MONTE CARLO STUDIES OF THE MECHANICAL PROPERTIES OF BIOPOLYMERS

by

Sara Sadeghi

M.Sc., Brock University, 2007

B.Sc., Sharif University of Technology, 2004

Thesis Submitted in Partial Fulfilment

of the Requirements for the Degree of

Doctor of Philosophy

in the

Department of Physics

Faculty of Science

SIMON FRASER UNIVERSITY

Fall 2011

APPROVAL

Name: **Sara Sadeghi**

Degree: **Doctor of Philosophy**

Title of Thesis: **Monte Carlo studies of the mechanical properties of biopolymers**

Examining Committee:

Chair: Dr. Andrei Frolov, Assistant Professor (Chair)

_____

Dr. Eldon Emberly
Senior Supervisor
Canada Research Chair Tier II, Associate Professor, Physics Department

_____

Dr. Martin Zuckermann
Supervisor
Adjunct Professor, Physics Department

_____

Dr. David Boal
Supervisor
Professor Emeritus, Physics Department

_____

Dr. Nancy Forde
Supervisor
Associate Professor, Physics Department

_____

Dr. Jenifer Thewalt
Internal Examiner
Professor, Physics Department

_____

Dr. Apichart Linhananta
External Examiner
Associate Professor, Lakehead University

Date Defended/Approved: October 14, 2011

Declaration of
Partial Copyright Licence

Abstract

Biopolymers are one of the main components of living systems. Their sequence dictates their structure that ultimately determines their function. Many factors play key mechanical roles in the cell and one of the most abundant biopolymers that is involved in such tasks is the class of coiled-coil proteins. Various theoretical and experimental studies have been done to explore the mechanical properties of these proteins and there are now a number of single molecule measurements that measure their force response characteristics, making coiled-coils an excellent model system to test folding models connecting sequence to structure to function. In this thesis we have developed a coarse-grained atomistic model to study coiled-coil formation and explore both mechanical and thermal properties. Our model is able to reproduce known coiled-coil structures using only a simple hydrophobic-polar (HP) representation of their sequence and is able to explain the observed mechanical response measured in single molecule experiments. To address how common coiled-coil formation is with respect to all possible helix packs, we have evaluated the designability of the space of possible helical folds, defined as the number of sequences that can fold into a particular structure. We find that left-handed coils emerge as one of the most highly designable structures. From the designability calculation we can identify sequence patterns that design particular coiled-coil folds and mutations that lead to their instability. We also predict that designable coiled-coil structures are more mechanically stable than less designable helical packs.

Keywords: Monte Carlo; coiled-coils; α-helices; transition force; transition temperature; designability

Acknowledgments

I am indebted to many for helping me during my PhD studies and development of this thesis.

My greatest gratitude goes to my senior supervisor, Dr . Eldon Emberly, whose patience and knowledge has tremendously helped me during my PhD studies. I greatly appreciate his thoughtful advices and fruitful interactions with me and I would always remember my PhD studies under his supervision as one of my best memories. I would also like to thank Dr. Martin Zuckermann, my co-supervisor. I have extremely benefited from his guidance and lively discussions.

I would like to thank Dr. David Boal and Dr. Nancy Forde, my committee members, for their invaluable ideas in the past few years, as well as supportive Physics Department staff, Dagni Lund and Rose Evans.

The friendly and energetic people of the Simon Fraser university has provided me with a wonderful opportunity to build my graduate life. Special thanks go to Benjamin Downing, Naghmeh Rezai, Marjan Shayegan, Laleh Samii, Peter Smith, Bitan Roy, Michelle Lee, Saeed Saberi, Simin Bagheri, Fatemeh Rostam Zadeh, Saheena Dsai, Zhang Tan, Suvayu Ali, Sara Taghipour, Michelle Lee, Alireza Hojjati and Azadeh Akhtari.

I would also like to thank Jafar Taghiyar for his valuable support and help during the past few years. I really acknowledge what he did specially he did more than 90% of all the graphics of this thesis except for the scientific results.

Finally, my warmest feelings and deepest thanks are for my parents and my only brother who have wholeheartedly supported me during my graduate studies. I would like to dedicate my thesis to them as the humblest gratitude.

Contents

List of Tables

List of Figures

Chapter 1

Introduction and Background

1.1 What Are Proteins?

Proteins are polymers built from amino acids which can fold into very specific three dimensional structures. The specific sequence of amino acids, the primary structure of a protein, determines the three dimensional configuration and the specific function (including mechanical function) of the protein in living organisms. Many proteins have mechanical functions that require being able to withstand or exert force within the cell. For example, kinesin, a molecular motor, is a protein that generates force and is involved in microtubule-based motility [1]. Some proteins are part of connective tissues like collagen and keratin that form fibrillar structures; collagen is the key component of extracellular matrix and keratin is the main structural material for the outer layer of human skin. Determining how the mechanical properties of a protein depend on its structure which in turn depends on sequence is an ongoing area of research, both experimentally and theoretically. By probing the mechanical properties of a protein, it is possible to learn about how the protein folded into its structure. Understanding how sequence leads to structure is one of the outstanding problems in biophysics [2]. With respect to mechanical properties, one of the aims of making progress on this problem is the hope of sequence engineering specific properties for protein based bio-materials [3, 4].

The primary structure of a protein is a polymer of amino acids. Amino acids contain an amino group and a carboxylic acid group plus a side-chain group (R) that is connected to the central carbon atom (See Fig. 1.1(a)). Amino acids are linked together by a peptide

bond, that leads to the repeating backbone of the protein that is illustrated in Fig. 1.1(b). The side-chain group is different from one amino acid to another and it dictates the physical size and chemical properties of each amino acid. Twenty different amino acids act as the major building blocks of proteins. The main classification of amino acids is according to their affinity to a polar solvent such as water. Roughly half of the amino acids have an affinity for polar solvent and are called polar (P) or hydrophilic amino acids, while the remaining are called hydrophobic (H) or non-polar amino acids. It is the hydrophobic effect that drives compaction in the folding process of the polymer with the hydrophobic side-chains in the sequence residing in the core of the final protein structure.

When there is no hydrophobic residue in water, the water molecules are able to form hydrogen bonds. A hydrophobic residues or region is not able to make these bonds which cause disruption of the network of hydrogen bond. To compensate for this effect the water molecules are making a cage around the hydrophobic region. To minimize the surface area of the hydrophobic region the hydrophobic residues aggregate to exposed less surface towards the polar water molecule. In another word the hydrophobic effect is due to the changes in entropy rather than attraction between the residues.

Moving beyond the primary sequence, the backbone of the protein can form secondary structure: such as α helices and β sheets (See Fig. 1.1(c)); β sheets are made of β strands that can have parallel or anti-parallel conformations. These structures are formed by hydrogen bonds between atoms that are part of the backbone of the protein structure. Certain amino acids have a propensity for forming specific secondary structures [6, 7]. This is predominantly an entropic effect. For instance, short peptides that contain only alanine and lysice, or alanine and glutamate are forming stable monomeric helices in the aqueous solutions [8, 9, 10]. On the other hand, lysine, which has a larger side chain, prefers to exist as part of β sheets rather than helices, since there is less entropic cost [7].

As mentioned above, secondary structures form due to the formation of hydrogen bonding networks between backbone atoms. α helices have right-handed spiral conformations where 3.6 residues are in each turn and each residue gives a 1.5Å rise to the helix. The radius of an α helix is 2.3Å. There are hydrogen bonds between each amino acids and its fourth neighbour residues in the primary strand of the protein which helps to keep the spiral shape of α helix. On the other hand β sheets contain strands of amino acids connected by hydrogen bonds. β sheets can come in different topologies depending on how the strands

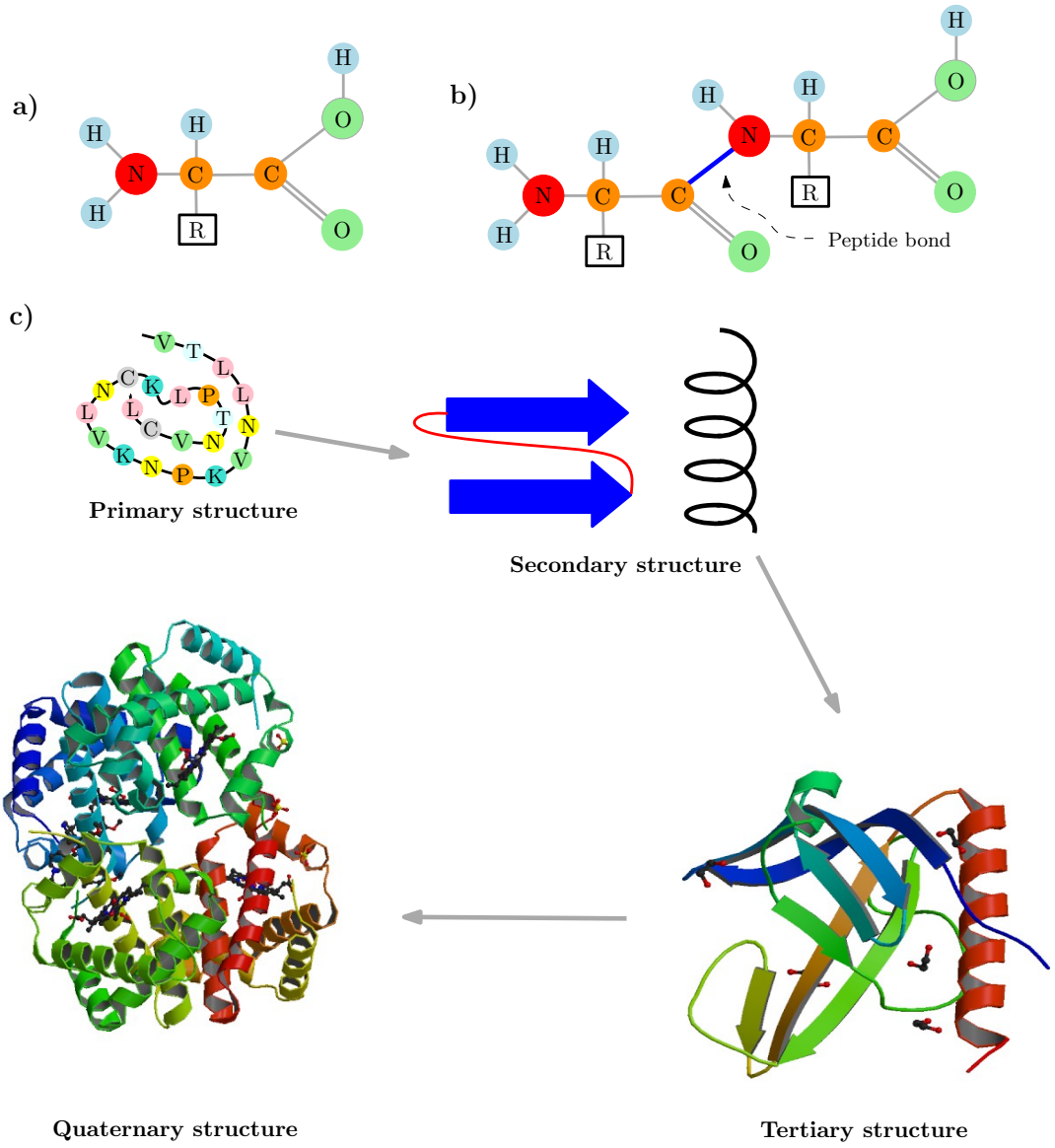

Figure 1.1: (a) Amino acid structure. (b) Peptide bond between two amino acids. (c) Protein structure from amino acid peptide to 3D conformation or quaternary structure [The tertiary and the quaternary structures are from [5]].

are connected and oriented in the sheet. The hydrogen bonding network of both alpha helices and beta sheets leads to relatively rigid structures, that nevertheless can deform, and these deformations can be crucial for the formation of the final protein structure.

As we said the amino acids in the primary sequence are playing the main role in protein folding by their hydrophobicities. For example when hydrophobic and polar amino acids are changing alternatively (HP), the β sheet structure is dictated; or a repeat of seven amino acids with hydrophobic amino acids at the first and the fourth position (HPPHPPP) will fold in to a coiled-coil conformation, which is the subject of this study. Specifically for coiled-coil structures, which are the main focus of this study, other factors such as salt-bridges or buried polar groups are important too [11].

The actual 3D structure of the proteins is called tertiary structure and is generated from a specific packing together of its secondary structural elements. For a given protein, the final fold consists of a unique packing of its secondary structural elements predominantly driven by the hydrophobic force. Thus the primary sequence dictates the type of secondary structure that forms that in turn leads to the unique tertiary structure that is formed. Understanding how sequence leads to the final structure is the protein folding problem. When more than one poly peptide present in a protein structure, it is called the quaternary structure. A protein like hemoglobin consists of the packing together of several repeated protein tertiary structures (See Fig. 1.1(c)).

1.2 Studying Mechanical Properties of the Proteins

Studying the mechanical response of a protein to an applied force provides insight into how the structure formed. Different experimental methods and theoretical models have been developed to study the mechanical properties of proteins. These studies have addressed two main questions: can we find the folding pathway of the proteins and how does sequence mutation affect the mechanical properties of a protein?

There have been experimental studies on understanding the mechanical properties of proteins using single molecule techniques [12, 13, 14, 15, 16, 17, 18, 19]. Single molecule approaches help us to understand the behaviour of an individual protein in the complex dynamics of a biological process [14] which may vary from the ensemble averaged behaviour of a collection of the molecules. There are a number of single molecule measurement tech-

niques exist for applying force: optical tweezers, magnetic tweezers, nanopores and the atomic force microscope (AFM) [20, 21]. Each of these experiments is used for a different range of required force, or for a different type of force. For example it is possible to apply longitudinal force by optical tweezers while magnetic tweezers are also used for applying torque (rotation). Using these methods it has been shown that unfolding of α-helices mechanically protects the myosin molecules from dissociation at the physiological relevant forces [16]. Furthermore it has been shown that the network of hydrogen bonds that makes up secondary structural elements has a strong influence on the protein's mechanical response [12].

The main experimental method which is used in studying the mechanical properties and unfolding of proteins is Atomic Force Microscopy (AFM). The same protein which is mentioned above, titin, was studied using AFM methods [22]. In this experiment the required force to unfold a single molecule of titin is reported to be in a range from 150 to 300 pico Newtons; the magnitude of the force depends on the pulling speed. Also it is indicated that immunoglobulin domain of titin refolds upon relaxation [22] as discussed above [23].

Other than AFM, one of the other main experimental methods that is employed in studying mechanical properties of proteins is optical tweezers. In one of the studies, the mechanical properties of a single molecule of the giant muscle protein titin have been explored using optical tweezers [23]. Titin, which is also called connectin, is responsible for the elasticity and scaffold of muscles. In this study, force was applied on two different time scales: fast and slow. On a fast time scale titin is elastic and force-extension data can be fitted with a standard random-coil polymer model. On the slow time scale, the molecule displays stress-relaxation which occurs in rapid steps. The authors proposed by the same study that the stress-relaxation derives from unfolding of immunoglobulin and fibronectin domains [23].

Using all these different experimental methods the protein folding pathway is explored by studying the unfolding pathway of a protein. Also comparing the mechanical properties of a protein with its native and the mutated sequences helps us to understand the differences between the mechanical properties of that protein and the factors that are the key points for those functions [24].

Besides the experimental studies the mechanical properties of these structures have

been studied using different theoretical methods [25, 26, 27, 28, 29, 30, 31]. The theoretical protein folding literature is vast. Molecular dynamics, that involves solving the equations of motion for a given potential, is one of the main approaches that has been used in exploring the folding pathways of proteins [26, 28, 29, 32]. Using this method, some of the properties of collagen were explained [29]. Collagen is a protein material that contains collagen fibrils which are composed of a staggered array of ultra-long tropocollagen. This study explained why collagen fibrils consist of tropocollagen molecules of length 300 nm and it has also reported that collagen disease changes the intermolecular adhesion properties that influence the mechanical properties [29]. It has been reported in another study that the interior part of a folded globular protein is more fluid-like and atoms follow a diffusional motion [32]. Molecular dynamics has been able to reveal pathways to folding and binding, but it still faces the challenge of not being able to fold large proteins or reach typical folding timescales of tens of seconds.

Instead of complex all-atom models, lattice models have also been used to study general mechanisms of the folding problem [33, 34, 35]. In a simple lattice model the amino acids are restricted to lattice sites and they are labelled either as hydrophobic or polar residues. When a protein goes through the folding process it tries to minimize sum of the contact energy between its own amino acids and the interaction energy between the amino acids and the solvent or exterior environment; thus the hydrophobic amino acids are buried in the middle of the protein and the polar amino acids are facing towards the solvent. It has been shown that using the lattice model only a small set of structures emerges [35]. In one study using this model, the protein finds a global minimum on the potential surface through a fast approach by collapsing into a local semi-compact globule and then it goes through searching the transition state to find the native state. The second step of the folding process is slow, hence it is the rate-determining step of the whole protein folding process [34]. Using the same model, it has been shown that the best-folding proteins are reported as those which have their native conformation as a pronounced energy minimum [36].

One of the main constraints of computational modelling of protein folding is the time of each simulation. Monte Carlo simulation and coarse-grained models are some other alternatives that are used to address this problem. One method that reduces the computational time is G$\bar{o}$ model, which combines all the forces experienced by a residue into a single potential [30, 37, 38, 39]. Also the G$\bar{o}$ model helps to find kinetic pathways of folding, since

the protein is forced to adopt its correct structure with the energy function. For example it has been reported that the folding of protein G is initiated by two nucleation contacts, one hydrophobic and one hydrophilic as shown in previous experimental studies [39]. Besides, it has been shown in another study that in the fast-folding pathway, partial α-helical regions form before hydrophobic core collapse while in the slow-folding pathway partial core collapse happens before helical formation [30].

One of the main barriers in computational protein folding studies, specifically molecular dynamic, is the duration of each simulation. In nature protein folding occurs in seconds to minutes, however the computer can simulate upto microseconds. As we mentioned earlier Monte Carlo (MC) simulation is another method to reduce the computational time of a protein folding process. In MC the different conformations of a protein are sampled statistically and the ones which have lower energy are selected [40, 41, 42]. Over all, in MC methods, minimizing hydrophobic energy is one of the key points to fold the proteins. A different sequence of amino acids results in a different secondary structure.

One of the main obstacles in the protein folding problem is the degeneracy of the protein landscape. This is one of the main challenges in this field. Different pathways such as simulated annealing methods are used to find the overall native state other than local minima but this is an ongoing challenge in this field.

1.3 Designability

The inverse problem to protein folding is protein design. Namely, given a specific tertiary structure, is it possible to find an amino acid sequence that will fold into it? Because structure space is highly constrained in the design problem, it makes this problem potentially more tractable than the folding problem. In fact, computational design of sequence followed by experiments has shown that design is indeed possible. The pioneering work from Steve Mayo's laboratory showed that it was indeed possible to take a specific backbone structure (a zinc-finger structure) and completely redesign the sequence [43]. This was then followed by work that showed that novel structures could be designed too, including a right-handed coiled-coil [44] and a unique beta-sheet topology not seen in nature [45]. Different computer programs, such as ELVES or SOCKET, have been developed to determine the structure of a whole protein or a specific domain of a protein during recent years

as well [46, 47]. Another database which specifically focused on coiled-coil structures is CC+ [48]. This database provides a periodic table of coiled-coil protein structure plus a dynamic interface that delivers higher complexity and details [48, 49]. Despite progress in this area, is it possible to design any structure that we want? This would allow in principle the engineering of bio-materials possessing a wide spectrum of mechanical properties.

An examination of known protein structures reveals that the number of folds, consisting of a specific packing of secondary elements, is much smaller than the actual number of known protein structures [11, 50]. To date a few hundred thousand protein sequences are known and this is just small fraction of possible protein sequences (an average protein contains 300 amino acids, so in principle there are 20^{300} possible sequences). The number of different known folds that these sequences fold into is even smaller, only a few thousand [51]. Thus there is a many-to-one mapping of sequence to structure. Different proteins with different biological functions can possess similar folds. In fact the number of possible folds for proteins is about a few hundred [52, 53]. Why does this occur and why are not all structures created equal? What is so specific about these sequences and structures?

Answers to some of these questions have been provided by some theoretical work that has introduced the idea of *designability*. Designability is defined as the number of different sequences that have the same ground state structure. A number of theoretical studies have shown [54, 55], that there are some structures with high designability and yet many structures in the structural space that are not the ground state of any sequence at all. This theory is illustrated in Fig. 1.2. The designability principle states that only a small subset of possible structures can be achieved by folding. Some of the studies focused on the energy of the most designable structures and why those are preferred over all the other folds [56]. Some studies insist on designability as the dominant factor in protein folding and the relation between the designability of the proteins and disease [57], while some studies think that designabilitiy may not be one of many factors in protein folding [58].

Thus the designability principle gives some insight into why only a few folds might emerge out of a large population of possible sequences. More importantly, structures that are highly designable have properties that are consistent with naturally occurring protein folds. Namely they have been shown to be thermally and mutationally stable; they are also able to find their tertiary structures faster than structures with low designability [55, 35].

However, how do mechanical properties relate to a structure's designability? Finding a realistic system to address this is one of the goals of this thesis.

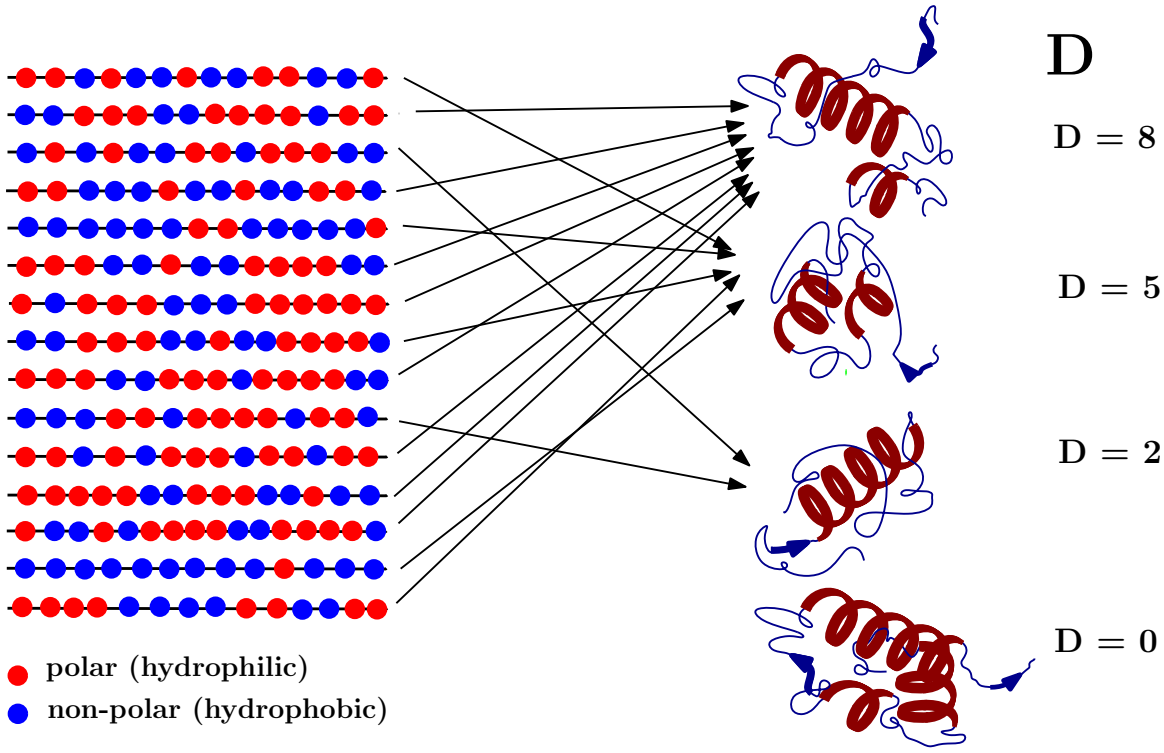

Figure 1.2: Some random sequences of polar (red) and hydrophobic (blue) residues are assigned to their ground state structure. Structures with high designability score (D) represent a small part of possible structural space.

