A Minimal Model for the Hydrophobic and Hydrogen Bonding Effects on Secondary and Tertiary Structure Formation in Proteins

by

Kyle Denison

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Science
in
Physics

Waterloo, Ontario, Canada, 2009

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

A refinement of a minimal model for protein folding originally proposed by Imamura [1] is presented. The representation of the α-helix has been improved by adding in explicit modelling of the entire peptide unit. A four-helix bundle consisting of four α-helices and three loop regions is generated with the parallel tempering Monte Carlo scheme. Six native states are found for the given sequence, four U-bundle and two Z-bundle states. All six states have energies of $E \approx -218\epsilon$ and all appear equally likely to occur in simulation. The highest probability of folding a native state is found to be at a hydrophobic strength of $C_h \approx 0.8$ which agrees with the value of $C_h = 0.7$ used by Imamura in his studies of α to β structural conversions.

Two folding stages are observed in the temperature spectrum dependent on the magnitude of the hydrophobic strength parameter. The two stages observed as temperature decreases are 1) the hydrophobic energy causes the random coil to collapse into a compact globule 2) the secondary structure starts forming below a temperature of about $T \approx 0.52\epsilon/k_B$. The temperature of the first stage, which corresponds to the characteristic collapse temperature T_θ, is highly dependent on the hydrophobic strength. The temperature of the second stage is constant with respect to hydrophobic strength. Attempts to measure the characteristic folding temperature, T_f, from the structural overlap function proved to be difficult due mostly to the presence of six minima and the complications that arose in the parallel tempering Monte Carlo scheme. However, a rough estimate of T_f is obtained at each hydrophobic strength from a native state density analysis. T_f is found to be significantly lower than T_θ.

Acknowledgements

I would like to give thanks to the people that have made this Masters thesis possible over the last two years. First and foremost I give thanks to my supervisor Dr. Jeff Chen for his guidance and support in regards to the coding, background research, and analysis necessary to complete this document. I would also like to thank my committee members Dr. Ha and Dr. Leonenko and my external examiner Dr. Bizheva for their constructive criticism of the thesis. This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET:www.sharcnet.ca).

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Proteins

Proteins are large macromolecular polymer chains that fold into unique three dimensional configurations and perform a variety of essential tasks in all living organisms. A proteins structure and thus its function is believed to be encoded in the sequence of amino acids that define the protein chain [2]. Haber and Anfinsen demonstrated this principle by experiment when they showed that ribonuclease can spontaneously regain its full function in vitro [3]. That is, the ribonuclease was found to fold and refold with no external stimulus to guide the folding process.

The importance of proteins in biological systems cannot be overstated. They are responsible for sustaining life through the processes of self-regulation, the regulation of all chemical, physical and biological processes within a system, and self-replication, the process by which DNA polymerase proteins oversee DNA replication during cell division [1]. Examples of the functions proteins perform include enzyme proteins which catalyse the chemical reactions in cells, cell signalling and signal transduction proteins such as insulin which transmit signals from one cell to another, antibodies which bind to foreign agents in an organism and target them for destruction, and ligand-binding proteins such as hemoglobin which transport oxygen from the lungs to other organs and tissues [1].

Despite the importance of proteins as a part of all living creatures the mechanisms by which a protein chain folds into its native state are still largely not understood. Although much effort has been expended to study them the relationships between a proteins basic building blocks, its structure, its dynamics, and its final function cannot yet be fully defined. Characterizing these elements defines the essence of the protein folding problem and is a critical step toward the ultimate goal of designing and manufacturing artificial proteins.

The protein folding problem is stated simply "given the sequence of a protein and its folding environment what is the folded structure?" At first glance the solution seems simple; find the configuration which minimizes the energy of the system and

you have the native state of the protein in a given environment. Closer examination reveals the immense complexity of this problem. Levinthal first elucidated this complexity when he proposed the Levinthal paradox which states that a normal size protein cannot find its native state on a timescale within the age of the universe by randomly sampling all possible states because the degrees of freedom available are truly immense [4]. Thankfully the search is far from random and is in fact guided by the free energy landscape of the system.

The free energy landscape of a protein suffers from the multi-minima problem. This is because the large number of possible configurations result in a wide variety of possible energies which in turn result from inappropriate contacts between residues and misoriented structures in the protein. The result is a very rugged energy landscape. The energy landscape is often viewed as a folding funnel where instead of a flat plain with mountains and valleys the landscape is sloped toward the global minimum, Figure 1.1. This sloping guides the configuration toward the minimum energy state without having to visit most of the largely unfavourable states available to it in a random search [5]. Another driving force that reduces the folding time is the hydrophobic collapse hypothesis [6, 7]. In this methodology hydrophobic interactions drive the folding of the native state by driving the collapse of a denatured random coil like protein into a random globule phase. This is predicted from the observation that "the buried interior regions and the peptide chain turns of the folded protein (i.e., inside and outside) are predicted solely by the hydrophobicity of the residues" [7]. Further, evidence suggests that chain regions rich in hydrophobic residues serve as small clusters that fold against each other. Secondary structure is formed either as or after the hydrophobic collapse occurs.

To gain insight into protein folding a minimal off lattice protein model is presented. The model is inspired by both the Imamura [1] and Thirumalai [9] models and represents a further refinement of the forcefields of those studies. The characteristics of the model and its suitability in modelling protein structure are examined through simulation of a four-helix bundle protein. The remainder of this study is structured in the following way. The next section details the important characteristics of protein structure from the basic building blocks to how they arrange themselves in nature. This is followed by a review of some computational techniques used to study proteins in the past. The next chapters provide a detailed description of the simulation methods and the model employed in this study. Finally the four-helix bundle analysis is presented and conclusions on the structure, energetics, and folding pathway are drawn.

1.2 An Overview of Protein Structure

This section provides an overview of protein structure. The discussion is mainly focused on properties of proteins which naturally lend themselves to modelling in computer simulation. Protein structure is classified according to a hierarchy of four levels [10]. The primary structure is the sequence of amino acids that

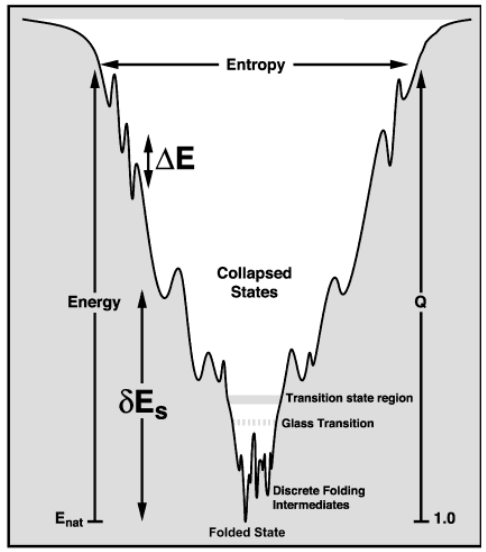
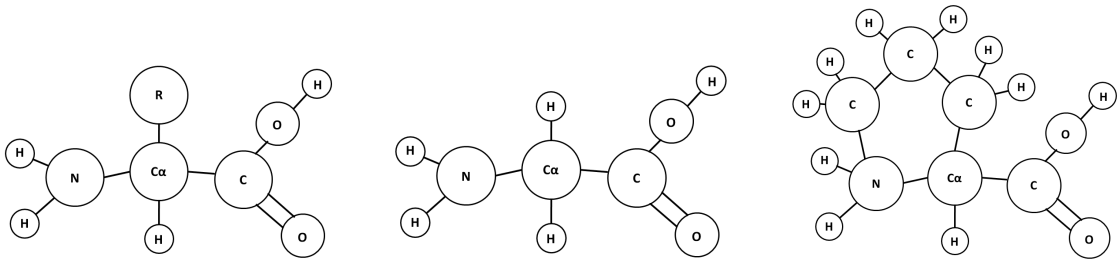


Figure 1.1: Shown is a graphical representation of the folding funnel. The height of the funnel corresponds to the energy of the folded state, E_{nat}, as referenced from the bottom of the funnel. The width of the funnel is approximately the entropy or the density of states at that level of the funnel. As the system progresses down the funnel, it has the possibility of passing through a collapsed state, a transition state, and a glass transition. The degree of ruggedness also changes as the system progresses down the funnel [8].

make up the protein backbone. Secondary structure is defined by repeating local conformations established by hydrogen bonding. Tertiary structure is defined by the relative orientation of the secondary structures within one protein molecule and quaternary structure is how tertiary structures bond to each other to form larger protein domains [10]. The details of these structural levels are discussed in the following sections.

1.2.1 Amino Acids and the Primary Structure

An amino acid consists of a central carbon atom (C_α) bonded to an amino functional group (NH_2), a carboxyl functional group (COOH), a hydrogen atom, and a side chain (R), Figure 1.2. There are 20 different side chains which define the type and properties of each amino acid [10]. The side chains are typically divided into three groups. These are the hydrophobic group, consisting of Alanine (Ala), Valine (Val), Leucine (Leu), Isoleucine (Ile), Phenylalanine (Phe), Proline (Pro), and Methionine (Met), the charged group consisting of Aspartic acid (Asp), Glutamic acid (Glu), Lysine (Lys), and Arginine (Arg), and the polar group consisting of Serine (Ser), Threonine (Thr), Cysteine (Cys), Asparagine (Asn), Glutamine (Gln), Histidine (His), Tyrosine (Tyr), and Tryptophan (Trp). Glycine (Gly) has only a single hydrogen atom as its side chain and is typically considered to be its own fourth



(a) General Form (b) Glycine (c) Proline

Figure 1.2: a) The general form of the 20 amino acids is shown where R indicates the side chain. b) Glycine is shown with its single H side chain. This gives Glycine a much large set of allowed conformations in protein structure, see Section 1.2.4. c) Proline is shown with its conformationally restrictive side chain.

group or as a member of the hydrophobic group. All amino acids exist in the L-form chirality state even though there is no known reason for any preference over the D-form state [10]. It is this variety of amino acids which gives proteins their immense structural and functional diversity.

A protein is a chain of amino acids, its primary structure, linked together by bonds between the carboxyl atom of one amino acid and the amino nitrogen of another. These bonds, shown in Figure 1.3, are known as peptide bonds [10]. The sequence of peptide bonds is referred to as the protein backbone and can be divided into convenient groups known as residues. A residue consists of the C_α, the amino group, the carboxyl group, the hydrogen atom, and the side chain [10]. The sequence order is defined as beginning at the open NH group on one end of the chain and terminating on the open C=O group at the other end.

Strong covalent bonds hold the protein backbone and side chains together. These bonds are permanent on the time scale of a protein's life. Of special note is the N-C bond which characterizes the peptide bond. This bond is planar due to a delocalization of the lone electron pair of the nitrogen atom onto the carboxyl oxygen [1]. The C-N bond is shortened by 10 percent and has a double bond character which is resistant to twisting.

