suggest a method by which the minimum number of required hypersurfaces can be determined (180). This factor precludes the overall trajectory from being split into a large number of small milestones in a bid to maximise the benefits described above. In their original paper, Faradjian and Elber point out that not all processes can be easily compartmentalised into the discrete sections required; indeed they suggest that "identifying reaction coordinates that satisfy the above requirement is far from obvious" and are often selected on an *ad hoc* basis (179). They also note that the hyperplanes which make up the milestone boundaries are only approximations, rather than true representations of the hypersurfaces. This may lead to inaccuracies in calculations, a phenomenon discussed later in this Thesis. The requirement for an initial reaction coordinate from which the individual milestoning simulations can be commenced is also a prerequisite not present in some other types of MD. In practice, a number of structures representing each hypersurface will suffice, rather than a full reaction coordinate. Vanden-Eijnden *et al.* detailed the assumptions and drawbacks of the method in 2008 (181).

Implicit Solvent Milestoning

Work by Toofanny in 2005 (182) was the first to apply the milestoning technique to the problem of forced protein unfolding. 21 μs of simulation were carried out using an implicit solvent model. The energy landscapes produced from these results showed encouraging agreement with more traditional molecular dynamics trajectories, including the presence of the intermediate and transition state structures. A further intermediate was seen at approximately 10 Å similar to that reported by Best *et al.* (183). This involves a rearrangement of the A' and G strands; rather than moving quickly apart following the concurrent breaking of six hydrogen bonds, as seen in AFM experiments, MD simulations show a slower "sliding" apart of the two strands. Employing an implicit solvent model does have the advantage of allowing long-timescale events to be observed more quickly due to decreased friction and damping of the system being studied. However, Toofanny acknowledges that the utilisation of a more detailed explicit solvent model would provide a more accurate model, albeit one beyond the scope of available computing power at that time.

3.3 Explicit Solvent Milestoning

The use of an explicit solvent model should, in the majority of cases, provide a superior and more detailed model than an implicit solvent approximation. The nature of molecular dynamics simulations mean that the number of calculations required to calculate one time step is approximately equal to the square of the number of particles in the system. While a number of strategies have been developed to reduce this number, the difference in required computational power between the two solvent treatments is still sizeable. This has traditionally resulted in short timescales and low numbers of re-

peats. The increased availability of computing power in recent years, along with a better understanding of how to utilise existing resources, has meant that long-timescale simulations in explicit solvent are now becoming more common. For the first time, it has been possible to study the forced unfolding of Titin I27 in explicit solvent at low force with a large ensemble, using the technique of milestoning.

3.3.1 Initial Simulations

Prior to starting milestoning simulations in explicit solvent, a number of calculations were run using an implicit solvent model. The aim of these was to obtain an approximation of the magnitude of force required to unfold the I27 domain, and see whether the extension of the molecule at different forces reproduced the results seen by other groups. Experimental work to examine the behaviour of mechanically stable proteins under force has shown that the higher the force applied to a molecule, the more resistant it becomes to that force (125).

The PDB structure for I27 had the N–terminal nitrogen and C–terminal carbon atoms pulled in opposing directions using the AFM function of CHARMM. A constant force was applied to each terminus, with force values from 200 pN to 500 pN. Fifty repeats of each force were carried out in implicit solvent using the CHARMM19 force field and EEF1 implicit solvent model. Each run lasted 10 ns, with a timestep of 2 fs using the CHARMM SHAKE algorithm. All simulations were carried out at a temperature of 300 K.

The results showed that, while the majority of runs showed full or partial extension of the protein at forces over 300 pN, there was very little or no extension at 200 pN and 250 pN (Table 3.1).

Force (pN)	Mean Extension (Å)
200	1.3
250	3.3
300	41.4
350	107.8
400	148.8
450	215.4
500	226.2

Table 3.1: Extension of the I27 domain at various forces in implicit solvent.

A typical extension curve for an unfolding trajectory (Figure 3.5) indicates that there are two main unfolding events, one to an intermediate at approximately 52 Å N-C terminus distance, with another intermediate at approximately 58 Å. Inspection of the trajectories for these simulations shows that mechanism of unfolding involves the separation of the A and B strands, followed later by the A′ and G strands (Figure 3.6).

These results are in agreement with the experimental and computational studies discussed earlier.

These implicit simulations were designed to give an initial idea of the force required to extend the domain. The absence of unfolding at lower forces led to the selection of forces from 300 pN to 500 pN for subsequent work. Although extension of around 40 Å was seen on average at 300 pN in implicit solvent, in the increased viscosity of an explicit solvent environment, less extension may be apparent.

In explicit solvent, three repeats of 1.5 ns of simulations were run at each of the selected forces. The decreased simulation time versus the implicit simulation was due to the

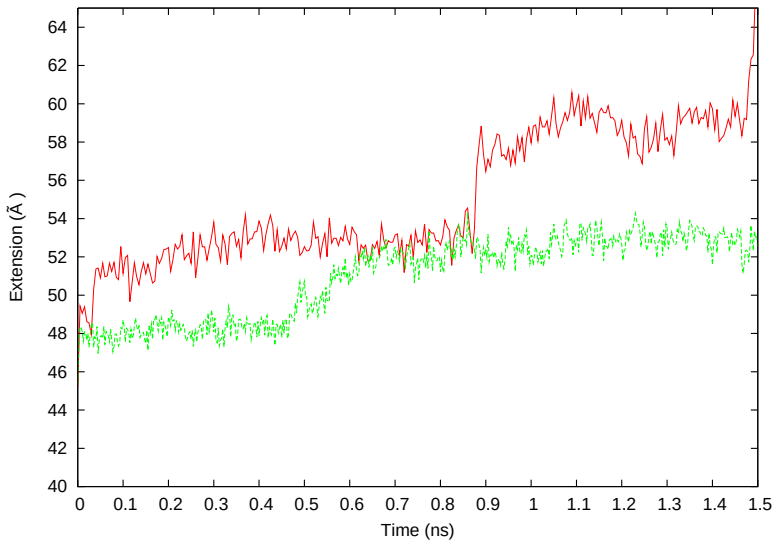

Figure 3.5: Extension of the I27 domain under 300 pN of force in implicit solvent (red) and explicit solvent (green).

computationally expensive nature of the explicit solvent model. All simulations used the CHARMM22 hydrogen parameters, with TIP3P solvent molecules in a 75.4 Å cubic solvent box. The system included 13275 water molecules. A 1 fs time–step was used. Electrostatics were treated with Particle Mesh Ewald summation, and non–bonded interactions cut off at 11 Å with a smoothing function between 7.5 Å and 11 Å. A constant temperature and pressure ensemble (NVT) was used, with the Nosé–Hoover thermostat maintaining the temperature at 298 K. Version 29b1 of the CHARMM software was used throughout.

The inherent friction and reduced simulation time compared with the implicit solvent simulations resulted in less terminus–to–terminus extension. As a result, even at 500 pN the maximum extension seen was around 55 Å. This includes the first unfolding event seen in the implicit studies, *i.e.* the separation of the A and B strands (Figure 3.7). This extension was noted with forces as low as 300 pN, and so this value was chosen for all

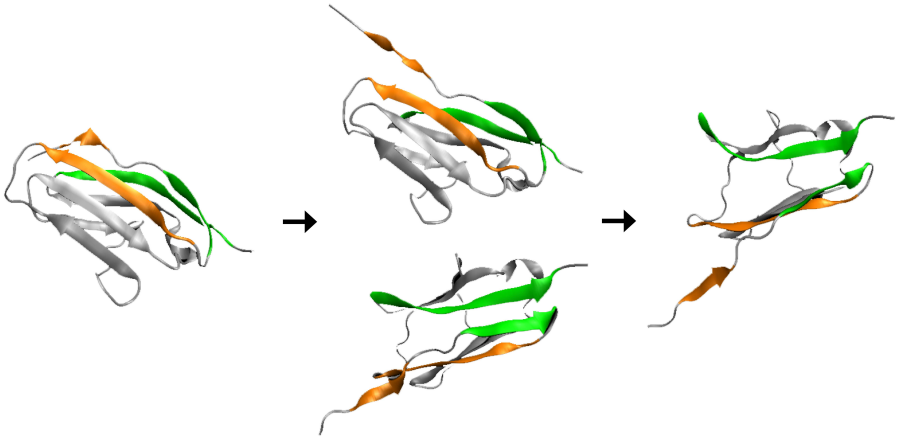

Figure 3.6: Unfolding events in implicit solvent: separation of A and B strands (orange) followed by separation of A′ and G strands (green).

subsequent simulations. The second unfolding event, separation of the A′ and G strands, was not noted in any of the short explicit simulations.

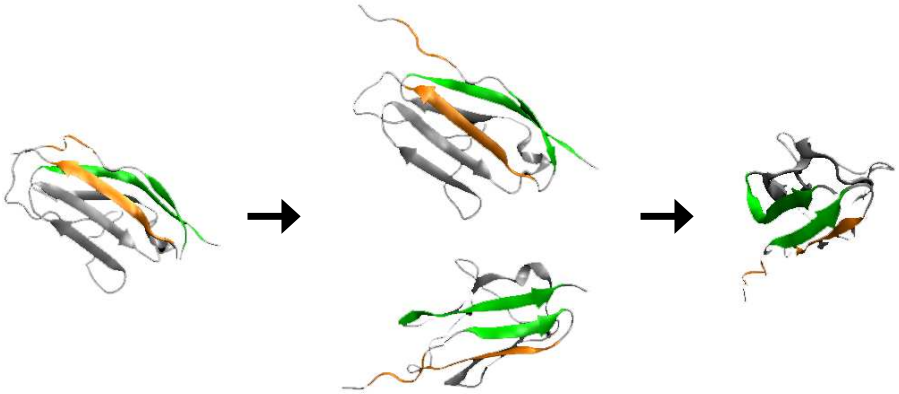

Figure 3.7: Unfolding events in explicit solvent: separation of A and B strands (orange) but no separation of A′ and G strands (green).

In both sets of simulations, the first unfolding event occurs at 7 Å extension, and in the implicit simulations the second event occurs at 14 Å extension. After this the protein extends rapidly. As these two events have been shown by this work and previous studies to be the important steps in the unfolding process, future work has focussed on the first 15 Å of extension. Although the implicit solvent simulations indicated that the

protein extended to around 300 Å, this is not physiologically relevant as the domain rarely unfolds to such an extent in muscle tissue.

3.3.2 Explicit Solvent Milestoning

The starting milestoning structures were obtained from work done by Toofanny *et al.* in 2005. The structures are taken from the first 15 Å of extension in a trajectory obtained using the Stochastic Difference Equation in Length (SDEL) (184, 185, 186, 187). SDEL approaches the protein folding issue as a boundary problem, rather than solving Newton's equations of motion. The boundaries consist of the folded and unfolded structures, and the search in this case is for a trajectory which makes the action stationary. From such an SDEL–derived trajectory, structures were originally extracted by Toofanny *et al.* at 1 Å intervals. For this work, the structures were solvated in an octahedral solvent box of edge length 80.655 Å. The solvated structures contained between 14,228 and 14, 272 atoms. The simulation methods and experimental conditions were the same as those used in the initial explicit simulations described previously. However, each starting structure ran for a period of 1 ns, and was then repeated. Although the 1.5 ns simulation time of previous work was not sufficient to give a full picture of the unfolding process, this piece of work resulted in multiple 1 ns snapshots. Each 1 ns trajectory can then be broken up according to the length of the structures contained within. Where two structures have the same N– to C– terminus length, they are considered to be equal, that is, they both lie at the same hyperplane. In this manner, it is possible to consider the multiple runs as a single ensemble. A total of fifty repeats were carried out for each milestone, as serial processes. This work was carried out with pulling forces of 200, 250 and 300 pN, giving a total of 750 ns simulation time at each force and 2.25 μs overall.

Analysis of the movement between milestones can be used to determine an energy landscape for the protein under force. A milestone distance is used to create the hyperplanes. The number of times simulations move between two milestones, and the time taken to do so, is counted. This can be a movement forwards, *i.e.* extension of the protein, or backwards, *i.e.* contraction. For example, if the milestone distance is set to 0.1 Å and a trajectory is started at milestone H, the structure could extend 0.1 Å to milestone H_{s+1} or contract 0.1 Å to H_{s-1}. From the new point, the structure may again extend or contract, thus passing another milestone. This process is repeated many times, to give multiple repeats. From the ratio of these folding and unfolding events, it is possible to determine the energy (Equation 3.3.2).

When a system is in thermodynamic equilibrium, the rate at which it traverses between two states is exponentially dependent on the difference in free energy of those states. From the ratio of the proportion of simulations moving forwards to the proportion of simulations terminating at the previous milestone, it is possible to estimate the free energy difference between the milestones, according to equation 3.3.1.

$$\Delta E_m = -ln\left(\frac{Proportion_m\ moving\ forwards}{Proportion_m\ moving\ backwards}\right) \tag{3.3.1}$$

The overall change in energy as the system traverses the milestones can then be calculated using equation 3.3.2:

$$E_m = \sum_{i=1}^{m} \Delta E_i \tag{3.3.2}$$

By plotting this information for each of the specified milestones, a free energy landscape can be generated; this is plotted in Figure 3.8.