1.4 Coiled-coils

One of the most ubiquitous folds is a coiled-coil which consists of two or more α helices that wrap around each other to form a superhelical structure. They have several biological functions, from mediating protein-protein interactions to forming mechanically important structural elements, such as the necks of molecular motors (e.g., kinesin and myosin). Because they are often subject to mechanical stresses inside cells, understanding how they fold

and respond to forces is of physical and biological interest. Thus they make an excellent model protein fold to study the bigger questions outlined above.

For a given coiled-coil, the superhelical structure and its stability are dictated by the underlying amino acid sequence of the helices forming the coil. The superhelix forms due to the packing of repeating hydrophobic residues that wind around each alpha helix. Naturally occurring coiled-coils form left-handed superhelical structures caused by a repeating seven amino acid motif called a heptad repeat, that has hydrophobic residues at specific positions in the 7mer leading to a hydrophobic strip that winds in a left-handed fashion around each helix. However, by varying the sequence, to an 11mer repeat with hydrophobic residues at first, fourth and the eight positions, right-handed coiled-coil structures are also possible too [44]. Thus coiled-coils provide a model system for studying the relationship between protein sequence and structure.

The folding kinetics of coiled-coils have been studied both in bulk using fluorescence techniques [59, 60, 61] and at the single-molecule level through the application of force to a single coiled-coil structure [62, 63, 64, 65]. These studies have elucidated the folding trajectory of coiled-coils and how stability depends on sequence. Particularly in one of the papers [64], it was shown the coiled-coil of myosin can unfold at relatively large forces, 20 to 25 pN, and extends 2.5 times of its original length in a reversible process; obviously the transition force is rate dependent. In another study [63], a high resolution map of the gradual unzipping of coiled-coils of differing lengths has been reported. In this experiment, a single coiled-coil domain (leucine zipper) was subject to a loading force using an atomic force microscope (AFM) that pulled apart the coiled-coil from one end, while the other end remained linked. They were able to see transitions corresponding to the sequential pulling apart of the superhelical turns. By fitting a multistate model to their data they were able to extract key folding parameters such as the energy to form a turn and nucleate the coiled-coil, and how these depend upon the sequence of the underlying heptad repeat.

Beside studying the mechanical properties of coiled-coils under an applied load, exploring the thermal stability of coiled-coils and its dependence on the original sequence is an area of active research. Mutations can stabilize or destabilize the coiled-coil structure and alter mechanical and thermodynamic properties of the structure. It has been reported that vimentin, which supports and anchors the organelles in the cytosol [66], is crystallized if there is one mutation at a specific location along the sequence from polar (P) to hydropho-

bic (H) residue [67]. Also in another study changing polar residue to a hydrophobic residue resulted in more robust structure of a coiled-coil protein with GCN4 sequence [63]. GCN4 stands for general control non-derepressible and it is a basic leucine zipper protein which is one of the regulatory proteins in yeast.One of the aims of this thesis will be to develop a model that can explain these observed mutational effects in coiled-coils.

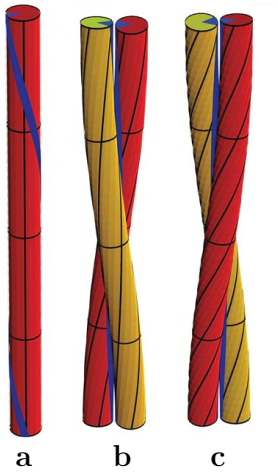

Figure 1.3: (a) α helix with heptad repeat sequence has a left-handed hydrophobic strip wrapping around the helix. (b) Two helices with heptad repeat in a left-handed coiled-coil structure. (c) Right-handed coiled-coil made with heptad repeat sequence. The energy cost to deform the helices is higher than for part (b).[Reproduced with permission by the authors of [68]]

Several theoretical methods for predicting coiled-coil structure from sequence have been provided ranging from detailed molecular dynamics simulations [69] to coarse-grained Monte Carlo approaches [62, 70]. Recent theoretical work, using a continuum model for the helices, has shown how the superhelical structure arises as a competition between the energy gained from packing the hydrophobic residues and the energetic cost of deforming the helices [68]. In that work they showed that the structure which minimizes the total energy is a superhelix whose twist is the same as the underlying twist of the hydrophobic strip of each helix. In the absence of any energetic cost of deformation, they predict that it should be possible for a variety of different structures to form all satisfying the criterion to pack the hydrophobic residues together as shown in Fig. 1.3. The main conclusion is that

incorporating deformation energy in an appropriate way is of utmost importance in any model for coiled-coil formation.

1.5 Thesis Outline

To complement the above experimental and theoretical work, we have developed a coarse-grained atomistic model to study different properties of coiled-coils. We have in particular studied the mechanical and the thermal stability properties of coiled-coils. In Chapter 2, we describe our coarse-grained model for coiled-coils. It is based on the packing together of flexible helical elements that are governed by a hydrophobic energy as well as a deformation energy associated with distorting the helical structure. We present our method for finding structures that minimize the energy, and allow for the inclusion of an applied force when looking for the ground state conformation. Lastly we also present methods that we use to connect our findings to experimental data, such as structural alignment and the calculation of thermodynamic quantities.

In Chapter 3 we show that our model is able to regenerate simulated structures close to the canonical left-handed coiled-coil, namely a leucine zipper (based on the yeast protein GCN4). We show results for mechanical response simulations that correspond well with experimental studies for leucine zippers of varying length. We also detail results that show that the model is capable of predicting the observed effects due to mutating the GCN4 sequence [63]. Furthermore we have shown that our model is able to make right-handed structures according to the sequence provided. Experimental studies on coiled-coils to date have focussed on applying forces perpendicular to the superhelical axis, and at the end of this Chapter we present results for pulling on coiled-coils using a parallel force.

Having been able to reproduce the mechanical properties of a single coiled-coil forming sequence, namely the leucine zipper, in Chapter 4, we extend our study to different coiled-coil structures allowing for different numbers of α helices and helices with various lengths. In it, we show that our model is able to reproduce natural coiled-coil structures ranging from 58 to 200 residues per structure. Also we have studied the mutation effects on thermal stability. In particular we have shown that mutated sequence of vimentin is more stable compared to the structure with the wild-type sequence. Besides, we have found that mutations at specific positions along the coiled-coils have bigger effects on destabilizing

the structure.

In Chapter 5 we have focused on the designability problem. We make structure space with 30,000 packs of double helices with specific length and we find the most desirable packs among all of them (more detail in Chapter 5). We show for the first time using an atomistic model for proteins, that high designable structures are more mechanically stable than low designable ones.

After discussions about the strengths and weaknesses of our model in Chapter 6, we talk about the future possible studies that we can pursue using our model.

Chapter 2

Methods

To complement the experimental and theoretical work mentioned in Chapter 1, we have developed a coarse-grained atomistic model to study the formation of coiled-coils and their mechanical properties under an applied load. Previous work explored the packing of rigid helices to study helix-bundle formation [71] and now we allow for the helices to have flexible degrees of freedom.

To model coiled-coil formation, we consider a simple energy function which consists of three contributions: a contribution from the hydrophobic energy gained by packing together the hydrophobic residues, an energetic cost for deforming each α helix, and finally the work done by applied forces to the coiled-coil. In particular it is computationally efficient, using the collective motions (normal modes) of each helix [31] to explore the space of possible structures, analogous to the backbone parameterization used in [72]. For a range of parameters, this simple model is able to produce coiled-coils as free-energy ground-state structures for sequences that have either left-handed or right-handed chirality. We now detail the model for coiled-coil formation and the methods we have used in analyzing the results throughout this thesis.

2.1 Energy Function

Coiled-coils consist of two or more α helices that wrap around each other to form a superhelical structure. To represent the structure of each helix we use the C_α atoms of the helical backbone. Each undistorted α helix has 3.6 residues per turn, with a rise of 1.5

Å per residue and a helix radius, r_h = 2.3 Å. The helices making up the coiled-coil interact via the energy function that we now detail.

There are 20 different amino acids and they interact with each other to drive the folding process. By analyzing favoured contacts between residues in protein structure Miyazawa and Jernigan derived a semi-empirical interaction matrix to describe the energy of interaction between amino acids. This matrix is the result of statistical analysis on the residue-residue contacts in the crystal structures of the proteins [73]. The inter-residue contact energy between each pair of 20 amino acids is summarized in the Miyazawa-Jernigan (MJ) matrix, which is a 20×20 matrix [73]. This matrix has been widely used in protein folding and design [74, 75]. By using the eigenvalue method decomposition, it is possible to decompose the MJ matrix into a form that is equivalent to a solvation model where the interactions between two residues only depend on their individual hydrophobicities. It has been shown that first two principal eigenvectors are the dominant ones [76]. Also the effective hydrophobicities found in the decomposition correlate well with experimentally determined hydrophobicity scales. For most of this thesis we simplify even further by considering amino acids as to be either hydrophobic (H) or polar (P) and use an interaction model where the energy only depends on the hydrophobicity of each amino acid.

Instead of considering all of the different interactions that exist between the various amino acids, we assume that the residues are either hydrophobic or polar since the superhelical structure is driven to form primarily by the hydrophobic energy. Besides the hydrophobic energy, there is an energetic cost associated with deforming each helix. In our model, these are the two dominant energetic terms which govern coiled-coil formation. In addition to these two energies, we consider the application of applied forces that do work on the coiled-coil. Putting these contributions together, we arrive at the total energy, E, of a coiled-coil,

$$E = E_H + W_D E_D - \sum_i \vec{F}_i \cdot \vec{R}_i \tag{2.1}$$

where E_H is the energy from hydrophobic contacts between the residues of helices in the structure, E_D is the associated energy due to deformations of the bonds of each helix, described by the spring model below, W_D is the associated weight of the deformation energy to the overall energy, $\vec{F}_i$ is an externally applied force (e.g., from an optical trap or an AFM), and $\vec{R}_i$ is the displacement between the first residues of the two helices when an external

force is applied perpendicular to the coiled-coil axis, or is the end-to-end displacement of the ith helix in the case of applying a parallel load.

The deformation and hydrophobic energies are not absolute but both of them are expressed in terms of k_BT unit. The weight of deformation allows us to find a balance between them and treat both of them in the same way.

2.1.1 Hydrophobic Energy

The hydrophobic energy, E_H, we use the Lennard-Jones 6-12 potential with three different contact energies which is given by:

$$E_H = \sum_{i\neq j}^{N_H} \sum_{k,l} \varepsilon(t_k^i, t_l^j) \left[\left(\frac{r_0}{r_{k,l}^{i,j}} \right)^{12} - 2 \left(\frac{r_0}{r_{k,l}^{i,j}} \right)^{6} \right] \tag{2.2}$$

where the sum is over all pairs of residues between the N_H helices, with t_k^i being the type of the jth residue of the ith helix, and $r_{k,l}^{i,j}$ is the distance between the kth and lth residues of helices i and j. The optimal distance between two residues of different helices is r_0 where the energy has its lowest possible value which is $-\varepsilon(t_k^i, t_l^j)$. Most of the results of this work are calculated with constant radius over all amino acids; but we have considered the cases of having different contact radius for different pairs of amino acids as well. The results have a slight improvement compared to the case of having a unique r_0 for all the amino acids. The energy of a given interaction is determined by $\varepsilon(t_k^i, t_l^j)$, where $t = H$ or $t = P$, and where we take $\varepsilon(H,H) > \varepsilon(H,P) > \varepsilon(P,P)$ with $\varepsilon(H,H) > 2\varepsilon(H,P)$ to favor separation of H and P.

2.1.2 Deformation energy

The helices in a coiled-coil are distorted from their ideal straight helical configuration and an energetic cost must be assigned to these distortions. We use a ball and spring model to represent the deformation energy of each helix. The parameters of the model (i.e., spring constants) were determined by fitting the resulting normal modes to the principal deformations of α-helices in the protein data bank [71]. The model contains four springs: nearest neighbors, next-nearest neighbors and next-next-nearest neighbors, and one for hydrogen

bonding between the i and $i+4$ residues. A schematic model is shown in Fig. 2.1; each set of springs with the same color corresponds to one type of the springs mentioned above. The energy E_D is given by

$$E_D = \sum_i \sum_j \sum_{l=1}^{4} \frac{1}{2} k_l (|r^i_{j,j+l}| - |r_l|)^2 \tag{2.3}$$

where $\vec{r}^{\,i}_{j,j+l}$ is the vector between the j and $j+l$ residues of the ith helix. The parameters for the model are: $k_1 = 100\,k_BT\,\text{Å}^{-2}$, $k_2 = 20\,k_BT\,\text{Å}^{-2}$, $k_3 = 20\,k_BT\,\text{Å}^{-2}$, and $k_4 = 7\,k_BT\,\text{Å}^{-2}$ and the spring rest lengths are $r_1 = 3.8$Å, $r_2 = 5.4$ Å, $r_3 = 5.0$ Å, and $r_4 = 6.2$ Å.

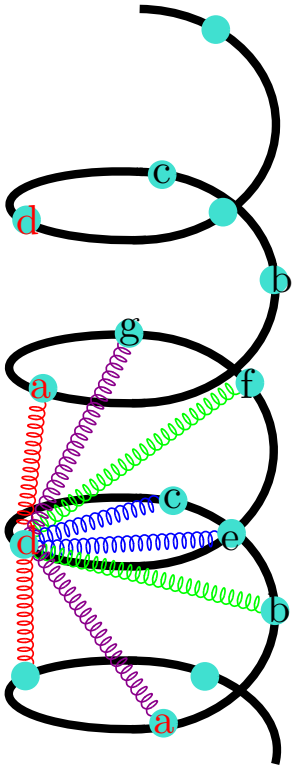

Figure 2.1: Each amino acid is connected to its neighbouring residues by four different types of springs. Springs are connecting each residue to its neighbours, next-nearest neighbours and next-next-nearest neighbours; the fourth spring represents the hydrogen bonds of an α-helix; each type of spring is shown by one color.

In the next section we describe how we find the structure that minimizes the total energy,

Eq. 2.1, using Metropolis Monte Carlo and a move set centered around the normal modes of each α helix.

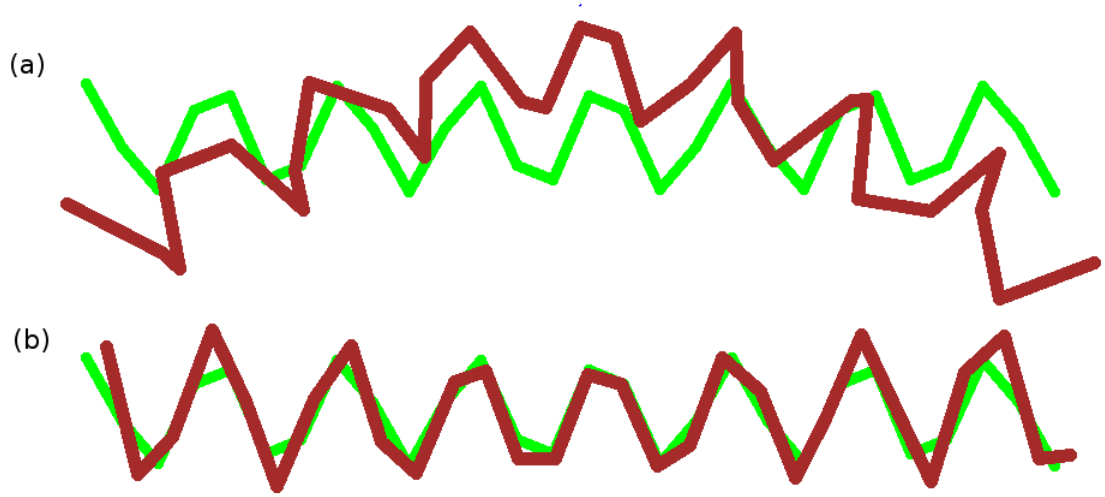

Figure 2.2: The deformations associated with the (a) bend and (b) twist normal modes of an α helix calculated from the spring model given in Eq. 2.3.

2.2 Generating Coiled-coil Structures: Using a Normal Mode Move Set

To study the formation and unfolding properties of coiled-coils using the energy function outlined in Sec. 2.1 we use Monte Carlo (MC) simulations. Molecular dynamics has been applied to study the force response of coiled-coils [77] but we have chosen to use MC, an approach that has been applied to study other unfolding and protein-protein interaction problems [78, 79, 80, 81]. The big disadvantage of the MC method compare to Molecular Dynamic (MD) method is choosing local minima over the global minimum in the energy landscape. In any MC simulation of protein structure, an appropriate move set must be chosen such that the space of possible structures can be fully sampled in a computationally reasonable amount of time [81]. Since moving each C_α atom individually would lead

to inordinately long folding times, we have chosen to use the normal modes and the six translational and rotational degrees of freedom of each helix as a move set [82].

Each mode has a corresponding eigenvector, $\delta\vec{x}_i$ that gives the unit displacement from the undistorted helix [see, for example, the bend and twist modes in Fig. 2.3]. These are the modes of the deformation model given in the previous section. Thus the coordinates of a deformed helix can be written as

$$\vec{x} = R(\phi,\theta,\psi)\cdot\left(\vec{x}_0 + \sum_i a_i\delta\vec{x}_i\right) + \vec{T} \tag{2.4}$$

where $\vec{x}$ contains the $3N$ coordinates of a N-residue helix, $\vec{x}_0$ are the coordinates of an undistorted helix, a_i is the amplitude of the ith normal mode, $R(\phi,\theta,\psi)$ is the rotation matrix corresponding to the rotation by the Euler angles, ϕ,θ and ψ, and $\vec{T}$ is the 3D translation of the helix. The $\{a_i\},\{\phi,\theta,\psi\}$, and $\vec{T}$ form the move set used in the Monte Carlo simulation, with corresponding step sizes δa_i. We adjust δa_i continually throughout the simulation to obtain a 50 % acceptance rate for each move. Instead of using all the normal modes, only a fraction of the entire set are used: the low-energy collective motions of the residues making up the helix. Dependence of the results on this mode cutoff will be discussed in the following chapters.

Starting from a non-interacting state [see Fig. 2.3] we obtain the lowest energy structures of Eq. 2.2 using simulated annealing. The simulation starts with the N_H helices randomly oriented in space such that they do not intersect and that their centers of mass reside within a sphere of radius ≈ 10Å. We use the following temperature schedule $T_{i+1} = T_i/\alpha$. We have found that by using $1.75 < T_0 < 2.5$ and $1.1 < \alpha < 1.2$ with eight or more temperature steps, we are able to consistently produce coiled-coil structures. It should be mentioned that the $T = 1.0$ is not the room temperature and the scale between the room temperature and the temperature in this simulation should be investigated in more details. We use on the order of 500 000 Monte Carlo steps at each temperature. For long helices $L > 70$ residues, we include a random kick that is implemented by adding a kick probability p_{kick} and a kick amplitude, a_{kick}. At a given MC step, if a kick is successful, then a randomly selected mode's $\delta a_i = a_{kick}$. Using a $p_{kick} = 0.1$ and $a_{kick} = 10$ is adequate to get the longer coiled-coils out of local energy minima.

In the first part of the folding the initial interactions between the residues of the two α-helices are formed. Usually the two helices are partly zipped and the other two ends are

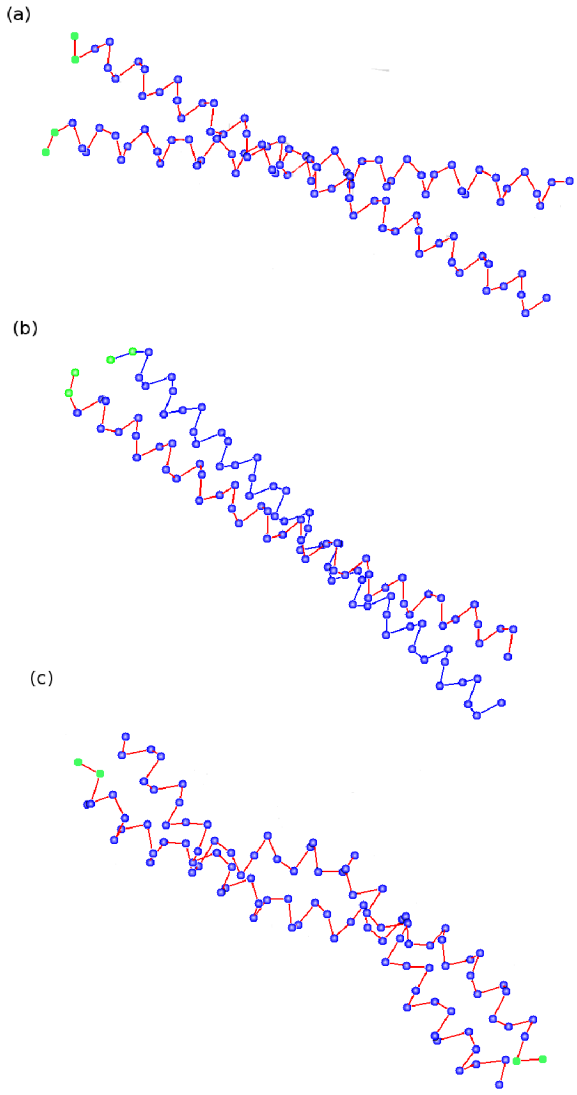

Figure 2.3: Structure of coiled-coils formed from two 50mer α helices. (a) Initial random configuration of two undeformed α helices. (b) Final left-handed coiled-coil structure using the heptad repeat sequence, HPPHPPP. (c) Final right-handed coiled-coil structure using a repeating HHPP sequence showing more than one full wind of the superhelix.

flexible. By minimizing the energy the free ends wrap around each other and the coiled-coil zips up.

2.3 Translating Sequence into an HP Pattern

For a given amino acid sequence of a coiled-coil, we convert it to a hydrophobic-polar sequence (HP) using the following translation: hydrophobic amino acids = CYS, ILE, PRO, PHE, GLY, LEU, TRP, ALA, VAL, MET and polar amino acids = ASP, SER, GLN, LYS, THR, ASN, HIS, ARG, GLU, TYR [76]. The sequences and the corresponding HP translations for the coiled-coil structures studied in this thesis can be found in the Appendix.