Proline is unique among the amino acids in that its side chain loops back and bonds to the nitrogen of its amino functional group, Figure 1.2c [11]. This fixes the ϕ dihedral angle to -65° (see section 1.2.4 for dihedral angle definitions) and consequently proline is rarely found at the centre of secondary structure elements. When it does occur in an α-helix proline is usually found in the first turn.

The primary structure is an unbranched sequence. The sequence must be unbranched because it is not possible for DNA, which stores the primary sequence, to store the information of a branched sequence [1]. Occasionally a sequence can be cross bridged when two cysteine residues oxidize to form a disulphide bridge (-S-S-). The main role of disulphide bridges is to stabilize 3D structure [10].

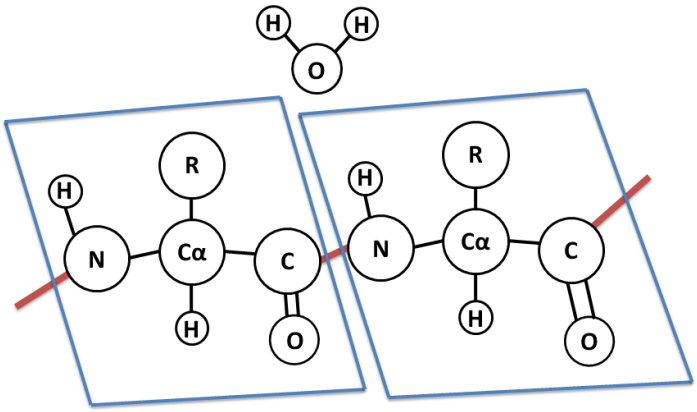


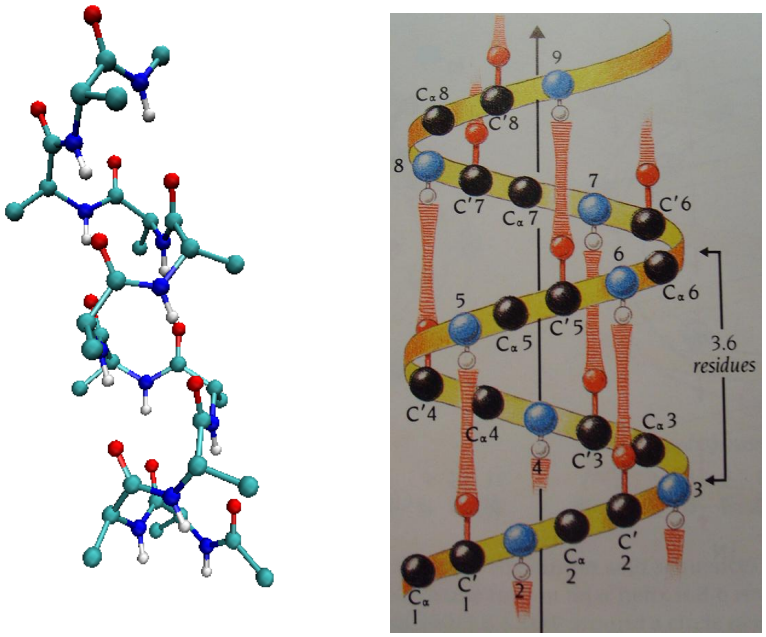
Figure 1.3: A peptide bond (red) between two amino acids is shown. A water molecule is ejected as the bond forms. A residue is defined as all the atoms enclosed in one of the two blue outlines.

1.2.2 Secondary Structure

The secondary structure of a protein is formed when the protein chain folds in such a way as to establish hydrogen bonding between the H of the NH group, a hydrogen bond donor, of one residue and the O of the C=O group, a hydrogen bond acceptor, of another residue [10]. By forming these hydrogen bonds the backbone can no longer form bonds with the folding medium (typically water) and so these bonds are said to have the effect of neutralizing the hydrophilic tendencies of the protein backbone. This allows entire molecules to fold so that they have a hydrophobic core and a hydrophilic surface which is typically dictated by the side chain properties and not the protein backbone [10]. The two main secondary structures observed in proteins are the α-helix and the β-sheet.

The α-helix is characterized by the spiral pattern shown in Figure 1.4. The helical structure can in theory be either a right-handed or left-handed helix depending on the screw direction but in nature the right-handed helix accounts for the vast majority of observed structures [10]. A typical α-helix structure will have 3.6 residues per turn around the helical axis. α-helices vary in length between five and forty residues with an average length of ten residues [10]. Figure 1.4b shows an idealized view of the structure which emphasizes the role of hydrogen bonding in maintaining the helical shape. It can be seen that the C=O group of residue i is bound to the NH group of residue i+4. Other types of helices exist such as the 3-10 helix which shows an i to i+3 bonding pattern and the π helix which shows an i to i+5 pattern. These helices are relatively rare. The first NH group in the chain and the last C=O group will not be bound making the structure polar [10]. α-helices are often amphiphilic with respect to their side chain properties. This property is critical in the formation of more complex structures.

The β-sheet structure is built from a collection of similar regions in the protein known as β-strands. Each β-strand consists of an almost fully extended protein



(a) Alanine helix (b) Hydrogen Bonding Pattern

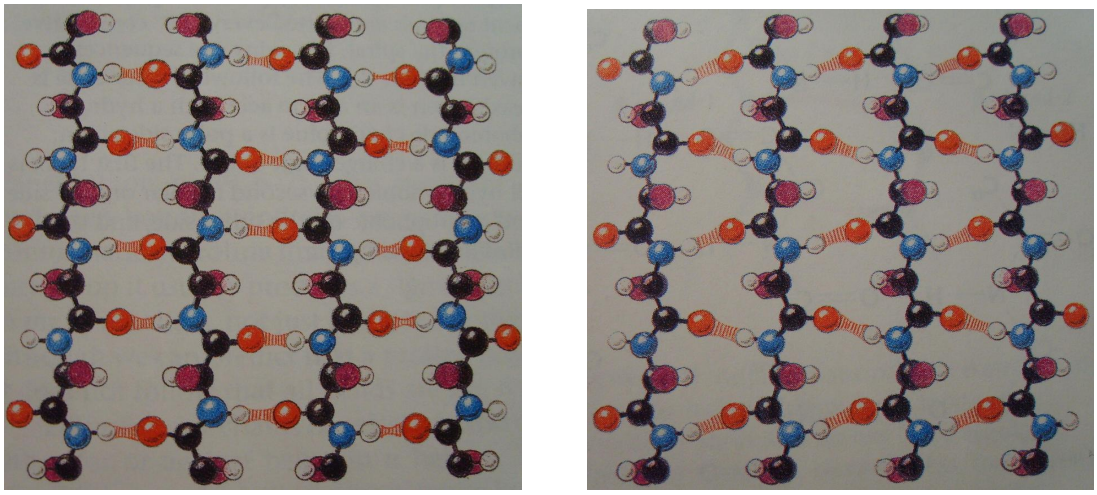
Figure 1.4: a)The α-helix for alanine is shown in ball and stick representation. b) The hydrogen bonding pattern in an α-helix is shown. Image taken from [10].

chain. These strands form a sheet by lining up parallel or anti-parallel to each other such that the C=O groups on one strand can hydrogen bond to the NH groups on an adjacent strand and vice versa, Figure 1.5 [10]. Each β-strand is typically five to ten residues long. In order for two β-strands to line up in parallel there must be an extended region such as an α-helix or a long loop region connecting them. A common example of an anti-parallel β structure is that of the β-hairpin, Figure 1.6. The hairpin structure consists of two anti-parallel β-strands joined by a hairpin loop [10].

Loop regions are irregularly shaped regions of the protein chain which connect the various secondary structures within a protein molecule together . An important aspect of these regions is that they do not generally form hydrogen bonds with other residues in the protein and instead are found at the surface of the molecule where they can bond with water [10]. It has been shown that insertions and deletions of residues happen almost exclusively within these loop regions.

1.2.3 Tertiary Structure

The tertiary structure of a protein chain refers to the way in which the secondary structures orient themselves within the same protein molecule. Tertiary structures have a large impact on the function of a protein. These structures come about due to the tendency for proteins to bury hydrophobic side chains in the interior of the



(a) Anti-Parallel (b) Parallel

Figure 1.5: The hydrogen bonding patterns in antiparallel and parallel β-sheets. Images taken from [10].

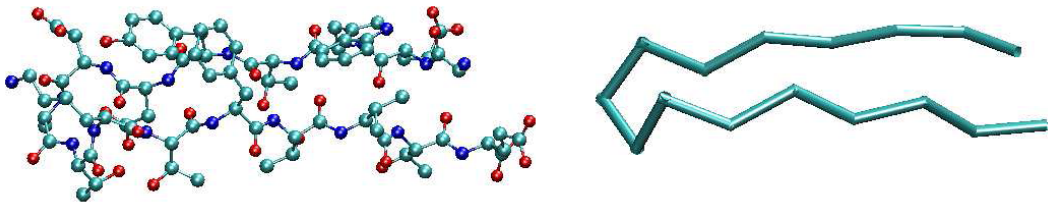


Figure 1.6: A β-hairpin of the GB1 protein. The ball and stick representation (left) and the backbone representation (right) [1].

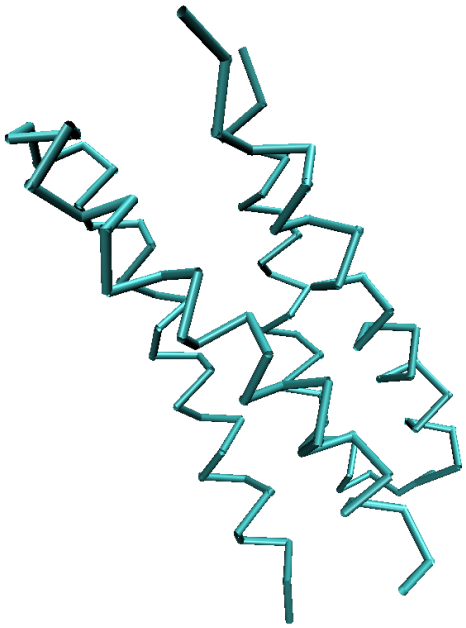


Figure 1.7: The Four-Helix Bundle with PDB structure code 1U7M is shown.

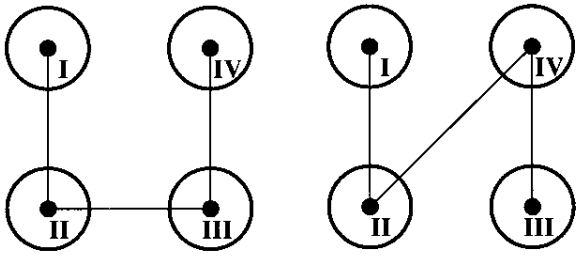


Figure 1.8: Helix Packing in the U (left) and Z (right) four-helix bundle structures.

molecule while presenting hydrophilic side chains to the proteins surroundings [10]. This tendency to bury hydrophobic side chains results in the residues associated with these side chains being grouped together in the core of the protein.