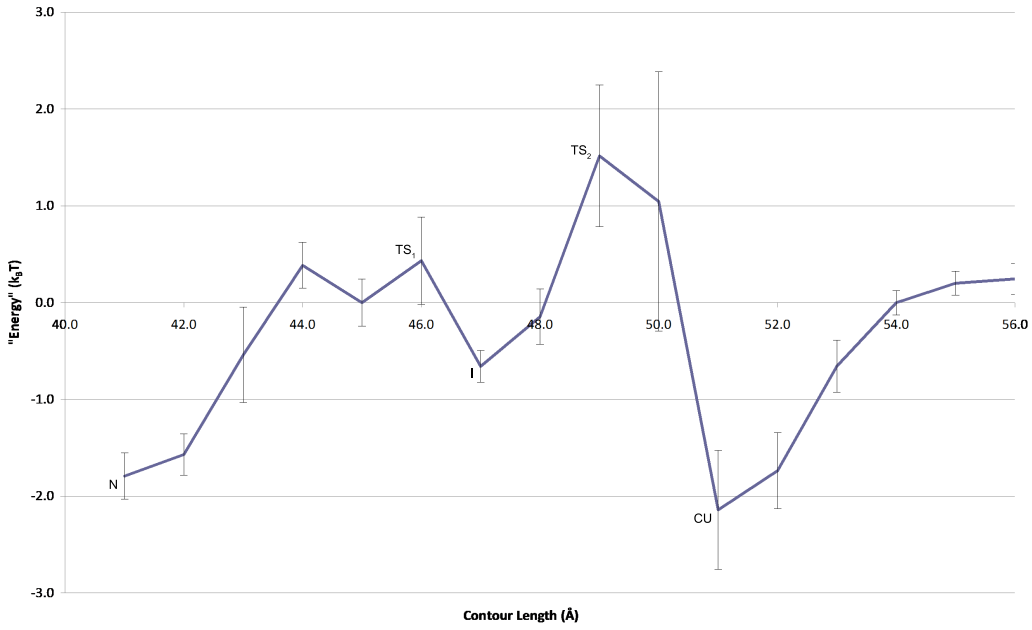

Figure 3.8: Energy landscape of Titin I27 under 200 pN of force. TS represents a Transition State, I an Intermediate and CU the Catastrophic Unfolding event.

The assumptions underlying the above are of a system in thermodynamic equilibrium, and that the time taken to traverse between milestones is significantly larger than the time taken to equilibrate at each milestone. Thus, the relative energies are shown as "energy" in Figure 3.8. While little reliance should be placed on the absolute values of the "energies" given, the underlying trend that relates to the shape of the energy landscape reveals valuable information and detail.

This landscape shares features with those proposed by a number of groups, suggesting that the important unfolding events noted in existing work may be also be seen in the trajectories obtained through these milestoning simulations. These landscapes all include an initial transition state, representing the energy barrier associated with the breaking of two hydrogen bonds between the A and B strands. A subsequent intermediate is the well-characterised meta-stable structure which is present prior to the concurrent breaking of six hydrogen bonds between the A′ and G strands. This breakage forms the large

peak in the energy landscape, and is the main barrier to unfolding of the molecule. It is this feature that provides the main mechanical strength to the protein. After this transition state is passed, the mechanical strength is lost, and the molecule proceeds to extend with little application of force.

The intermediate configuration is noted at approximately 6.5 Å extension; this is in agreement with a number of previous studies. The structure was first noted by Marszalek *et al.* (166), with a distance of 6.6 Å obtained from AFM work, and 6 Å from steered molecular dynamics, at a force of 100 pN. Lu and Schulten calculated the extension to be between 6 Å and 12 Å depending on the force applied (126). Figure 3.9 shows the average structure across all runs at the intermediate (46.8 Å N:C distance). The A strand has become detached from the B strand. The A′ and G strands are still aligned together, as expected.

The progression from the intermediate to the transition state has been described as "the key event in force-induced unfolding" of I27 (127) by Lu and Schulten, who found this burst phase occurring after 3 Å of further extension from the intermediate. From figure 3.8, a distance of 2.8 Å can be derived from this explicit milestoning. This work agrees with that of Williams *et al.* (188), who also proposed a 3 Å extension for the event based on AFM studies. Previous milestoning work in implicit solvent provides a third source of agreement for the proposed distance (182). Best *et al.* found the two features to be 3 Å apart (183).

The well-characterised "all or nothing" nature of individual I27 domain unfolding is thought to be due to the rapid unfolding after separation the A′ and G strands at the transition state. The average structure of the transition state conformations show that these strands are still adjacent to one another at the transition state (Figure 3.10).

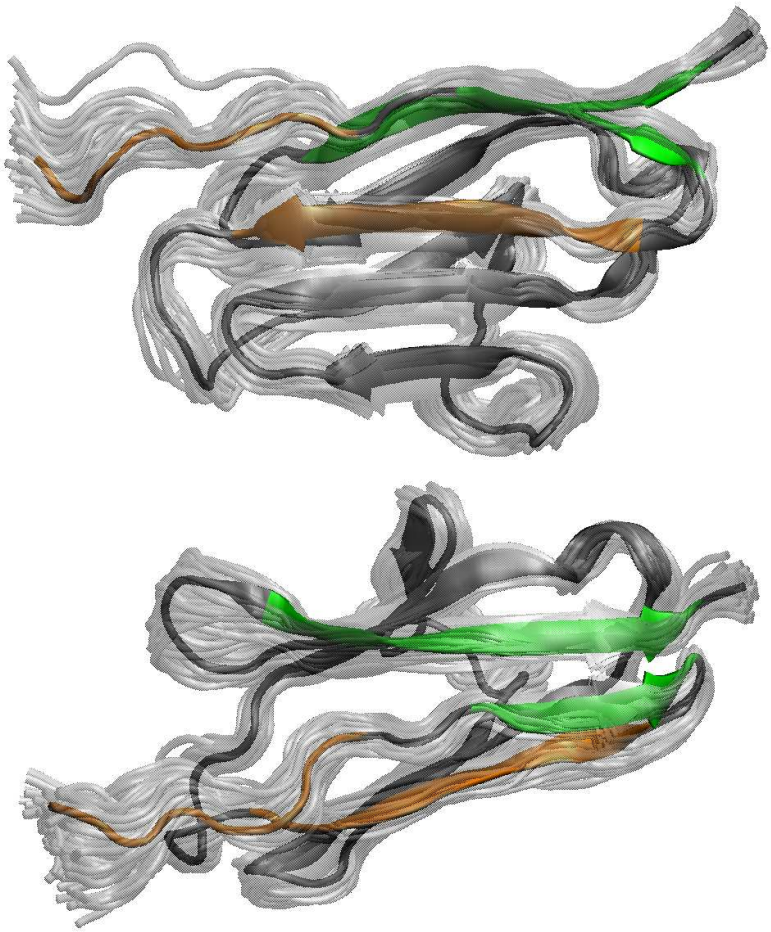

Figure 3.9: The average structure (dark grey) and all structures (light grey) at the 6.5 Å intermediate, 200 pN. Top oriented to show separation of A and B strands (orange). Bottom oriented to show A′ and G β–sheet contact remains (green).

An angstrom further along the reaction coordinate, the A′ and G strands have separated. These strands are responsible for the majority of the structural strength of the molecule, and their separation results in rapid unfolding of the protein. Figure 3.11 shows these two strands.

The landscapes at 250 and 300 pN show the same features (Figure 3.12), albeit at extended distances due to the larger forces being applied.

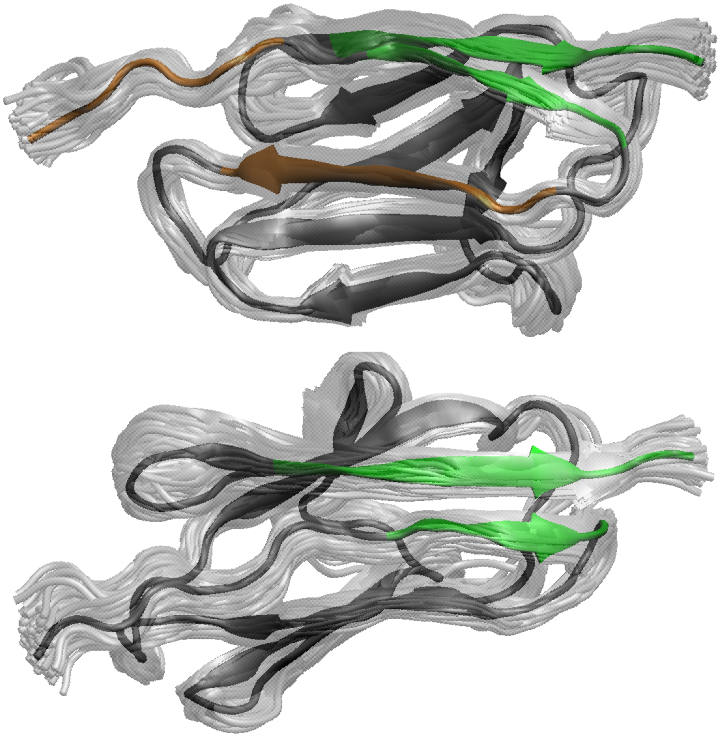

Figure 3.10: The average structure (dark grey) and all structures (light grey) at the main transition state, 200 pN. Top oriented to show continued separation of A and B strands (orange). Bottom oriented to show A′ and G β–sheet contact remains (green).

The ability to run repeats of this work at a variety of forces is a consequence of the efficiency of the milestoning technique. By comparing the relative positions of features on the landscape at different forces, it is possible to calculate the Young's modulus, E, of the molecule (Equation 3.3.3).

$$E = \frac{F/A_0}{\Delta L/L_0} \tag{3.3.3}$$

where E is potential energy, F is force, A_0 is initial cross–sectional area, ΔL is change in length and L_0 is initial length.

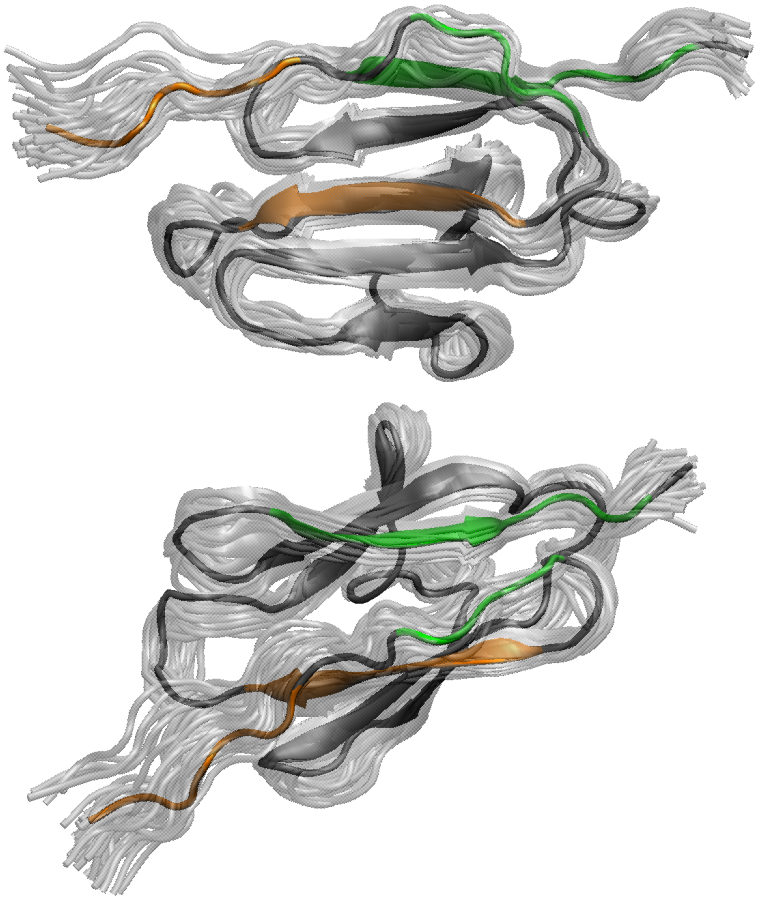

Figure 3.11: The average structure (dark grey) and all structures (light grey) after the main unfolding event, 200 pN. Top oriented to show continued separation of A and B strands (orange). Bottom oriented to subsequent separation of A′ and G strands and resulting loss of β–sheet (green).

I27 is an elastic domain, and the Young's modulus provides an indication of the stiffness of the protein. Using figures from both landscapes generated above, an original molecule length (L_0) of approximately 7.3 Å is derived. This equates to an Young's modulus for the domain of 0.7GPa, which is in the order of those noted for polymers. The calculation of the Young's modulus for biological molecules is complicated by the value's relationship with the cross–sectional area of the molecule in question. In calculating the value quoted above, a cross–sectional area of 1 nm^2 has been assumed.

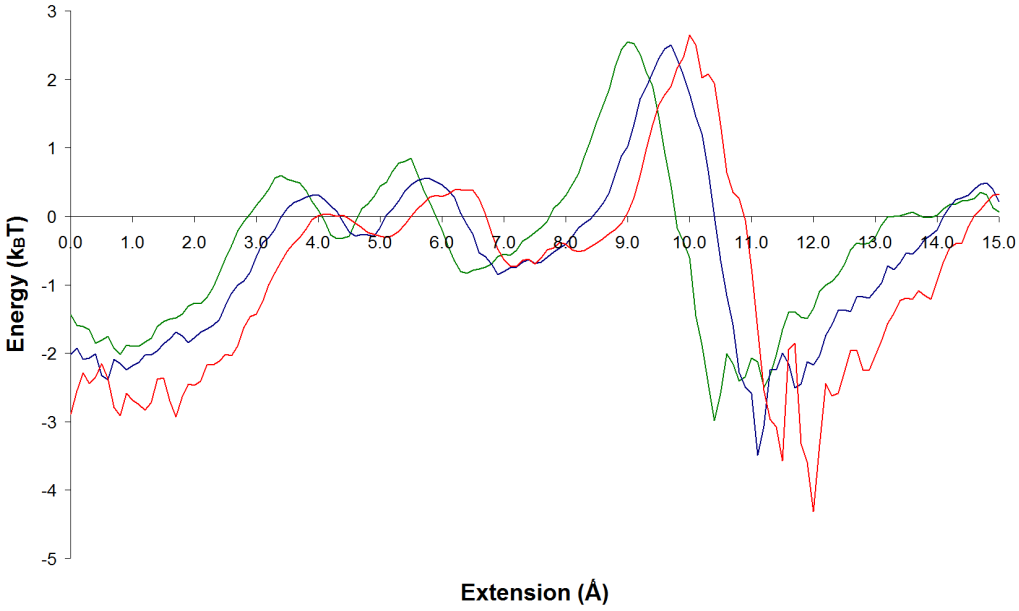

Figure 3.12: The energy landscape of I27 under 200 pN (green), 250 pN (blue) and 300 pN (red) of force.