2.4 Aligning Protein Structures

Throughout this thesis it will be necessary to compare how similar two structures are. A common way to do so is minimizing the root-mean-squared distance (RMSD) between the backbone coordinates of the two structures [83]. Briefly, the method calculates the necessary rotation and shift that minimize the following metric,

$$r^{\mu,\nu}_{RMSD} = \sqrt{\frac{1}{N}\sum_i (\vec{C^{\mu}_{\alpha,i}} - \vec{C^{\nu}_{\alpha,i}})^2} \tag{2.5}$$

between the C_α coordinates of structures μ and ν, where both structures have N backbone atoms. Since the sequences for each helix are identical, and the simulated structure is generated starting from random initial positions for each helix, we consider the possible permutations of labelling the helices, and compute the RMSD for each permutation. We take the lowest RMSD as the best matching fit of the simulated structure to the experimentally determined structure.

2.5 Applying Force to Coiled-coils

To study mechanical properties of coiled-coils we apply force in two different ways: (i) transverse force and (ii) longitudinal force.

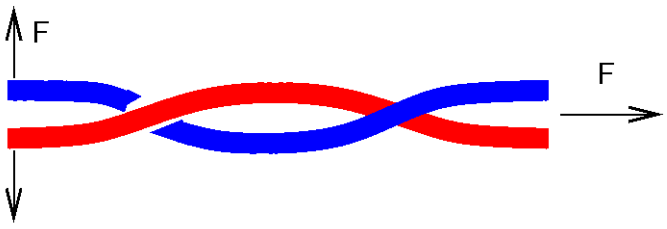

Figure 2.4: Schematic of a coiled-coil consisting of two α helices that wrap around each other to form a superhelical structure. We consider cases where the force is applied perpendicular or along the long axis of the coiled-coil.

2.5.1 Transverse Force

We apply transverse force, a force perpendicular to the long axis of the superhelix, in two different ways. One is applying force that leads to a *constant velocity* of separation and the other is applying *constant force*. In both cases there is a spring between the third last residues of each helix holding the end of the coiled-coil together.

Constant Velocity

For the constant velocity simulation, the separation between the first two residues is increased gradually in steps of $\Delta\vec{d}$, and at each separation both helices are equilibrated by MC. After equilibration we calculate the average energy of the structure and the average force from

$$|\bar{F}_{i+1}| = \frac{\langle E_{i+1}\rangle - \langle E_i\rangle}{|\Delta\vec{d}|} \tag{2.6}$$

where $\langle\bar{E}_i\rangle$ is the average energy at step i. The results presented in the following chapters are for a choice of $\Delta d = 1.0$ Å. Changing $\Delta\vec{d}$ at each step by the same amount is equivalent to pulling the helices at constant velocities. At each step the structure was relaxed for $\approx 2\times 10000\times N_{mode}$ MC steps. This works out to be $\approx 500\,000$ steps for the lengths of coils studied (i.e. less than 100 residues per each helix). We find that our unzipping results become insensitive to the number of MC steps used.

Force Increasing at a Constant Rate

For the second method of applying transverse force, that is increased at a constant rate, increments of $\Delta\vec{F}$ were added to the force gradually. After each increment the helices of the coiled-coil are equilibrated through MC. The energy of the applied transverse force is calculated by

$$E_F = -\sum_i \vec{F}_i \cdot \vec{R}_i \tag{2.7}$$

which should be added to the hydrophobic and deformation energy [See Eq. 2.1]. As explained earlier, $\vec{F}_i$ is the applied force and $\vec{R}_i$ is the displacement between the first residues of the two helices in the coiled-coil. The simulation starts at $\vec{F} = 0$ and continues until coiled-coil unzipped and all hydrophobic contacts between the residues of the two helices are broken. As in the other parts of this study for each step structures are relaxed for $\approx 2 \times 10000 \times N_{mode}$ MC steps.

By increasing the force gradually, we would like to find the minimum force required to unzip the coiled-coil and break all the hydrophobic contacts between two helices, which we call the transition force. At the transition force, the unwrapping of the coiled-coil starts at the ends of the α-helices. This process is reversible if some of the residues from both helices still be connected to each other (i.e. the coiled-coil is partly unwrapped). We calculate the thermodynamic susceptibility of the hydrophobic energy, defined as:

$$\chi_H = \frac{d\langle E_H \rangle}{|\Delta\vec{F}|} \tag{2.8}$$

where E_H is hydrophobic energy after helices are equilibrated at each step and $|\Delta\vec{F}|$ is the magnitude of each increment of the applied force. The transition force corresponds to the force where χ_H has its maximum value. There is only a single transition force: once this force is met, the entire coil unzips since all the forces required to unzip the rest of the interactions between the two helices are less than the first transition force [See the following chapters for the results].

2.5.2 Longitudinal Force

In many biological situations, such as molecular motors whose necks consist of a coiled-coil, the load is often exerted parallel to the axis of the coiled-coil. The force is applied

according to Fig. 2.4. We consider the possibility of applying a force to just one or both of the helices. We use the Monte Carlo approach of unzipping the coiled-coil structures as described for transverse force, except now the force is applied parallel to the coiled-coil main axis. Under tension, the helices try to keep their hydrophobic contacts, but when the force is large enough their hydrophobic contacts break. As mentioned above this force is called the transition force.

2.6 Heat Capacity of Simulated Structures

We characterize the transition from the folded coiled-coil state to the unfolded state by examining the specific heat capacity, defined by $C_v = \left(\frac{\langle E(T)^2\rangle - \langle E(T)\rangle^2}{k_B T}\right)_v$. We use the average total energy at a given temperature, $\langle E(T)\rangle$ to define this quantity. At this temperature the hydrophobic energy, E_H, rapidly changes from low energy to zero as all the contacts break. We define the unfolding temperature as the temperature at which this transition occurs, or correspondingly as the largest change in the heat capacity versus heat plot.

The unfolding simulations are carried out by starting with the structure in a folded configuration and then gradually increasing the temperature in a steps of ΔT, allowing the structure to equilibrate at each temperature. For the results presented we used $\Delta T = 0.1$. Different ΔT results in different transition temperatures. The simulation pathway for unfolding the coiled-coils are happening through a non-equilibrium process so it is possible for a coiled-coil to unfold if it is kept for an infinite time at any temperature. Obviously the probability of unfolding a coiled-coil increases with raising the temperature.

According to Arrehenius equation the rate of any reaction is $\sim exp(-E_a/k_B T)$; where E_a is the activation energy of a reaction, k_B is Boltzmann constant and T is the temperature. We can assume that the folded coiled-coil is the initial state and the unwrapped coiled-coil is the final state of the unfolding process either by an external force or increments in the temperature. Unfolding rate increases at higher temperatures, because $E_a/k_B T$ decreases and $exp(-E_a/k_B T)$ gets bigger. Similarly for unzipping a coiled-coil by an external force, the force should be equal or bigger than the transition force. This provides enough activation energy to change the energy state of the structure.

2.7 Designability

In the part of the thesis we focus on the designability problem. In some of previous studies on the designability problem, different methods were applied such as, HP-Lattice model and MJ-Lattice model [35]. In the former one there are just three different contact energies between the residues in the protein as explained earlier (i.e. HH, HP and PP) and in the latter one the MJ-matrix has been applied for contact energies. Besides, the off-lattice models also applied to study the designability of the proteins [84, 35].

In this thesis, we used off-lattice model using HP contact energies to study coiled-coils. To address the designability problem, we generate the space of all different folds of double helices packed with random sequences. Then, to find the designability score of each pack in the structure space we find the number of sequences which choose one specific structure as their best structure in terms of energy. Then, structures with similar backbone are clustered together and we studied the properties of the clusters. We compare mechanical and thermal stability of high designable and low designable clusters. More details along with the results are presented in Chapter 5.

Chapter 3

Length-dependent Force Characteristics of Coiled-coils

In this chapter we present results for the mechanical unfolding properties of coiled-coils generated using the model outlined in Chapter 2. To study the mechanical unfolding properties as a function of length we used a fixed set of parameters that can produce coiled-coil structures. Recent experimental work has looked at the length dependence of coiled-coil unfolding [63]. We apply a force perpendicular to the long axis of left-handed coiled-coils to reproduce the unzipping results of single-molecule experimental work [63, 62, 65]. Our results are able to capture the essential features seen experimentally, including the effects of sequence mutation. The experimental results are shown in Fig. 3.1 for three different length of the helices [63]. The mean pulling velocities are 460, 510 and 690 nm/s for the LZ10, LZ18 and LZ26 coiled-coils respectively. More detailed experiment will be explained in the following section and will be compared to our results using our model. We also consider applying a load parallel to the axis of right-handed coiled-coils to see when the coiled-coil will unwind as it is pulled along its length. For parallel load, we consider two possible experimental situations: (i) both helices are being pulled, (ii) one of the helices is being pulled and the other one is free to move. Most of the time we see a collective unfolding of the coiled-coil where the unzipping or untwisting of the ends leads rapidly to the unwinding of the entire coil. In the case of parallel applied force for the right-handed coiled-coils, we also find that the unfolding force is smaller for shorter coils that possess less than two superhelical turns in comparison to longer coiled-coil structures where it be-

comes roughly length independent. We find that the changes in the force required to unzip the coiled-coils is not significant for different lengths. These predictions should be readily testable experimentally.

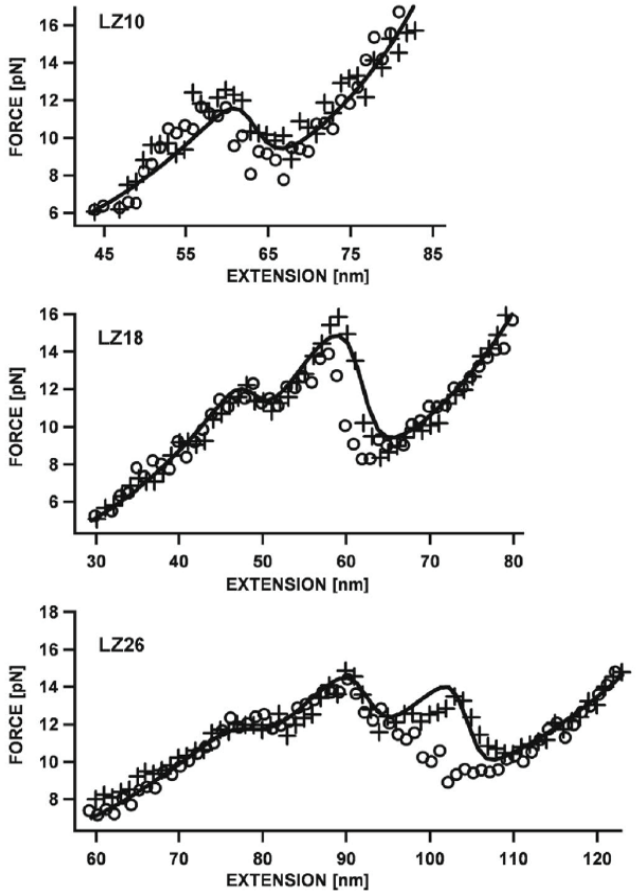

Figure 3.1: Experimental results for applying force on different length of GCN4. The solid line is calculated force-extension trace for each specific length and the crosses and the circles are showing the experimental results for stretching and relaxation traces respectively [63]

3.1 Generating Naturally Occurring Left-handed Coiled-coils

Most naturally occurring coiled-coils are left-handed, such as those found kinesin and myosin. A majority of left-handed coiled-coils form due to a heptad repeat that has hydrophobic residues at the first and the fourth positions. To date, single molecule measurements have been performed on left-handed coiled-coils. We have chosen the sequence repeat of 2ZTA that is one of the members of the canonical leucine zipper GCN4 family [85].The sequence can be found in appendix under the name of 2ZTA which is one of the GCN4 family members. The sequence of GCN4 along with its HP translation can be found in the Appendix under the name of 2ZTA which is one of the proteins of GCN4 Superfamily. Structurally, each α-helical turn contain 3.5 residues and so each heptad contains two turns. In a leucine zipper the first and the fourth positions of the heptad repeat are occupied by valine (V) and leucine (L) and every eight α-helical turns (28 residues), hydrophobic valine is replaced by asparagine (N) which is a polar amino acid. Given the leucine zipper sequence, we used our model to generate left-handed structures that could be compared to the naturally occurring one. The energies of interactions between residues were chosen to be $\varepsilon(H,H) = 3k_BT$, $\varepsilon(H,P) = 0.25k_BT$, and $\varepsilon(P,P) = 0.1k_BT$. [Making $\varepsilon(H,P)$ as large as $1k_BT$ had little effect on the final coiled-coil structure]. We set the interaction distance between two residues to be $r_0 = 5.2$Å. The parameters for the deformation energy are as in chapter 2, and it is scaled by W_D, which is a unitless weighting function.

To get insight into how the weight of deformation W_D changes the resulting coiled-coil structure, we calculated the root-mean-square-distance (RMSD) between the simulated structures and the crystal structure for GCN4 from the protein data bank [85]. Specifically, for each value of W_D we generated ten structures and computed the average RMSD to the GCN4 structure (see Fig. 3.2). We found that using only 28 normal modes was sufficient to generate reproducible RMSD results. Higher modes had negligible amplitudes-this will cease to be the case when loads are applied to the coiled-coils. As can be seen in Fig. 3.2 at low W_D, the cost of deformation is low, resulting in structures with high pitch and hence a poor RMSD. At high W_D, the cost of deformation is so high that helices behave as rigid rods, resulting in a poor RMSD that plateaus above a certain W_D. However, there is a broad range of W_D where the RMSD is low (below 1.5Å) showing that our simple energy model

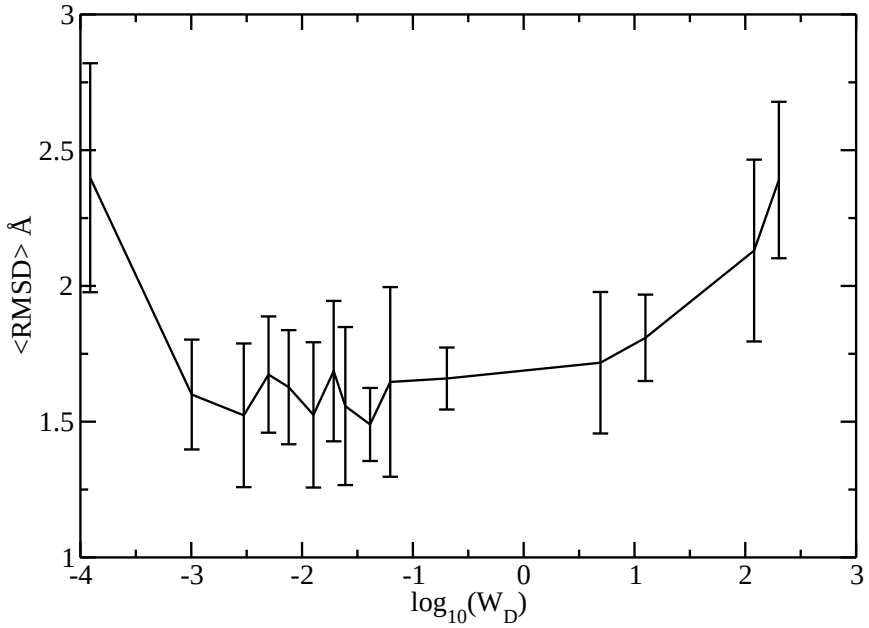

Figure 3.2: Root-mean-square-distance (RMSD) between simulated structures and crystal structure for GCN4 [85] as a function of deformation weight W_D. Shown are the average RMSDs for ten structures generated at each value of W_D.

is capable of reproducing the shape of the naturally occurring leucine zipper.

3.2 Unzipping Left-handed Coiled-coils

Recently, the mechanical properties of leucine zippers have been studied by unzipping them using single molecule approaches [64, 63, 62, 65]. In particular, the force-extension curves of leucine zippers of different lengths as well as leucine zippers with point mutations have been measured [63]. These measurements revealed that the coil would unzip in stages under constant velocity pulling measurements and that point mutations where the asparagine (P) is substituted with valine (H) would lead to an overall stabilization of the coil [63]. Using our model, we simulated the experiments above by unzipping left-handed coils of differing

lengths as well as coiled-coils that possess point mutations.

In our simulation, we apply a load perpendicular to the axis of the coiled-coil at the first residues of each helix in the coiled-coil. We consider two different ways of applying the external force: one is by increasing the distance between the first residues of the helices using a constant displacement (similar to the AFM experiments) and the other is by using an incremental force as explained in Section 2.5.

As in the experiments [63] we chose different lengths of leucine zipper, with 10, 18 and 26 α-helical turns referred to as LZ10, LZ18, and LZ26, respectively. The sequences and their HP translations are in the Appendix. We used 24, 26, and 28 modes for each length, respectively. For the case of perpendicular pulling, adding more modes did not change the results of the unzipping simulations as the amplitudes of higher-order modes are negligible. We used $W_D = 0.2$, $T = 1.0$ and $\Delta d = 1.0$ Å in all simulations.

In Fig. 3.3, we show the calculated force-extension curves for the LZ10, LZ18 and LZ26 sequences. For each length, ten structures were generated and each structure was pulled ten times; what is shown is the average force-extension curve from these runs. We first considered the situation of an "ideal" leucine zipper sequence where every N that occupies the first position of a heptad repeat is replaced by hydrophobic V [Fig. 3.3(a)]. As in the experiment, we see a staged unzipping, marked by rises and falls in the force as a function of length. Initially there is an increase in force required to break the hydrophobic contacts at the end, which then decreases as the coil becomes more unzipped. The number of stages increases with length, with the stages of the shorter coils occurring at the same locations in coils of longer length. LZ10 only displays a single transition, LZ18 has two clear stages, and LZ26 has three. For the longer helices, the sequence possesses several runs of hydrophobic contacts that produce the transitions seen around 4 and 9 nm.

We next consider the effects of changes in sequence. The wild-type sequences possess Asn residues in place of Val at certain heptads. We treat Asparagine as a *P*-type residue, though it is capable of forming hydrogen bonds which we neglect. Nevertheless, because of its polar character it is experimentally seen to have a destabilizing effect on the force-extension curves. As can be seen in Fig. 3.3(b) the substitution of Asn decreases the initial required force to unzip the first section of the coil, consistent with what was seen experimentally [63]. Its presence does not seem to change the number of stages in the unzipping process. Experimentally, only one N was changed to V in LZ18, so it will be interesting to

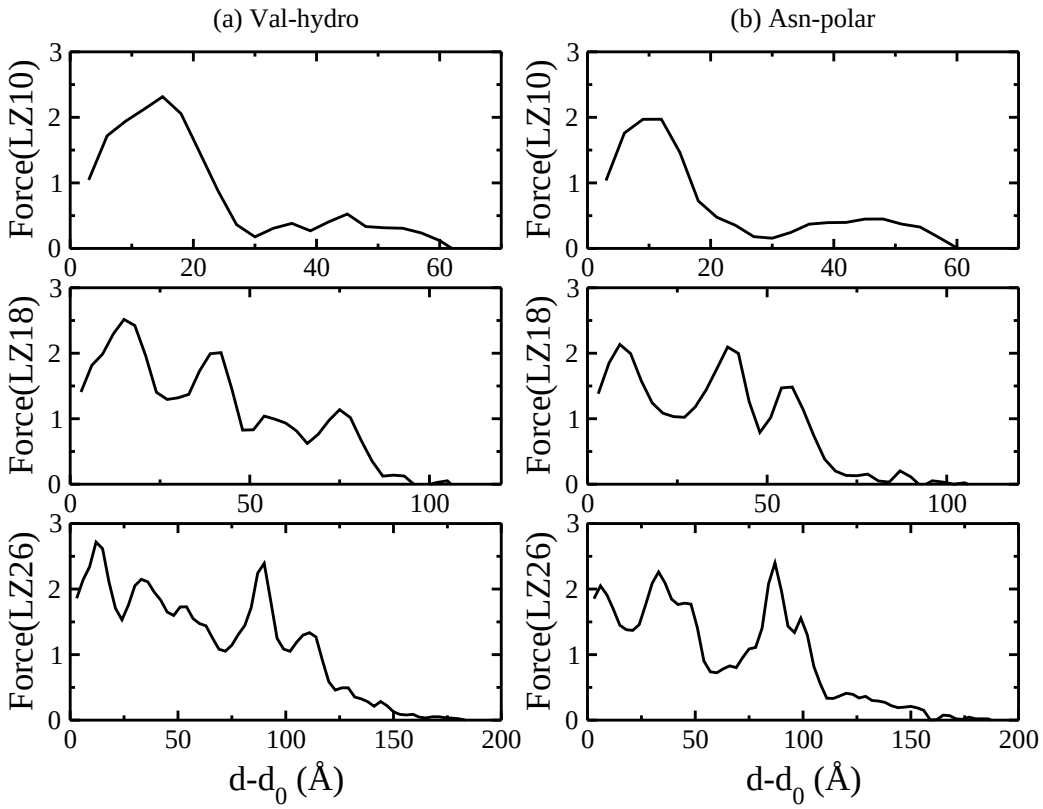

Figure 3.3: Simulated force-extension curves for leucine zippers of differing lengths, LZ10, LZ18 and LZ26. (a) Results for sequences where the Asn residue at the first position of the heptad repeats is replaced by Val. (b) Results for wild-type sequences that contain Asn residues at the first site of selected heptad repeats.

see whether changing all the Ns to Vs has any effect on the number of unzipping stages.

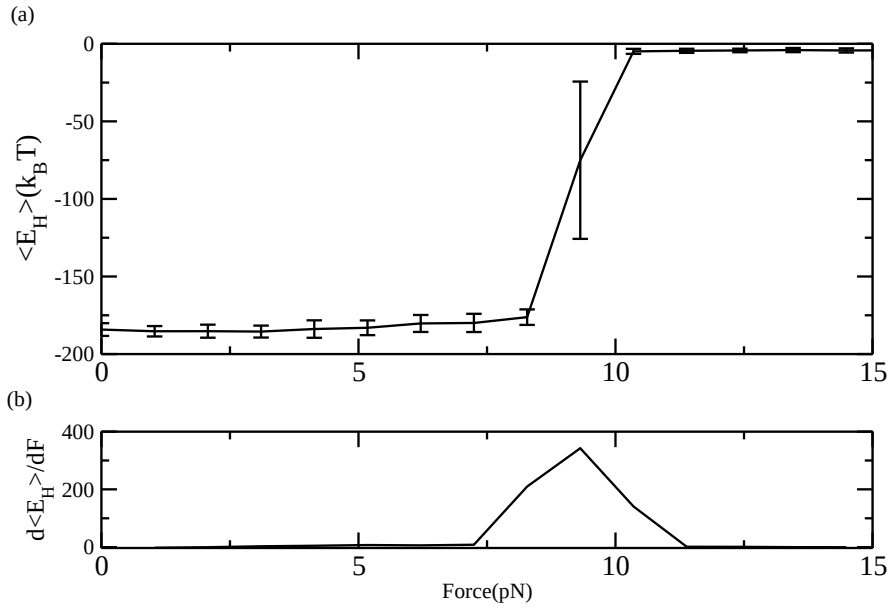

Figure 3.4: (a) Average hydrophobic energy vs applied force for LZ18. The error bars are the standard deviation of hydrophobic energy over some specific number of MC steps after the structure is equilibrated. (b) The transition force is defined to be the force corresponding to the peak in the susceptibility as defined in Eq. 2.8.