A very common tertiary structure is the four-helix bundle which consists of four α-helices lined up next to each other either in parallel or anti-parallel orientation, Figure 1.7. The bundles are typically formed from either one or two protein molecules. The helices are packed to bury hydrophobic residues in the core of the bundle. Typically four-helix bundles are found in either the U-bundle or Z-bundle helix packing arrangement, Figure 1.8. The four-helix bundle "occurs in several widely different proteins, such as myohemerythrin, a non-haem iron containing oxygen transport protein in marine worms; cytochrome c' and cytochrome b_{562}, which are haem containing electron carriers; ferritin, which is a storage molecule for Fe atoms in eucaryotic cells; and the coat protein of tobacco mosaic virus" [10].

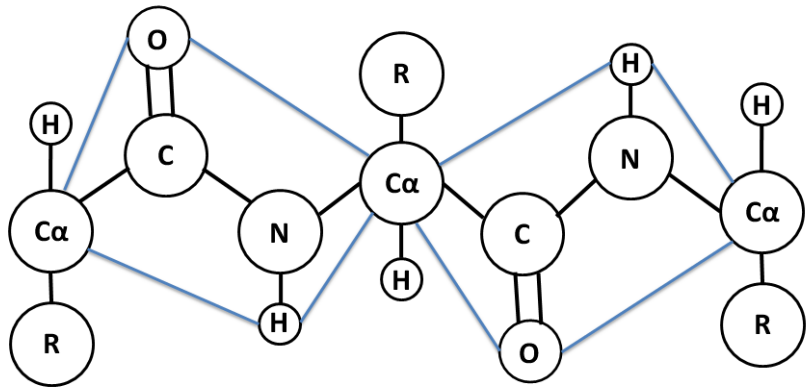


Figure 1.9: The Peptide Unit. Two peptide units are displayed (blue lines). Atoms enclosed in a peptide unit exist in a rigid plane.

1.2.4 The Peptide Unit and Ramachandran Angles

The peptide unit methodology groups the backbone atoms from one C_α to the next as the basic unit of protein structure, Figure 1.9. The peptide unit then exists in a plane where the bond lengths and bond angles are nearly the same for all peptide units in all proteins [10, 12]. The peptide units are essentially rigid and so the only degrees of freedom available to the protein backbone are those associated with rotations about the covalent bonds to the C_α on either end of the peptide unit. The hydrogen atom and the side chain associated with each C_α are not included in the peptide unit [10].

Since the peptide unit is essentially rigid the backbone atom positions can be completely specified by a set of three dihedral angles for each residue. Conventionally the three dihedral angles for a residue i are defined as ϕ_{i} centred on the $N_{(\mathrm{i})}$-$C_{\alpha(\mathrm{i})}$ bond, ψ_{i} centred on the $C_{\alpha(\mathrm{i})}$-$C_{(\mathrm{i})}$ bond, and ω_{i} centred on the $C_{(\mathrm{i})}$-$N_{(\mathrm{i+1})}$ bond, Figure 1.10 [13]. ω can be either $^{+}/_{-}180^o$ for the trans-state or 0° for the cis-state due to the twist resistant nature of the peptide bond, discussed in Section 1.2.1. The trans-state is sterically favorable over the cis-state and thus ω is overwhelmingly found to be in the trans-state in nature [14].

The remaining dihedral angles, ϕ and ψ, are named the Ramachandran angles after G.N. Ramachandran who first conceived of using them to characterize protein structure. Although in theory the Ramachandran angles can each vary between $+180^o$ and -180° in actual protein structures steric restrictions significantly reduce the number of possible angles [14]. A Ramachandran plot is a plot of the allowed regions in ϕ and ψ space, Figure 1.11. Glycine is less constrained because it has only a hydrogen atom as a side chain. Secondary structures are characterized by sections of the protein chain with repeating values of the Ramachandran angles, Table 1.1 [12].

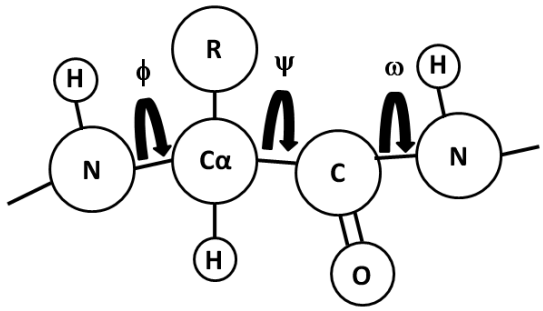


Figure 1.10: The three dihedral angles that define the protein backbone are shown.

Table 1.1: Ramachandran angles are listed for some common secondary structures.

Structure	ϕ	ψ
extended chain	180	180
α-Helix	-57	-47
α-Helix left-handed	57	47
β-sheet antiparallel	-139	135
β-sheet parallel ideal	-120	120
β-sheet parallel	-120	113

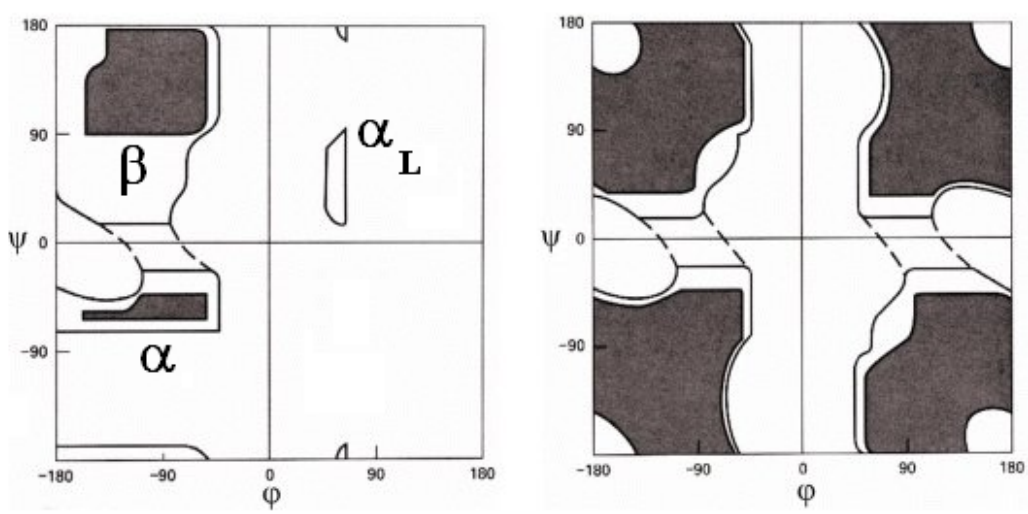


Figure 1.11: The Ramachandran Plot for non-glycine amino acids (left) show three main areas of allowed structure corresponding to α-helices, left handed helices, and β structures. The glycine plot (right) shows several more allowed regions due to the simplicity of the glycine side chain. The fully allowed regions are shaded; the partially allowed regions are enclosed by a solid line. The connecting regions enclosed by the dashed lines are permissible with slight flexibility of bond angles [15].

1.3 Previous Research

The following section will provide a brief sampling of some of the models used to study protein folding. These models are typically grouped into several categories including all atoms models, off lattice minimal models, and lattice models. The advantages and disadvantages of each will be discussed.

It seems appropriate to begin the discussion with a look at all atom simulation models [16, 17]. As the name implies in these simulations all of the atoms of a protein chain are explicitly represented. The forcefields typically include potentials for local bonding including terms for bond length, bond angle, and dihedral angle interactions as well as long range potentials including Lennard-Jones and electrostatic interactions. These typically take the form

$$V_{\text{Total}} = \sum_{bonds} K_r \left(r - r_{eq}\right)^2 + \sum_{angles} K_\theta \left(\theta - \theta_{eq}\right)^2 + \sum_{dihedrals} \frac{V_n}{2} \left(1 + cos\left(n\phi - \delta\right)\right) + \sum_{i<j} \left(\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{D R_{ij}}\right). \quad (1.1)$$

The first terms is the bond length potential which sets the bond lengths, r, for all bonds in a structure to their corresponding bond lengths, r_{eq}, through a harmonic potential of force constant K_r. The deviation from r_{eq} is typically small and occurs on a short timescale. In course grained models this potential is often set to a constant. The second term does much the same for the bond angles where the force constant K_θ is typically large to enforce only slight deviations from the equilibrium angle θ_{eq}. The third term is used to describe the energy change as part of a molecule undergoes rotation about one of its bonds. This dihedral angle potential depends on the dihedral angle ϕ, the force constant V_n, the multiplicity n, and a phase shift δ. This potential is not harmonic since it is observed that for many molecules ϕ can assume any value between 0° and 360° with no large differences in energy [16]. The last term contains the Lennard-Jones interaction, $\sum_{i<j} \left(A_{ij}/R_{ij}^{12} - B_{ij}/R_{ij}^{6}\right)$, and the electrostatic interaction, $\sum_{i<j} \left(q_i q_j / D R_{ij}\right)$. The Lennard-Jones energy mimics the long range dispersion and short range repulsion interactions between atoms i and j with separation distance R_{ij}. A_{ij} and B_{ij} are positive constants that depend on the atoms. The electrostatic term treats each of the explicitly modelled atoms as point charges with a fractional charge q_i and q_j. D is the dielectric constant of the medium. For a system which explicitly models the solvent atoms $D = 1.0$ whereas in a model with no solvent D is set appropriately to mimic a solvent, $D = 80$ for water [1]. The Lennard-Jones and electrostatic interactions typically ignore interactions between atoms involved in local bonding because local bonding interactions are much stronger than the long range interactions.

The explicit modelling of the structure combined with an extremely detailed forcefield allow the all atom models to accurately predict protein folding and struc-

ture. Unfortunately because of the complexity and computational demand of all atom models they are typically only able to simulate proteins on the order of ten to 100 nanosecond timescales which is insufficient to examine the microsecond or millisecond timescales in real protein folding [18]. In addition the various models such as CHARM [19] and AMBER [20] are developed from experimental datasets and tend to be targeted to particular physical systems, which limits their predictive power.

The various non all atom models all contain some form of course graining of the energies. Course graining is accomplished by reducing the structural and/or energetic detail in a model through such techniques as representing each residue as a single point mass or imposing structurally biased potentials to guide the folding process [18]. These approximations greatly reduce the computational demand and allow for detailed examination of longer timescale processes and larger structures but come at the cost of transferability and increased complexity in the parameter space. This is because as "the graining becomes coarser more specific interactions must be included in fewer parameters and functional forms" [18]. Before discussing some of these course grained models it is appropriate to examine the rational for the course graining of two of the most important factors in protein folding, the hydrogen bonding and hydrophobic effects.

Hydrogen bonding is a bond established between a hydrogen atom and an electronegative atom such as nitrogen, oxygen, or fluorine. In protein chains the oxygen of the carboxyl group of one residue forms a hydrogen bond with the hydrogen of the amino group of another residue. As discussed in Section 1.2.2 these asymmetric bonds have the effect of neutralizing the hydrophilic tendencies of the protein backbone [10].