The sequence of unfolding events described above has been previously elucidated as a result of a number of *in vitro* and *in silico* studies (126, 166). Some work has also involved exposing the roles of specific residues within the protein (178, 188). Contact maps derived from the milestoning trajectories show that residues identified in those studies also seem to be important in this work. For example, Li *et al* created point mutations in the A' strand of I27, disrupting hydrogen bonding and therefore destabilising the molecule. Their work showed residues in the A' strand to be important in retaining the structure of the molecule, which is supported by interactions in these milestoning simulations. Figure 3.13 shows that these residues remain in contact with those making up the G strand until the main transition state is reached. The protein loses its stability only after the contacts between these two strands are lost.

Further analysis shows that, using a 6 Å cutoff to define 'contact' between residue side–chains, there were nine A' and G strand residue pairs in contact at the transition state.

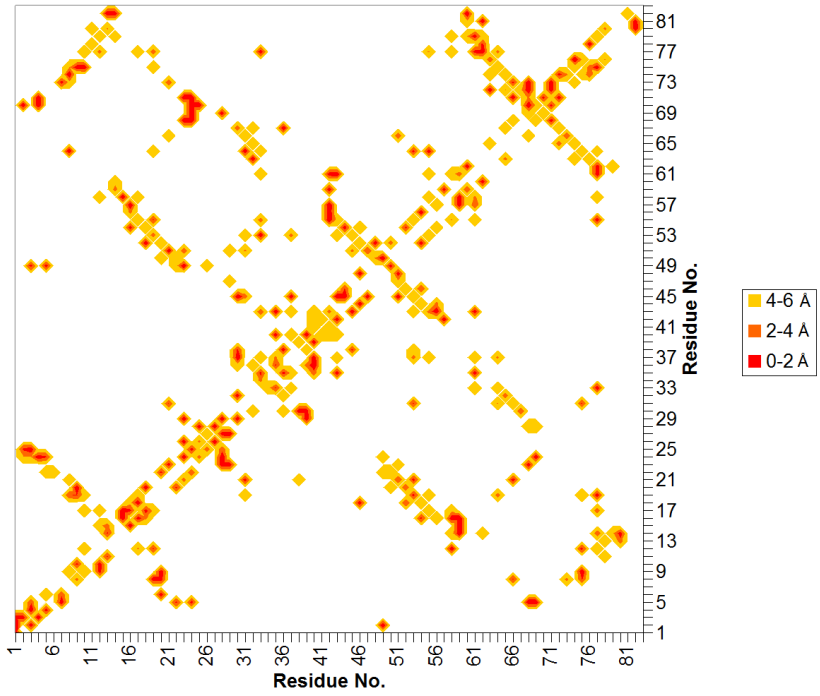

Figure 3.13: Contact maps for I27 at 200 pN (6 Å cutoff). Top–left shows contacts at the intermediate, bottom–right shows contacts after the transition state has been passed.

At 1 Å extension past the transition state, only four of these pairs remained in contact. The contact maps also show the breaking apart of the A and B strands, *i.e.* the hydrogen bonds between residues 4 to 7 and 18 to 25.

The developers of the milestoning method, Elber *et al.*, propose a more detailed analysis method in their original paper (179, 189). The "*QK* picture" entails using the individual times taken for each forward or backward movement. A histogram of transitional probability and waiting time is produced from this data for each milestone.

These are then used to effectively re–run the simulation. A set number of repeats start at each milestone, and after one timestep the relevant proportions are moved to the previous or next milestone 'bin' according to the probabilities calculated from the raw data. At the next timestep, moves are again made according to the probabilities for each separate

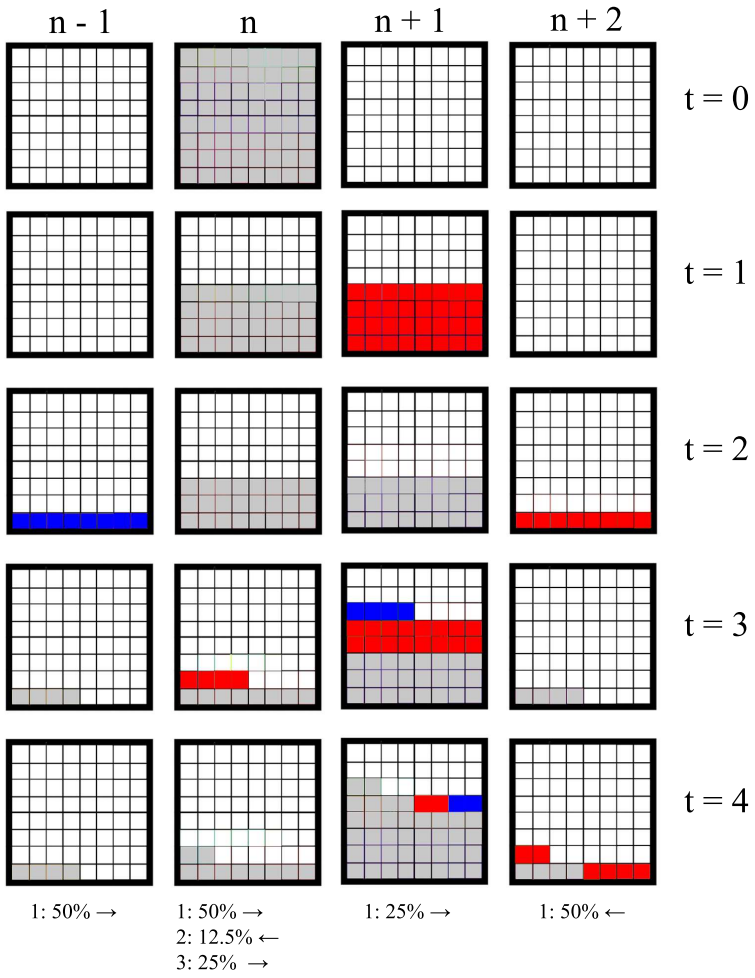

Figure 3.14: A diagrammatic example of the "QK" method. Repeats moving forwards at each timestep are shown in red, while those moving backwards are shown in blue. n is the milestone number. t is time.

combination of waiting times and milestone position. Each milestone position may include repeats which have waited at that point for differing lengths of time; each of these different waiting time sets will move forwards or backwards in different proportions. This process is repeated for each milestone and each timestep until there is no further movement between bins (Figure 3.14). The final distributions of the repeats amongst the milestones represent the Boltzmann weight of each. From these, it is possible to calculate

the energy of the milestones in a similar fashion to Equation 3.3.2. Figure 3.15 shows the energy landscape determined using this method, at a pulling force of 200 pN.

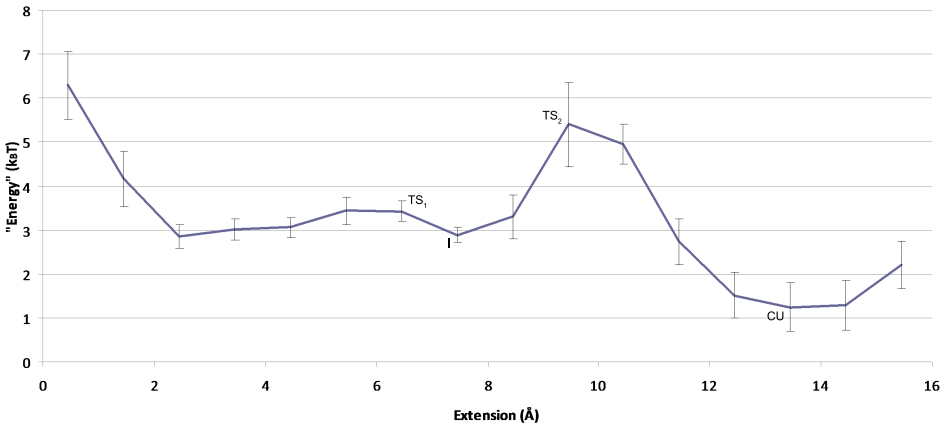

Figure 3.15: Energy landscape of Titin I27 at 200 pN, as determined by QK calculations. TS represents a Transition State, I an Intermediate and CU the Catastrophic Unfolding event.

Using this *QK* method to investigate the explicit solvent milestoning results, it has been possible to reconstruct energy landscapes at each of the different forces, and compare their effects (Figure 3.16). In this analysis, there was a separation distance of 0.1 Å between each milestone. The landscapes show the predominant features seen in the initial analysis, *i.e.* the 6 Å transition state and 9 Å intermediate.

Previous work on the forced unfolding of multiple proteins has shown that the application of force can tilt the energy landscape (190), and in some cases even alter the unfolding pathway (178). Given that the effect of force is to lower energy barriers on a landscape, the subsequent tilting is the logical modification. The increased efficiency of the milestoning method has allowed increased numbers of repeats to be carried out at a variety of forces in this work. This in turn has again demonstrated that force can tilt the

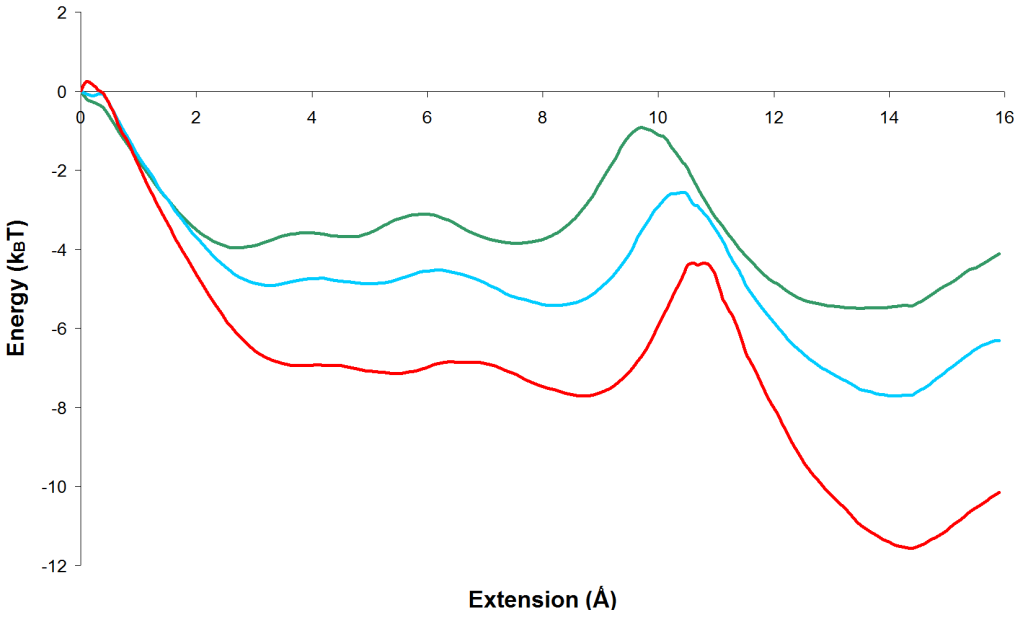

Figure 3.16: Energy landscapes of Titin I27 at 200 pN (green), 250 pN (blue) and 300 pN (red), as determined by QK calculations.

energy landscape (Figure 3.16). The smaller barriers mean that the protein is more likely to reach the transition state and therefore also more likely to unfold.

It should be noted that the energy values given in the landscapes resulting from this work are much smaller than expected. For example, the energy difference between the intermediate and the second transition state is approximately $3k_BT$. While the relative positions of the features on the landscape are correct, as are the relative heights (*e.g.* the second transition state is a much larger barrier than the first), the energies cannot be considered as absolute figures. This may be due to the nature of the hyperplanes, which are defined based on N–C terminal distance. As the distances seen within results do not exactly match up with the hyperplane distances, it is possible that there is some inaccuracy around the subsequent energy difference. Instead of being calculated exactly between hyperplanes, the energy differences calculated may be between points either side of the two hyperplanes in question. This presents the possibility of underestimating

the energy due to effectively omitting the energy associated with certain parts of the landscape. However, a more likely explanation than this effective "rounding" error is one discussed by Faradjian *et al.* (179). They point out that the time taken for a simulation to travel between milestones must be greater than that required for equilibration at each hyperplane. It is possible that the timescales in this simulation are not sufficiently great to ensure such integrity. As the two times become less disparate, the equilibration time at each milestone becomes significant. This, however, is neglected in these calculations and may account for the difference between the expected energies and those obtained using the method.

Work by Marszalek *et al.* was among the first to examine the unfolding mechanism of I27. It showed an elongation of the protein which was only noted at a force of 108 $\pm$ 19 pN (166). The energies of the intermediate at the forces undertaken in this study indicate that the predicted value for unfolding to the intermediate is 105 pN (Figure 3.17). This is therefore in agreement with the work of Marszalek *et al.* and also that of Williams *et al.* (188).

3.4 Conclusions

Computational power has long been the restricting factor in the simulation of biological systems. For over a decade, the I27 domain of Titin has been the focus of a large number of experimental and computational studies. As the processing power available to researchers has increased, so the *in silico* studies of I27 have become more detailed. Here we present the most fine-grained simulations to date. The milestoning method has allowed multiple repeats to be performed, and in a manner which tends to increase the frequency of occurrence of traditionally rare events. This method provides the benefits

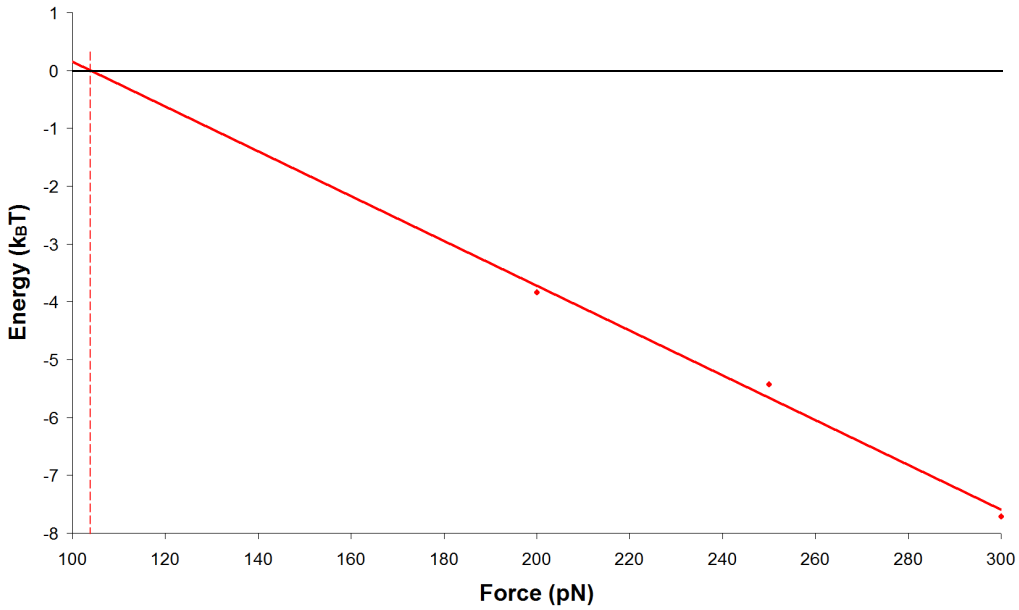

Figure 3.17: Explicit milestoning simulations suggest that I27 could unfold to the intermediate structure at a force of approximately 105 pN.

of parallel computing, without the time-consuming requirement to customise software or communicate between processes during the simulations. In combination with the availability of a large pool of compute resource, this provided a method by which an explicit water model could be used to study the effects of very low levels of force on the molecule.