We now move on to consider unzipping the coiled-coils by applying a constant force perpendicular to the coil at the first residues of each helix. The force is gradually increased until the coil completely unzips [see Fig. 3.4(a)]. We find a single transition force for each coil, since all the forces required to unzip the later stages are less than the first transition [see Fig. 3.3(b)]. For the results presented below, we used a force increment of $\Delta F = 0.25 k_B T/\text{Å}$.

In Fig. 3.5 we show the calculated transition force as a function of leucine zipper length, from the smallest LZ10 to the longest LZ26. What is shown is the average transition force found from pulling on ten different structures of each length, ten times. As can be seen

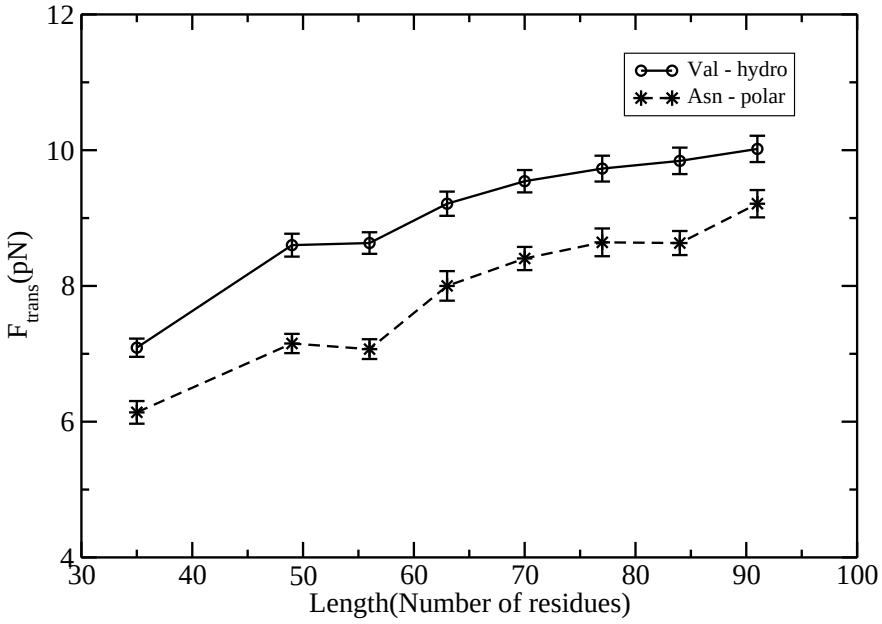

Figure 3.5: Transition force for leucine zippers as a function of length. (solid line) Sequences with select Asn replaced with Val. (dashed line) wild-type leucine zipper sequences (see the appendix).

the force needed to unzip coiled-coil structures that have Val substituted for Asn is higher, consistent with the observation that Val substitutions produce more stable structures [63]. The transition force is also seen to gradually increase with length for both wild-type and mutated sequences. This could have been inferred from the constant displacement results (Fig. 3.3) as the force of the first transition is seen to increase with length. This was also seen experimentally [63]. Hence our model confirms the experimentally observed increase in stability of longer leucine zippers.

We now move on to consider the mechanical properties of coiled-coils under a parallel load (see Sec. 2.5) and in particular right-handed coiled-coils that form more superhelical turns per unit length.

3.3 Dependence of Right-handed Coiled-coil Structure on Deformation Energy

As mentioned in Chapter 2 most naturally occurring coiled-coils are left handed although right-handed coiled-coils have been rationally designed [44]. For coiled-coils based on the heptad repeat sequence, the periodicities of the superhelical structures are on the order of 80-100 residues for the completion of one full wind. We have chosen to study helices patterned with a right-handed hydrophobic strip that has a periodicity of 32 residues so that several wrappings are possible for lengths $N < 100$ residues (superhelical structures such as collagen, although not built from α helices, have multiple windings over a span of 100 residues). The sequence repeat that we chose to use which leads to a right-handed hydrophobic strip is a tetramer given by HHPP. The parameters such as the interaction energies between the residues or the stiffness constant for the springs connecting the residues are the same as left-handed coiled-coil parameters as explained in Sec. 2.1.

Using the above parameters and the simulated annealing approach outlined in Chapter 2, we generated sets of structures at different values of W_D. The resulting average energy of each set of structures is shown in Fig. 3.6. For $W_D > 0.1$ the helices are essentially rigid rods and the resulting packs have higher energy. In this case the deformation energy E_D dominates and the stiff helices are not able to satisfy all their hydrophobic contacts. For long helices with 50 or more residues and at $W_D > 0.2$, helices behave as rigid rods and the total energy no longer varies as W_D increases. For shorter helices, the energy does not plateau because due to their short length, they are able to make contacts between the hydrophobic residues at both ends of the helices. If W_D is made too small, then the hydrophobic residues cause unphysical compactification of the α helices. Hence in our simulations, we only ever see coiled-coil structures that adopt the same chirality as the hydrophobic strip that winds around each helix. In principle, if W_D was able to be made arbitrarily small, then it should be possible to see multiple structures emerge that satisfy the packing of the hydrophobic strip, as predicted in [68].

As mentioned in Chapter 2, we do not use all the $3N$ normal modes of a helix of length N. In Fig. 3.8 we show the amplitudes of the normal modes in the final coiled-coil configuration at zero applied force, for the situation where 90 modes were used to make the structure. For modes above 30 the amplitudes are negligible and we find that we do not

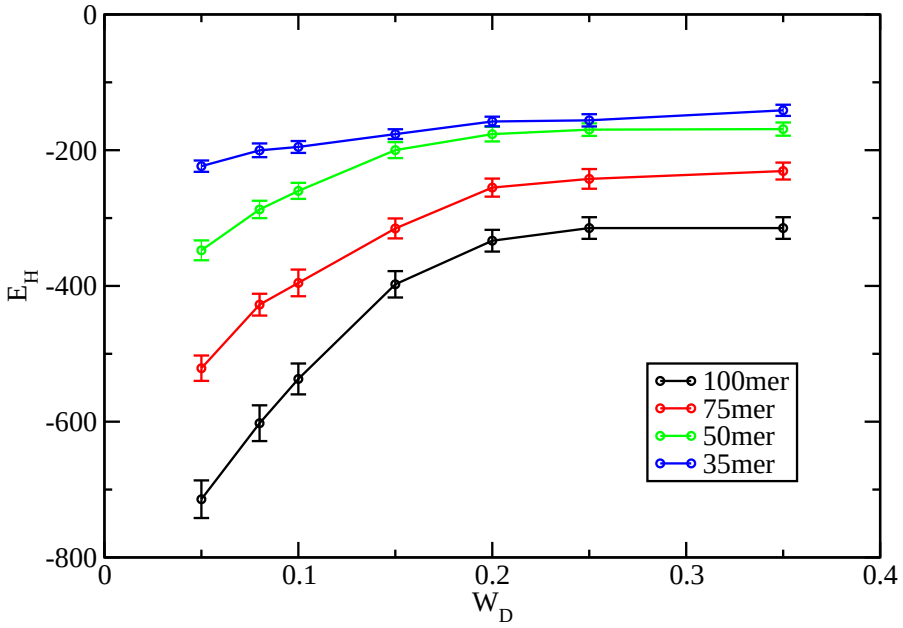

Figure 3.6: Dependence of the average hydrophobic energy E_H on the weight of the deformation energy W_D for right-handed coiled-coils patterned with HHPP sequence repeat. For a given W_D the average energy was computed using 15 different coiled-coil structures generated from the simulated annealing approach.

need to use them in generating the coiled-coil structures starting from an unfolded configuration. When the coiled-coil is under an applied force, then this restriction is lifted as there are higher-order modes which do end up contributing to the unfolding process. This will be discussed in the following section.

To get further insight into how differences in the structure of coiled-coils of different lengths affect the total energy, we compare the normalized bending and hydrophobic energy as a function of helix length (see Fig. 3.7) for coils generated using fixed $W_D = 0.08$. For helices with $N < 50$, the resulting coiled-coils have less hydrophobic contact energy per residue than those which are longer. They also are less deformed from their ideal helical configuration and so have overall less deformation energy. Above $N = 50$, the coiled-

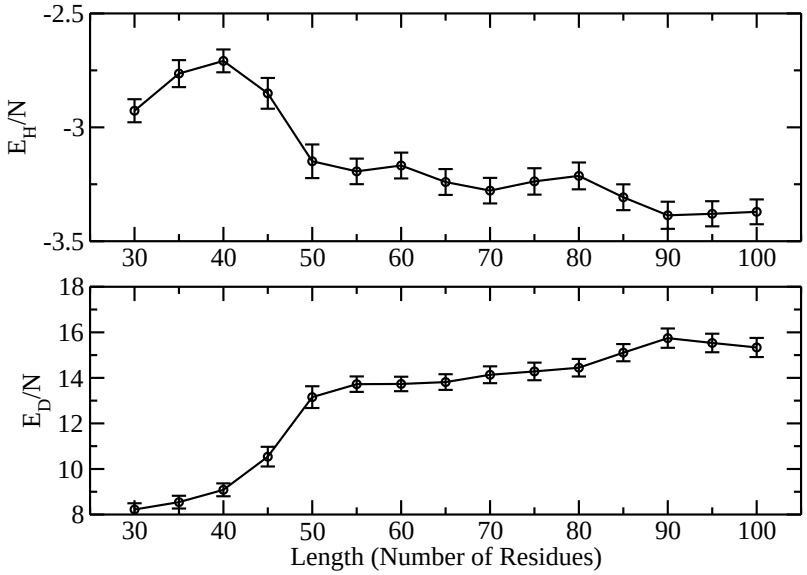

Figure 3.7: Plot of (a) hydrophobic energy E_H and (b) deformation energy E_D as a function of length N. Each energy is normalized by N so that the y axis represents the energy per residue.

coils form more than one complete superhelical wrapping and are able to satisfy more hydrophobic contacts. However, the better packing corresponds to larger deformations and hence an overall higher deformation energy than that of shorter helices. Nevertheless, at larger lengths the contact and deformation energy become length independent (see Fig. 3.7) showing that the resulting superhelical structure is periodic. Thus adding more residues just continues the established periodic structure.

3.4 Mechanical Unfolding Properties of Right-handed Coiled-coils

In this section we consider the unfolding properties of the coiled-coil structures generated in section 3.3 subject to a force applied parallel to their long axis. We consider the possibility of applying a force to just one or to both of the helices. We use the Monte Carlo approach of unzipping the coiled-coil structures as described in Chapter 2, except now the force is applied parallel to the coiled-coil.

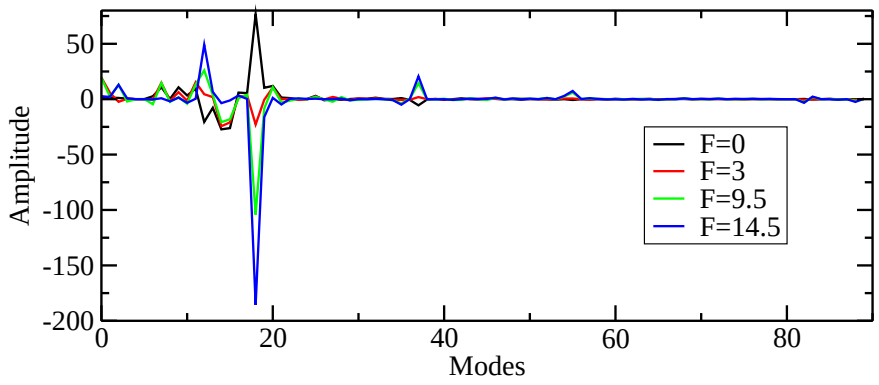

Figure 3.8: Amplitudes of normal modes at different applied forces for right-handed coiled-coil with $N = 100$. For this plot both helices were being pulled simultaneously.

Under tension, the helices try to keep their hydrophobic contacts, but when the force is large enough, their hydrophobic contacts break. As we discussed in Chapter 2 this force is called the transition force. For most of the unfolding studies, just as for leucine zippers, we tend to see one large transition where the entire coiled-coil unfolds; there are smaller

transitions, especially at initially low applied forces ($F < 5k_BT/$Å) when the ends of the coiled-coil fray slightly. The force corresponding to the peak in the thermodynamic susceptibility [Eq. 2.8] we define as the transition force, F_{trans}. The transition force varies with length, and we now show how it changes depending on whether only one helix or both are pulled.

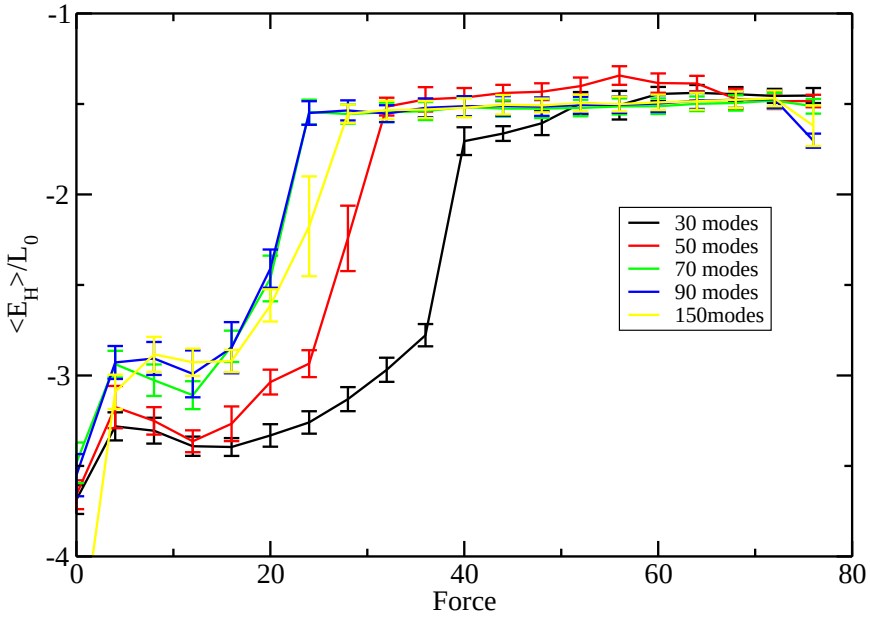

Figure 3.9: Dependence of unfolding as a function of the number of modes. Shown is the average hydrophobic energy per unit length as a function of force for $N = 90$ using different normal mode cutoffs.

Before moving on to discuss the results of the pulling studies, we again comment on the number of normal modes used in carrying out the Monte Carlo simulations. When generating the coiled-coil structures in Section 3.3, we found that only ≈ 30 modes were necessary. Now when there is an applied load, we find that significantly more modes need to be included in order to generate consistent results. This is because there are higher-energy stretch modes which contribute to the unfolding process. This is shown in Fig. 3.8,

where we show the mode amplitudes at different applied force. Some of the modes decrease in amplitude due to the unwinding of the coil (e.g., several of the modes between 10 and 20), whereas a number them increase (e.g., mode 19, 36, and 55) due to the stretching of the coil. We have found that using more than 30% of the modes for a given length produces consistent transition forces (see Fig. 3.9). Using fewer modes causes the transition force to be higher for all lengths, and result in strong variation as a function of length.

We now discuss the results for the case of pulling on just one of the helices. For small coiled-coils with 45 or fewer residues, there is no transition force and the coiled-coil smoothly unwinds. For longer lengths the helix on which no force is applied stretches along with the other, and then at the transition force, it springs free, returning to its relaxed undeformed configuration. In Fig. 3.10, we plot the transition force as a function of length. For coiled-coils above $N = 60$, the structure consists of two or more full windings, and we see an increase in the force needed to unfold the coiled-coil. There is some variation in the force as a function of length, and we attribute this to the shearing that happens for some structures since residues on the ends that are being pulled do not line up perfectly along the long axis of the coil. Thus there will be periodic changes with length as the end-to-end vector departs from being parallel to the long axis.

For the case when both helices are being pulled, we applied an equal force to both. Thus the total load is twice that of the case when one helix was being pulled. Because both helices are being pulled, they tend to stretch in unison, thereby being able to maintain their hydrophobic contacts. Now for all studied lengths, $N \geq 30$, we find that there is a transition force where the coil unfolds. The transition force as a function of length for the case where both helices are pulled is shown in Fig. 3.10. Similar to that found for the case where only one of the helices was pulled, we find that smaller coils unfold at lower forces than longer coils ($N > 60$). It also takes more than twice the force of that applied to a single helix to unfold the coil. For lengths greater than 60 residues, the force required to unfold the coil is essentially length independent. Some variation with length is again observed, and as before we attribute this to the change in the amount of shear that occurs as the length of the coil changes.

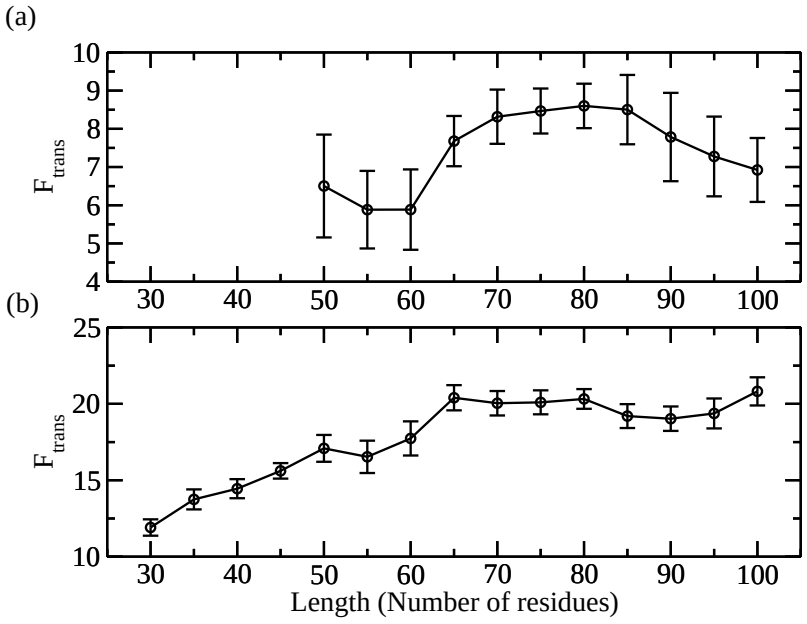

Figure 3.10: Transition force, corresponding to the force at which the coiled-coil loses most or all of its superhelical structure, as a function of right-handed coiled-coil length for pulling on (a) one helix or (b) both helices.

3.5 Conclusion

In summary, we have presented a simple energy model for coiled-coil formation and a computational approach that allows for the efficient generation of coiled-coil structures. Because of its simplicity and dependence on relatively few parameters, it can be adjusted to accommodate a variety of different coiled-coil structures.

We first used the model to study the folding properties of leucine zippers. Using just two types of residues, hydrophobic and polar, the model was able to fit the naturally occurring leucine zipper GCN4. We then simulated the unzipping of leucine zippers of differing lengths and with point mutations. Our model was able to capture all of the essential results seen experimentally [63] such as staged unfolding when pulled at constant velocity,

increased stability of coils when Val is substituted for Asn, and the overall increase in stability with coiled-coil length. The effects of pulling at different velocities can easily be incorporated into the model and this will be a subject of future research.

With respect to the prediction that it is the deformation energy that selects for the chirality of the resulting coiled-coil structure [68], our work has made some progress at addressing this prediction using an atomistic model. At a particular cost for the deformation energy, the model transitions from generating non-coiled structures as the α helices are too rigid, to coiled-coil structures, where the superhelical twist matches that of the underlying hydrophobic strip that patterns each helix. However, when the deformation energy was given a small weight, we were unable to generate structures whose twist departed from that of the hydrophobic strip with the resulting structures possessing unphysical distortions. Future work will focus on exploring the possibility of generating coiled-coil structures with chiralities that are different from that of the hydrophobic strip when there is little or no deformation cost.

Our results from studying the mechanical unfolding properties of right-handed coiled-coils make several interesting predictions. First, when pulled along the long axis, we consistently find that in nearly all unfolding simulations, the coiled-coils unfold together in an abrupt transition. We also found that coils that possess less than two full superhelical wrappings unfold at lower force than those that possess more. Additionally, pulling on both helices simultaneously will require a higher force to unfold the coil than the situation when only one helix is pulled. Current experiments have applied forces perpendicular to the long axis of the coil, but we expect that it should be relatively straightforward to pull along the long axis as has been done for other coiled-coil structures such as fibrin [86]. Thus these predictions should be readily tested.

Chapter 4

Generating Experimental Coiled-coil Structures Using a Binary Patterning Model

Several methods for predicting coiled-coil structure from sequence have been provided, ranging from detailed molecular dynamics simulations [69] to coarse-grained Monte Carlo approaches [62, 70]. Our model as explained in Chapters 2 and 3 greatly reduces the search space of possible structures by utilizing only a limited number of the flexural modes of each helix. We showed that using a Monte Carlo approach we could generate leucine zipper structures [70] and reproduce mechanical results from single molecule measurements [63]. Given the simplicity of the model and its ability to capture the physical properties of one particular type of coiled-coil, we now explore whether it can be generalized further so as to predict the structures of other coiled-coil sequences beyond the leucine zipper. Recent work on fitting coiled-coil structures to the Crick parametrizations for coiled-coils showed small variation in parameter values from one structure to the next [87]. Our model possesses even fewer fit parameters and we wish to test whether the values that give the best fit are consistent across all the naturally occurring structures.

The canonical sequence of naturally occurring coiled-coils is based on a heptad repeat with the *a* and *d* positions being hydrophobic since they form the interface between helices. Amongst the coiled-coil families, significant variation in sequence is observed [88, 89]. Such variations have both thermodynamic and kinetic consequences to folding of the re-

sulting structure. One aim of the models is to predict the consequences of such sequence variation and identify sequence features that are important for driving the folding process. In the context of binary patterning, the sequence that is consistent with the exposure pattern of the structure may not necessarily have the best folding properties due to degeneracies. The simplicity of our model allows us to explore the consequences of varying the underlying hydrophobic pattern for coiled-coil structures.

In this chapter we first present the set of coiled-coil structures that we have chosen to analyze from the Structural Classification of Proteins (SCOP) database [51]. SCOP, which first was generated manually, is a database that provides a detailed and comprehensive description of the structural and evolutionary relationship of the proteins of known structures [90, 91]. The SCOP hierarchy contains Species, Proteins, Family, Superfamily, Fold and Class. Species are distinct protein sequences, Proteins are similar sequences that essentially have the same function, Families contain proteins with similar sequences but different functions, Superfamilies group Families with common function; Folds are the structurally similar superfamilies with different characteristics and Classes categorize folds based solely on their secondary structure content [90, 91].

Using the naturally occurring structures we determine an optimal residue radius that maximizes the hydrophobic energy of each pack. In the next section we present results of generating computational structures and by fitting the simulated structures to the natural structures we constrain the model parameters. Our fitting procedure is based on minimizing the root mean squared distance (RMSD) between computational and experimental coiled-coil structures. In the last section we use the constrained model to predict the thermodynamic properties of the selected structures by calculating their unfolding temperatures. Besides capturing known thermodynamic properties, we use the model to predict the effect of sequence mutations.