Course grained models will typically use knowledge based potentials, potentials derived from features of experimentally determined protein structures, in place of physics based potentials, potentials like those described in the all atom discussion above [21]. To this end hydrogen bonds are typically modelled with Lennard-Jones 6-12 or 10-12 interactions between the doner and acceptor atoms. Other more course grained models such as the early Thirumalai work [2, 9, 22] treat hydrogen bonding through a dihedral potential between any four successive C_α atoms. The Imamura minimal model uses a shifted 6-12 Lennard-Jones potential with virtual interaction centres to mimic hydrogen bonding [1].

The hydrophobic effect is characterized by three main properties. The insertion of non-polar solutes into water is strongly unfavourable, strongly opposed by entropy at room temperature, and accompanied by a large positive heat capacity [23]. These properties arise because water is essentially a dynamic loose network of hydrogen bonds. Insertion of a non-polar solute such as a protein causes a local rearrangement of the network to preserve the number of hydrogen bonds by lining up around the solute. Enthalpy is mostly unchanged while entropy of the solute is decreased due to an increase in local order. The entropy decrease is balanced by a corresponding entropy increase in the surrounding water caused by the water

molecules coalescing around the protein to preserve their bonds. This coalescing of water drives the hydrophobic components of the solvent (the hydrophobic side chains) into the centre of the structure. This is why hydrophobic residues are found in the centre of protein bundles [1]. In real protein the hydrophobic effect is temperature dependant. It is weak at low temperature and gets stronger at high temperature. The weakness of the interaction at low temperatures causes denaturation of proteins at cold temperatures [1]. In simple models the temperature dependence is usually ignored [2].

To mimic this behaviour in simple protein models knowledge based potentials are once again employed. Since the solvent which drives the hydrophobic packing is rarely modelled due to computational limitations a less direct but equivalent method is utilized. Both the early Thirumalai [2, 9, 22] model and the Imamura [1] model employ a simple 6-12 Lennard-Jones potential between hydrophobic residues. With proper implementation this potential has the effect of encouraging these residues to group together in the centre of the protein. Thus a repulsive entropic force is replaced with a position dependant attractive potential.

In lattice based models the protein chain is represented by placing residues on a 2D or 3D lattice. Depending on the model various interaction potentials are defined between the residues. In the typical HP model each conformation is specified by assigning the lattice points as hydrophobic or polar. Any two hydrophobic beads that are topological neighbours then contribute an energy of -1 to the system while all other pairings have no effect on the energy [24]. The native state is found by searching through the possible conformations. The search can be guided by the energy or simply brute forced via a sequential search. Although simple lattice models cannot be used to generate realistic conformations the HP model is invaluable as a toy model to investigate new concepts as they are proposed. The main advantages of the model are [25]

1. All of the conformations can be enumerated.

2. The exact nature of the model is well defined. This allows one to investigate the effects of the model and its approximations rather than the effects of parameter choices in the model.

3. Because of their simplicity these models can reveal properties of chain like molecules other than proteins.

4. Simple exact models are useful for testing new conformational search algorithms since all of the states are exactly defined and typically known from previous searches. Thus the efficiency of a new algorithm can be gauged.

5. Simple lattice models explicitly account for specific monomer sequences, chain connectivity, and excluded volume and are useful for testing analytical theories, such as mean-field treatments of heteropolymer collapse and spin-glass models. For instance simple models show how the rugged energy landscape can arise in a protein model.

Gō models are a class of model that contains only native contact potentials. For example, in an α-helix hydrogen bonds takes an i to i+4 bonding pattern between residues while the 3-10 helix takes an i to i+3 bonding pattern. A Gō model would explicitly ignore any interactions not found in the native state so that when simulating the α-helix the i to i+3 interactions would not be calculated and a 3-10 helix would not be possible. These models are applied to lattice, off-lattice and all-atom simulations. The original Gō model was a 3D lattice model that included interactions for short range local conformational propensities, long range native biased potentials, and hydrophobic potentials [26]. The main attractiveness of Gō models is the large decrease in folding time over other models. This is a direct result of excluding any energy interaction not found in the native state. Statistically sufficient sampling is possible even in all atom models. The flip side to this is that a priori knowledge of the native state is required and as such Gō models have no predictive power. Also there is little to no physical justification for ignoring the non-native contacts.

The final set of course grained models to discuss are the bead models. These models are based on a united atom formulation of protein structure where there are typically one to six interaction centres per residue [18]. One bead models tend to be evolutions of Gō models with a more complicated forcefield. They retain a partial bias toward a reference configuration due to the inherent difficulty of including the effects of amino acid size, geometry and conformation in the model. This dependence is typically expressed through a dihedral potential which must be changed depending on the desired secondary structure. Two bead models are one bead models that represent the side chain with a second bead at the appropriate centre of mass position. Four to six bead models increase the complexity by explicitly modelling the backbone C_α positions, the heavier O, N, and C atoms, and the H atoms. As in two bead models the side chain is represented with a bead at its centre of mass. Since the atoms involved in hydrogen bonding are explicitly represented in these models hydrogen bonding interactions can be much more accurately defined. These more complicated models do not usually need any a priori knowledge of the native state.

Chapter 2

Simulation Methods

2.1 Introduction

Computer simulations are an important complement to the theoretical and experimental techniques used in the sciences. Simulations serve a twofold purpose in science. Firstly they can offer insight into the characteristics of a system for which there is no experimental data available such as the folding pathways of proteins. Secondly simulation can be used to solve otherwise impractical theoretical models such as complex quantum systems. By taking advantage of the ability to explicitly model every facet of a system a simulation can probe a model under any set of assumptions [27]. For example the balance of forces which drive secondary and tertiary structure formation in proteins can be examined by varying the strength of the interactions driving the folding process [28].

Monte Carlo techniques were first developed as a way to study systems with many degrees of freedom [27]. The method takes its name from the use of random numbers and probability distributions to explore the properties of a system through a governing expression like a Hamiltonian. The Metropolis Monte Carlo method has been used extensively to study the protein folding problem [16, 17]. The method can be used to study both equilibrium and stochastic processes of a non-equilibrium system if a physically realistic move set is chosen. Monte Carlo methods have and continue to provide great insight into the protein folding problem. The sections of this chapter detail the theory of the Metropolis Monte Carlo method and the parallel tempering Monte Carlo method, an efficiency algorithm used to overcome deficiencies in standard Metropolis Monte Carlo simulations.

2.2 The Metropolis Monte Carlo Method

The Metropolis Monte Carlo method generates new configurations by treating the trajectory of the system as a Markov chain. That is, each new state m is randomly

generated from the previous state n and a suitable transition probability without consideration of how state n arose [27, 29]. In this model time is represented through the Monte Carlo steps and the time evolution of the system is governed by

$$\frac{\partial P_n(t)}{\partial t} = -\sum_{n \neq m} \left[P_n(t) W_{n \to m} - P_m(t) W_{m \to n} \right], \tag{2.1}$$

where $P_n(t)$ is the probability of the system being in state n at time t and $W_{n \to m}$ is the transition rate for the $n \to m$ transition. When the system reaches equilibrium the time derivative $\partial P_n(t)/\partial t$ will be zero and the system evolution is governed by the detailed balance relationship

$$P_n(t) W_{n \to m} = P_m(t) W_{m \to n}. \tag{2.2}$$

By taking the probability of the nth state in a classical system as the Boltzmann weight with partition function Z

$$P_n(t) = \frac{1}{Z} e^{\frac{-E_n}{k_B T}}, \tag{2.3}$$

it is easy to show that the ratio of the transition probabilities will be dependent only on the energy difference of the two states and the simulation temperature through the relation

$$\frac{W_{n \to m}}{W_{m \to n}} = \frac{P_n(t)}{P_m(t)} = e^{\frac{-\Delta E}{k_B T}}, \tag{2.4}$$

where the partition function has been cancelled in the ratio. Equation 2.4 does not specify the transition rate uniquely. The Metropolis method chooses the transition rate to be of the form

$$W_{n \to m} = \begin{cases} \tau_o^{-1} e^{\frac{-\Delta E}{k_B T}} & \Delta E > 0 \\ \tau_o^{-1} & \Delta E < 0 \end{cases}, \tag{2.5}$$

where τ_o is an arbitrary factor that can be related to Monte Carlo time. This has the effect of driving the system to lower energy over time.

After the system is initialized the Metropolis Monte Carlo routine is applied through the following algorithm.

1. Perform a local move on one component of the system.
2. Calculate the energy change from the old to the new state ΔE.
3. Generate a random number r in the range $0 < r < 1$.
4. If $r < exp(-\Delta E/k_B T)$ accept the move.
5. Go to the next component of the system and go to step 2. Repeat for each component of the system until all moves have been attempted.

By repeating this procedure over multiple time steps the system evolves according to the weighting function. Eventually the system will reach its native state at which time equilibrium properties can be measured. Statistical time averages of a system evolving under the Metropolis method are calculated with the simple relationship

$$\langle A \rangle = \frac{1}{M} \sum_{i=1}^{M} A_i, \tag{2.6}$$

where M is the number of independent measurements taken over the course of a simulation. Measurements should only be taken after the system is allowed to equilibrate from its initial state for a sufficient time so that any non-equilibrium characteristics of the initial configuration are removed.

One of the more serious deficiencies of the Metropolis method is the tendency for systems to become stuck in local minimum energy wells at low temperature. If the temperature is too low the system will be incapable of jumping over the energy barriers of the local minimum so that it can reach the native state. This is especially prevalent in systems with very frustrated energy phase spaces such as proteins. The parallel tempering method discussed in the next section is an optimization technique used to counteract this deficiency.

2.3 Parallel Tempering Monte Carlo

In the parallel tempering or replica exchange method M non-interacting replicas of a system are simulated in parallel [30, 31]. In this ensemble of systems each replica m shares a common Hamiltonian and evolves according to the Metropolis Monte Carlo method outlined in Section 2.2. Each replica is simulated at a different temperature T_m and inverse temperature β_m. A state of the ensemble is specified by $\{X\} = \{X_1, X_2, \ldots, X_M\}$ and the probability of finding the system in state $\{X\}$ with temperatures $\{\beta\}$ is

$$P(\{X, \beta\}) = \prod_{m=1}^{M} P_{\text{eq}}(X_m, \beta_m), \tag{2.7}$$

where the probability of replica m being in state (X_m, β_m), $P_{\text{eq}}(X_m, \beta_m)$, is

$$P_n(t) = \frac{1}{Z(\beta_m)} e^{-\beta_m E_m}. \tag{2.8}$$

By imposing the detailed balance relationship, Equation 2.2, on the ensemble of systems the ratio of the configuration transition rates between replicas is

$$\frac{W_{X,\beta_m \to X',\beta_n}}{W_{X',\beta_m \to X,\beta_n}} = e^{-\Delta\beta\Delta E}, \tag{2.9}$$

where

$$\Delta\beta\Delta E = (\beta_n - \beta_m)(E_X - E'_X). \tag{2.10}$$

By adopting the Metropolis method the replica-exchange probability is then defined to be

$$W_{X,\beta_m \to X',\beta_n} = \begin{cases} e^{-\Delta\beta\Delta E} & \text{if} - \Delta\beta\Delta E > 0 \\ 1 & \text{if} - \Delta\beta\Delta E < 0 \end{cases}. \tag{2.11}$$

The parallel tempering algorithm consists of the following steps.