Previous work in this area has involved simulated forced unfolding by molecular dynamics, by the application of a constant force or constant velocity (*i.e.* steered molecular dynamics, or SMD) and the subsequent study of the time evolution of one or more simulations. Presented here is the first simulation of forced unfolding through the analysis of kinetics. Molecular dynamics simulations are undertaken to reveal the kinetics alone of transit across the lanscape. Analysis of these kinetics permitted an energy landscape of the unfolding protein to be produced, and features in this related to the observed experimental results.

The results of the simulations have shown good agreement with a number of previous studies. As well as reinforcing the results of this earlier work, this also suggests that the forced unfolding of I27 is a suitable model for the application of the milestoning method. The simulations, at 200, 250 and 300 pN, identified the main features previously proposed on the forced unfolding pathway. These include the separation of the A and B strands, which provide less structural stability than the later separation of the A' and G strands. The results also suggest that, in line with predictions by other groups, I27 may unfold to the main intermediate structure at forces of approximately 105 pN under physiological conditions.

Chapter 4

Least Action Dynamics

4.1 Background

The previous Chapter discussed a method by which a most likely, *i.e.* lowest–energy, reaction coordinate could be determined. While traditional molecular mechanics approaches rely on the integration of Newton's laws of motion to describe the movement of a molecule over time, there are alternative methods which can be used to describe motion. One such method is considered in this Chapter, and developed in a novel manner utilising the the lattice model introduced in Chapter 2.

4.1.1 Principle of Least Action

An object moving from one point to another over a defined period of time can follow a number of possible paths. For example, an object thrown straight up into the air could, hypothetically, return to its start point by moving only upwards and then downwards, or by oscillating up and down before eventually returning to its starting position (Figure 4.1).

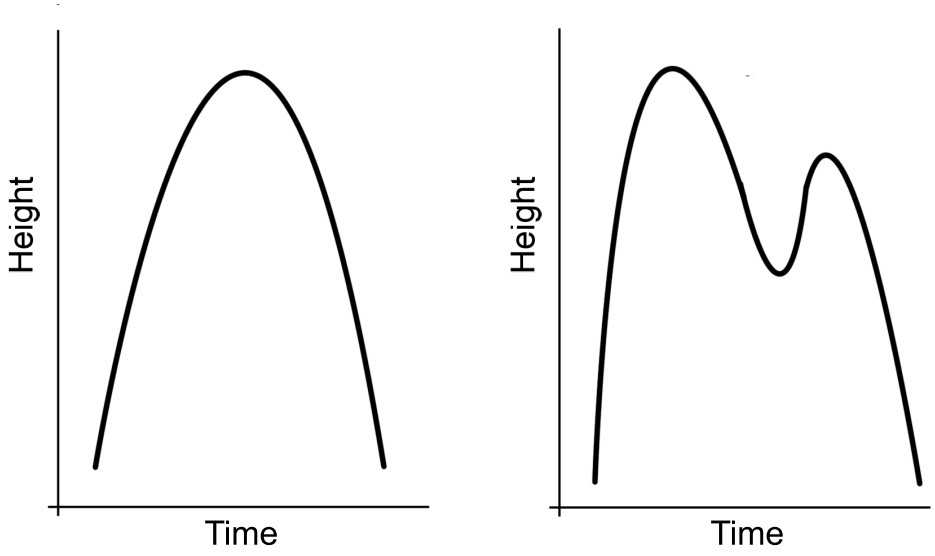

Figure 4.1: Possible paths for an object thrown straight up into the air returning to its original position. The true path, on the left, has the smallest action.

The Principle of Least Action states that, for these two alternative paths, the actual one taken will be that where the average kinetic energy minus the average potential energy is minimised. When applied to a single point in time, this term is known as the Lagrangian (Equation 4.1.1).

$$L = T - V \tag{4.1.1}$$

where L is the Lagrangian, T is average kinetic energy, and V is average potential energy.

To determine the 'action', S, for an object over time, we integrate the Lagrangian over the duration of the event, t (Equation 4.1.2).

$$S = \int_{t_1}^{t_2} L \, dt \tag{4.1.2}$$

where S is the action, t is time, and L is the Lagrangian.

It follows, therefore, that by optimising the above function over multiple paths, where the start and end configurations form the boundaries, which can determine which one is most probable.

The example given above involves a single particle moving in one dimension; when this principle is applied to a large system such as a protein, moving in three dimensions, it becomes more complex but equally applicable. The kinetic energy, for example, can be calculated by considering the movement of each particle in all three dimensions (Equation 4.1.3).

$$K_E = \frac{m}{2}\left[\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2\right] \tag{4.1.3}$$

where K_E is kinetic energy, m is mass, and t is time.

It should be borne in mind that t has no real meaning in this instance and can be treated simply as a parameter. In addition, moves undertaken do not have to be physically possible in the manner required by traditional molecular dynamics; the separation over time means that each new structure discovered does not necessarily have to be immediately accessible from the current structure. There are several methods by which the functional may be minimised, and thus the most likely path determined. For example, it is possible to generate large numbers of candidate paths, and compare their actions to determine which has the lowest action. While this is a trivial computation *in silico* for one path, assessing a large number of paths becomes more demanding. The methods used to generate the paths may vary, being either refinements of a previous path (as used in this Thesis) or generated at random by Monte Carlo, for example. Alternatives have been proposed which seek to find a stationary point on the function representing the action (191), but such minimisations often entail the costly calculation of second derivatives.

One major advantage of least action calculations over molecular dynamics is that they effectively increase the frequency with which rare events are observed. When moving between two stable states on an energy landscape, for example, traditional molecular dynamics simulations are reliant on probability, running until the energy barrier between the states is overcome. By its nature, least action moves towards the most likely path without having the requirement of overcoming this barrier. In addition, it is also possible to find least action pathways in a manner which does not necessarily require second derivatives to be calculated.

4.1.2 Action–Derived Molecular Dynamics (ADMD)

The idea of using least action in the realm of molecular dynamics was first explored by Elber and Karplus in 1987 (192). They used a modified version of CHARMM to search for the least action paths of conformational changes in cyclohexane, dialanine and myoglobin. Their choice of examples raises an important point in consideration of this method, which is that both the start and end conformations of the system must be known. They noted that the system was efficient at finding local minima, and could also be run in parallel. Olender and Elber developed the model in the context of a very large time step (relative to traditional MD) in 1996 (193), using a variety of methods to determine the best action. These included conjugate gradient minimisation and straight line interpolation. Despite some success searching for local minima, they found these methods did not efficiently search for the lowest global action. By employing other variations, such as a variety of initial guess trajectories and altering the resolution of the intial guess path, they were able to improve the performance of the model. Lee *et al.* also found the former to be important in their work (194). The technique subsequently saw relatively little use until it was revisited in 2001 (191) and 2003 (195).

In a computational setting, the time between the start and end configurations, t, is actually discretised into a number of intervals, P. Each interval represents a small proportion of the total time (Equation 4.1.4).

$$\Delta t = \frac{t}{P} \tag{4.1.4}$$

where t is time, and P is the number of intervals.

The calculations remain applicable because the pathway with the lowest action remains the same regardless of whether it is calculated alone or as part of a larger path (Figure 4.2). It is this discretisation which also makes ADMD so suitable for being run as a parallel process.

The total action is then calculated over the sum of these intervals (Equation 4.1.5).

$$S^h = \sum_{0}^{P-1} L_j \, \Delta \tag{4.1.5}$$

where S is the action, P is the number of intervals, and L_j is the Lagrangian.

This is repeated for a number of different paths, with the one corresponding to the lowest action being the most likely.

4.2 Action–Derived Lattice Dynamics (ADLD)

4.2.1 Introduction

Work to date involving ADMD has used all-atom representations of target systems. While these undoubtedly provide a higher resolution and accuracy in their use, they

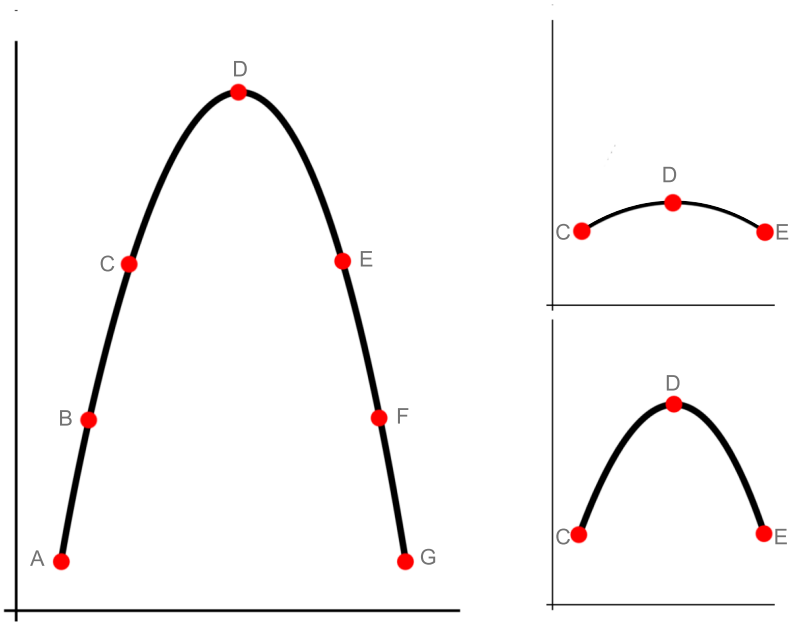

Figure 4.2: Discretising a least action pathway. The pathway calculated between points C, D and E alone is the same as that obtained as part of the pathway from A to G; calculating the small section in isolation does not affect the action. For this reason, the pathway shown top-right is not the one with the least action; the pathway shown bottom-right remains the pathway with the lowest action, despite being calculated separately.

also require much more computational power and computing time than lower resolution models. The coarse–grain nature and reduced search space of lattice models make them ideal tools with which to quickly generate large numbers of candidate structures for a system. The efficiency of the lattice search allows the conformational space to be explored in a much more comprehensive manner than would be possible with molecular dynamics. This in turn means that many more trajectories can be investigated. While a lattice could be applied to atomistic simulations, integration of the laws of motion on a lattice are problematic; the lack of such integration in least action also represents another advantage of the model.

The LaMP lattice model discussed in Chapter 2 has been employed in order to compare actions and find least action pathways. A simple example was considered first, to verify that the model functioned in the expected manner. The method was then been applied to the unfolding trajectory of Titin I27. This allowed the sampling of high numbers of candidate structures and actions, and resulted in the creation of a dynamic trajectory through a means other than the standard integration of Newton's equations of motion.

4.2.2 Validation

The standard LaMP model detailed earlier in this work was used as the basis for the new model. As previously, the less accurate α–helix term was de–activated for use on I27, which contains only β–sheet. All searches were carried out using the Replica Exchange algorithm, with no restraints on the molecules.

The simplest possible example of a least action calculation was chosen, using only three structures. The beginning and end structures were the same as those used in the previous section; the middle structure was the fifth intermediate. In each run, a lattice search was carried out to generate new structures. All searches used the Replica Exchange algorithm for 10^7 steps, with new candidate structures saved every 10^5 steps. The timestep was varied for each run, as described in Table 4.1.

At the end of each run, the middle structure representing the least action was saved. The five runs were carried out in series, that is, the end result from Run 1 formed the start structure for Run 2. This ensured that different start structures were used to initiate each run.

The structures obtained at each of the separate temperatures were compared. It can be seen from Figure 4.3 that the structures run using different timescales are markedly

Run	Timestep (ns)
1	10^{-6}
2	10^{2}
3	10^{-5}
4	10^{2}
5	10^{-5}

Table 4.1: Timesteps used in ADMD validation.

different. However, repeats using the same timestep were found to be the same. The two runs using a 100 ns timestep generated identical structures, as did the two runs using a 10^{-5} timestep.

Altering the timestep in the model clearly affects the structures generated, as expected. The use of a very small timestep infers that a structure needs to move from a start conformation to an end conformation very quickly. There is little time for movement, and so the middle structure could be expected to be close to an interpolation between the start and end structures. When a large timestep is used, there is more time for the structure to visit different conformations which may have a smaller action.