4.1 Properties of Naturally Occurring Coiled-coils

We extracted a set of naturally occurring coiled-coils and attempted to generate them using our model. We chose a representative from each Superfamily of parallel coiled-coils from the SCOP database [51] (See Table 4.1 for the list of selected structures). Some of the families were not chosen because they had experimental structures involving the interaction

of the coiled-coil with DNA. We also excluded coiled-coil superfamilies whose structures are stabilized due to other interactions such as salt bridges or disulfide bonds which are not currently included in our model.

As we explained in Chapter 2 the total energy in our model consists of only hydrophobic and deformation energies for the α-helices. The model contains only a few parameters: three energies of interaction in the HP model, contact radii between residues $R_{i,j}$ and spring constants for the deformation energy. The HP energies are chosen to be consistent with a HP decomposition of the Miyazawa-Jernigan matrix [54], and the spring constants are obtained from an analysis of the bending of α-helices [71]. That leaves the $R_{i,j}$, contact radius and the weight of the bending energy, W_D, as unconstrained.

Given the sequences and coordinates of the natural structures, we evaluated the hydrophobic and bending energy of each structure to examine the degree of homogeneity in energies across the different coiled-coils. Since we use a HP-model for our hydrophobic energy, we translated the amino acid sequence into a corresponding HP sequence for each coiled-coil (see Appendix for translations). In the model, the contact radius $R_{i,j}$ between residues making up the separate helices set the superhelical radius. We first considered the case where the contact radius between interacting residues was constant, and we adjusted it for each natural structure to determine the radius that gave the lowest hydrophobic energy. The results are shown in Fig. 4.1(a) for several natural structures. There is not much variation in the optimal contact radius over the natural structures with an average of $R_{i,j} = R_0 = 5.5$. Thus we use $R_0 = 5.5$ in the latter sections of this thesis when generating structures with a constant contact radius. With the contact radius fixed, we show the average hydrophobic energy per hydrophobic residue as a function of coiled-coil length for the experimentally determined structures in Fig. 4.1(b). For the most part, the hydrophobic energy per unit length is fairly uniform, consistent with our previous findings for leucine zippers [70]. Also the graph shows that coiled-coils with three helices have higher hydrophobic energy per residue as they are able to make better contacts between the hydrophobic residues of all three different helices.

With respect to the bending energy, in Fig. 4.1(c) we show the deformation energy per residue for the experimental structures. Most coiled-coil helices show nearly the same degree of deformation energy, with the triple helices (1JCC and 2BA2) showing less deformation; in the case of 1DEB one helix is far more deformed than the other. The average

SCOP family	Protein	N_{res}	N_{helix}	E_D	E_D/N_{res}	E_H	E_H/N_H	$H\%$	R_0
Leucine zipper domain	2ZTA	29	2	29.4	0.51	-60.51	-2.75	37.9%	5.7
N-terminal domain from apc	1DEB	48	2	58.0	0.60	-58.33	-1.82	33.3%	5.8
Outer membrane lipoprotein	1JCC	44	3	44.2	0.33	-175.48	-3.44	38.6%	4.8
FYVE/PHD zinc finger	1JOC	56	2	98.8	0.88	-65.17	-1.55	37.5%	5.2
Geminin domain	1T6F	37	2	95.9	1.29	-61.75	-2.57	32.4%	5.5
G protein binding domain	1UIX	63	2	61.7	0.48	-40.06	-0.83	38.1%	5.4
MPN010-like	2BA2	60	3	74.8	0.42	-441.57	-5.66	43.3%	7.4
Myosin rod fragment	2FXM	100	2	122.8	0.61	-115.86	-1.70	34.0%	5.4
Myosin rod fragment	3BAS	73	2	74.6	0.51	-108.04	-2.25	32.9%	5.7
Vimentin (mutated)	3G1E	37	2	61.1	0.82	-111.2	-3.47	43.5%	6.3

Table 4.1: List of the experimental structures studied. The first column is the name of the superfamily in the SCOP database; the second column is the PDB ID of the protein. The third and the fourth columns are the number of residues in each helix and the number of helices in the coiled-coil structure. In the following columns the bending and the normalized bending energy (E_D and E_D/N_{res} respectively) and by total number of amino acids in each structure followed by hydrophobic energy, E_H, and hydrophobic energy per total number of hydrophobic residues, E_H/N_H, in the whole protein at the optimized value of contact radius, R_0. Bending and hydrophobic energies are calculated using C_α coordinates. The second last column has the percent of hydrophobic residues out of the total number of residues and in the last column the optimized radius (R_0) for each structure has been shown.

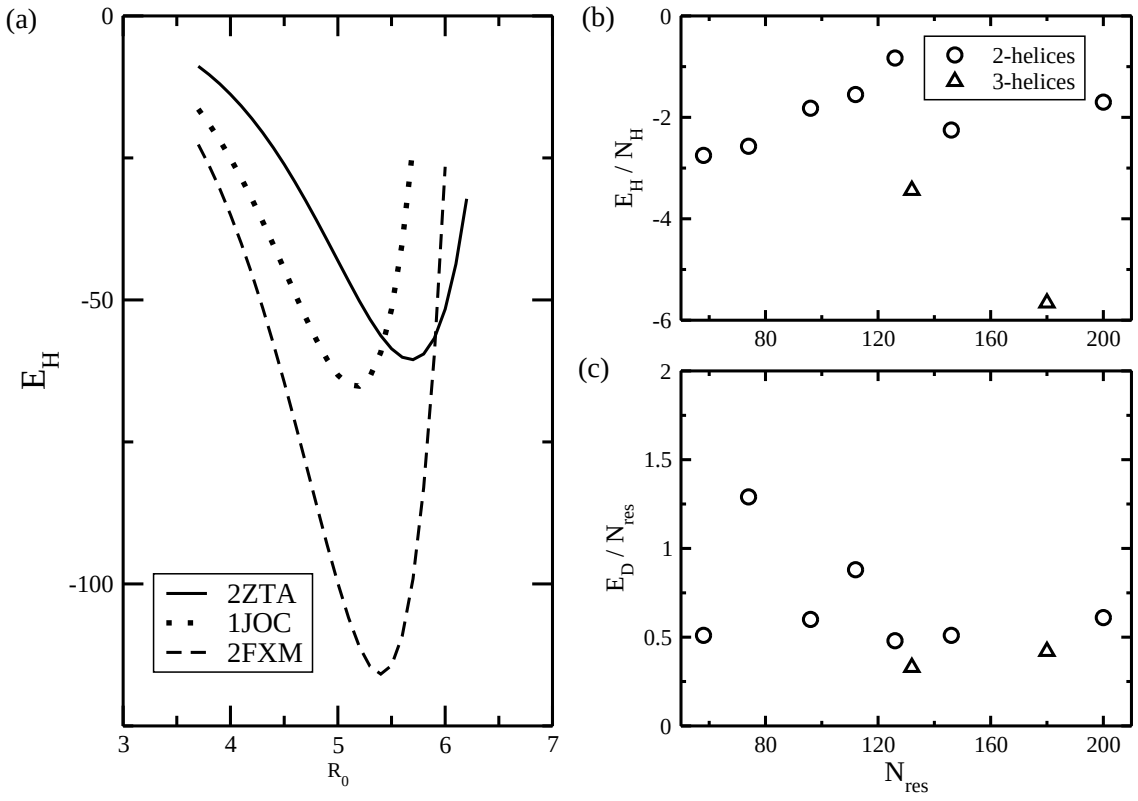

Figure 4.1: (a) Hydrophobic energy as a function of contact radius for three double helix coiled-coils with different lengths. The average optimized contact radius over all the double helix coiled-coils is 5.5. (b) Hydrophobic energy per hydrophobic residue as a function of coiled-coil length. (c) Bending energy per residue as a function of coiled-coil length.

bending energy per residue in all the helices over all the structures is $E_{bend}/N_{res} = 0.59$ and we will use this to constrain the bending energy weight when generating the simulated structures below.

4.2 Fitting the Model to Experimental Coiled-coil Structures

Using the HP sequence for each coiled-coil we use our model to generate a set of simulated structures. The type of structures that emerge from the model depend strongly on the weight of the deformation energy, W_D. For each coiled-coil sequence we made 15 different simulated structures at each of 12 different W_D's starting from $W_D = 0.05$ to $W_D = 10.0$. Our goal is to determine the range of W_D values that produces structures that best match the experimentally determined ones. We consider two criteria for finding the best W_D: (i) the W_D that produces the lowest average root mean squared distance(RMSD) between the simulated and naturally occurring coiled-coil, or (ii) the W_D that produces the same average deformation energy for the simulated structures as that of the naturally occurring one.

We made structures using the fixed contact radius R_0 determined above from the experimental structures as well as using non-uniform $R_{i,j}$ from Ref.[84]. At a given W_D we aligned each simulated structure to the experimental structure by minimizing the RMSD as discussed in Methods section, and then calculated the average RMSD. We show the results of aligning several naturally occurring coiled-coil structures (a double and triple helix coiled-coil) to their lowest RMSD simulated counterparts in Fig. 4.2. The red line is the backbone of the natural protein and the blue line is the structure simulated using our coarse grained model. The double helix structure is a myosin rod fragment (2FXM) with 100 residues in each helix, and the RMSD for the shown best alignment is 2.16 Å. The trimeric coiled-coil is from an outer membrane lipoprotein (1JCC) and has 44 residues in each helix. The RMSD between it and its best simulated structure is 1.24 Å.

Fig. 4.3(a) shows the average RMSD as a function of W_D for two representative coiled-coil structures. The graph shows that there is an interval of W_D that yields a minimum in RMSD values. Indeed, for all the different coiled-coils studied we find an interval of deformation energy weight that yields low RMSD. At low W_D's helices are able to make better

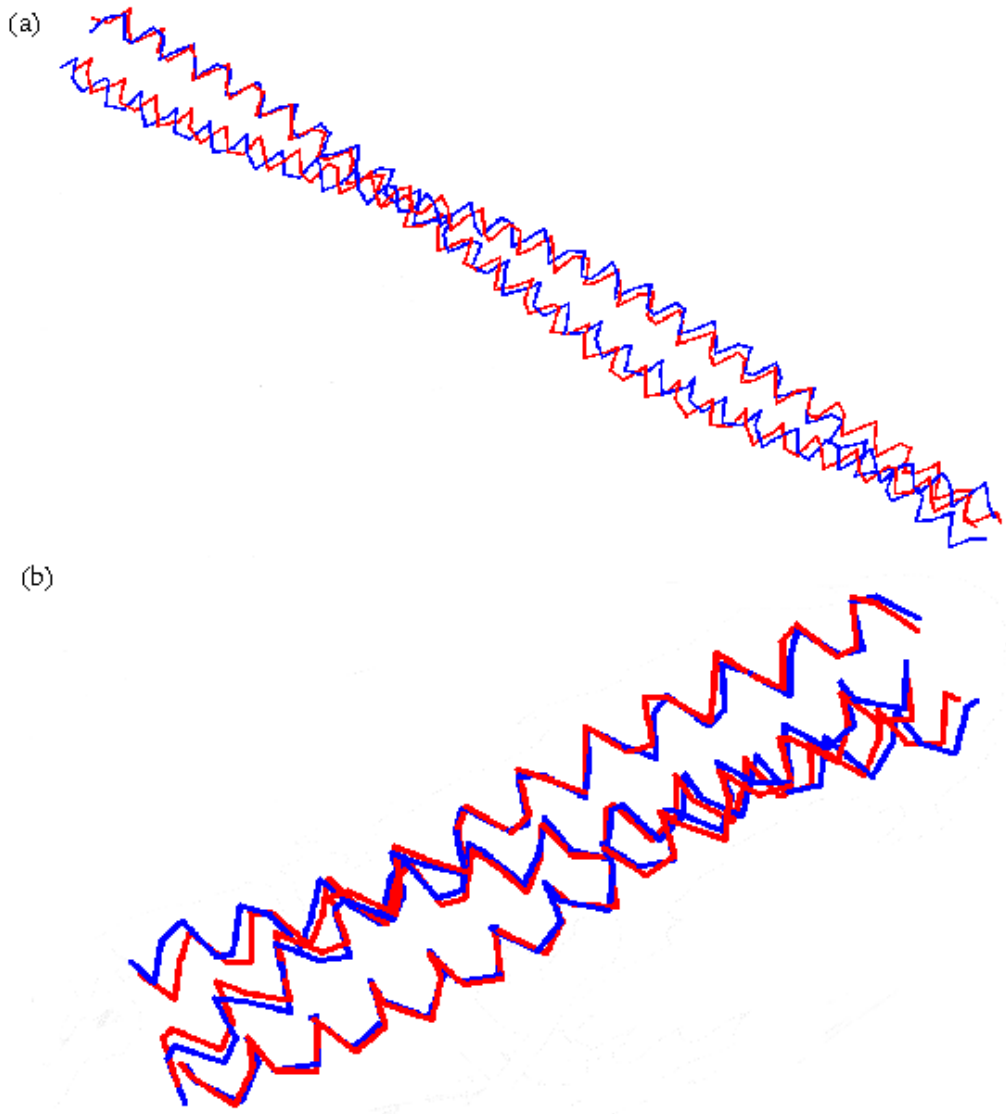

Figure 4.2: Alignments of C_α coordinates between experimentally determined coiled-coil structures and simulated ones. In both alignments the red line is the experimentally determined structure and the blue line is the simulated structure using our model. (a) 2FXM (myosin rod fragment) is shown which is a dimeric coiled-coil with 100 residues in each helix. The lowest average RMSD as 2.16 Å between the natural and simulated structure using our model. (b) 1JCC (outer membrane lipoprotein) is a trimeric coiled-coil with 44 residues in each helix. The lowest average RMSD for this structure was 1.24 Å.

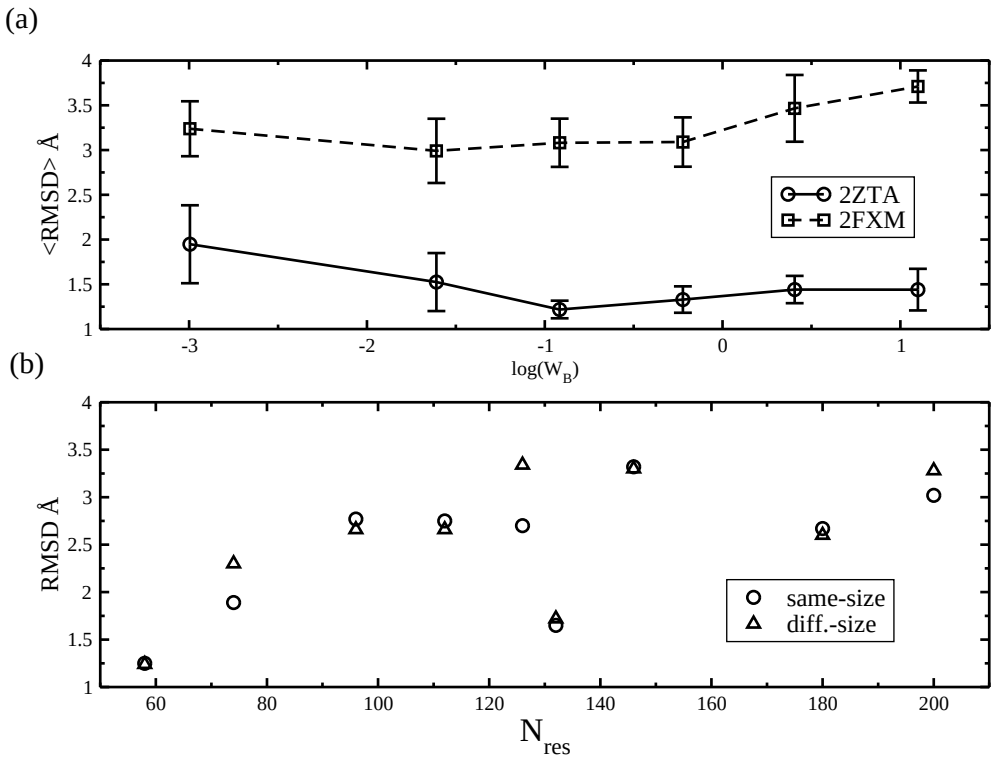

Figure 4.3: (a) The open squares show the average RMSD at different W_D for 2FXM (myosin rod fragment) which is a dimeric coiled-coil containing 100 residues in each helix. The open circles is for 2ZTA (leucine zipper domain), a dimeric coiled-coil with 29 residues in each helix. Lines are a guide to the eye. (b) The lowest average RMSD over all different W_D's as a function of their length calculated for same contact radius, circles, and for different contact radius, triangles, are shown in the figure.

hydrophobic contacts due to their softer backbone, but the structures are non-physically compressed and this results in higher RMSD between the simulated and the experimental structures. The average RMSD increases at higher W_D because the helices behave as rigid rods and therefore have less chirality with respect to the experimental structures. In Table 4.2 we summarize the results from fitting the structures, giving the W_D that yields the lowest average RMSD between the simulated structures and the experimental structure. From Fig. 4.3(a) and Table 4.2 we conclude that using W_D in the range $0.2 - 0.8$ yields simulated structures that fit well to naturally occurring coiled-coils over a range of helix lengths and HP patterns. In Fig. 4.3(b) we plot the minimum average RMSD as a function of helix length and not surprisingly we see that RMSD tends to go up with the length of the helices. For all of the structures considered, the hydrophobic content is roughly the same($\sim 36\%$), so we see no clear dependence of our fits on this value. In Fig. 4.3(b) and Table 4.2 we compare the results from generating simulated structures, using constant contact radii, R_0 or non-uniform contact radii. There is not much difference, though the case of uniform radii tends to do slightly better in terms of RMSD.

To get a sense for these RMSD numbers we aligned several known structures from the Leucine-zipper superfamily in SCOP containing 29 residues in each helix. All the members of this superfamily have sequences built from the heptad repeat (hydrophobic residues at the first and fourth positions of a 7mer repeat) and possess strong sequence similarity. We extracted the C_α coordinates of all the parallel dimeric coiled-coils of this family and performed an all-against-all RMSD alignment. We found an average RMSD of 2.0 ± 1.3 Å across the family. Comparing this number to the fits of the model structures to known coiled-coils, the model is able to produce structures that are within the alignments of natural structures in a given family.

As a second check to constrain W_D, we next compared the deformation energies of our simulated structures at different W_D to those of the naturally occurring coiled-coils. We find that this yields similar constraints on W_D as that found for the RMSD fit. For each set of simulated structures at a given W_D we calculate the average E_H and E_D. When we compare the hydrophobic energy of the simulated structures to the naturally occurring one we consistently find that the simulated structures have lower E_H due to a better packing of the model's residues. We show in Fig 4.4 the average deformation energy per unit length of the simulated structures at different values of W_D. The horizontal lines correspond to

Protein	RMSD$_{same-size}$	W_D	$\overline{RMSD}_{same-size}$	W_D	RMSD$_{diff-size}$	W_D	$\overline{RMSD}_{diff-size}$	W_D
2ZTA	0.88	3.0	1.25	0.8	0.86	1.5	1.25	0.8
1DEB	2.28	0.2	2.77	0.4	2.28	0.2	2.66	0.2
1JCC	1.24	0.2	1.65	0.4	1.49	0.8	1.82	0.4
1JOC	2.12	0.2	2.90	0.2	2.13	0.05	2.84	0.2
1T6F	1.39	0.4	1.90	0.4	1.80	0.2	2.30	0.2
1UIX	1.50	0.2	2.70	0.2	2.13	0.2	3.34	0.4
2BA2	2.12	0.4	2.67	0.4	2.22	0.4	2.60	0.4
2FXM	2.16	0.2	3.02	0.2	2.71	0.2	3.28	0.4
3BAS	2.56	0.2	3.32	0.2	2.47	0.2	3.30	0.2

Table 4.2: The best RMSD between simulated and natural structures. The first column is the PDB ID of all the proteins that we studied in this chapter. The second column is the best RMSD using constant contact radius $R_{i,j} = R_0 = 5.5$ for all the residues in the model; followed by the W_D which gives that RMSD. The next column shows the lowest average value for the RMSD and the corresponding W_D. The following columns are the best RMSD, W_D for the best RMSD, the average RMSD and the W_D related to the average RMSD. These values are calculated between simulated structures with different contact radii $R_{i,j}$ [84], and the natural structures.

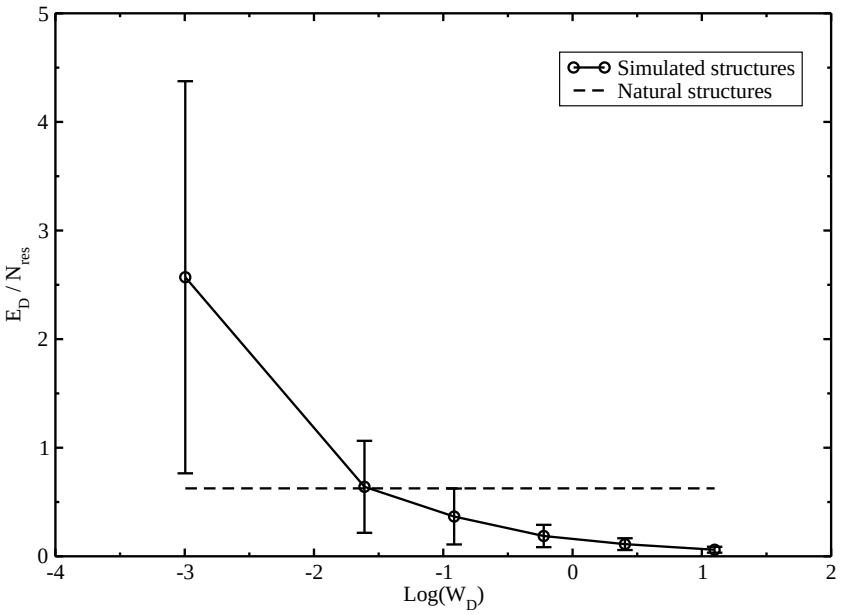

Figure 4.4: The dashed line gives the average deformation energy per residue over all the natural structures including dimeric and trimeric coiled-coils. Open circle denote the average of normalized bending energy E_{bend}/N_{tot} of simulated structures at different values of W_D over all the coiled-coils. The solid line is the guide to the eye. These two data setsintersect close to -1.5 which corresponds to $W_D = 0.2$.

the previously calculated average deformation energy of the natural coiled-coils. Where it intersects with the deformation energy of the simulated structures gives a corresponding W_D that produces simulated structures with similar deformation energies as the naturally occurring ones. Over all the structures these two graphs intersect at $W_D = 0.2$ and this is in agreement with RMSD results.

4.3 Thermal Stability of Coiled-coil Structures

The energy landscape for a protein is funnelled shape with one global minimum and many local minima. To have a successful folding process we need to avoid these local minima; it is hard to scape these local minima once the protein is trapped in them. On the reverse side, these local minima do not interfere with the unfolding process. In order to explore the effect of sequence variation on thermal stability, we find the unfolding temperature of all the simulated structures. We have also tried to refold the unfolded structure. The energy of the folding process from a simulation perspective needs to be funnelled (i.e. the helices need to be aligned) in order for the process to proceed but the reverse process, unfolding of a protein, is starting from a well defined configuration and it goes to an unfolded one. This explains why we were not successful in refolding the unfolded structures all the time. If the refolding process was happening while some of the hydrophobic interactions remain in the structure, then the refolding process was successful but when all the hydrophobic interactions are broken the refolding process was a challenge.