1. Initialize and run m Metropolis Monte Carlo replicas on m processors each with a temperature T_m.
2. Periodically attempt to exchange replicas between systems X_m and X_{m+1} with the Metropolis acceptance scheme and the probability defined in Equation 2.11.

The temperatures are typically arranged in an ascending or descending order and exchanges are only attempted between adjacent temperature values because the exchange rate decreases exponentially with increasing $\Delta\beta$. The temperatures are chosen such that the lowest temperature will allow the system to freeze in its native state while the maximum temperature is high enough to allow the system total freedom in energy phase space. This allows the high temperature replicas to sample the entire configurational space. When a state is found to have a preferred energy it will be shifted to processors with lower temperature to be folded in a less free environment. Eventually the lowest energy state will migrate to the lowest temperature processor.

Since the energy of the system is dependent on temperature it is difficult to determine the optimal temperature distribution between the chosen maximum and minimum values. It has been shown both theoretically [31] and empirically [32] that an average swap rate of twenty percent for all adjacent temperatures is the optimal value. Using this property and the desired temperature extremes the number of required replicas can be determined. Typically an exponential temperature distribution is best as a first guess and can then be refined through trial and error measurements. Additionally it is found that sampling efficiency is increases with increasing exchange attempt frequency [33].

Chapter 3

The Model

3.1 Introduction

The following chapter describes the structure, dynamics, and energetics of the minimal model developed in this study. Presented is a five-bead course grained off-lattice minimal model that uses knowledge based potentials to mimic interactions in real proteins. The model is a refinement of a previous minimal model developed by Imamura [1]. The chapter finishes with a discussion of the reduced unit scales used in simulation and how they relate to real proteins.

3.2 Structure and Dynamics

A protein is modelled as a series of repeating peptide units (sometimes known as peptide planes or amide planes) each containing six atoms, Figure 3.1. Each peptide unit contains two C_α's as well as one each of C, N, O, and H. Any two peptide units adjacent in the sequence will share an C_α. Each C_α is separated from its neighbours by a fixed length of l and is surrounded by a spherical hard boundary of radius $1.2l$. The C and N atoms in each peptide unit are explicitly modeled in their proper positions as defined in the Pauling Corey model [12]. The O and H atoms are placed to act as hydrogen bonding interaction centres and as such are shifted from their true positions in a real peptide unit. The positions of the four non-C_α atoms relative to the C_α-C_α bond in each peptide unit are summarized in Table 3.1. The distances in Table 3.1 are expressed in a reduced length scale as compared to real proteins. This is accomplished by scaling the separation distances of the Pauling Corey model down by 3.81Å, the typical separation distance between nearest neighbour C_α atoms in real proteins. As the backbone C_α positions are typically initialized randomly and the other peptide unit atom positions are defined in relation to the backbone the only input necessary for simulation is then the classification of each of the C_α atoms.

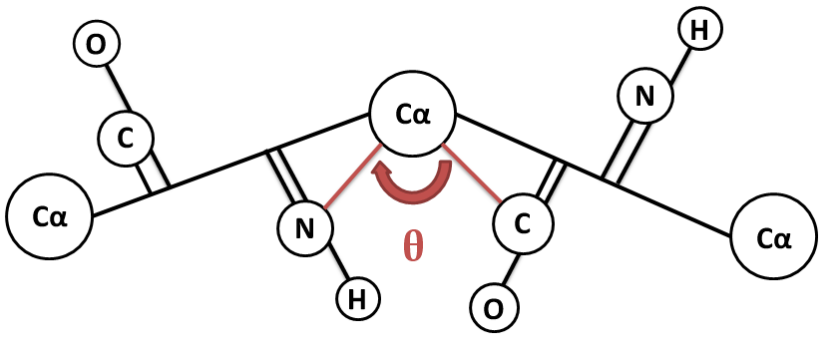


Figure 3.1: Two peptide units that share an C_α atoms are shown along with the definition of the bond angle, θ, from the bond angle potential energy.

C_α atoms in this model are classified in one of three ways. These classes control how the potential energy terms described in Section 3.3 are applied to each C_α and its associated peptide units. The three classes are non-hydrophobic, hydrophobic, and loop region. The non-hydrophobic class indicates that the corresponding residue in a real protein is either a polar or a charged residue. The hydrophobic class indicates a corresponding hydrophobic residue in real proteins while the loop region class indicates the residue is normally found in a loop region in real proteins. The bond angle potential energy is applied to each bond angle in the system independent of the associated C_α atoms class. Hydrophobic C_α's are the least spatially free structural units in a system. These C_α atoms are attracted to other hydrophobic atoms through the hydrophobic potential energy. The O and H atoms in each adjacent peptide unit are attracted to H and O atoms, respectively, in other peptide units through the hydrogen bonding potential energy and the peptide unit orientations are restricted through the Ramachandran angle potential energy. Non-hydrophobic C_α atoms experience the same interactions as hydrophobic C_α's except for hydrophobic potential energy interactions with other C_α atoms which are set to zero. Loop region C_α atoms and their peptide units are allowed much more spatial freedom than the other classes. Their C_α atoms do not participate in hydrophobic interactions. The OH pair in each of the adjacent peptide units does not participate in hydrogen bonding interactions and the Ramachandran angle potential energy is not considered for any Ramachandran angle pair associated with a loop region C_α.

The backbone atoms and their peptide units are folded with one of two move sets depending on their placement in the sequence. The first move set applies to terminal C_α atoms. A terminal C_α and its single associated peptide unit are rotated away from their current positions with the three Euler angle rotation matrices, Equation 3.1 [34]. The rotations are centred around the terminal C_α's nearest neighbour C_α.

$$A = \begin{bmatrix} cos\psi & sin\psi & 0 \\ -sin\psi & cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos\theta & sin\theta \\ 0 & -sin\theta & cos\theta \end{bmatrix} \begin{bmatrix} cos\phi & sin\phi & 0 \\ -sin\phi & cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.1}$$

Table 3.1: The positions of the N, C, O, and H atoms in each peptide unit are listed relative to the position of the first C_α atom in the peptide unit. The parallel and perpendicular components represent how far along and away from the C_α-C_α bond each atom is placed. A negative indicates the atom is on the opposite side of the bond from its counterparts. All distances are expressed in units of l.

Atom	Parallel Distance	Perpendicular Distance
C	0.3727447	-0.1520979
N	0.6234216	0.0883107
O	0.4226724	-0.77
H	0.5741070	0.65

The second move set applies to non-terminal or central C_α atoms. In this move set a central C_α denoted i and its associated peptide units, denoted (i-1) and i for the left and right peptide units respectively, are rotated around an axis defined between $C_{\alpha(\text{i-1})}$ and $C_{\alpha(\text{i+1})}$ with Equation 3.2 [34]. Peptide unit i is then rotated around the axis defined between $C_{\alpha(\text{i})}$ and $C_{\alpha(\text{i+1})}$ using Equation 3.2.

$$\vec{r'} = \vec{r}cos(\theta) + \vec{n}\,(\vec{n}\cdot\vec{r})\,(1 - cos(\theta)) + (\vec{r}\times\vec{n})sin(\theta) \tag{3.2}$$

3.3 Energetics

This model employs four potential energy expressions to drive the folding process. Potentials are defined to enforce a bond angle between adjacent peptide units, to mimic steric restrictions, and to mimic hydrogen bonding and hydrophobic effects. The basic unit of distance is l, as defined in the previous section, and the basic unit of energy is ϵ.

The protein chain folds under a stiff harmonic potential energy which enforces a bond angle between the vectors defined by the N-C_α and C-C_α atoms where the C_α atom is shared between the vectors, Figure 3.1. The potential takes the form

$$V_{\text{BA}} = \frac{K_{\text{BA}}}{2}(R^2 - R_0^2)^2, \tag{3.3}$$

where $K_{\text{BA}} = 450\epsilon$, R is the separation distance of the N and C atoms and R_0 is the preferred separation distance. Using the magnitudes of the N-C_α and C-C_α vectors the preferred separation distance is set through the cosine law to enforce a bond angle of 111° as dictated by real protein structure [35]. Separation distances are used instead of the bond angles to improve computational efficiency by avoiding expensive calls to the cosine and square root functions.

By freeing the peptide units to rotate around their associated C_α-C_α bond instead of fixing their orientation as in the Imamura model secondary structure can be more accurately modelled. This structural improvement comes with a large increase in conformational freedom which greatly increases the number of non-native

states a system can take. To improve computational efficiency and avoid sterically unfavourable conformations in the folding pathway this increase in freedom is counteracted by imposing an energy penalty on the Ramachandran angles in the system. A preferred set of Ramachandran angles is established with a simple well potential defined by

$$V_{\text{Ram}} = \begin{cases} -2\epsilon & \text{if } |\phi - (\phi_o)_{i\ \text{Cen}}| < (\phi_o)_{i\ \text{Tol}} \text{ and } |\psi - (\psi_o)_{i\ \text{Cen}}| < (\psi_o)_{i\ \text{Tol}} \\ 0 & \text{Otherwise} \end{cases}, \quad (3.4)$$

where a pair of Ramachandran angles ϕ and ψ within an angular region i centred on $(\phi_o)_{i\ \text{Cen}}$ and $(\psi_o)_{i\ \text{Cen}}$ and with a tolerance $(\phi_o)_{i\ \text{Tol}}$ and $(\psi_o)_{i\ \text{Tol}}$ will be energetically preferred. This potential significantly reduces the number of energetically favorable configurations. The Ramachandran potential energy is not applied to any Ramachandran angles which involve a neutral C_α in their calculation. For the purposes of this study the left handed α-helix region is defined by $(\phi_o)_{i\ \text{Cen}} = -57^o$ and $(\psi_o)_{i\ \text{Cen}} = -47^o$ with the tolerance $(\phi_o)_{i\ \text{Tol}} = (\psi_o)_{i\ \text{Tol}} = 15^o$. In the future this potential could be generalized for many regions of the Ramachandran plot to allow multiple types of secondary structure to form.

Hydrogen bonding is established between the O and H atoms in each peptide unit through the shifted Lennard-Jones potential

$$V_{\text{OH}} = 4\epsilon \left(\left(\frac{l}{r + r_0} \right)^{12} - \left(\frac{l}{r + r_o} \right)^{6} \right), \quad (3.5)$$

where r is the distance between any pair of oxygen and hydrogen atoms not in the same peptide unit and $r_o = 2^{1/6} l$ sets the minimum energy to occur at an OH separation distance of zero. This energy will contribute $-\epsilon$ to the total energy for each fully realized bond. If either of the peptide units in an interaction contain a loop region classed C_α then V_{OH} is zero.