The similarity between the structures obtained using identical timesteps is encouraging. The two structures resulting from the 100 ns timestep are identical, as are those from the 10^{-5} ns runs. In itself, this suggests that the model consistently finds the lowest action path within those sampled on a given landscape. However, the difference between the timestep results is also of interest. With the shortest timestep, 10^{-6} ns, the structure representing the least action has an RMSD of 2.018 Å from the start structure. The RMSD increases to 3.458 Å when the timestep is increased to 10^{-5} ns. The conformations re-

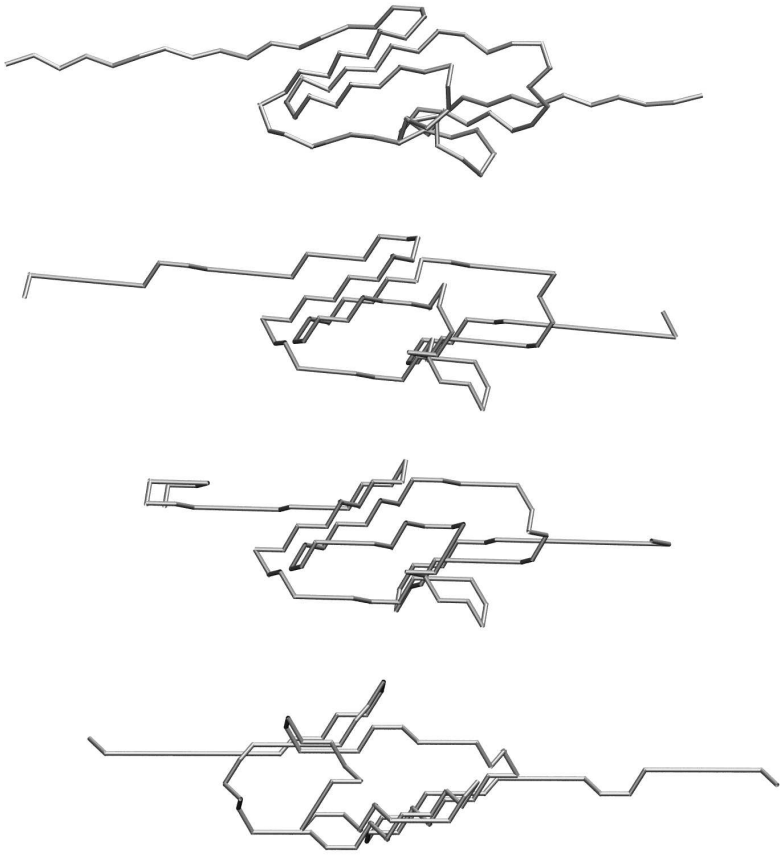

Figure 4.3: Structures obtained from ADMD runs using varying timesteps. Top - bottom: Original conformation, 10^{-6} ns, 10^{-5} ns, 10^{2} ns.

sulting from the 100 ns runs are very different to the results from the shorter timestep runs; the RMSD between the 100 ns structures and the starting structure is 8.046 Å. These results reflect the expected behaviour of the model when applied to this example, and support the conclusion that the method is an efficient way of finding the pathway of least action.

4.2.3 Forced Unfolding of Titin I27

The start configuration for the overall calculation set was defined as the native state (N–C terminal distance 40.1 Å), while the end configuration was an extended conformation

(N–C distance 132.6 Å) taken from a previous all-atom forced unfolding trajectory. Ten other conformations along this trajectory were also chosen, resulting in a total of twelve structures. Each was converted to a lattice approximation using the *PDB2Lattice* package of LaMP.

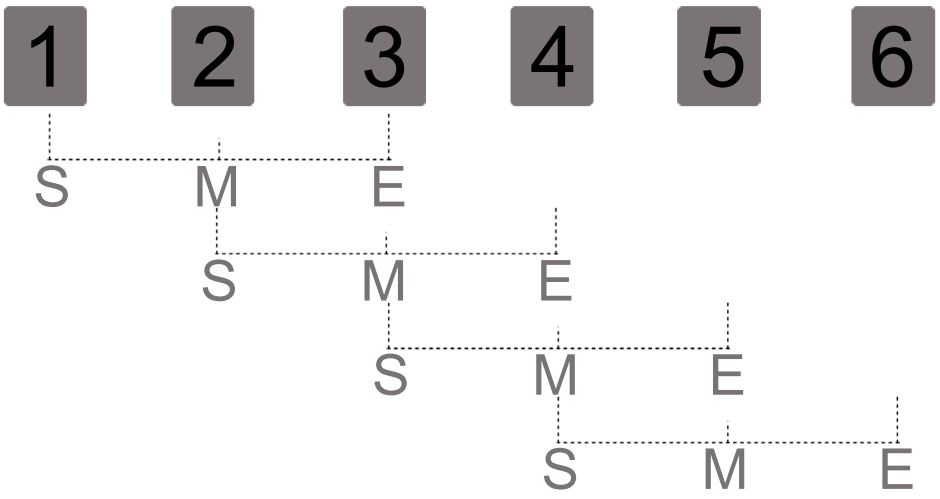

Figure 4.4: Schematic showing the configuration of the least action lattice dynamics setup. Each triplet represents a start, a middle and an end structure. Structures 7 - 12 and triplets starting from structure 4 are omitted for clarity.

As discussed previously, calculation of an action requires three structures; two boundary structures at the start and end, and a middle structure. It is this middle structure which is altered to create the different paths, each of which has a different action. The twelve structures were therefore considered as a set of ten triplets. For each triplet, the start and end configurations were fixed, and the middle structure altered by LaMP to generate new candidate structures. A search was carried out on the first triplet, *i.e.* structures 1, 2 and 3. This generated 1,000 potential 'middle' structures. The action for each of these structures was evaluated, and the conformation resulting in the lowest action then replaced the existing structure. This process was repeated for the next triplet, consisting

of structures 2, 3 and 4, and so on for the remaining eight triplets. Each time, the structure with the least action was carried through to the next calculation. When the final triplet was completed, the temperature of the search was reduced by 10% and the procedure started again. This ran for over 400 repeats, resulting in the action being continually refined from start to finish across the ten triplets.

The method described above replaces the ten original intermediate structures with those found to represent the least action pathway. By combining these final structures, it is possible to create a trajectory showing the predicted most probable path from the overall start structure to the end structure. Figure 4.5 shows the twelve frames in the trajectory resulting from the work. In these calculations, a timestep of 1.0 ns was used. Over 412,000 structures and actions were generated and evaluated during the calculations.

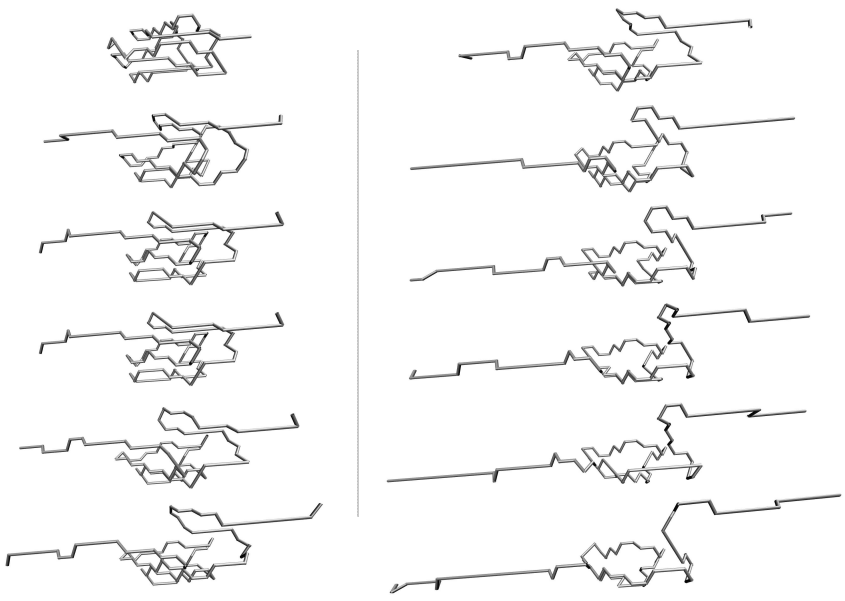

Figure 4.5: Twelve structures making up the trajectory resulting from Least Action Lattice Dynamics (1 ns timestep).

The calculations were also repeated using the same starting conformations, but a much larger 100 ns timestep. As noted in the initial validation, the structures obtained with the longer timestep showed much more variation (Figure 4.6).

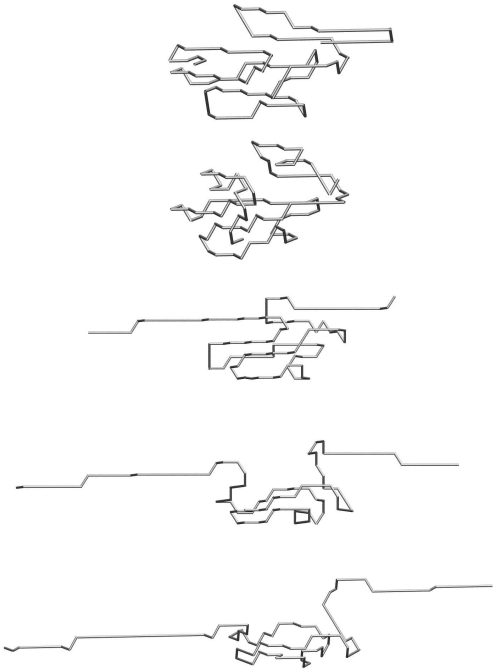

Figure 4.6: A selection of the structures from the I27 unfolding trajectory (100 ns timestep).

Conclusions

This method has successfully produced a series of structures representing the transition from a folded to unfolded state in Titin I27. This pathway has the lowest action of those sampled over the entire trajectory. The structures demonstrate a continued extension of the molecule, with no artifacts or obvious deficiencies. While the final structures are broadly similar to the starting structures provided to the model, this is to be expected. The initial structures were not a disparate set of random structures, but already represented a trajectory thought to be the likely unfolding pathway.

It should be noted that the model is dependent on the underlying search strategy, and the potential used to assess the energy of structures. Any further development of the LaMP model would also benefit this work through more accurate calculation of energies.

While the model is an efficient one, the fidelity could be improved by the introduction of more points between the overall start and end structures. However, this effectively restricts the trajectory in its earliest phase and would therefore require increased computation time to allow for this. It should also be borne in mind that all structures generated for the final trajectory are still united–atom residues bound to the lattice. The next logical step is to convert these back into all–atom representations, and consider running small atomistic simulations with mild backbone restraints. This may allow the structures to relax from the imposed lattice positions towards truer conformations.

The choice of timestep is also an important factor in the quality of the structures produced. Varying the timestep can cause a large variation in the resulting structures, a phenomenon which is discussed in more detail below. Further work should therefore be undertaken to determine the optimal timestep.

Chapter 5

Conclusions

The use of computers to simulate biological systems has been developed over several decades. From the simplest polypeptides to large macromolecular systems, these simulations have developed in parallel with computing power. Over this time our understanding of the processes governing protein folding has also developed, and the simulations have progressed as a result of this.

Despite this progression, it is still not possible to accurately simulate large systems on a relatively short timescale. This is due in part to the size of the systems under scrutiny, and the inherent complexity of the forces governing their behaviour. It is also a consequence of the systems being studied; these tend to involve relatively rare processes whereby a system overcomes an energetic barrier. The nature of these events means that even when simulations can be run, they do not necessarily demonstrate the event of interest.

As well as providing useful insight in isolation, MD data provides a useful tool to complement experimental data. For these reasons, work continues to enhance the capabilities of models. The ability to run simulations in a shorter amount of time has been

approached from many angles; parallelisation, coarse-grain models, and alterations to experimental conditions have all been utilised.

This thesis has investigated two such methods. The first involved the development of a coarse–grain lattice model requiring less computational power than all–atom off–lattice models. The second technique, milestoning, facilitated better utilisation of existing resources, along with increased probability of observing rare events. Finally, the coarse–grain model was used to efficiently generate dynamics data, using the principle of Least Action as an alternative to traditional Newtonian mechanics.

5.1 LaMP - A Lattice Model of Protein Folding

Coarse–grain models reduce the number of particles within the system being studied. This reduces the number of pairwise contacts, and therefore the number of calculations required to determine movement at each timestep. By fitting particles within a system to a grid or mesh, it is possible to reduce the search space for a given problem. Assuming a suitable grid, or lattice, is chosen, the search may still be of sufficient fidelity to provide useful structural information. These techniques have been combined into the LaMP program, which can be used to discover low–energy protein structures.

A standard PDB structure is converted to a reduced–atom structure on a lattice framework. A Monte Carlo search scheme is used to probe the lattice search space to find conformations with the lowest energy, as defined by the model's energy terms. These terms have been modified to include reproduction of hydrogen bonding in protein secondary structure elements. The β–sheet term has been shown to facilitate the production of sheet elements, whereas the previous version of the model tended towards compact globular structures. The reproduction of α–helices proved more complex due to the coarse–grain

nature of the lattice. A number of metrics were investigated and combined to produce a set of criteria for forming helix. Although this definition also forced the model away from globular structures, it was not possible to reproduce longer sections of helix. A variety of other features were added to or changed within the package, including more accurate conversion of PDB structures to the lattice format, the ability to restrain residues during a lattice search, and much greater customisation of output.

While β–sheet has been reproduced successfully, the continued development of the α–helix term is an obvious target for future work. The parameters defined so far have started to push structures into a relevant conformation, but only over very short peptide lengths within a larger structure. The application of both terms in a single model will also prove a crucial point to be addressed. Each term must exert sufficient influence to cause the formation of the appropriate secondary structure element where appropriate, but one must not dominate over the other. The model developed so far has provided a sound base on which these developments can be made.

This model represents the first time that a coarse–grain on–lattice model has been used in conjunction with Monte Carlo sampling of conformational space to find low–energy structures. Previous lattice models have, for example, required prior knowledge of the native or folded state, meaning that they are less applicable to problems where elucidation of this structure has not been possible experimentally. Traditional molecular dynamics have suffered from the requirement for lengthy run times. This has been minimised in this model by the use of a reduced–atom model, requiring the calculation of fewer interactions between particles, and the implementation of a lattice system, reducing the conformational search space. The development of hydrogen bonding within the model, in order to promote the formation of protein secondary structure, has given initially

promising results. There is now the potential for further development of the model to study protein folding in peptide aggregates and other systems.

5.2 Explicit Solvent Milestoning

While it is possible to use ever more powerful processors, or parallelise software and simulations, milestoning seeks an alternative solution in dealing with large particle numbers. By splitting the simulation into a number of discrete parts prior to it beginning, it is possible not only to effectively parallelise the simulation, but also to increase the chances of witnessing rare events such as protein unfolding. This dual effect makes milestoning a suitable choice for running more accurate simulations using the high numbers of atoms found in explicit solvent models.