The unfolding temperature is calculated as explained in the Methods chapter. From the results of the previous section we use $W_D = 0.2$ to generate structures. As an example of how we calculate the unfolding temperature, we find the specific heat capacity for 3BAS with 73 residues as explained before, Fig 4.5(a). The corresponding hydrophobic energy for the same simulated structure is shown in Fig 4.5(b). The graph shows one clear transition, which corresponds to the unfolding temperature where all the hydrophobic bonds between the two helices are broken. We then find an average unfolding temperature for eacj of the simulated structures as displayed in Fig. 4.5(c). It can be seen that it increases with the length of the helices making up a coiled-coil. Even though the hydrophobic energy per residue is nearly constant for coiled-coils of different lengths, the increase in unfolding temperature arises because of the requirement to break more contacts in order to unfold.

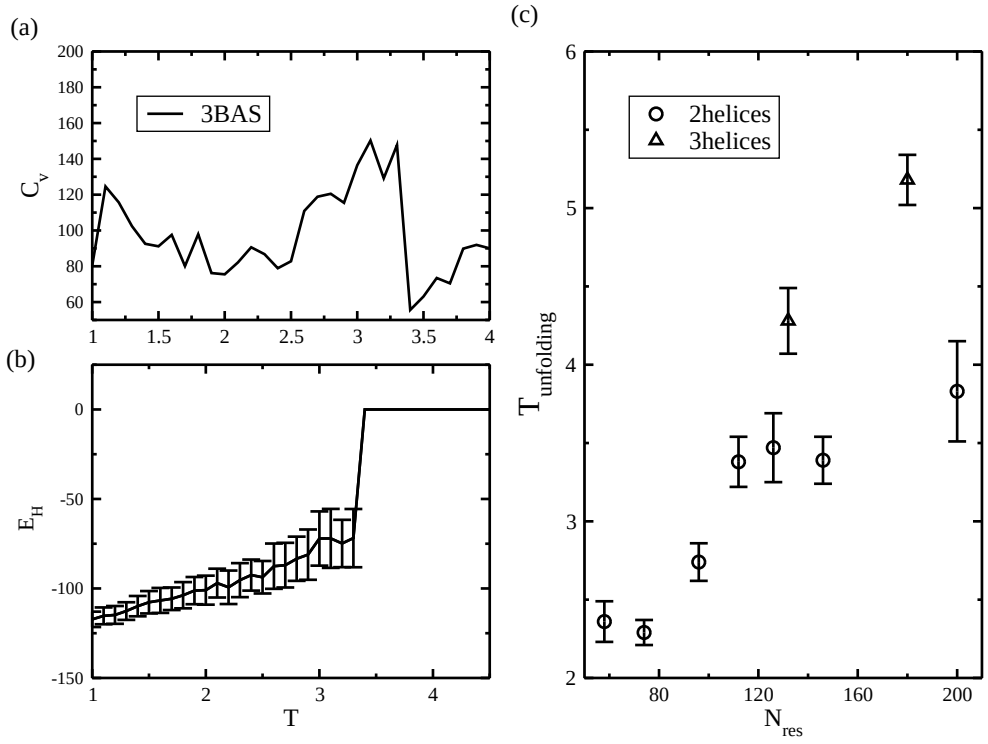

Figure 4.5: (a) The specific heat capacity for 3BAS (myosin rod fragment) which has two helices in its structure with 73 residues in each helix. At unfolding temperature there is a big change in the total energy. (b) The hydrophobic energy as a function of temperature. As temperature increases the hydrophobic energy eventually becomes zero when all the hydrophobic contacts break and the helices separate. (c) The unfolding temperature as a function of the total number of residues for each simulated coiled-coil structure is shown in this graph.

This also explains why coiled-coils made of three helices consistently had higher unfolding temperatures, as again, to unfold, more contacts must be broken.

Next we used the model to investigate the consequences of mutating each coiled-coil sequence on the unfolding temperature. We mutated sequences by changing each hydrophobic residue in a sequence to a polar residue, one at a time. To find the unfolding temperature of the mutant we started with the simulated structures generated by the unmutated sequences and we then equilibrate the structure with the mutated sequence at a fixed low temperature. We then found the unfolding temperature using the same method as in the Methods section. We categorized the effect of the mutated residues on the unfolding temperature according to their positions *a-g* in the heptad sequence by aligning the binary pattern of each coiled-coil to the ideal heptad sequence. Fig. 4.6(a)-(g) shows the difference between the unfolding temperatures of the wild-type, T^{WT}, and the mutant,T^{M}, for each of the positions (*a-g*) in the heptad repeat respectively. Among all the mutations at *a-g* positions the ones at the *a* and *d* positions led to the greatest differences in unfolding temperature compared to those that were at other positions *b-c-e-f-g*.

Most of these mutations are at *a* and *d* positions so just comparing the mean and the variance is not a good tool for comparison. To have a better view of how similar these results are, we compare the distributions of the results using the *p*-value for each pair of them. The results of these calculations are presented in Table 4.3. These values are calculated using R, a programming language and software environment for statistical computing and graphics [92]. The smaller the *p*-value, the fewer similarities between the distributions. The *p*-value confirmed most similarity between the distribution of the results of the mutations at *a* and *d* positions.

In addition to the positions of the mutations with respect to the alignment to the heptad repeat, the types of the neighbouring residues are also important. If one or both of the neighbours are hydrophobic then the change in T^{M} upon mutations of H to P would be greater. We also found that the position of the mutated residue along the protein may influenced the change of unfolding temperature. To show this, we divided each protein into three equal segments and we calculated the average unfolding temperature over all the mutated sequences where the mutations were at the *a* and *d* positions. We found that all the structures that have helices longer than 50 residues showed less stability when the mutations occurred in the middle segment, see Fig. 4.6(h). Having different results for

Mutated position	g	f	e	d	c	b	a
a	1.1e-4	3.0e-7	0.26	0.79	7.2e-3	1.2e-7	1
b	0.66	0.19	6.1e-3	7.2e-8	0.50	1	
c	0.76	0.14	0.075	0.010	1		
d	1.6e-4	4.4e-7	0.32	1			
e	2.2e-2	1.0e-3	1				
f	0.14	1					
g	1						

Table 4.3: Higher p-value shows more similarity between two distributions.

shorter helices could be due to their different deformations for each of the helices in their coiled-coil structure (1DEB), the high bending energy (1T6F) or the lower numbers of hydrophobic residues (2ZTA).

We now consider our results in the light of one particular structure, namely Vimentin (3G1E). Experimentally it was found that vimentin could only be crystalized if a single mutation of polar TYR117 to hydrophobic LEU117 is made [67]. This results in an increased structural stability of the mutant over the wild type sequence, which possesses a polar residue making it unstable under crystallizing conditions. With respect to calculating the effect of the mutation of the Vimentin sequence in our model, we used the optimal contact radius, $R_{i,j} = 6.3$, to generate structures for the mutated and unmutated sequences. In our own model we see a significant difference in unfolding temperatures when we change TYR117 (polar residue) to LEU117 (hydrophobic residue) in the vimentin sequence; their unfolding temperatures are 3.04 and 3.78 respectively. This means the unfolding temperature is increased 24% by mutation. Hence our model confirms the increased stability observed from this single mutation. Fig. 4.7 shows the RMSD between the mutated vimentin (3G1E) and the simulated structures with natural sequence versus different W_D's. As shown in the figure the RMSD of the structure with LEU117 (hydrophobic residue) is lower than that of the structures with TYR117 (polar residue). This shows the structures with hydrophobic residue at position 117 are more stable and they are closer to the crystal structure of this mutated sequence.

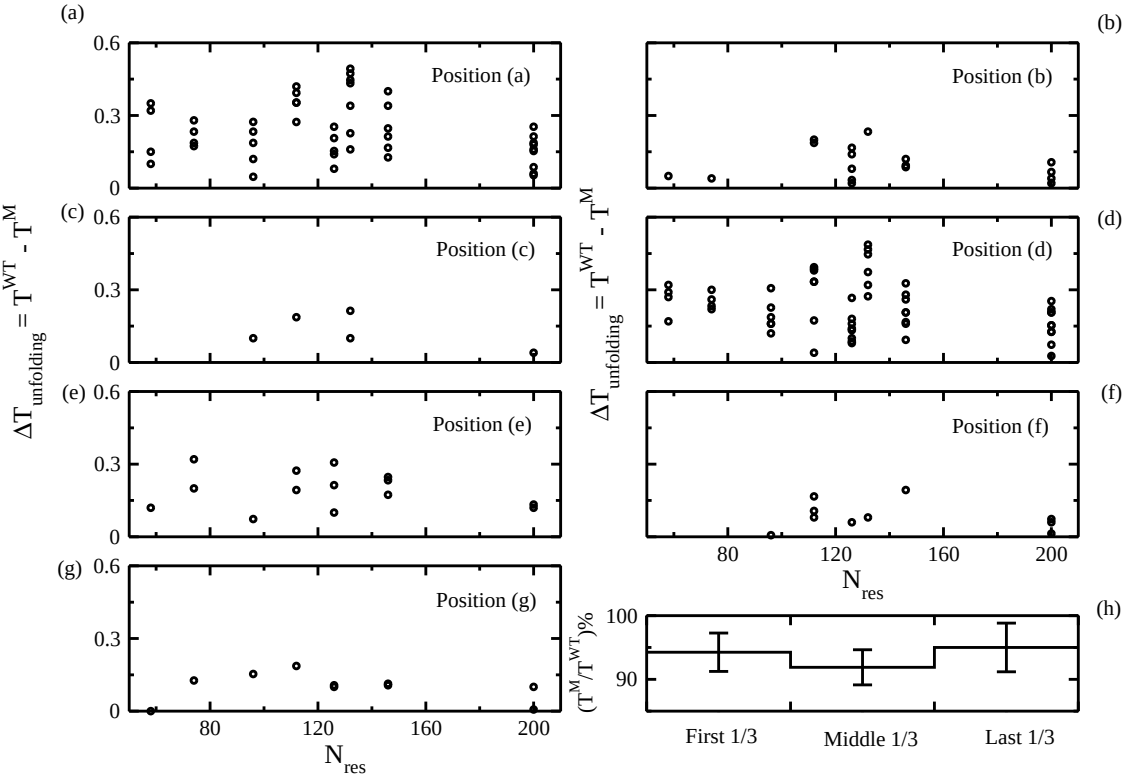

Figure 4.6: (a)-(g) The differences between the unfolding temperature of the wild-type, T^{WT}, and the mutated type, T^{M}, for each of the *a-g* positions in the heptad repeat over all the structures. (h) The average percentage of unfolding temperature for each segment over all the structures longer than 50 residues per helix.

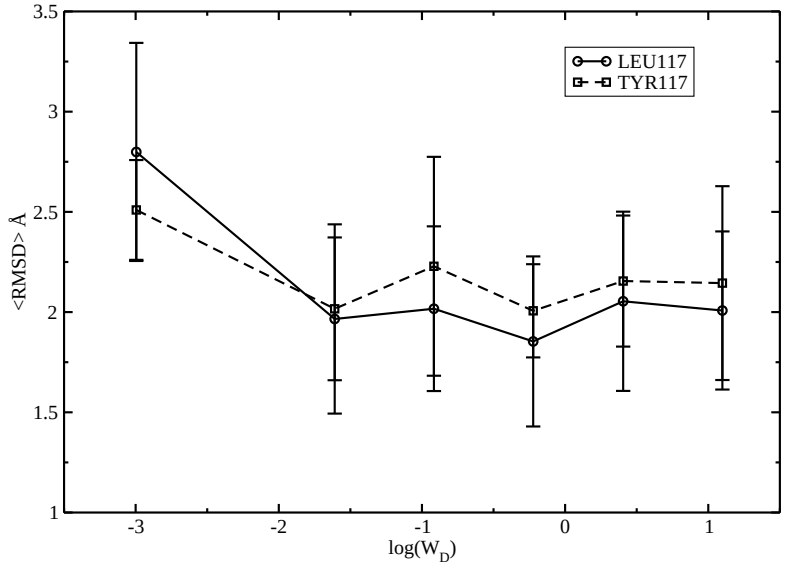

Figure 4.7: Average RMSD for vimentin (3G1E) is shown in the graph for unmutated and mutated sequences. Structures with polar TYR117 (unmutated) in their sequence have larger RMSD compared to those with a mutation to hydrophobic LEU117 (mutated).

4.4 Conclusions

In this chapter we have shown that a simple energy function built from a competition between the hydrophobic energy gained by burying patterned hydrophobic residues and the cost of deforming helices from their ideal structure can be constrained to generate known coiled-coil structures. By fitting known coiled-coils, ranging from long dimeric coiled-coils to those consisting of three helices we found that there was a consistent weight for the contribution of the deformation energy to the overall energy of the coiled-coil structure. Thus the individual helices making up a broad range of different coiled-coils all possess similar intrinsic mechanical properties independent of sequence, arising solely from the hydrogen-bonded backbone of the α-helix that is the common building block among all coiled-coils. Remarkably, using only a coarse grained HP model with only a limited num-

ber of attractive spring and HP interactions for the sequence was sufficient to encode the chirality of the coil. Side-chain structure was explored and little improvement over using a constant contact radius was observed; more detailed model of side-chain structure is one of the future goals of this model that is not explored in this thesis.

Previously we have shown that the model can reproduce the known mechanical properties of leucine zipper coiled-coils [70], and here we show that it can capture the thermal stability of the selected structures. The unfolding temperature of coiled-coils increases with length for sequences that possess roughly equivalent hydrophobic content. Also we have studied the unfolding temperature for the same selected structures with mutated sequences. Our results confirmed that mutations are important, and that mutations at *a* and *d* positions have the most effect on the unfolding temperature for each structure. To compare our model to experimental results, we considered one particular case where mutating only a single residue can lead to significant changes in thermal stability, namely vimentin, where it could be crystallized by changing a single polar residues to hydrophobic. Our simulations confirmed this, showing that the mutated sequence possessing the hydrophobic residue had a higher unfolding temperature. So the model is able to predict if a given HP pattern can produce a coiled-coil structures and it also can explore the mechanical and thermal properties of the coiled-coil. In the next chapter we explore the designabilities and stability characteristics of structures by folding random HP-patterns, examining what sequence features are required to produce stable coiled-coils of arbitrary chirality.

Chapter 5

Designability of Coiled-coils

As explained earlier in Chapter 1, it is possible for vastly different sequences to adopt the same structure, yet have vastly different function. For example we know there are many different amino acid sequences that fold into left-handed coiled-coils. Why are left-handed coiled-coils are chosen by these sequences? Are they more stable compared to the other possible conformations for these sequences? To answer these questions we generated all possible structures that a set of flexible helices can form, in order to explore the selectivity of coil-forming sequences. Then, we explored different properties of the most desirable conformations versus the low designable ones.

Previous theoretical work has shown that in the space of possible structures only a few emerge as designable. The designability score of a structure is the number of sequences that have that structure as their energetic ground state. Those few structures that are highly designable have been shown to possess properties that are characteristic of naturally occurring proteins: mutationally stable, thermally stable and fast folders [55, 93, 94].

In the following sections we explain how we generated the structure space formed by flexible helices and how we compared different folds with high designability scores with the ones with the low designability scores. In addition to confirming that designable coiled-coil structures are thermally stable we show that mechanical stability also correlates strongly with designability. Our results show that coiled-coils emerge naturally as a highly designable fold, possessing qualities that would allow them to be one of the most ubiquitous folds seen in nature.

5.1 Generating Structure Space

To explore the designability of coiled-coil structures, we chose to generate the space of possible structures that two helices of a specific length can form. We are interested in the differences between the designability scores and the thermal and mechanical stabilities of the conformations in the structure space. To do so, we need to have a complete set of different configurations such as left handed, right handed and two straight helices. Through the rest of the chapter we use structure and conformation terms interchangeably.

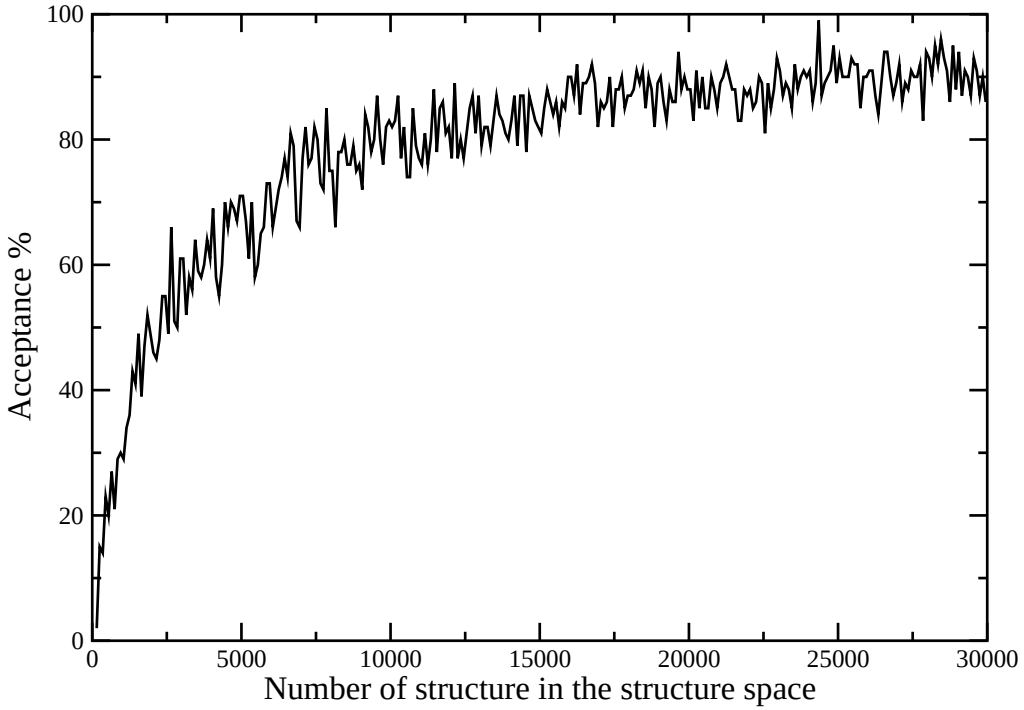

Figure 5.1: The completion percentage is defined as the fraction of newly generated structures with RMSD<1.5Å compared to all the structures in the structure space

Here we focus on structures with two helices with 49 resides each. We chose this length because at this length, as shown in Chapter 3 and 4, left-handed coiled-coils show a clear

chirality. Right-handed coiled-coils also emerge at this length. This length also allows for helices to pack in a nearly straight configuration. Shorter than this length it can be difficult to assess chirality of the helical pack and longer lengths lead to computational challenges due to the size of the space that needs to be sampled. The key point of generating structures using our model is minimizing the ground state energy for a conformation by burying the hydrophobic residues between the helices and bending the helices. To make the structures in the structure space, we made random HP sequences where the hydrophobic content of the helices varies between 25% and 55%. This range covers the hydrophobic content of the experimentally determined structures. Then we used our simulated annealing method to generate a structure. The key is to have a structure generation method that will broadly sample low energy configurations.

As structures are generated and added to the structure space, we assess its completeness in the following way: 1) We find the smallest RMSD for each of the most recently generated 100 structures to all previous structures in the structure space. For a given newly generated structure, if RMSD<1.5Å we assume that there is a similar structure to the new one in the set, and if RMSD>1.5Å the newly generated structure has no similar structure. 2) For each 100 new structures we count the number of structures with RMSD<1.5Å and report this number as completion percentage. The newly generated structures become part of the structure space for the next step.

As shown in Fig. 5.1, the completion percentage increases when there are more structures in the structure space. New structures are generated until we get completion percentages higher than 85%. This means more than 85% of the newly generated configurations have at least one similar structure in the structure (RMSD<1.5Å) space and fewer than 15% of them are new. We assume the structure space is complete when that completion percentage is higher than 85%. Using this method we needed 30,000 structures to complete the structure space.

The structure space contains a variety of double helix conformations; right-handed coiled-coils, left-handed coiled-coils and two straight helices are the major categories in the structure space. Fig. 5.2 shows one representative for each of these categories both from side view and central axis view. When helices have higher hydrophobic content (more than 50%) or all the hydrophobic residues are in one region, the helices tend to make straight conformations (See Fig. 5.2(c)). In these cases hydrophobic residues are facing towards

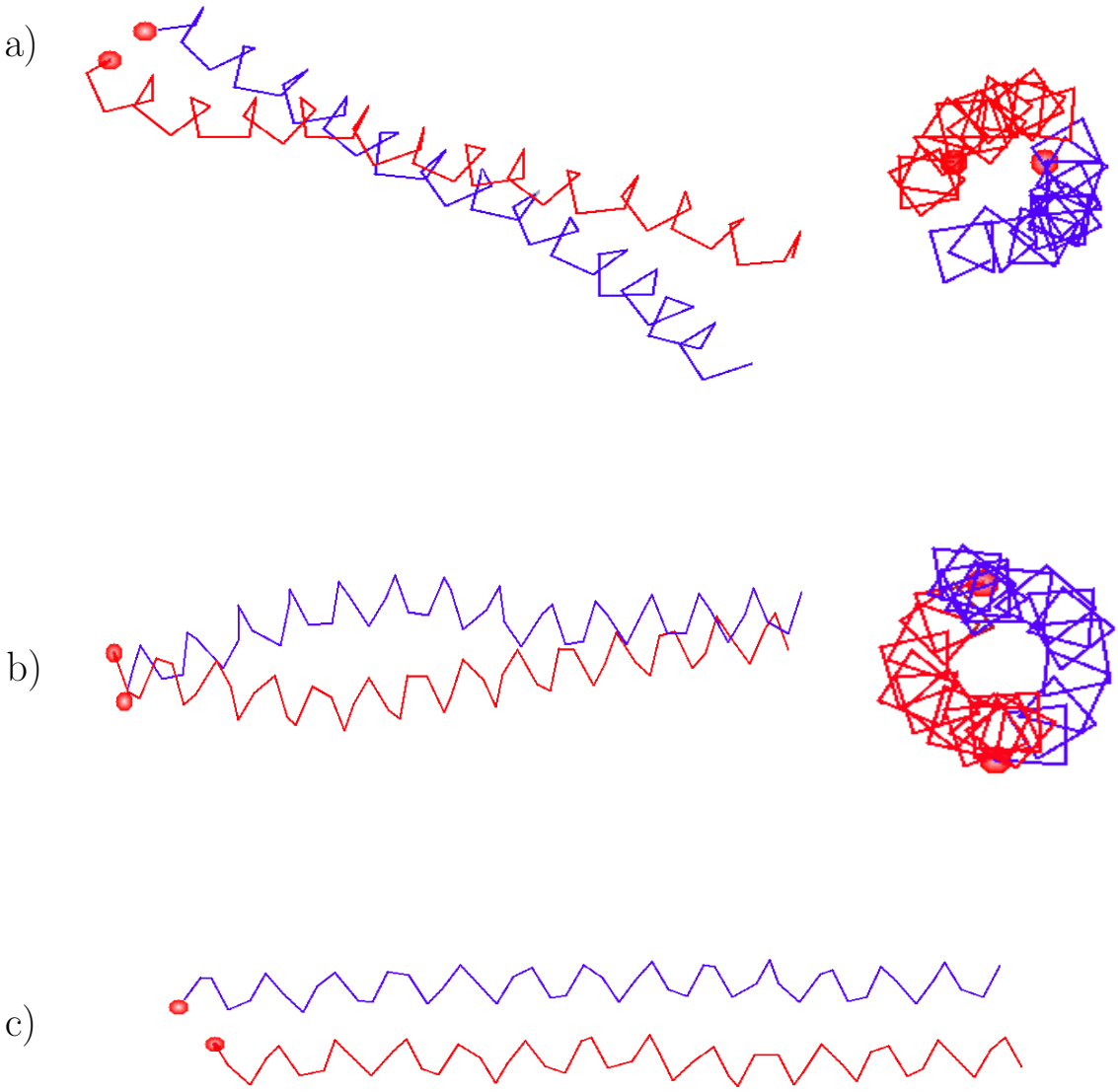

Figure 5.2: Some representative conformations are shown in this Figure. To have a clear view both side view and the central axis view are shown here for (a) left-handed coiled-coils, (b) right-handed coiled-coils, and the side view of (c) two straight helices

each other so the helices do not bend to pack them.