The hydrophobicity of the sequence is modelled with another Lennard-Jones potential. Any two non-nearest neighbour C_α atoms classed as hydrophobic interact through the relation

$$V_{\text{Hy}} = 4C_h\epsilon \left(\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^{6} \right), \quad (3.6)$$

where C_h controls the strength of the hydrophobic interaction, $\sigma = 1.2(2^{-1/6})l$ scales the minimum energy to a separation distance equal to the diameter of the hard spherical boundary around each C_α, and r is the separation distance of the two hydrophobic C_α's. This potential will contribute $-C_h\epsilon$ to the energy whenever two hydrophobic beads come to a separation distance of $1.2l$.

3.4 Reduced Units and Real Proteins

The output generated from simulations using this model is presented in reduced units as defined in equation 3.7. The length scale is set through the constant l

and the energy scale is set through the constant ϵ. As mentioned previously the length scale is set by comparing a fixed distance in the model with its respective distance in a real protein. The distance chosen is the separation between C_α atoms in the backbone chain. In simulation this distance is unity while in real proteins it is 3.81Å. Thus any distances in simulation output should be scaled accordingly for comparison to real structures. The energy scale is set by choosing an appropriate value for epsilon. For this model a reasonable value is $\epsilon \approx 1kcal/mol$, the energy of a typical hydrophobic bond [36]. From this epsilon it is a simple matter to calculate real world temperatures. For example at a simulation temperature of $\tilde{T} = 0.6$ the corresponding real temperature is $T \approx 302K$.

$$\tilde{r} = \frac{r}{l}, \quad \tilde{V} = \frac{V}{\epsilon}, \quad \tilde{T} = \frac{k_B T}{\epsilon} \tag{3.7}$$

Chapter 4

Results and Discussion

4.1 Introduction

This chapter presents the results of simulations performed with the minimal model discussed in Chapter 3. The first section of the chapter defines the thermodynamic quantities that are measured and discusses how they relate to the folding of proteins. Focus is placed on measuring the characteristic collapse and folding temperatures. The next section details the structure of an α-helix in the minimal model and compares it to real helix structural properties. Finally, the results of four-helix bundle simulations are reported and the characteristic temperatures are determined.

4.2 Thermodynamic Properties

4.2.1 Characteristic Temperatures

The phases of protein folding are typically characterized by two temperatures [2, 9, 22]. At sufficiently high temperature the protein is in the unfolded state (U) and is expected to behave as a random coil. In practice there can be remnants of secondary structure present even in this phase. As the temperature decreases to the collapse temperature T_θ the chain folds into a compact phase known as the intermediate (I) or molten globule phase. The phase change can be first or second order depending on the nature of the interactions driving it. The different properties of the twenty amino acid side chains cause a subset of possible conformations to have lower energies. This defines a second phase transition temperature, the folding temperature T_f, below which the molten globule folds into the native state. These two temperature obey the relation $T_f \leq T_\theta$.

4.2.2 Measuring The Characteristic Collapse Temperature

The characteristic collapse temperature is typically measured in two ways [9, 22]. Since this phase transition is (usually) second order T_θ can be determined from a plot of heat capacity versus temperature. The heat capacity is defined as

$$C_v = \frac{\langle E^2 \rangle - \langle E \rangle^2}{T^2}, \tag{4.1}$$

where E is the total energy of the system, T is the simulation temperature, and $\langle\rangle$ is a time average. The peak in the specific heat is taken to be T_θ. An alternate methodology to measure T_θ utilizes the radius of gyration of the system. The radius of gyration is defined as

$$R_g^2 = \frac{1}{N} \sum_{i=1}^{N} (\vec{r}_i - \vec{r}_{\text{mean}}) \tag{4.2}$$

where $\vec{r}_i$ is the position vector of the ith C_α and $\vec{r}_{\text{mean}}$ is a vector to the average C_α position in the current system. R_g^2 provides a measure of the effective size of the system. At high temperatures R_g^2 is maximal because the random coil is typically in an extended conformation. As temperature lowers the coil collapses and R_g^2 drops suddenly. This drop is indicative of the collapse transition. T_θ is measured as the point on a plot of R_g^2 versus temperature where R_g^2 first stops decreasing and starts to plateau at the value for the compact globule.

4.2.3 Measuring The Characteristic Folding Temperature

The folding transition temperature can be measured in a variety of ways [9, 22]. The simplest method for measuring T_f is through fluctuations in the structural overlap function. The structural overlap function measures the degree to which one structure is like a given reference structure within a tolerance and is defined as

$$\chi = 1 - \frac{1}{N'} \sum_{i=1}^{N-3} \sum_{j=i+3}^{N} \Theta\left(0.2l - \left|r_{ij} - r_{ij}^o\right|\right), \tag{4.3}$$

where r_{ij} is the distance between $C_{\alpha(\text{i})}$ and $C_{\alpha(\text{j})}$, r_{ij}^o is the corresponding distance in the native state, N' is a normalization which sets χ to 0 for a perfect structural match and 1 for no match, and $\Theta(x)$ is the Heaviside function. The sums in the χ calculation skip atomic placements where the indexes i and j are within three of each other because atoms which are close to each other in the sequence will tend to be within the tolerance of the χ equation and counting them provides no relevant indication of whether the structure matches the native state. Since loop region C_α's in the minimal model are nearly free in energy space, see Section 3.3, they are excluded from the i and j loops in the calculation of χ.

Fluctuations in χ are defined as

$$\Delta\chi = \left\langle \chi^2 \right\rangle - \langle \chi \rangle^2, \tag{4.4}$$

where $\langle\rangle$ is a time average. For a sequence with a unique native state $\Delta\chi$ will exhibit a peak at T_f corresponding to a sudden drop in χ. For a system with multiple minimum energy states a different reference conformation must be defined for each minima. Accurate measurements of $\Delta\chi$ are further complicated in the parallel tempering method since multiple native states are typically exploring energy and temperature space in the same simulation. Thus as the configurations swap between T_i and T_{i+1} the time averages average measurements from different native states in the same calculation.

4.2.4 Potential Energy Fluctuations

Although the fluctuation in the total energy is encapsulated in the specific heat measurements it is still useful to examine the fluctuation of each type of energy on its own. The fluctuation in energy type i is defined by

$$\Delta E_i = \left\langle E_i^2 \right\rangle - \left\langle E_i \right\rangle^2 . \tag{4.5}$$

where $\langle\rangle$ is a time average. By studying the fluctuation of each energy term individually the effects of the potentials driving protein folding in the minimal model can be examined in detail. The primary drivers of secondary structure formation are the Ramachandran angle potential and the hydrogen bonding potential energies while tertiary structure formation is driven by the hydrophobic potential energy. Plots of the energy fluctuation versus temperature allow the determination of which interactions drive the folding at different temperatures and thus the characteristic temperatures of secondary and tertiary structure formation can be determined.

4.3 The Alpha Helix of the Minimal Model

Several simulations of a sequence consisting of fourteen non-hydrophobic C_α atoms were run. As expected the native state is an α-helix, Figure 4.1. The helix in the minimal model displays the correct number of residues per turn around the helical axis, 3.6 in real α-helices. By explicitly calculating all of the hydrogen bonding interactions in the sequence the expected i to (i+4) oxygen to hydrogen bonding pattern is observed, Figure 4.2. The Ramachandran angles for this structure are shown in Table 4.1. The angles vary only slightly around the expected values for an ideal α-helix. From these structural properties it is concluded that the minimal model accurately represents the α-helix secondary structure.

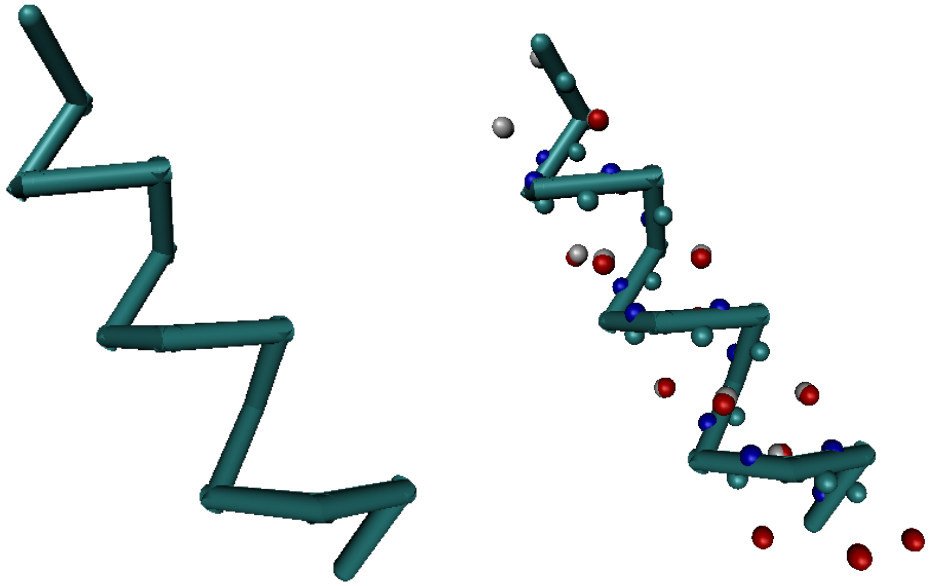


Figure 4.1: A Typical α-helix in the minimal model is shown. a) Only the backbone C_α atoms and the bonds between them are shown. b) The backbone and the peptide unit atoms are shown. C, N, O, and H are represented as green, blue, red, and white spheres respectively.

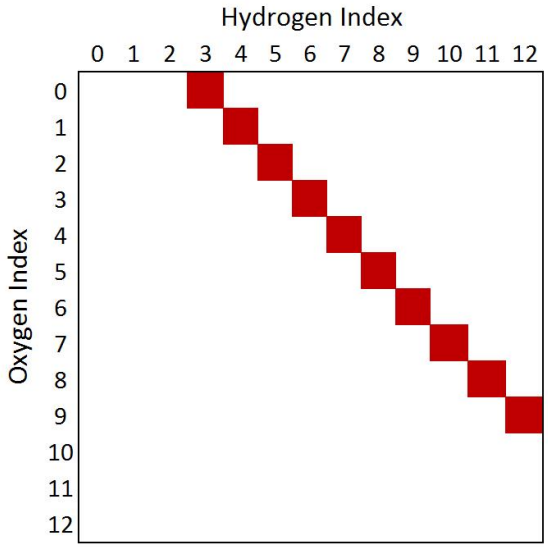


Figure 4.2: α-Helix Hydrogen Bonding Contact Map. Red squares indicate a bond with energy at most -0.9ϵ.

Table 4.1: The Ramachandran angles for an $N = 14$ monomer α-helix in the minimal model are listed. The ideal values for ϕ, -47°, and ψ, -57°, are taken from Table 1.1.