The forced unfolding of the I27 domain of Titin has been studied using the milestoning technique. This is a well-characterised event which has previously been investigated both *in silco* and *in vitro*. Simulations followed an experimental procedure previously used in both domains, *i.e.* the application of force to pull the N– and C–termini of the molecule in opposite directions. The use of milestoning not only allowed multiple repeats to be carried out, but also facilitated the use of multiple very low forces of 200 - 300 pN. These are much more physiologically relevant than the extreme conditions which researchers have previously been forced to use to facilitate unfolding. The experiments showed good agreement both with other MD simulations and with AFM work. The energy landscape produced at all forces demonstrated features seen by other groups, including the production of an intermediate followed by the crossing of a transition state to rapid unfolding. The results obtained at different forces also agree with the concept of force tilting an energy landscape. Finally, extrapolation from the low forces used sug-

gests that the I27 intermediate may unfold *in vivo* under approximately 100 pN of force. These results show encouraging agreement with a number of other studies, and indicate that milestoning is an efficient method for increasing simulation efficiency without reducing the fidelity of the model being used.

The ability to model events over a physiologically lengthy timescale is a valuable tool, and one that can be applied across a number of fields. Milestoning has been shown to reduce the amount of time required to simulate such biologically–relevant timescales, thus increasing the relevance of simulation data in relation to such process. With particular reference to forced protein unfolding, any technique which allows the application of less extreme simulation conditions is particularly useful. Milestoning could therefore provide an accurate way of validating in detail previous simulations undertaken in implicit solvent, or of further reducing the force which is applied to molecules to demonstrate unfolding. It has also been possible, for the first time, to elucidate an energy landscape by studying the reaction kinetics alone. The resulting landscape has then been matched to features predicted through experimental work and traditional molecular dynamics simulations. This indicates that the method can provide an alternative to the study of structural data when studying energy landscapes, and the key features of interest therein.

5.3 Least–Action Lattice Dynamics

Increased efficiency in computer simulations can come from a variety of sources. The Least Action method has been shown in the past to be a way of generating trajectories for proteins moving from one state to another. In this Thesis, the method has been enhanced to also take advantage of the benefits of coarse–grain modelling. The LaMP model dis-

cussed in Chapter 2 was adapted to generate large numbers of candidate structures from a starting conformation. A start and end conformation were also defined, forming the boundaries to the problem. For each of the generated candidate conformations, it was possible to use LaMP to calculate energies and therefore an action. The pathway with the lowest action value is the most likely of those sampled.

This model was first tested using a simple example. The aim of this was to ensure that the software did sample a larger number of structures when using a larger timestep. As expected, a small timestep resulted in structures with little deviation from the start. Using a larger timestep permitted the model to explore more conformations within the search space.

Following this validation, the method was applied to a much larger problem. The boundaries of this problem were formed by the start and end conformations of a forced unfolding trajectory for Titin I27. A set of ten structures within this were also selected, and the least action calculated for each individual structure and its neighbours. This was refined through several hundred repeats, and thus produced a trajectory deemed to be the one with the least action. Again, this is subject to the caveat that it is the least action pathway *of those sampled*, and the action is also dependant on the energy terms being used to assess the potential energy (in this case, the LaMP package). Nevertheless, the model produced a trajectory showing gradual unfolding of the molecule, and was able to produce such results in a much shorter timescale than would be required by traditional molecular dynamics simulations.

This work represented the novel application of the newly–developed LaMP package to a problem which has been studied many times before using existing off–lattice methods. The initial results are promising, and refinement of parameters such as the timestep – and indeed the energy functions within the LaMP model itself – may provide valuable

results to be considered alongside conventional atomistic simulations. The process of generating pathways using an iterative process, rather than by more traditional function minimisation or Monte Carlo for example, has also been shown to be a valid method for refinement of the action towards the most probably unfolding pathway.

References

[1] J. T. P. DeBrunner and E. Munck (Editors). *Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois.* University of Illinois Press, 1969.

[2] P. J. Thomas, B. H. Qu and P. L. Pedersen. Defective protein folding as a basis of human disease. *Trends Biochem Sci*, 20 (1995): 456–459.

[3] N. C. Inestrosa and C. Soto. Molecular biology of the amyloid of Alzheimer's disease. An overview. *Biol Res*, 25 (1992): 63–72.

[4] J. C. Scheinost, H. Wang, G. E. Boldt, J. Offer and P. Wentworth. Cholesterol seco-sterol-induced aggregation of methylated amyloid-beta peptides–insights into aldehyde-initiated fibrillization of amyloid-beta. *Angew Chem Int Ed Engl*, 47 (2008): 3919–3922.

[5] C. B. Anfinsen. The formation and stabilization of protein structure. *Biochem J*, 128 (1972): 737–749.

[6] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181 (1973): 223–230.

[7] C. A. Smith. How do proteins fold? *Biochemical Education*, 28 (2000): 76–79.

[8] C. Levinthal. Are there Pathways for Protein Folding? *Extrait du Journal de Chimie Physique*, 65 (1968): 44–5.

[9] K. A. Dill and H. S. Chan. From Levinthal to pathways to funnels. *Nat Struct Biol*, 4 (1997): 10–19.

[10] L. Mirny and E. Shakhnovich. Protein folding theory: from lattice to all-atom models. *Annu Rev Biophys Biomol Struct*, 30 (2001): 361–96.

[11] O. B. Ptitsyn. Protein folding: Hypotheses and experiments. *J. Prot. Chem.*, 6 (1987): 273.

[12] P. S. Kim and R. L. Baldwin. Intermediates in the folding reactions of small proteins. *Annu Rev Biochem*, 59 (1990): 631–660.

[13] K. A. Dill, K. M. Fiebig and H. S. Chan. Cooperativity in protein-folding kinetics. *Proc Natl Acad Sci U S A*, 90 (1993): 1942–1946.

[14] J. L. Bobadilla, M. Macek, J. P. Fine and P. M. Farrell. Cystic fibrosis: a worldwide analysis of CFTR mutations–correlation with incidence data and application to screening. *Hum Mutat*, 19 (2002): 575–606.

[15] B. Winchester, A. Vellodi and E. Young. The molecular basis of lysosomal storage diseases and their treatment. *Biochem Soc Trans*, 28 (2000): 150–154.

[16] J. A. Johnston, W. W. Liu, S. A. Todd, D. T. R. Coulson, S. Murphy, G. B. Irvine and A. P. Passmore. Expression and activity of beta-site amyloid precursor protein cleaving enzyme in Alzheimer's disease. *Biochem. Soc. Trans.*, 33 (2005): 1096–1100.

[17] M. Stefani and C. M. Dobson. Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution. *J Mol Med*, 81 (2003): 678–99.

[18] V. S. Pande, I. Baker, J. Chapman, S. P. Elmer, S. Khaliq, S. M. Larson, Y. M. Rhee, M. R. Shirts, C. D. Snow, E. J. Sorin and B. Zagrovic. Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. *Biopolymers*, 68 (2003): 91–109.

[19] M. Bucciantini, E. Giannoni, F. Chiti, F. Baroni, L. Formigli, J. Zurdo, N. Taddei, G. Ramponi, C. M. Dobson and M. Stefani. Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature*, 416 (2002): 507–11.

[20] S. B. Prusiner. Prions. *Proc Natl Acad Sci U S A*, 95 (1998): 13,363–13,383.

[21] E. Maşolepsza, M. Boniecki, A. Kolinski and L. Piela. Theoretical model of prion propagation: a misfolded protein induces misfolding. *Proc Natl Acad Sci U S A*, 102 (2005): 7835–7840.

[22] L. G. Goldfarb, R. B. Petersen, M. Tabaton, P. Brown, A. C. LeBlanc, P. Montagna, P. Cortelli, J. Julien, C. Vital and W. W. Pendelbury. Fatal familial insomnia and familial Creutzfeldt-Jakob disease: disease phenotype determined by a DNA polymorphism. *Science*, 258 (1992): 806–808.

[23] R. B. Petersen, M. Tabaton, L. Berg, B. Schrank, R. M. Torack, S. Leal, J. Julien, C. Vital, B. Deleplanque and W. W. Pendlebury. Analysis of the prion protein gene in thalamic dementia. *Neurology*, 42 (1992): 1859–1863.

[24] R. Gabizon, H. Rosenmann, Z. Meiner, I. Kahana, E. Kahana, Y. Shugart, J. Ott and S. B. Prusiner. Mutation and polymorphism of the prion protein gene in Libyan Jews with Creutzfeldt-Jakob disease (CJD). *Am J Hum Genet*, 53 (1993): 828–835.

[25] F. Chiti, P. Webster, N. Taddei, A. Clark, M. Stefani, G. Ramponi and C. M. Dobson. Designing conditions for in vitro formation of amyloid protofilaments and fibrils. *Proc Natl Acad Sci U S A*, 96 (1999): 3590–3594.

[26] M. Fändrich and C. M. Dobson. The behaviour of polyamino acids reveals an inverse side chain effect in amyloid structure formation. *EMBO J*, 21 (2002): 5682–5690.

[27] M. Kidd. Paired helical filaments in electron microscopy of Alzheimer's disease. *Nature*, 197 (1963): 192–193.

[28] R. D. Terry. The Fine Structure of Neurofibrillary Tangles in Alzheimer's Disease. *J Neuropathol Exp Neurol*, 22 (1963): 629–642.

[29] R. Tycko. Molecular structure of amyloid fibrils: insights from solid-state NMR. *Q Rev Biophys*, 39 (2006): 1–55.

[30] G. G. Glenner and C. W. Wong. Alzheimer's disease: initial report of the purification and characterization of a novel cerebrovascular amyloid protein. *Biochem Biophys Res Commun*, 120 (1984): 885–890.

[31] E. D. Eanes and G. G. Glenner. X-ray diffraction studies on amyloid filaments. *J Histochem Cytochem*, 16 (1968): 673–677.

[32] H. Inouye, P. E. Fraser and D. A. Kirschner. Structure of beta-crystallite assemblies formed by Alzheimer beta-amyloid protein analogues: analysis by x-ray diffraction. *Biophys J*, 64 (1993): 502–519.

[33] L. C. Serpell, C. C. Blake and P. E. Fraser. Molecular structure of a fibrillar Alzheimer's A beta fragment. *Biochemistry*, 39 (2000): 13,269–75.

[34] R. Tycko. Progress towards a molecular-level structural understanding of amyloid fibrils. *Curr Opin Struct Biol*, 14 (2004): 96–103.

[35] M. Sunde, L. C. Serpell, M. Bartlam, P. E. Fraser, M. B. Pepys and C. C. Blake. Common core structure of amyloid fibrils by synchrotron X-ray diffraction. *J Mol Biol*, 273 (1997): 729–739.

[36] P. E. Fraser, L. K. Duffy, M. B. O'Malley, J. Nguyen, H. Inouye and D. A. Kirschner. Morphology and antibody recognition of synthetic beta-amyloid peptides. *J Neurosci Res*, 28 (1991): 474–485.

[37] R. Tycko. Insights into the amyloid folding problem from solid-state NMR. *Biochemistry*, 42 (2003): 3151–9.

[38] R. G. Spencer, K. J. Halverson, M. Auger, A. E. McDermott, R. G. Griffin and P. T. Lansbury. An unusual peptide conformation may precipitate amyloid formation in Alzheimer's disease: application of solid-state NMR to the determination of protein secondary structure. *Biochemistry*, 30 (1991): 10,382–10,387.

[39] J. Lansbury, P. T., P. R. Costa, J. M. Griffiths, E. J. Simon, M. Auger, K. J. Halverson, D. A. Kocisko, Z. S. Hendsch, T. T. Ashburn, R. G. Spencer and et al. Structural model for the beta-amyloid fibril based on interstrand alignment of an antiparallel-sheet comprising a C-terminal peptide. *Nat Struct Biol*, 2 (1995): 990–8.

[40] T. L. Benzinger, D. M. Gregory, T. S. Burkoth, H. Miller-Auer, D. G. Lynn, R. E. Botto and S. C. Meredith. Propagating structure of Alzheimer's beta-amyloid(10-

35) is parallel beta-sheet with residues in exact register. *Proc Natl Acad Sci U S A*, 95 (1998): 13,407–13,412.

[41] H. Sticht, P. Bayer, D. Willbold, S. Dames, C. Hilbich, K. Beyreuther, R. W. Frank and P. Rösch. Structure of amyloid A4-(1-40)-peptide of Alzheimer's disease. *Eur J Biochem*, 233 (1995): 293–298.

[42] S. Zhang, K. Iwata, M. J. Lachenmann, J. W. Peng, S. Li, E. R. Stimson, Y. Lu, A. M. Felix, J. E. Maggio and J. P. Lee. The Alzheimer's peptide a beta adopts a collapsed coil structure in water. *J Struct Biol*, 130 (2000): 130–141.

[43] C. J. Barrow and M. G. Zagorski. Solution structures of beta peptide and its constituent fragments: relation to amyloid deposition. *Science*, 253 (1991): 179–182.

[44] S. Baglioni, F. Casamenti, M. Bucciantini, L. M. Luheshi, N. Taddei, F. Chiti, C. M. Dobson and M. Stefani. Prefibrillar amyloid aggregates could be generic toxins in higher organisms. *J Neurosci*, 26 (2006): 8160–8167.

[45] M. Balbirnie, R. Grothe and D. S. Eisenberg. An amyloid-forming peptide from the yeast prion Sup35 reveals a dehydrated beta-sheet structure for amyloid. *Proc Natl Acad Sci U S A*, 98 (2001): 2375–80.

[46] R. Diaz-Avalos, C. Long, E. Fontano, M. Balbirnie, R. Grothe, D. Eisenberg and D. L. Caspar. Cross-beta order and diversity in nanocrystals of an amyloid-forming peptide. *J Mol Biol*, 330 (2003): 1165–75.

[47] J. C. Chan, N. A. Oyler, W. M. Yau and R. Tycko. Parallel beta-sheets and polar zippers in amyloid fibrils formed by residues 10-39 of the yeast prion protein Ure2p. *Biochemistry*, 44 (2005): 10,669–80.