5.2 Designability Scores

As mentioned before, the designability score of a structure is defined as the number of sequences that have that structure as the ground state over all the different conformations in a structure space. To find the designability scores for all the different conformations in the structure space, we generated 5,000,000 random sequences of length 49 with 25% to 55% hydrophobic content. Then for each random sequence we used our energy function to search structure space for the one that has the lowest energy for that sequence. The number of sequences assigned to each structure is called the designability score.

More than half of the conformations in the structure space were not chosen by any sequence among all 5,000,000 sequences. Just 15 conformations out of 30,000 conformations in the structure space had designability scores higher than 15,000 and the highest designable structure was the ground state of more than 38,000 sequences. Overall these 15 structures were designed by $\sim 6\%$ of the sequences.

Since many of the structures in the structure space are highly similar, and therefore compete for sequences, in order to more accurately assess designability it is necessary to perform clustering. By clustering we will assign all similar structures to a single cluster and add up their designabilty scores to arrive at a score representative of that particular fold. There are many different methods to cluster these conformations; here we use a simple clustering method that we will discuss in the next section.

5.3 Clustering the Conformations

We use a greedy algorithm to cluster structure space. To cluster the structures we start from the conformations with the highest designability and include all similar conformations (RMSD<1.5Å) in that cluster. The new designability score for the whole cluster is the sum of all the designability scores of each individual conformation in that cluster. All the conformations in the first cluster are excluded from the structure space and we move on to the next most designable structure and repeat the process, creating the next cluster. This is done until all the conformations are clustered.

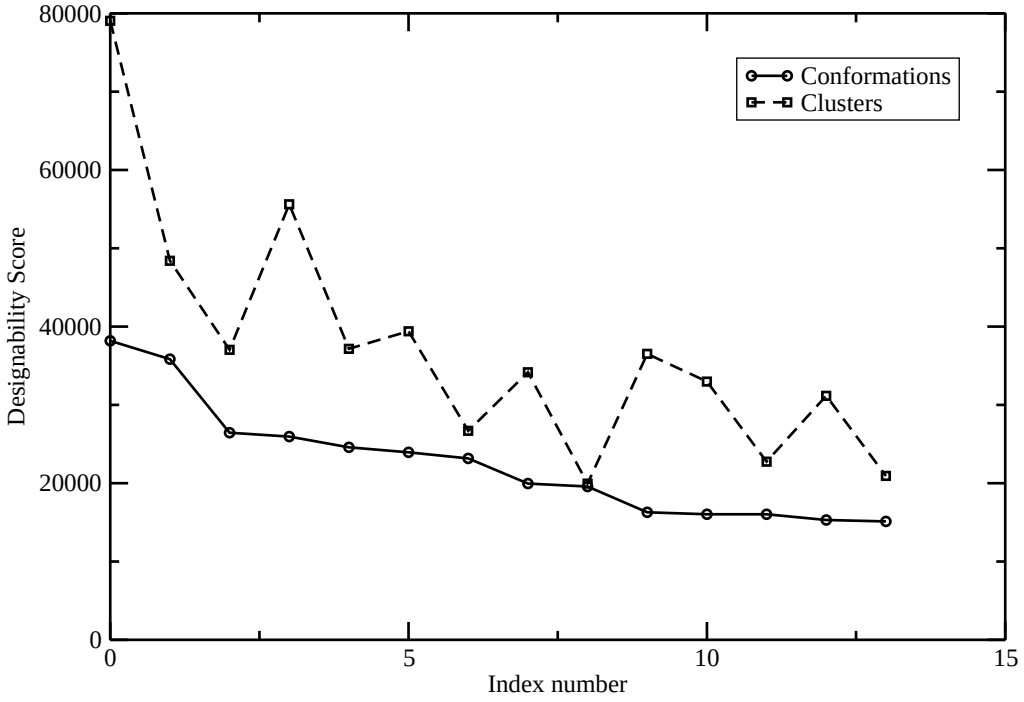

Figure 5.3: The designability scores for some of the high designable structures and the designability scores of the clusters that they represent.

The number of conformations with at least one sequence (designability score=1) is $\sim$ 15,000 which resulted in $\sim$ 4,500 clusters. Obviously the designability scores of the clusters are higher than the designability score of each conformation in that cluster. Fig. 5.3 shows the 15 highest designable conformations and the designability scores of the clusters that they belong to. As shown in Fig. 5.3 the order of the highest designable clusters is not the same as the order of highest designable conformations. As a test of the clustering method, we generate the average HP sequence for all the sequences in the cluster, use simulated annealing to fold it, and check that the resulting structure falls within RMSD of 1.5Å of the cluster. More than 98% of the new structures belong to the clusters that the average sequence was calculated from.

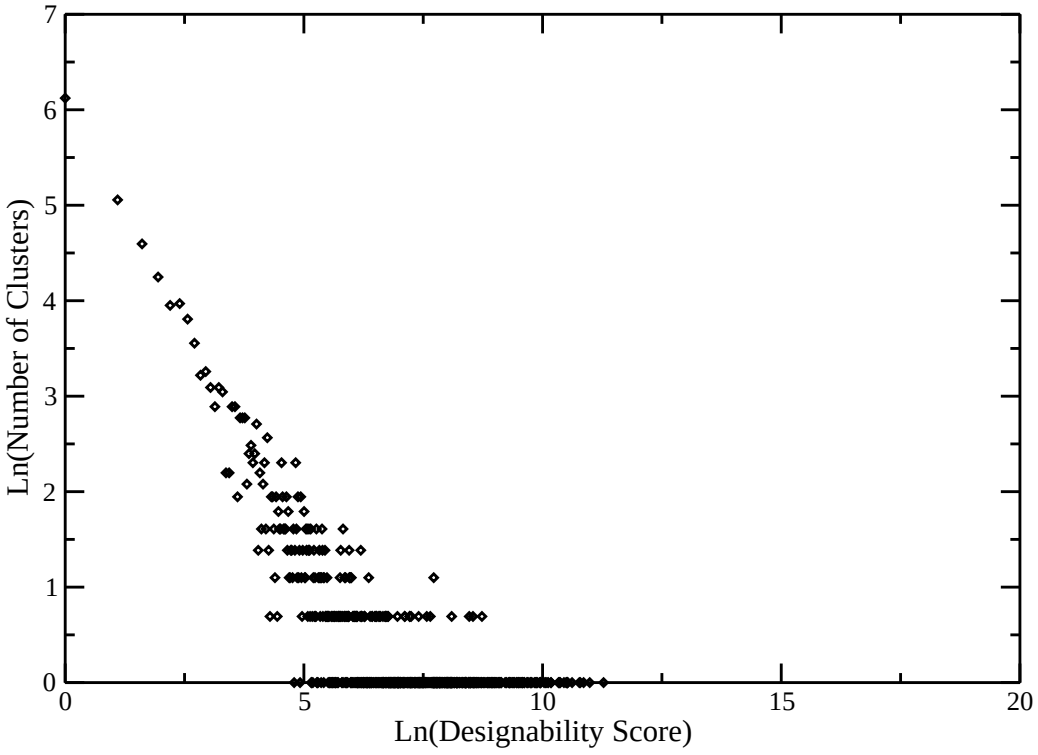

Figure 5.4: The histogram of clusters according to their designability scores in a $\ln - \ln$ graph.

To have a better understanding of designability scores and the number of conformations with high designability scores, we plot the histogram of the structures according to their designability scores on a double ln scale graph (See Fig. 5.4). Only a few structures are highly designable and the number of structures with lower designability scores are higher than the structures with higher designability scores. We consider the structures with ln (Designability Score)> 9 as highly designable structures. This elaborate on 'power law' behaviour has been observed repeatedly in studies of the designability of protein structures using a range of different structure and folding models [94].

5.4 Experimentally Determined Structures in the Structure Space

Each cluster has an average sequence over all the sequences that were assigned to each of the conformations in that cluster. To find the average sequence for a cluster we determine the probability of finding a hydrophobic residue at each of the 1-49 positions using the sequences that designed that structure. The 19 residues with the highest probability are assumed to be hydrophobic and the remaining residues are assumed to be polar; the resultant sequence is the average sequence for the cluster. We choose 19 residues because with 19 hydrophobic residues we ended up with 38% hydrophobic content which is the average hydrophobic content over the experimentally determined structures that we studied in previous chapters and it is close to the hydrophobic content of GCN4.

To check the correlation between the highly designable clusters and the experimentally determined structures we make two comparisons: first we compare the RMSD between the structures and the clusters and find the best matching clusters, and then we compare the average sequence of that cluster with the sequence of the experimentally determined structures. All the results showed the selected clusters by RMSD which are close to the experimentally determined structures have similar average sequence.

5.4.1 GCN4 and Heptad repeat in the Structure Space

GCN4 is the canonical leucine zipper protein whose sequence is based on the heptad repeat. The first question that we would like to answer here is 'what is the designability score of the similar structure to the heptad repeat or GCN4?'. Also we would like to find 'How similar are the average sequences of the similar clusters to the sequence of the heptad repeat or GCN4?'

As explained in Chapter 3 we simulated several GCN4-like coiled-coil structures. Here we compare simulated structures for a GCN4 sequence made with 49 residues to the structure space. The results for some of the representatives of simulated structures with GCN4 sequence are presented in the Table 5.1. These sequences are listed in the Appendix under the name of 2ZTA which is one of the proteins of GCN4 family.

Sim. Struc.	$RMSD_{min}$	Number of Seq.	Alignment score to GCN4
1	0.9358	26525	89%
2	0.9850	14599	89%
3	0.9060	8514	91%

Table 5.1: Three different representatives of simulated structures with GCN4 sequence are listed in this table followed by the minimum RMSD between the simulated structure and the conformations in the cluster. The rest of the columns give the overall designability score for the cluster and the alignment score between the average sequence of the cluster and the GCN4 sequence.

As it is shown in Table 5.1 all the clusters which are similar to GCN4 sequence are among the highly designable clusters. We define the high designable clusters as clusters that are chosen by more than 5000 sequences as their ground state conformation. This shows the left-handed coiled-coil structure is one of the stable conformations among all different conformations in the structure space. In the last column of Table 5.1 the alignment score between the average sequence of each cluster and the sequence of GCN4 by ClustalW2 is given [95]. This shows that the clusters that have structural similarity to GCN4 also have sequences that have strong similarity to the naturally occurring heptad repeat sequence.

We did the same analysis for the coiled-coil structures with the ideal heptad repeat sequence and we figured out the closest clusters to the heptad repeat are among the highly designable structures as well. The results for these calculations are presented in Table 5.2.

Sim. Struc.	$RMSD_{min}$	Number of Seq.	Alignment score to ideal heptad repeat
1	0.9218	26968	89%
2	0.9624	17176	89%
3	1.183	12456	91%

Table 5.2: Three different representatives of simulated structures with ideal heptad sequence are listed in this table followed by the minimum RMSD between the simulated structure and the conformations in the cluster. The rest of the columns give the overall designability score for the cluster and the alignment score between the average sequence of the cluster and the ideal heptad repeat.

5.5 Comparing Mechanical and Thermal Stability of the Conformations

To check the stability of highly designable conformations in the structure space versus the stability of the low designable conformations we use our methods as explained in Chapter 2 to find the transition temperature and the transition force.

Ten different conformations were chosen as representatives for each category of high designable, average designable and low designable clusters. We define high designable clusters as the ones with ln(Designability Score) > 9 , average designable as the ones with 2 < ln(Designability Score) < 9 and the low designable ones as the ones with ln(Designability Score) < 2. Low designable structures are defined as the structures that are chosen by the random sequences less than five times. The limit for the high designable structure is defined using the structures which are close to the natural left handed helices [see Table 5.3]. The designablity scores reported in the Table 5.3 is for the clusters that these structures belong to; these structures are the highest designable structures out of their own cluster. Then to find the transition force, we applied constant force perpendicular to the central axis of the structure. The sequence we chose to study for each unfolding simulation was the average sequence for that structure. The average HP sequence for each chosen structure was calculated as described above. To be consistent, for all the conformations we looked for low designable structures with the same hydrophobic content (19 out of 49 or 38%), since the melting temperature and the transition force have direct dependency on the hydrophobic content of the helices.

The transition forces for each of the conformations was found over ten times pulling and the average transition force for each category is given in Table 5.3. We found that the transition force for the designable structures, that includes high designable and the average designable structures is higher than that of the low designable structures. Obviously other factors play roles in determining the transition force such as the HP pattern of the sequence for each conformation. When the hydrophobic residues are accumulating in one part of the helices, higher force is required to break the hydrophobic bonds; specially when this

happens at the end where the force is applied.

To find the unfolding temperature or melting temperature, the temperature is increased until all the hydrophobic bonds are broken. As in Chapter 3, the temperature at which this occurs is recorded as the unfolding temperature of the conformation. A trend was observed for the unfolding temperature that indicates that the unfolding temperature of the high designable structures is higher than that of the average designable structures which is higher than that of low designable structures. This proves that the structures with higher designability scores are more thermally stable compare to the ones with low designablity scores. All these results are reported in Table 5.3.

The transition forces for the high and the average designable structures are higher compared to the low designable conformations. This shows the low designable conformations are less stable compared to the other two categories. The unfolding temperature analysis shows that the low designable category is also thermally less stable compared to the average designable category which itself is less stable compared to the highly designable category.

5.6 Discussion

In this chapter we have extended our study from known structures to the structure space that includes the known flexible conformations, coiled-coils. We have shown that the designability for this very specific protein fold emerges as the top designable fold of possible two helix conformations. In our control checks we found that two experimentally known structures are among the highly designable structures; besides, the sequence of each of these structures was similar to the average sequence of the cluster that the experimentally determined structure belongs to.

We should emphasize that more than half of the conformations in the structure space do not correspond to a ground state of even one random sequence out of 5,000,000. Also we showed the highly designable structures are both thermally and mechanically more stable compared to conformations with lower designability scores. Obviously the mechanical properties of a protein is not the only reason to have the left-handed coiled-coils but our results showed that is one of the main constraints.

Conformations			Trans. Force		Trans. Temperature	
Design. Category I	Design. Category II	Design. Score of Clusters	F	<F>	T	<T>
Designable structure	High Designable	76427	6.7 ± 0.3	7.0 ± 2.0	3.5	3.5± 0.2
		37039	7.6 ± 0.1		3.6	
		39420	4.7 ± 0.2		3.4	
		34168	10.9± 1.5		3.8	
		59143	8.1 ± 0.9		3.9	
		37161	6.0 ±0.2		3.5	
		48545	10.1 ± 0.2		3.7	
		26688	5.5 ± 0.6		3.7	
		19944	5.2 ± 0.5		3.0	
		36519	5.7 ± 0.2		3.5	
	Average Designable	12	8.5 ± 0.7	7.8 ± 1.8	2.8	3.1±0.3
		2929	11.5 ± 0.9		2.8	
		181	8.1 ± 0.5		3.7	
		620	5.1 ± 0.4		2.9	
		51	8.1 ± 1.9		3.2	
		89	8.2 ± 0.3		3.2	
		3711	5.8 ± 0.3		3.4	
		1042	8.1 ± 0.1		3.1	
		756	8.7 ± 0.2		3.1	
		80	6.2 ± 0.2		2.6	
Low Designable		1	6.1 ± 0.9	5.7 ± 1.41	2.3	2.67±0.22
		1	7.7 ± 0.5		2.7	
		1	4.9 ± 1.0		2.7	
		1	5.9 ± 0.3		2.9	
		1	3.9 ± 0.3		2.3	
		1	8.0 ± 1.2		2.6	
		1	5.5 ± 0.93		2.7	
		1	5.6 ± 0.3		2.7	
		1	5.2 ± 0.2		2.8	
		1	4.6 ± 0.3		3.0	

Table 5.3: Transition force and unfolding temperature value for three different categories of conformations are listed here. All the sequences of all the conformations have the same hydrophobic content.

Furthermore the relation between the sequences and the designability scores of the structures should be explored. We are looking to explore the properties of the sequences of the high designable structures. Why do the structures made of these sequences end up with similar conformations? For example we have never seen that we start from a heptad repeat and we get to a right handed coiled-coil. Besides the conformations made of the sequences of the low designable structures are not similar (large RMSD).

We have mentioned earlier that the generated structure space covers all variety of the conformations made of two helices including right handed and left handed. Also we have seen that each of these two category has conformations which are among the high designable ones. What is so special about the left handed sequences that nature prefers to choose those ones?

For future studies we need to do more quantitative analysis on the correlation between the complexity of a sequence and the structure that is chosen by that sequence among all the structures in the structure space. For example the sequences with many hydrophobic residues within the neighbourhood of each other (like HHHHPPPPP) maps to structures with lower designability scores while the more complex sequences (like HPPHHPHHP) is choosing higher designable structures. Quantifying the complexity in a sequence may be predictive of the designability of the structure it folds into

Chapter 6

Conclusions

In this thesis we have introduced a new simple model to generate one specific fold, coiled-coils, by using the HP translation of their sequence through simulated annealing MC. We are able to generate these structures just by defining two terms in our energy function: deformation energy and hydrophobic energy. We have also determined how to weight these two energies with respect to each other. We have shown that our model is able to regenerate the experimental results [63] of applying transverse load to the axis of the coiled-coil.

Because of the chiral nature of coiled-coil structure, using experimental approaches such as magnetic tweezers that allow for the application of torque will be of interest. Magnetic tweezers have been used to apply torque to study the winding of DNA [96, 97], and they could be applied to coiled-coils to study both the uncoiling and supercoiling of the superhelical structure. Incorporating torque is relatively straightforward in our model by adding the work done by the torque to the energy function.

Naturally occurring coiled-coils possess sequence disorder, namely, there are positions along the sequence that differ from the repeating heptad unit. Prior work has made progress on identifying how substitutions affect the stability of the resulting fold in specific situations [67]. Using our model it should be possible to quantify how the distribution of disorder affects the stability and kinetics of coiled-coil formation. Preliminary results suggest that strategically placed sequence disorder can drastically speed up the kinetics of folding, allowing the coiled-coil to escape otherwise low-lying energetic traps. It will be interesting to see if naturally occurring coiled-coils exploit sequence disorder to alter the kinetics, besides just affecting stability.

We have also extended our studies from one specific natural structure to several representative of different superfamilies of the SCOP database. By fitting known coiled-coils, ranging from dimeric coiled-coils to those consisting of three helices we found that there was a consistent weight for the contribution of the deformation energy to the overall energy of the coiled-coil structure. Thus the individual helices making up a broad range of different coiled-coils all possess similar intrinsic mechanical properties independent of sequence arising solely from the hydrogen bonded backbone of the α-helix that is the common building block among all coiled-coils. Remarkably, using only a coarse hydrophobic-polar pattern with only a limited number of attractive interactions for the sequence was sufficient to encode the chirality of the coil. Side-chain size was explored and little improvement over using a constant contact radius was observed.

We showed that our model can capture the thermal stability of the selected structures. The unfolding temperature of coiled-coils increases with length for sequences that possess roughly equivalent hydrophobic content. Also we have studied the unfolding temperature for the same selected structures with mutated sequences. Our results confirmed that mutations are important and that mutations at a and d positions have the most effect on the unfolding temperature for each structure. To compare our model to experimental results, we considered one particular case where mutating only a single residue can lead to significant changes in thermal stability, namely vimentin, where it could be crystallized by changing a single polar group. Our simulations confirmed this, showing that the mutated sequence possessing the hydrophobic residue had a higher unfolding temperature. Thus the model can rapidly assess whether an HP pattern can produce a coiled-coil structure and what thermal, and as shown previously, mechanical properties such a sequence may possess.

Future work will consider how changes in sequence can affect the kinetics of folding, as too little or too much hydrophobic content on top of the chiral hydrophobic pattern may serve to create kinetic barriers. Indeed, recent experimental work has examined the effect on the kinetics of coiled-coil formation based on sequence changes [65]. The model will also be applied to address the question of whether there is a chirality for patterning hydrophobic residues on each helix surface that leads to the greatest thermal stability. We also will use the model to explore the assembly of multi-helix solutions to address the important question of dimeric to trimeric or multi-meric assembly. We expect the model

to be able to identify the key sequence requirements for both fast kinetics and thermal stability.

In the last chapter, we focused on more general questions of the folding of coiled-coils. In particular we calculated the designability of two helix packs and we showed some structures are more designable compared to the rest of the structures in the structure space. In other words some structures are the ground state of more sequences than other structures. We showed that the ideal heptad repeat and the GCN4 sequence are among the high designable structures. We also shown that highly designable structures are more statistically stable both mechanically and thermally. This is a novel finding, for it is the first time that mechanical stability has been related to designability of a structure for a realistic off-lattice model. Thus in the case of coiled-coils, their mechanical stability and ability to withstand loads would have naturally emerged as a result of them being highly designable. Future work will look at the kinetics of the folding of flexible helical structures as a function of their designability, as prior work has shown that highly designable folds are also fast folders [55, 98, 35]. This thesis has laid the foundation to explore some of these relationships given the ability to quickly generate a vast space of structures and assess the folding of sequences into their respective targets.

Bibliography

[1] R. Vale, T. Reese, and M. Sheetz, "Identification of a novel force-generating protein, kinesin, involved in microtubule-based motility," *Cell*, vol. 42, no. 1, pp. 39–50, 1985.

[2] P. Chou and G. Fasman, "Prediction of protein conformation," *Biochemistry*, vol. 13, no. 2, pp. 222–245, 1974.

[3] M. Liang and S. Andreadis, "Engineering fibrin-binding tgf-β1 for sustained signalling and contractile function of msc based vascular constructs," *Biomaterials*, vol. 32, no. 33, pp. 8684–8693, 2011.