Bond Index	ϕ	ψ	% from Ideal ϕ	% from Ideal ψ
1	-50.75	-59.33	7.97	4.09
2	-49.68	-57.14	5.70	0.25
3	-49.84	-60.06	6.04	5.37
4	-49.79	-60.74	5.93	6.57
5	-45.76	-59.16	2.64	3.80
6	-52.85	-61.14	12.44	7.27
7	-43.59	-61.53	7.25	7.95
8	-48.88	-61.49	3.99	7.87
9	-48.26	-61.19	2.69	7.36
10	-46.43	-61.56	1.21	8.01
11	-47.35	-60.90	0.75	6.84
12	-48.36	-59.89	2.89	5.08

4.4 Four-Helix Bundles in the Minimal Model

4.4.1 Primary Sequence and Simulation Details

A sequence of sixty five C_α atoms is used to generate the conformations discussed in the following sections. The sequence is composed of four α-helix structural units, each of fourteen C_α atoms, connected by three loop regions, each of three loop region class C_α atoms. The α-helix region C_α atom classification follows a repeating pattern of two non-hydrophobic followed by two hydrophobic C_α atoms. This has the effect of making one side of the helix hydrophobic in nature. Similar sequences were studied by Thirumalai [36] and Rey and Skolnick [28] using different minimal models.

The simulations were carried out such that the hydrophobic strength factor C_h was varied between 0.0 and 1.5 in steps of 0.1. Ten parallel tempering simulations were carried out at each value of C_h for 175 million Monte Carlo steps. It is convenient for the following discussion to define a low hydrophobic strength regime as $0.0 \leq C_h \leq 0.5$, a medium strength regime as $0.6 \leq C_h \leq 1.2$, and a high strength regime as $C_h \geq 1.3$. The simulation temperatures are distributed between $T = 0.025\epsilon/k_B$ and $T = 12.8\epsilon/k_B$. The structure and energy measurements discussed below are taken at the lowest temperature in the spectrum.

4.4.2 Characterizing the Native States

In the following discussion six native state four-bundles will be identified. A misfold state is defined as any structure that does not fold into one of the six native states.

The number of simulations which resulted in the folding of a native state or a misfold are recorded in Table 4.2 for each hydrophobic strength. The average energies of the six native states and the misfold states are recorded in Table 4.3 for each hydrophobic strength. A scaled conformational energy is calculated by dividing the hydrophobic energy contributions to each structures total energy by the strength of the hydrophobic interactions, C_h. This puts the energies of different hydrophobic strengths on an equal footing for easier comparison. The average scaled energies for each of the native states are listed in Table 4.4.

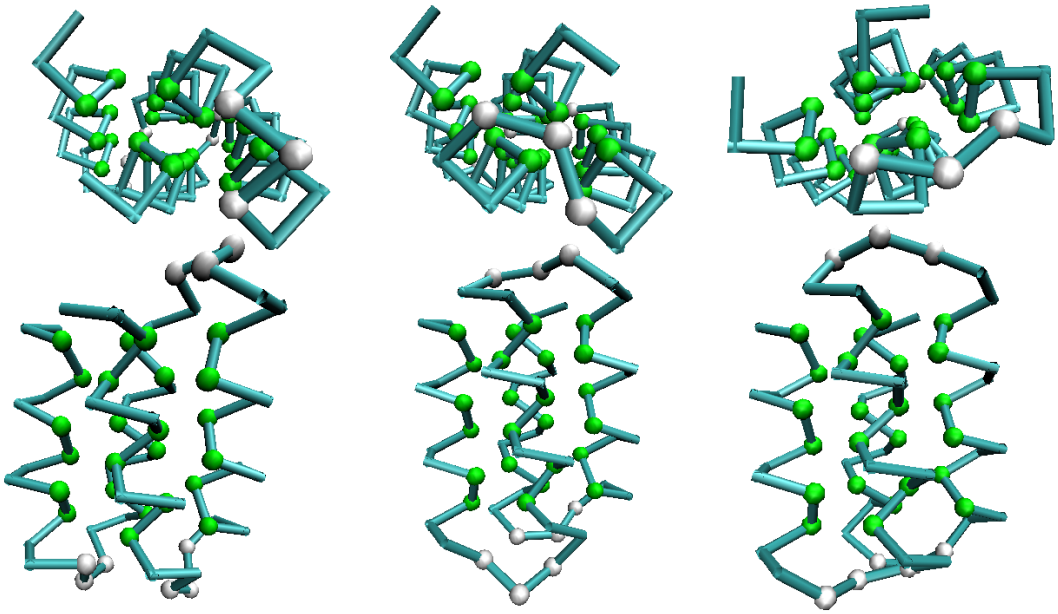
The six native states of this sequence are energetically degenerate and are found in the medium strength regime. The hydrophobic strength in this regime is strong enough to bring the four α-helices together to form the four-bundle conformations but not so strong that it overpowers the secondary structure formation and breaks the α-helices. The native states are shown in Figure 4.3. Four of these states are classified as U-bundle conformations and two as Z-bundle conformations. The U-bundles in this model are typified by hydrophobic bonding inducing an α-hairpin structure with helices one and three or two and four. The non-hairpin helices then line up along either side of the hairpin to cover the exposed hydrophobic C_α atoms on each side. The Z-bundles are arranged in the same way except the hairpin is formed with helices two and three. Figure 4.4 shows a simple representation of the helix packing in these six structures.

Other less common structures include a four-bundle composed of two α-hairpin structures that have come together to share some hydrophobic bonds and a three-helix bundle with a fourth helix sitting away from the bundle. These are typically found at the lower end of the regime and have higher energy than the six native states. It is expected that allowing the simulation to run longer will eventually produce one of the six native states. Limits on computational power have prevented verification of this expectation.

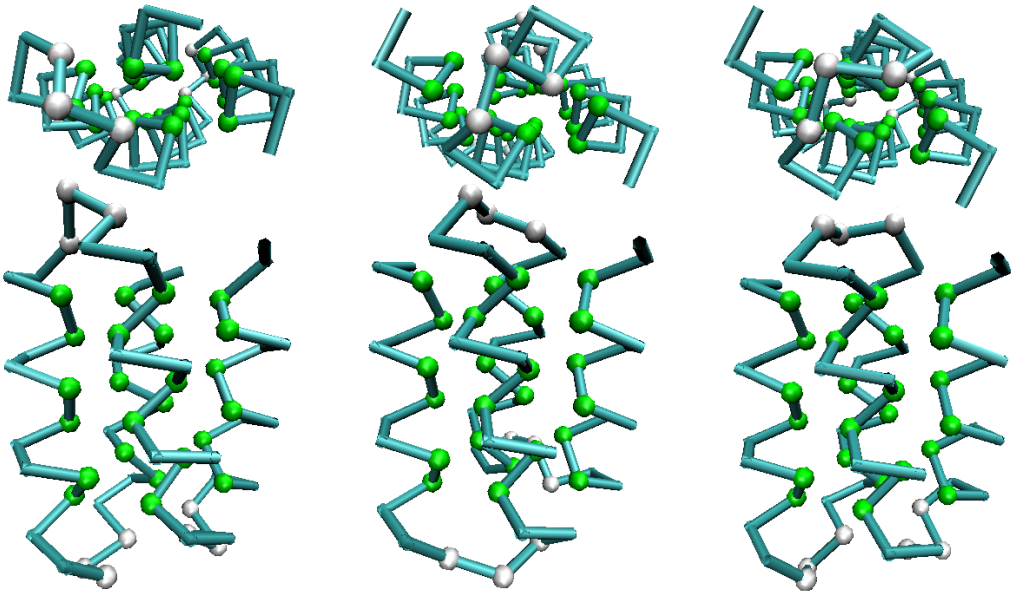
As expected the energy of the bundles decreases as the hydrophobic strength increases, Table 4.3. The energies of the native states are very similar to each other and are lower than the energies of the misfold states. This is easier to see by examining the scaled bundle energies, Table 4.4. These scaled energies are all very similar for the six bundle types while the average misfold energy is significantly higher.

The minimum energy states in the low strength regime are composed of four perfectly formed α-helices with no preferred tertiary structure. The four helices orient themselves randomly with respect to the other helices since in this regime the weak hydrophobic interactions provide little to no drive for the helices to assume a unique tertiary structure, Figure 4.5. Occasionally if two helices line up end-to-end hydrogen bonding will form between the O atoms in one helix and the H atoms in the other. At the high end of the regime α-hairpins occasionally form since the hydrophobic interactions are just strong enough to be preferred over the end-to-end hydrogen bonding.

The energies in the low strength regime are much higher than in the medium

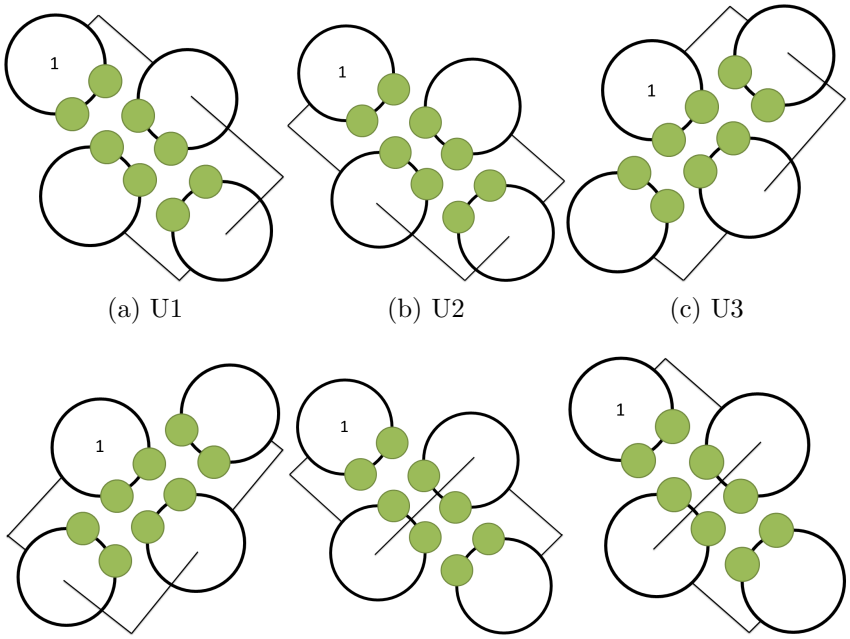


(a) U1 (b) U2 (c) U3



(d) U4 (e) Z1 (f) Z2

Figure 4.3: Shown are the six native states for the 65 monomer sequence considered. Each state is shown in a side and top profile. Hydrophobic atoms are bright green spheres and loop region atoms are large white spheres.



(a) U1 (b) U2 (c) U3

(d) U4 (e) Z1 (f) Z2

Figure 4.4: U and Z packing arrangements in the minimal model. Large white circles are α-helices as seen by looking down the helical axis. Helix one is labelled in each diagram. Black lines represent the loop regions. Green circles represent hydrophobic backbone atoms.