[48] F. Shewmaker, R. B. Wickner and R. Tycko. Amyloid of the prion domain of Sup35p has an in-register parallel beta-sheet structure. *Proc Natl Acad Sci U S A*, 103 (2006): 19,754–9.

[49] Z. Zhang, H. Chen, H. Bai and L. Lai. Molecular dynamics simulations on the oligomer formation process of the GNNQQNY peptide from yeast prion protein Sup35. *Biophys J*, (2007).

[50] B. J. Alder and T. E. Wainwright. Phase Transition for a Hard Sphere System. *J Chem Phys*, 27 (1957): 1208–1209.

[51] A. Rahman. Correlations in the Motion of Atoms in Liquid Argon. *Phys Rev*, 136 (1964): A405–A411.

[52] J. A. McCammon, B. R. Gelin and M. Karplus. Dynamics of folded proteins. *Nature*, 267 (1977): 585–590.

[53] P. van der Ploeg and H. J. C. Berendsen. Molecular dynamics simulation of a bilayer membrane. *J. Chem. Phys*, 76 (1982): 3271.

[54] W. F. van Gunsteren, H. J. Berendsen, R. G. Geurtsen and H. R. Zwinderman. A molecular dynamics computer simulation of an eight-base-pair DNA fragment in aqueous solution: comparison with experimental two-dimensional NMR data. *Ann N Y Acad Sci*, 482 (1986): 287–303.

[55] Y. Duan and P. A. Kollman. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*, 282 (1998): 740–4.

[56] G. Jayachandran, V. Vishal and V. S. Pande. Using massively parallel simulation and Markovian models to study protein folding: examining the dynamics of the villin headpiece. *J Chem Phys*, 124 (2006): 164,902.

[57] P. L. Freddolino, A. S. Arkhipov, S. B. Larson, A. McPherson and K. Schulten. Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure*, 14 (2006): 437–449.

[58] B. Brooks, R. Bruccoleri, D. Olafson, D. States, S. Swaminathan and M. Karplus. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J Comp Chem*, 4 (1983): 187–217.

[59] P. K. Weiner and P. A. Kollman. Assisted model building with energy refinement. A general program for modeling molecules and their interactions. *J Comput Chem*, 2 (1981): 287–303.

[60] M. Nelson, W. Humphrey, A. Gursoy, A. Dalke, L. Kalé, R. D. Skeel and K. Schulten. NAMD—A parallel, object-oriented molecular dynamics program. *International Journal of Supercomputer Applications and High Performance Computing.*, 10 (1996): 251–268.

[61] H. Berendsen, D. van der Spoel and R. van Drunen. GROMACS: A message-passing parallel molecular dynamics implementation. *Comp. Phys. Comm.*, 91 (1995): 43–56.

[62] J. E. Lennard-Jones. Cohesion. *Proc. Phys. Soc.*, 43 (1931): 461–482.

[63] T. Darden, D. York and L. Pedersen. Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. *J Chem Phys*, 98 (1993): 10,089.

[64] R. Hockney and J. Eastwood. *Computer Simulation Using Particles*. IOP, Bristol, 1988.

[65] W. Rocchia, E.Alexov and B.Honig. Extending the Applicability of the Nonlinear Poisson-Boltzmann Equation: Multiple Dielectric Constants and Multivalent Ions. *J Phys Chem B*, 105 (2001): 6507.

[66] C. W. Still, A. Tempczyk, R. C. Hawley and T. Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J Am Chem Soc*, 112 (1990): 6127–6129.

[67] T. Lazaridis and M. Karplus. Effective energy function for proteins in solution. *Proteins*, 35 (1999): 133–52.

[68] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J Chem Phys*, 79 (1983): 926–935.

[69] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren and J. Hermans. *Intermolecular Forces*. Reidel, Dordrecht, 1981.

[70] M. W. Mahoney and W. L. Jorgensen. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J Chem Phys*, 112 (2000): 8910–8922.

[71] H. Nada and J. P. J. M. van der Eerden. An intermolecular potential model for the simulation of ice and water near the melting point: A six-site model of H_2O. *J Chem Phys*, 118 (2003): 7401–7413.

[72] S. Nosé. A Unified Formulation of the Constant Temperature Molecular Dynamics Methods. *J. Chem. Phys.*, 81 (1984): 511–519.

[73] W. G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Phys Rev A*, 31 (1985): 1695–1697.

[74] H. C. Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *J Chem Phys*, 72 (1980): 2384–2393.

[75] D. E. Neves and R. A. Scott. Monte Carlo calculations on polypeptide chains. VIII. Distribution functions for the end-to-end distance and radius of gyration for hard-sphere models of randomly coiling poly(glycine) and poly(L-alanine). *Macromolecules*, 8 (1975): 267–271.

[76] S. Tanaka and H. A. Scheraga. Model of protein folding: inclusion of short-, medium-, and long-range interactions. *Proc Natl Acad Sci U S A*, 72 (1975): 3802–3806.

[77] G. M. Crippen. Topology of globular proteins. II. *J Theor Biol*, 51 (1975): 495–500.

[78] W. Kwak and U. H. E. Hansmann. Efficient sampling of protein structures by model hopping. *Phys Rev Lett*, 95 (2005): 138,102.

[79] J.-P. Ryckaert, C. G and H. J. Berendsen. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comp. Phys.*, 23 (1977): 327–341.

[80] W. B. Streett, D. J. Tildesley and G. Saville. Multiple time-step methods in molecular dynamics. *Mol. Phys.*, 35 (1978): 639–648.

[81] M. E. Tuckerman, B. J. Berne and G. J. Martyna. Molecular dynamics algorithm for multiple time scales: Systems with long range forces. *J Chem Phys*, 94 (1991): 6811–6815.

[82] W. F. van Gunsteren and H. J. C. Berendsen. Computer Simulation of Molecular Dynamics: Methodology, Applications, and Perspectives in Chemistry. *Angew. Chem. Int. Ed. Engl.*, 29 (1990): 992–1023.

[83] P. Ahlstrom, A. Wallqvist, S. Engstrom and B. Jonsson. A molecular dynamics study of polarizable water. *Mol. Phys.*, 68 (20 October 1989): 563–581.

[84] J. W. Halley, J. F. Rustad and A. Rahman. A polarizable, dissociating molecular dynamics model for liquid water. *J Chem Phys*, 5 (1993): 4110–4119.

[85] E. Harder, J. D. Eaves, A. Tokmakoff and B. J. Berne. Polarizable molecules in the vibrational spectroscopy of water. *Proc Natl Acad Sci U S A*, 102 (2005): 11,611–11,616.

[86] W. Xie, J. Pu, A. D. Mackerell and J. Gao. Development of a polarizable intermolecular potential function (PIPF) for liquid amides and alkanes. *J Chem Theory Comput*, 3 (2007): 1878–1889.

[87] O. Khoruzhii, A. G. Donchev, N. Galkin, A. Illarionov, M. Olevanov, V. Ozrin, C. Queen and V. Tarasov. Application of a polarizable force field to calculations of relative protein : ligand binding affinities. *Proc Natl Acad Sci U S A*, 105 (2008): 10,378–10,383.

[88] M. Souaille, H. Loirat, D. Borgis and M. Gaigeot. MDVRY: a polarizable classical molecular dynamics package for biomolecules. *Computer Physics Communications*, 180 (2009): 276 – 301.

[89] S. Patel and C. L. Brooks. CHARMM fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations. *J Comput Chem*, 25 (2004): 1–15.

[90] S. Patel, A. D. Mackerell and C. L. Brooks. CHARMM fluctuating charge force field for proteins: II protein/solvent properties from molecular dynamics simulations using a nonadditive electrostatic model. *J Comput Chem*, 25 (2004): 1504–1514.

[91] S. Patel and C. L. Brooks. Fluctuating charge force fields: recent developments and applications from small molecules to macromolecular biological systems. *Mol. Sim.*, 32 (2006): 231–249.

[92] ChemTips.com. Internet, 2008.

[93] N. Go. Theoretical Studies of Protein Folding. *Annu Rev Biophys Bioeng*, 12 (1983): 183–210.

[94] P. G. Wolynes. Recent successes of the energy landscape theory of protein folding and function. *Q Rev Biophys*, 38 (2005): 405–410.

[95] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas and H. S. Chan. Principles of protein folding–a perspective from simple exact models. *Protein Sci*, 4 (1995): 561–602.

[96] H. Li, R. Helling, C. Tang and N. Wingreen. Emergence of Preferred Structures in a Simple Model of Protein Folding. *Science*, 273 (1996): 666–669.

[97] A. Dinner, A. Sali, M. Karplus and E. Shakhnovich. Phase diagram of a model protein derived by exhaustive enumeration of the conformations. *J Chem Phys*, 101 (1994): 1444–1451.

[98] S. Miyazawa and R. L. Jernigan. Estimation of effective interresidue contact energies from protein crystal structures. *Macromolecules*, 18 (1985): 534–552.

[99] S. Miyazawa and R. L. Jernigan. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol*, 256 (1996): 623–44.

[100] J. Skolnick and A. Kolinski. Simulations of the Folding of a Globular Protein. *Science*, 250 (1990): 1121–1125.

[101] M. Levitt and A. Warshel. Computer simulation of protein folding. *Nature*, 253 (1975): 694–698.

[102] E.-H. Yap, N. L. Fawzi and T. Head-Gordon. A coarse-grained alpha-carbon protein model with anisotropic hydrogen-bonding. *Proteins*, 70 (2007): 626–638.

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

[104] B. P. Blackburne and J. D. Hirst. Three-dimensional functional model proteins: Structure function and evolution. *J Chem Phys*, 119 (2003): 3453–3460.

[105] S. Decatur and S. Batzoglou. Protein folding in the Hydrophobic-Polar model on the 3D triangular lattice, 1996.

[106] G. M. S. de Mori, G. Colombo and C. Micheletti. Study of the Villin headpiece folding dynamics by combining coarse-grained Monte Carlo evolution and all-atom molecular dynamics. *Proteins*, 58 (2005): 459–471.

[107] M. Neri, C. Anselmi, M. Cascella, A. Maritan and P. Carloni. Coarse-grained model of proteins incorporating atomistic detail of the active site. *Phys Rev Lett*, 95 (2005): 218,102.

[108] E. Lyman, F. M. Ytreberg and D. M. Zuckerman. Resolution exchange simulation. *Phys Rev Lett*, 96 (2006): 028,105.

[109] G. Binnig, C. F. Quate and C. Gerber. Atomic force microscope. *Phys Rev Lett*, 56 (1986): 930–933.

[110] G. Binnig and H. Rohrer. Scanning tunneling microscopy—from birth to adolescence. *Rev. Mod. Phys.*, 59 (1987): 615–625.

[111] T. E. Fisher, A. F. Oberhauser, M. Carrion-Vazquez, P. E. Marszalek and J. M. Fernandez. The study of protein mechanics with the atomic force microscope. *Trends Biochem Sci*, 24 (1999): 379–384.

[112] M. Radmacher, M. Fritz, H. G. Hansma and P. K. Hansma. Direct observation of enzyme activity with the atomic force microscope. *Science*, 265 (1994): 1577–1579.

[113] E. L. Florin, V. T. Moy and H. E. Gaub. Adhesion forces between individual ligand-receptor pairs. *Science*, 264 (1994): 415–417.

[114] S. Allen, J. Davies, A. C. Dawkes, M. C. Davies, J. C. Edwards, M. C. Parker, C. J. Roberts, J. Sefton, S. J. Tendler and P. M. Williams. In situ observation of streptavidin-biotin binding on an immunoassay well surface using an atomic force microscope. *FEBS Lett*, 390 (1996): 161–164.

[115] G. U. Lee, L. A. Chrisey and R. J. Colton. Direct measurement of the forces between complementary strands of DNA. *Science*, 266 (1994): 771–773.

[116] H. G. Hansma, K. J. Kim, D. E. Laney, R. A. Garcia, M. Argaman, M. J. Allen and S. M. Parsons. Properties of biomolecules measured from atomic force microscope images: a review. *J Struct Biol*, 119 (1997): 99–108.

[117] A. Soteriou, A. Clarke, S. Martin and J. Trinick. Titin folding energy and elasticity. *Proc Biol Sci*, 254 (1993): 83–86.

[118] H. P. Erickson. Reversible unfolding of fibronectin type III and immunoglobulin domains provides the structural basis for stretch and elasticity of titin and fibronectin. *Proc Natl Acad Sci U S A*, 91 (1994): 10,114–10,118.

[119] M. Rief, M. Gautel, F. Oesterhelt, J. M. Fernandez and H. E. Gaub. Reversible unfolding of individual titin immunoglobulin domains by AFM. *Science*, 276 (1997): 1109–12.

[120] P. F. Lenne, A. J. Raae, S. M. Altmann, M. Saraste and J. K. Hörber. States and transitions during forced unfolding of a single spectrin repeat. *FEBS Lett*, 476 (2000): 124–128.

[121] A. F. Oberhauser, P. E. Marszalek, H. P. Erickson and J. M. Fernandez. The molecular elasticity of the extracellular matrix protein tenascin. *Nature*, 393 (1998): 181–185.

[122] A. Imparato, S. Luccioli and A. Torcini. Reconstructing the free-energy landscape of a mechanically unfolded model protein. *Phys Rev Lett*, 99 (2007): 168,101.

[123] A. Li and V. Daggett. Characterization of the transition state of protein unfolding by use of molecular dynamics: chymotrypsin inhibitor 2. *Proc Natl Acad Sci U S A*, 91 (1994): 10,430–10,434.

[124] R. Day, B. J. Bennion, S. Ham and V. Daggett. Increasing temperature accelerates protein unfolding without changing the pathway of unfolding. *J Mol Biol*, 322 (2002): 189–203.

[125] E. Paci and M. Karplus. Unfolding proteins by external forces and temperature: the importance of topology and energetics. *Proc Natl Acad Sci U S A*, 97 (2000): 6521–6526.