[4] T. Miyata, T. Taira, and Y. Noishiki, "Collagen engineering for biomaterial use," *Clinical materials*, vol. 9, no. 3-4, pp. 139–148, 1992.

[5] "Protein data bank."

[6] M. Blaber, X. Zhang, and B. Matthews, "Structural basis of amino acid alpha helix propensity," *Science*, vol. 260, no. 5114, p. 1637, 1993.

[7] D. Minor Jr and P. Kim, "Measurement of the β-sheet-forming propensities of amino acids," *Nature*, vol. 367, no. 6464, pp. 660–663, 1994.

[8] A. Chakrabartty, T. Kortemme, and R. Baldwin, "Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions.," *Protein science: a publication of the Protein Society*, vol. 3, no. 5, p. 843, 1994.

[9] A. Chakrabartty, J. Schellman, and R. Baldwin, "Large differences in the helix propensities of alanine and glycine," *Nature*, vol. 351, pp. 586–588, 1991.

[10] S. Marqusee, V. Robbins, and R. Baldwin, "Unusually stable helix formation in short alanine-based peptides," *Proceedings of the National Academy of Sciences*, vol. 86, no. 14, p. 5286, 1989.

[11] D. Woolfson, "The design of coiled-coil structures and assemblies," *Advances in protein chemistry*, vol. 70, pp. 79–112, 2005.

[12] M. Lantz, S. Jarvis, H. Tokumoto, T. Martynski, T. Kusumi, C. Nakamura, and J. Miyake, "Stretching the α-helix: a direct measure of the hydrogen-bond energy of a single-peptide molecule," *Chemical Physics Letters*, vol. 315, no. 1-2, pp. 61–68, 1999.

[13] M. Rief, J. Pascual, M. Saraste, and H. Gaub, "Single molecule force spectroscopy of spectrin repeats: low unfolding forces in helix bundles," *Journal of Molecular Biology*, vol. 286, no. 2, pp. 553–561, 1999.

[14] X. Zhuang and M. Rief, "Single-molecule folding," *Current opinion in structural biology*, vol. 13, no. 1, pp. 88–97, 2003.

[15] J. Petersen, J. Skalicky, L. Donaldson, L. McIntosh, T. Alber, and B. Graves, "Modulation of transcription factor Ets-1 DNA binding: DNA-induced unfolding of an alpha helix," *Science*, vol. 269, no. 5232, p. 1866, 1995.

[16] F. Berkemeier, M. Bertz, S. Xiao, N. Pinotsis, M. Wilmanns, F. Gräter, and M. Rief, "Fast-folding α-helices as reversible strain absorbers in the muscle protein myomesin," *Proceedings of the National Academy of Sciences*, 2011.

[17] K. Svoboda and S. Block, "Biological applications of optical forces," *Annual Review of Biophysics and Biomolecular Structure*, vol. 23, no. 1, pp. 247–285, 1994.

[18] K. Svoboda, C. Schmidt, B. Schnapp, and S. Block, "Direct observation of kinesin stepping by optical trapping interferometry," *Nature*, vol. 365, no. 6448, pp. 721–727, 1993.

[19] G. Oukhaled, J. Mathe, A. Biance, L. Bacri, J. Betton, D. Liarez, J. Pelta, and L. Auvray, "Unfolding of proteins and long transient conformations detected by single nanopore recording," *Physical Review Letters*, vol. 98, no. 15, p. 158101, 2007.

[20] K. Neuman and A. Nagy, "Single-molecule force spectroscopy: optical tweezers, magnetic tweezers and atomic force microscopy," *Nature Methods*, vol. 5, no. 6, pp. 491–505, 2008.

[21] S. Weiss, "Fluorescence spectroscopy of single biomolecules," *Science*, vol. 283, no. 5408, p. 1676, 1999.

[22] M. Rief, M. Gautel, F. Oesterhelt, J. Fernandez, and H. Gaub, "Reversible unfolding of individual titin immunoglobulin domains by afm," *Science*, vol. 276, no. 5315, p. 1109, 1997.

[23] L. Tskhovrebova, J. Trinick, J. Sleep, and R. Simmons, "Elasticity and unfolding of single molecules of the giant muscle protein titin," *Nature*, vol. 387, no. 6630, pp. 308–312, 1997.

[24] J. Zhou, J. Hertz, A. Leinonen, and K. Tryggvason, "Complete amino acid sequence of the human alpha 5 (iv) collagen chain and identification of a single-base mutation in exon 23 converting glycine 521 in the collagenous domain to cysteine in an alport syndrome patient," *Journal of Biological Chemistry*, vol. 267, no. 18, p. 12475, 1992.

[25] C. Wolgemuth and S. Sun, "Elasticity of α-helical coiled-coils," *Physical Review Letter*, vol. 97, no. 24, p. 248101, 2006.

[26] W. Humphrey, A. Dalke, and K. Schulten, "VMD: visual molecular dynamics," *Journal of Molecular Graphics*, vol. 14, no. 1, pp. 33–38, 1996.

[27] J. Onuchic and P. Wolynes, "Theory of protein folding," *Current Opinion in Structural Biology*, vol. 14, no. 1, pp. 70–75, 2004.

[28] M. Karplus, "Molecular dynamics simulations of biomolecules," *Accounts of Chemical Research*, vol. 35, no. 6, pp. 321–323, 2002.

[29] M. Buehler, "Nature designs tough collagen: explaining the nanostructure of collagen fibrils," *Proceedings of the National Academy of Sciences*, vol. 103, no. 33, p. 12285, 2006.

[30] A. Linhananta, J. Boer, and I. Mackay, "The equilibrium properties and folding kinetics of an all-atom Gō model of the Trp-cage," *Journal of Chemical Physics*, vol. 122, p. 114901, 2005.

[31] E. Emberly, R. Mukhopadhyay, N. Wingreen, and C. Tang, "Flexibility of α-helices: results of a statistical analysis of database protein structures," *Journal of Molecular Biology*, vol. 327, no. 1, pp. 229–237, 2003.

[32] J. McCammon, B. Gelin, and M. Karplus, "Dynamics of folded proteins," *Nature*, vol. 267, no. 5612, pp. 585–590, 1977.

[33] A. Sali, E. Shakhnovich, and M. Karplus, "Kinetics of protein folding a lattice model study of the requirements for folding to the native state," *Journal of Molecular Biology*, vol. 235, pp. 1614–1636, 1994.

[34] A. Sali, E. Shakhnovich, and M. Karplus, "How does a protein fold?," *Nature*, vol. 369, no. 6477, pp. 248–251, 1994.

[35] R. Helling, H. Li, R. Mélin, J. Miller, N. Wingreen, C. Zeng, and C. Tang, "The designability of protein structures," *Journal of Molecular Graphics and Modeling*, vol. 19, no. 1, pp. 157–167, 2001.

[36] E. Shakhnovich and A. Gutin, "Engineering of stable and fast-folding sequences of model proteins," *Proceedings of the National Academy of Sciences*, vol. 90, no. 15, p. 7195, 1993.

[37] Y. Ueda, H. Taketomi, and N. Gō, "Studies on protein folding, unfolding, and fluctuations by computer simulation. II. a. three-dimensional lattice model of lysozyme," *Biopolymers*, vol. 17, no. 6, pp. 1531–1548, 1978.

[38] A. Linhananta and Y. Zhou, "The role of side chain packing and native contact interactions in folding: discontinuous molecular dynamics folding simulations of an all-atom Gō model of fragment B of staphylococcal protein A," *Journal of Chemical Physics*, vol. 117, p. 8983, 2002.

[39] Y. Zhou and A. Linhananta, "Role of hydrophilic and hydrophobic contacts in folding of the second β-hairpin fragment of protein G: Molecular dynamics simulation studies

of an all-atom model," *Proteins: Structure, Function, and Bioinformatics*, vol. 47, no. 2, pp. 154–162, 2002.

[40] Z. Li and H. Scheraga, "Monte Carlo-minimization approach to the multiple-minima problem in protein folding," *Proceedings of the National Academy of Sciences*, vol. 84, no. 19, p. 6611, 1987.

[41] A. Kolinski and J. Skolnick, "Monte Carlo simulations of protein folding. i. lattice model and interaction scheme," *Proteins: Structure, Function, and Bioinformatics*, vol. 18, no. 4, pp. 338–352, 1994.

[42] U. Hansmann and Y. Okamoto, "New Monte Carlo algorithms for protein folding," *Current Opinion in Structural Biology*, vol. 9, no. 2, pp. 177–183, 1999.

[43] B. Dahiyat and S. Mayo, "De novo protein design: fully automated sequence selection," *Science*, vol. 278, no. 5335, p. 82, 1997.

[44] P. Harbury, J. Plecs, B. Tidor, T. Alber, and P. Kim, "High-resolution protein design with backbone freedom," *Science*, vol. 282, no. 5393, p. 1462, 1998.

[45] B. Kuhlman, G. Dantas, G. Ireton, G. Varani, B. Stoddard, and D. Baker, "Design of a novel globular protein fold with atomic-level accuracy," *Science*, vol. 302, no. 5649, p. 1364, 2003.

[46] J. Holton and T. Alber, "Automated protein crystal structure determination using ELVES," *Proceedings of the National Academy of Sciences*, vol. 101, no. 6, p. 1537, 2004.

[47] J. Walshaw and D. Woolfson, "SOCKET: a program for identifying and analysing coiled-coil motifs within protein structures," *Journal of Molecular Biology*, vol. 307, no. 5, pp. 1427–1450, 2001.

[48] O. Testa, E. Moutevelis, and D. Woolfson, "Cc+: a relational database of coiled-coil structures," *Nucleic Acids Research*, vol. 37, no. suppl 1, p. D315, 2009.

[49] E. Moutevelis and D. Woolfson, "A periodic table of coiled-coil protein structures," *Journal of Molecular Biology*, vol. 385, no. 3, pp. 726–732, 2009.

[50] H. Li, R. Helling, C. Tang, and N. Wingreen, "Emergence of preferred structures in a simple model of protein folding," *Science*, vol. 273, no. 5275, p. 666, 1996.

[51] "http://scop.mrc-lmb.cam.ac.uk/scop/index.html."

[52] X. Liu, K. Fan, and W. Wang, "The number of protein folds and their distribution over families in nature," *Proteins*, vol. 54, no. 3, pp. 491–499, 2004.

[53] Y. Wolf, N. Grishin, and E. Koonin, "Estimating the number of protein folds and families from complete genome data," *Journal of Molecular Biology*, vol. 299, no. 4, pp. 897–905, 2000.

[54] H. Li, C. Tang, and N. Wingreen, "Designability of protein structures: A lattice-model study using the Miyazawa-Jernigan matrix," *Proteins: Structure, Function, and Bioinformatics*, vol. 49, no. 3, pp. 403–412, 2002.

[55] R. Mélin, H. Li, N. Wingreen, and C. Tang, "Designability, thermodynamic stability, and dynamics in protein folding: a lattice model study," *Journal of chemical physics*, vol. 110, p. 1252, 1999.

[56] N. Buchler and R. Goldstein, "Effect of alphabet size and foldability requirements on protein structure designability," *Proteins: Structure, Function, and Bioinformatics*, vol. 34, no. 1, pp. 113–124, 1999.

[57] P. Wong, A. Fritz, and D. Frishman, "Designability, aggregation propensity and duplication of disease-associated proteins," *Protein Engineering Design and Selection*, vol. 18, no. 10, p. 503, 2005.

[58] E. Kussell, "The designability hypothesis and protein evolution," *Protein and Peptide Letters*, vol. 12, no. 2, pp. 111–116, 2005.

[59] H. Wendt, C. Berger, A. Baici, R. Thomas, and H. Bosshard, "Kinetics of folding of leucine zipper domains," *Biochemistry*, vol. 34, no. 12, pp. 4097–4107, 1995.

[60] L. Moran, J. Schneider, A. Kentsis, G. Reddy, and T. Sosnick, "Transition state heterogeneity in GCN4 coiled-coil folding studied by using multisite mutations and

crosslinking," *Proceedings of the National Academy of Sciences*, vol. 96, no. 19, p. 10699, 1999.

[61] D. Talaga, W. Lau, H. Roder, J. Tang, Y. Jia, W. DeGrado, and R. Hochstrasser, "Dynamics and folding of single two-stranded coiled-coil peptides studied by fluorescent energy transfer confocal microscopy," *Proceedings of the National Academy of Sciences*, vol. 97, no. 24, p. 13021, 2000.

[62] T. Bornschlögl and M. Rief, "Single-molecule dynamics of mechanical coiled-coil unzipping," *Langmuir*, vol. 24, no. 4, pp. 1338–1342, 2008.

[63] T. Bornschlögl and M. Rief, "Single molecule unzipping of coiled-coils: sequence resolved stability profiles," *Physical Review Letters*, vol. 96, no. 11, p. 118102, 2006.

[64] I. Schwaiger, C. Sattler, D. Hostetter, and M. Rief, "The myosin coiled-coil is a truly elastic protein structure," *Nature Materials*, vol. 1, no. 4, pp. 232–235, 2002.

[65] T. Bornschlögl, J. Gebhardt, and M. Rief, "Designing the folding mechanics of coiled-coils," *Chem. Phys. Chem.*, vol. 10, no. 16, pp. 2800–2804, 2009.

[66] T. Katsumoto, A. Mitsushima, and T. Kurimura, "The role of the vimentin intermediate filaments in rat 3y1 cells elucidated by immunoelectron microscopy and computergraphic reconstruction," *Biology of the Cell*, vol. 68, no. 1-3, pp. 139–146, 1990.

[67] M. Meier, G. Padilla, H. Herrmann, T. Wedig, M. Hergt, T. Patel, J. Stetefeld, U. Aebi, and P. Burkhard, "Vimentin coil 1a–a molecular switch involved in the initiation of filament elongation," *Journal of Molecular Biology*, vol. 390, no. 2, pp. 245–261, 2009.

[68] S. Neukirch, A. Goriely, and A. C. Hausrath, "Chirality of coiled-coils: Elasticity matters," *Physical Review Lett.*, vol. 100, p. 038105, Jan 2008.

[69] J. Apgar, S. Hahn, G. Grigoryan, and A. Keating, "Cluster expansion models for flexible-backbone protein energetics," *Journal of Computational Chemistry*, vol. 30, no. 15, pp. 2402–2413, 2009.

[70] S. Sadeghi and E. Emberly, "Length-dependent force characteristics of coiled-coils," *Physical Review E*, vol. 80, p. 061909, Dec 2009.

[71] E. Emberly, N. Wingreen, and C. Tang, "Designability of α-helical proteins," *Proceedings of the National Academy of Sciences*, vol. 99, no. 17, p. 11163, 2002.

[72] P. Harbury, B. Tidor, and P. Kim, "Repacking protein cores with backbone freedom: structure prediction for coiled-coils," *Proceedings of the National Academy of Sciences*, vol. 92, no. 18, p. 8408, 1995.

[73] S. Miyazawa and R. Jernigan, "Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation," *Macromolecules*, vol. 18, no. 3, pp. 534–552, 1985.

[74] R. Jernigan and I. Bahar, "Structure-derived potentials and protein simulations," *Current opinion in structural biology*, vol. 6, no. 2, pp. 195–209, 1996.

[75] E. Shakhnovich, "Proteins with selected sequences fold into unique native conformation," *Physical Review Letters*, vol. 72, no. 24, pp. 3907–3910, 1994.

[76] H. Li, C. Tang, and N. Wingreen, "Nature of driving force for protein folding: a result from analyzing the statistical potential," *Physical Review Letters*, vol. 79, no. 4, pp. 765–768, 1997.

[77] R. Akkermans and P. Warren, "Multiscale modelling of human hair," *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 362, no. 1821, p. 1783, 2004.

[78] H. Abe and H. Wako, "Analyses of simulations of three-dimensional lattice proteins in comparison with a simplified statistical mechanical model of protein folding," *Physical Review E*, vol. 74, p. 011913, Jul 2006.

[79] K. Eom, D. E. Makarov, and G. J. Rodin, "Theoretical studies of the kinetics of mechanical unfolding of cross-linked polymer chains and their implications for single-molecule pulling experiments," *Physical Review E*, vol. 71, p. 021904, Feb 2005.

[80] A. Kleiner and E. Shakhnovich, "The mechanical unfolding of ubiquitin through all-atom Monte Carlo simulation with a Gō-type potential," *Biophysical Journal*, vol. 92, no. 6, pp. 2054 – 2061, 2007.

[81] J. Shimada, E. Kussell, and E. Shakhnovich, "The folding thermodynamics and kinetics of crambin using an all-atom Monte Carlo simulation," *Journal of Molecular Biology*, vol. 308, no. 1, pp. 79–95, 2001.

[82] H. Goldstein, C. P. Poole, and L. Safko, *Classical Mechanics*. Addison-Wesley, 2001.

[83] A. Ortiz, C. Strauss, and O. Olmea, "MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison," *Protein Science*, vol. 11, no. 11, pp. 2606–2621, 2002.

[84] B. Park and M. Levitt, "Energy functions that discriminate X-ray and near-native folds from well-constructed decoys," *Journal of Molecular Biology*, vol. 258, no. 2, pp. 367–392, 1996.

[85] E. O'Shea, J. Klemm, P. Kim, and T. Alber, "X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled-coil," *Science*, vol. 254, no. 5031, p. 539, 1991.

[86] W. Liu, L. Jawerth, E. Sparks, M. Falvo, R. Hantgan, R. Superfine, S. Lord, and M. Guthold, "Fibrin fibers have extraordinary extensibility and elasticity," *Science*, vol. 313, no. 5787, p. 634, 2006.

[87] G. Grigoryan and W. DeGrado, "Probing designability via a generalized model of helical bundle geometry," *Journal of Molecular Biology*, vol. 405, pp. 1079–1100, 2011.

[88] M. Hicks, D. Holberton, C. Kowalczyk, and D. Woolfson, "Coiled-coil assembly by peptides with non-heptad sequence motifs," *Folding and Design*, vol. 2, no. 3, pp. 149–158, 1997.

[89] M. Hicks, J. Walshaw, and D. Woolfson, "Investigating the tolerance of coiled-coil peptides to nonheptad sequence inserts," *Journal of Structural Biology*, vol. 137, no. 1-2, pp. 73–81, 2002.

[90] A. Murzin, S. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," *Journal of Molecular Biology*, vol. 247, no. 4, pp. 536–540, 1995.

[91] A. Andreeva, D. Howorth, J. Chandonia, S. Brenner, T. Hubbard, C. Chothia, and A. Murzin, "Data growth and its impact on the SCOP database: new developments," *Nucleic Acids Research*, vol. 36, no. suppl 1, p. D419, 2008.

[92] http://www.r project.org/.

[93] N. Wingreen, H. Li, and C. Tang, "Designability and thermal stability of protein structures," *Polymer*, vol. 45, no. 2, pp. 699–705, 2004.

[94] H. Li, C. Tang, and N. Wingreen, "Are protein folds atypical?," *Proceedings of the National Academy of Sciences*, vol. 95, no. 9, p. 4987, 1998.

[95] M. Larkin, G. Blackshields, N. Brown, R. Chenna, P. McGettigan, H. McWilliam, F. Valentin, I. Wallace, A. Wilm, R. Lopez, *et al.*, "Clustal w and clustal x version 2.0," *Bioinformatics*, vol. 23, no. 21, p. 2947, 2007.

[96] A. Celedon, I. Nodelman, B. Wildt, R. Dewan, P. Searson, D. Wirtz, G. Bowman, and S. Sun, "Magnetic tweezers measurement of single molecule torque," *Nano letters*, vol. 9, no. 4, pp. 1720–1725, 2009.

[97] R. Fulconis, A. Bancaud, J. Allemand, V. Croquette, M. Dutreix, and J. Viovy, "Twisting and untwisting a single DNA molecule covered by RecA protein," *Biophysical Journal*, vol. 87, no. 4, pp. 2552–2563, 2004.

[98] J. Miller, C. Zeng, N. Wingreen, and C. Tang, "Emergnce of highly-designable protein-backbone conformations in an off-lattice model," *Arxiv preprint cond-mat/0109305*, 2001.

Appendix A

Sequence translation for experimentally determined structures

The HP translation of the experimentally determined structures in Chapter 3 and 4,

Leucine zipper domain(2ZTA)

MKQLEDKVEELLSKNYHLENEVARLKKLV
PPPHPPPHPPHHPPPPPHPPPHHPHPPHH

N-terminal domain from apc(1DEB)

DQLLKQVEALKMENSNLRQELEDNSNHLTKLETEASNMKEVLKQLQGS
PPHHPPHPHHPHPPPPHPPPHPPPPPPHPPHPPPHPPHPPHHPPHPPP

Outer membrane lipoprotein(1JCC)

KIDQLSSDAQTANAKADQASNDANAARSDAQAAKDDAARANQRL
PHPPHPPPHPPHPHPHPPHPPPHPHHPPPHPHHPPPHHPHPPPH

FYVE/PHD zinc finger(1JOC)

QDERRALLERCLKGEGEIEKLQTKVLELQRKLDNTTAAVQELGRENQS
LQIKHTQA
PPPPPHHHPPPHPPPPPHPPHPHPHHPHPPPHPPPPHHHPPHPPPPPP
HPHPPPPH

Geminin domain(1T6F)

TLYEALKENEKLHKEIEQKDNEIARLKKENKELAEVA
PHHPHHPPPPPHPPPHPPPPPPHHPHPPPPPPHHPHH

G protein binding domain(1UIX)

TSDVANLANEKEELNNKLKEAQEQLSRLKDEEISAAAIKAQFEKQLLT
ERTLKTQAVNKLAEI
PPPHHPHHPPPPPHPPPHPPHPPPHPPHPPPPHPHHHHPHPHPPPHHP
PPPHPPPHHPPHHPH

MPN010-like(2BA2)

KTEFKEFQTVVMESFAVQNQNIDAQGEQIKELQVEQKAQGKTLQLILE
ALQGINKRLDNL
PPPHPPHPPHHHPPHHHPPPPHPHPPPPHPPHPHPPPHPPPPHPHHHP
HHPPHPPPHPPH

Myosin rod fragment(2FXM)

FTRLKEALEKSEARRKELEEKMVSLLQEKNDLQLQVQAEQDNLADAEE
RCDQLIKNKIQLEAKVKEMNERLEDEEEMNAELTAKKRKLEDECSELK
RDIDDLELTL

HPPHPPHHPPPPHPPPPHPPPHHPHHPPPPPHPHPHPHPPPPHHPHPP
PHPPHHPPPHPHPHPHPPHPPPHPPPPPHPHPHPHPPPPHPPPHPPHP
PPHPPHPHPH

Myosin rod fragment(3BAS)

EEEMKEQLKQMDKMKEDLAKTERIKKELEEQNVTLLEQKNDLFGSMKQ
LEDKVEELLSKNYHLENEVARLKKL
PPPHPPPHPPHPPHPPPHHPPPPHPPPHPPPPHPHHPPPPPHHPPHPP
HPPPHPPHHPPPHPHPPPHHPHPPH

Intermediate filament protein(3G1E)

NEKVELQELNDRFANLIDKVRFLEQQNKILLAELEQL
PPPHPHPPHPPPHHPHHPPHPHHPPPPPHHHHPHPPH