[126] H. Lu and K. Schulten. The key event in force-induced unfolding of Titin's immunoglobulin domains. *Biophys J*, 79 (2000): 51–65.

[127] H. Lu, A. Krammer, B. Isralewitz, V. Vogel and K. Schulten. Computer modeling of force-induced titin domain unfolding. *Adv Exp Med Biol*, 481 (2000): 143–60; discussion 161–2.

[128] M. Gao, H. Lu and K. Schulten. Unfolding of titin domains studied by molecular dynamics simulations. *J Muscle Res Cell Motil*, 23 (2002): 513–21.

[129] R. Kaijmierkiewicz, A. Liwo and H. A. Scheraga. Energy-based reconstruction of a protein backbone from its alpha-carbon trace by a Monte-Carlo method. *J Comput Chem*, 23 (2002): 715–723.

[130] Y. Iwata, A. Kasuya and S. Miyamoto. An efficient method for reconstructing protein backbones from alpha-carbon coordinates. *J Mol Graph Model*, 21 (2002): 119–128.

[131] J. D. Honeycutt and D. Thirumalai. Metastability of the folded states of globular proteins. *Proceedings of the National Academy of Sciences*, 87 (1990): 3526–3529.

[132] D. J. Wales and P. E. J. Dewsbury. Effect of salt bridges on the energy landscape of a model protein. *The Journal of Chemical Physics*, 121 (2004): 10,284–10,290.

[133] S.-Y. Kim. An off-lattice frustrated model protein with a six-stranded beta-barrel structure. *The Journal of Chemical Physics*, 133 (2010): 135102.

[134] T. Hoque, M. Chetty and A. Sattar. Extended HP model for protein structure prediction. *J Comput Biol*, 16 (2009): 85–103.

[135] J. Song, J. Cheng, T. Zheng and J. mao. A Novel Genetic Algorithm for HP Model Protein Folding. In *Proceedings of the Sixth International Conference on Parallel and Distributed Computing Applications and Technologies*, PDCAT '05, 935–937. IEEE Computer Society, Washington, DC, USA, 2005.

[136] A. S. M.T. Hoque, M. Chetty. Protein folding prediction in 3D FCC HP lattice model using genetic algorithm. In *IEEE Congress on Evolutionary Computation, 2007.* 2007.

[137] J.-M. Shin and W. S. Oh. Study of Move Set in Cubic Lattice Model for Protein Folding. *J Phys Chem*, 102 (1998): 6405 – 6412.

[138] L. Toma and S. Toma. Folding simulation of protein models on the structure-based cubo-octahedral lattice with the Contact Interactions algorithm. *Protein Sci*, 8 (1999): 196–202.

[139] P. Pokarowski, A. Kolinski and J. Skolnick. A minimal physically realistic protein-like lattice model: designing an energy landscape that ensures all-or-none folding to a unique native state. *Biophys J*, 84 (2003): 1518–26.

[140] Z. Sun, X. Xia, Q. Guo and D. Xu. Protein Structure Prediction in a 210-Type Lattice Model: Parameter Optimization in the Genetic Algorithm Using Orthogonal Array. *Journal of Protein Chemistry*, 18 (1999): 39–46. 10.1023/A:1020643331894.

[141] Y. Ponty, R. Istrate, E. Porcelli and P. Clote. LocalMove: computing on-lattice fits for biopolymers. *Nucleic Acids Res*, 36 (2008): W216–W222.

[142] B. A. Patel, P. G. Debenedetti, F. H. Stillinger and P. J. Rossky. A Water-Explicit Lattice Model of Heat-, Cold-, and Pressure-Induced Protein Unfolding. *Biophys J*, (2007).

[143] R. B. Pandey and B. L. Farmer. Conformation of a coarse-grained protein chain (an aspartic acid protease) model in effective solvent by a bond-fluctuating Monte Carlo simulation. *Phys Rev E Stat Nonlin Soft Matter Phys*, 77 (2008): 031,902.

[144] P. Pokarowski, K. Droste and A. Kolinski. A minimal proteinlike lattice model: an alpha-helix motif. *J Chem Phys*, 122 (2005): 214,915.

[145] Y. Chen, Q. Zhang and J. Ding. A coarse-grained model for the formation of alpha helix with a noninteger period on simple cubic lattices. *J Chem Phys*, 124 (2006): 184,903.

[146] J. Krawczyk, A. L. Owczarek, T. Prellberg and A. Rechnitzer. Lattice model for parallel and orthogonal beta sheets using hydrogenlike bonding. *Phys Rev E Stat Nonlin Soft Matter Phys*, 76 (2007): 051,904.

[147] A. A. Albrecht, A. Skaliotis and K. Steinhöfel. Stochastic protein folding simulation in the three-dimensional HP-model. *Comput Biol Chem*, 32 (2008): 248–255.

[148] J. D. Westbrook and P. M. D. Fitzgerald. The PDB format, mmCIF, and other data formats. *Methods Biochem Anal*, 44 (2003): 161–179.

[149] N. Metropolis and S. Ulam. The Monte Carlo method. *J Am Stat Assoc*, 44 (1949): 335–341.

[150] A. Sali, E. Shakhnovich and M. Karplus. How does a protein fold? *Nature*, 369 (1994): 248–251.

[151] L. Zhang, D. Lu and Z. Liu. How native proteins aggregate in solution: a dynamic Monte Carlo simulation. *Biophys Chem*, 133 (2008): 71–80.

[152] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, E. Teller and A. H. Teller. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.*, 21 (1953): 1087.

[153] S. Kirkpatrick, C. D. Gelatt and M. P. Vecchi. Optimization by Simulated Annealing. *Science*, 220 (1983): 671–680.

[154] R. H. Swendsen and J.-S. Wang. Replica Monte Carlo simulation of spin glasses. *Phys Rev Lett*, 57 (1986): 2607–2609.

[155] C. L. Pierri, A. D. Grassi and A. Turi. Lattices for ab initio protein structure prediction. *Proteins*, 73 (2008): 351–361.

[156] C. Thachuk, A. Shmygelska and H. Hoos. A replica exchange Monte Carlo algorithm for protein folding in the HP model. *BMC Bioinformatics*, 8 (2007): 342.

[157] Z. Bagci, R. L. Jernigan and I. Bahar. Residue coordination in proteins conforms to the closest packing of spheres. *Polymer*, 43 (2002): 451–459.

[158] Z. Bagci, R. L. Jernigan and I. Bahar. Residue packing in proteins: Uniform distribution on a coarse-grained scale. *J Chem Phys*, 116 (2002): 2269–2276.

[159] P. G. de Gennes. Reptation of a Polymer Chain the Presence of Fixed Obstacles. *J Chem Phys*, 55 (1971): 572.

[160] M. T. Oakley, J. M. Garibaldi and J. D. Hirst. Lattice models of peptide aggregation: evaluation of conformational search algorithms. *J Comput Chem*, 26 (2005): 1638–46.

[161] K. Leonhard, J. M. Prausnitz and C. J. Radke. 3D-Lattice Monte Carlo simulations of model proteins. Size effects on folding thermodynamics and kinetics. *Biophysical Chemistry*, 106 (2003): 81 – 89.

[162] D. De Sancho and A. Rey. Evaluation of coarse grained models for hydrogen bonds in proteins. *J Comput Chem*, 28 (2007): 1187–99.

[163] D. K. Klimov, M. R. Betancourt and D. Thirumalai. Virtual atom representation of hydrogen bonds in minimal off-lattice models of alpha helices: effect on stability, cooperativity and kinetics. *Fold Des*, 3 (1998): 481–496.

[164] Y. Fujitsuka, S. Takada, Z. A. Luthey-Schulten and P. G. Wolynes. Optimizing physical energy functions for protein folding. *Proteins*, 54 (2004): 88–103.

[165] H. Lu, B. Isralewitz, A. Krammer, V. Vogel and K. Schulten. Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation. *Biophys J*, 75 (1998): 662–71.

[166] P. E. Marszalek, H. Lu, H. Li, M. Carrion-Vazquez, A. F. Oberhauser, K. Schulten and J. M. Fernandez. Mechanical unfolding intermediates in titin modules. *Nature*, 402 (1999): 100–3.

[167] R. D. Toofanny and P. M. Williams. Simulations of multi-directional forced unfolding of titin I27. *J Mol Graph Model*, 24 (2006): 396–403.

[168] L. Tskhovrebova, J. Trinick, J. A. Sleep and R. M. Simmons. Elasticity and unfolding of single molecules of the giant muscle protein titin. *Nature*, 387 (1997): 308–12.

[169] L. Tskhovrebova and J. Trinick. Muscle disease: a giant feels the strain. *Nat Med*, 11 (2005): 478–9.

[170] S. Lange, F. Xiang, A. Yakovenko, A. Vihola, P. Hackman, E. Rostkova, J. Kristensen, B. Brandmeier, G. Franzen, B. Hedberg, L. G. Gunnarsson, S. M. Hughes, S. Marchand, T. Sejersen, I. Richard, L. Edström, E. Ehler, B. Udd and M. Gautel. The kinase domain of titin controls muscle gene expression and protein turnover. *Science*, 308 (2005): 1599–1603.

[171] S. Labeit, B. Kolmerer and W. A. Linke. The giant protein titin. Emerging roles in physiology and pathophysiology. *Circ Res*, 80 (1997): 290–294.

[172] H. L. Granzier and S. Labeit. The giant protein titin: a major player in myocardial mechanics, signaling, and disease. *Circ Res*, 94 (2004): 284–95.

[173] E. Lübke, A. Freiburg, G. O. Skeie, B. Kolmerer, S. Labeit, J. A. Aarli, N. E. Gilhus, R. Wollmann, M. Wussling, J. C. Rüegg and W. A. Linke. Striational autoantibodies in myasthenia gravis patients recognize I-band titin epitopes. *J Neuroimmunol*, 81 (1998): 98–108.

[174] K. Wang, R. McCarter, J. Wright, J. Beverly and R. Ramirez-Mitchell. Viscoelasticity of the sarcomere matrix of skeletal muscles. The titin-myosin composite filament is a dual-stage molecular spring. *Biophys J*, 64 (1993): 1161–1177.

[175] J. M. Squire. Architecture and function in the muscle sarcomere. *Curr Opin Struct Biol*, 7 (1997): 247–57.

[176] W. A. Linke, M. R. Stockmeier, M. Ivemeyer, H. Hosser and P. Mundel. Characterizing titin's I-band Ig domain region as an entropic spring. *J Cell Sci*, 111 (Pt 11) (1998): 1567–74.

[177] A. S. Politou, M. Gautel, M. Pfuhl, S. Labeit and A. Pastore. Immunoglobulin-type domains of titin: same fold, different stability? *Biochemistry*, 33 (1994): 4730–7.

[178] S. B. Fowler, R. B. Best, J. L. Toca Herrera, T. J. Rutherford, A. Steward, E. Paci, M. Karplus and J. Clarke. 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*, 322 (2002): 841–9.

[179] A. K. Faradjian and R. Elber. Computing time scales from reaction coordinates by milestoning. *J Chem Phys*, 120 (2004): 10,880–10,889.

[180] A. M. A. West, R. Elber and D. Shalloway. Extending molecular dynamics time scales with milestoning: example of complex kinetics in a solvated peptide. *J Chem Phys*, 126 (2007): 145,104.

[181] E. Vanden-Eijnden, M. Venturoli, G. Ciccotti and R. Elber. On the assumptions underlying milestoning. *J Chem Phys*, 129 (2008): 174,102.

[182] R. D. Toofanny. *Investigations into the Mechanical Unfolding of Proteins in silico.* Ph.D. thesis, University of Nottingham, 2005.

[183] R. B. Best, S. B. Fowler, J. L. Herrera, A. Steward, E. Paci and J. Clarke. Mechanical unfolding of a titin Ig domain: structure of transition state revealed by combining atomic force microscopy, protein engineering and molecular dynamics simulations. *J Mol Biol*, 330 (2003): 867–77.

[184] A. E. Cárdenas and R. Elber. Atomically detailed simulations of helix formation with the stochastic difference equation. *Biophys J*, 85 (2003): 2919–2939.

[185] A. E. Cárdenas and R. Elber. Kinetics of cytochrome C folding: atomically detailed simulations. *Proteins*, 51 (2003): 245–257.

[186] A. Ghosh, R. Elber and H. A. Scheraga. An atomically detailed study of the folding pathways of protein A with the stochastic difference equation. *Proc Natl Acad Sci U S A*, 99 (2002): 10,394–10,398.

[187] K. T. Atsushi Uchida, Chieko Totsuji and H. Totsuji. Stochastic-Difference-Equation Method for Long Time-scale Molecular Dynamics Simulations. *Mem. Fac. Eng. Oka. Uni.*, 40 (2006): 36–39.

[188] P. M. Williams, S. B. Fowler, R. B. Best, J. L. Toca-Herrera, K. A. Scott, A. Steward and J. Clarke. Hidden complexity in the mechanical properties of titin. *Nature*, 422 (2003): 446–9.

[189] R. Elber. A milestoning study of the kinetics of an allosteric transition: atomically detailed simulations of deoxy Scapharca hemoglobin. *Biophys J*, 92 (2007): L85–L87.

[190] R. Merkel, P. Nassoy, A. Leung, K. Ritchie and E. Evans. Energy landscapes of receptor-ligand bonds explored with dynamic force spectroscopy. *Nature*, 397 (1999): 50–53.

[191] D. Passerone and M. Parrinello. Action-derived molecular dynamics in the study of rare events. *Phys Rev Lett*, 87 (2001): 108,302.

[192] R. Elber and M. Karplus. Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin. *Science*, 235 (1987): 318–321.

[193] R. Olender and R. Elber. Calculation of classical trajectories with a very large time step: Formalism and numerical examples. *J. Chem. Phys.*, 105 (1996): 9299–9315.

[194] I.-H. Lee, J. Lee and S. Lee. Kinetic energy control in action-derived molecular dynamics simulations. *Physical Review*, 68 (2003): 064,303.

[195] R. Crehuet and M. J. Field. Comment on "Action-derived molecular dynamics in the study of rare events". *Phys Rev Lett*, 90 (2003): 089,801; author reply 089,802.