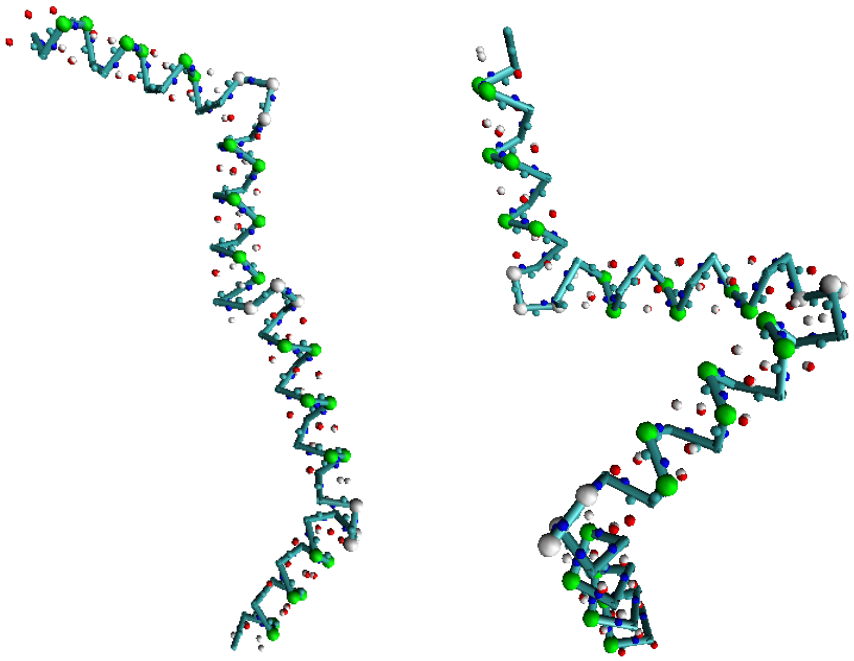
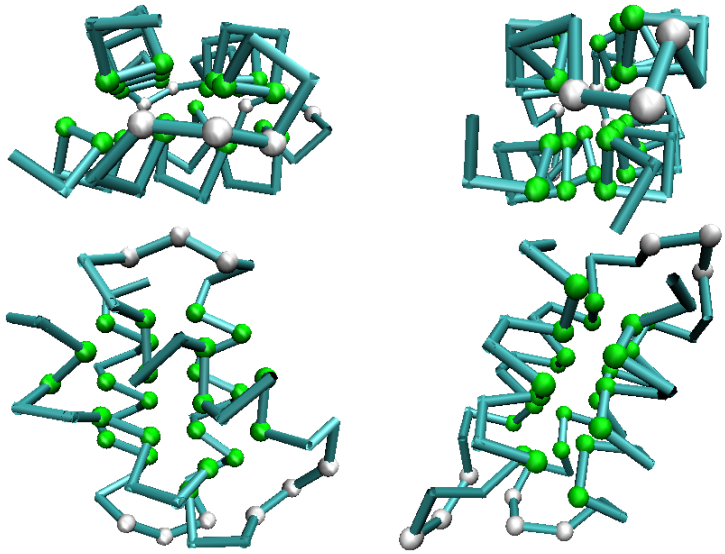


Figure 4.5: Shown are two typical structures from the low strength regime where the α-helices did not fold to a compact tertiary structure.

strength regime. This is primarily caused by the low hydrophobic strength being unable to promote tertiary structure formation. The scaled energies in the low strength regime are typically at least 5ϵ higher than the bundle energies and can be as much as 20ϵ higher.

In the high strength regime the propensity for misfolds sharply increases over that of the medium regime due to two main factors. Firstly, the increased hydrophobic strength increases the chances of helix pairs one-two and three-four forming separate α-hairpins. These hairpins then come together in a four bundle but do not form as many hydrophobic bonds as the minimum energy states in the medium regime, Figure 4.6. Secondly, the tendency of the hydrophobic side of the α-helix to compress so that hydrophobic C_α atoms are brought closer together bends the helices and breaks the hydrogen bonds along the compressed side. There is also a higher occurrence of kinks in a helix at the extreme end of the high strength regime as hydrophobic bonding overpowers the forces driving secondary structure. These kinks are typified by Ramachandran angles outside the potential well.

Energies for the few bundle states seen in the high strength regime are comparable with energies in the medium strength regime after scaling, Table 4.4. As expected the bundle state energies are lower than the energies of the misfold states indicating a set of local minima. It is likely that these local minima, such as the two hairpin structure, are more kinetically favoured since it is easier to form the hairpins from pairs of adjacent helices. The helix bending is a direct result of the hydrophobic interactions within a helix overpowering the hydrogen bonding inter-



(a) A Generic Misfold State (b) A Helix Hairpin Structure

Figure 4.6: Two sample structures from the high strength regime are shown with a side and top profile. Hydrophobic C_α atoms are bright green while loop region atoms are white. a) shows a misfolded state where two of the helices are are malformed. b) shows a helix hairpin structure.

actions. Although not explored here due to limits on computational power it is expected that as the hydrophobic energy well depth continues to increase the helical structure will increasingly breakdown until the structures resemble a compact but random globule.

4.4.3 The Folding Path and Characteristic Temperatures

The following section presents an analysis of the folding path of the four-helix bundle in temperature space and details the methods used to obtain the characteristic temperatures. Before continuing it is prudent to describe the methods used to process the simulation data to make the necessary plots. For each hydrophobic strength ten simulations were run with thirty two temperatures in each. Each simulation performs a time average at each temperature for each quantity measured. The specific heat, radius of gyration, and energy fluctuation plots are obtained by averaging the ten time averages with the same hydrophobic strength at each temperature. For each simulation at each temperature there are six measurements of the fluctuation of the structural overlap function. A plot is obtained by first averaging these six values to generate one value for each temperature and simulation. These values are then averaged in the same way as the other three measurements to generate a curve.

Table 4.2: Structure Count by Conformation and Hydrophobic Strength

C_h	Misfold	U1	U2	U3	U4	Z1	Z2
0.0	10	0	0	0	0	0	0
0.1	10	0	0	0	0	0	0
0.2	10	0	0	0	0	0	0
0.3	10	0	0	0	0	0	0
0.4	10	0	0	0	0	0	0
0.5	10	0	0	0	0	0	0
0.6	3	2	1	1	0	1	2
0.7	1	1	0	2	2	3	1
0.8	0	0	0	3	5	0	2
0.9	0	0	1	3	1	2	3
1.0	1	1	0	3	2	3	0
1.1	2	1	1	1	1	2	2
1.2	2	1	1	1	1	1	3
1.3	5	0	0	1	1	2	1
1.4	9	0	0	0	1	0	0
1.5	10	0	0	0	0	0	0

Table 4.3: Average Bundle Energy by Hydrophobic Strength. Energies are reported in units of ϵ. Values of NA indicate there is no available data for the specified hydrophobic strength and state combination.

C_h	Misfold	U1	U2	U3	U4	Z1	Z2
0.0	-156.69	NA	NA	NA	NA	NA	NA
0.1	-158.47	NA	NA	NA	NA	NA	NA
0.2	-162.07	NA	NA	NA	NA	NA	NA
0.3	-167.42	NA	NA	NA	NA	NA	NA
0.4	-174.35	NA	NA	NA	NA	NA	NA
0.5	-181.57	NA	NA	NA	NA	NA	NA
0.6	-188.79	-191.22	-190.82	-192.28	NA	-190.33	-191.47
0.7	-194.66	-200.22	NA	-197.69	-199.87	-199.76	-200.17
0.8	NA	NA	NA	-205.98	-205.45	NA	-205.89
0.9	NA	NA	-213.00	-212.78	-212.65	-212.49	-212.16
1.0	-215.73	-217.26	NA	-219.51	-219.56	-219.13	NA
1.1	-224.43	-226.43	-225.50	-226.82	-224.83	-225.49	-225.88
1.2	-231.70	-232.35	-233.49	-234.04	-232.31	-230.94	-232.96
1.3	-236.61	NA	NA	-235.70	-239.45	-238.81	-239.99
1.4	-243.49	NA	NA	NA	-245.64	NA	NA
1.5	-250.35	NA	NA	NA	NA	NA	NA

Table 4.4: The average of the scaled bundle energies are listed by bundle type. Energies are reported in units of ϵ.

Structure	Energy
U1	-217.84
U2	-218.33
U3	-218.84
U4	-218.67
Z1	-218.39
Z2	-218.67
Misfold	-203.51

The Characteristic Collapse Temperature

The characteristic temperatures measured from the peak of the specific heat plots at each hydrophobic strength are listed in Table 4.5. Specific heat plots in the low and medium strength regimes show a single large peak indicative of a phase transition, Figure 4.7. As the hydrophobic strength is increased the amplitude of the peak decreases while its width increases and its centre shifts upward in the temperature spectrum. In the high strength regime the peak continues to shift to higher temperatures as hydrophobic strength increases. There is a second peak in the high strength regime centred around $T \approx 0.52\epsilon/k_B$ which remains fixed as hydrophobic strength increases. The amplitude of this stationary peak initially decreases as the moving peak separates from it. At C_h 1.4-1.5 it remains constant in amplitude and width. The peak which shifts to higher temperatures indicates a phase transition from random coil to compact globule and coincides with the peaks in the hydrophobic and hydrogen bonding energy fluctuation curves discussed below. The second peak comes from the constant peaks in the Ramachandran angle and hydrogen bonding energy curves. All of the specific heat curves show a sharp increase as the temperature decreases to very low values. These specific heat spikes are a consequence of dividing the total energy fluctuation by T^2. The energy fluctuations in this region are on the order of unity and T is on the order of $0.01\epsilon k_B$.

All of the radius of gyration curves plateau at $R_g^2 \approx 39l^2$ at high temperature. This is the size of the random coil for this sequence. As T decreases R_g^2 drops suddenly indicating the random coil has collapsed into a compact state. This typifies a phase transition from random coil to compact globule. The temperature at each of these drops are listed in Table 4.5. R_g^2 plots are shown for each hydrophobic strength in Figure 4.7. Temperatures could not be measured in the low strength regime because as temperature decreases R_g^2 fluctuates and there is no clear compact state plateau. In addition R_g^2 in the low strength regime never reaches as low a values as in the medium and high strength regimes. This is a result of the four helices in this region spreading out in space since there is not enough hydrophobic force to drive them together. In the medium and high strength regimes there is a

more uniform decrease until $R_g^2 \approx 5.5l^2$ and then the value plateaus, Figure 4.7. These temperatures show excellent agreement with the centre of the specific heat peaks.

The characteristic collapse temperatures measured from the radius of gyration and the specific heat are largely the same for each hydrophobic strength. The temperatures increase as hydrophobic strength increases since the deeper hydrophobic well depth drives the system to collapse at higher temperature.

The Characteristic Folding Temperature

The average of the structural overlap functions is consistently low at high temperatures and increases as temperature decreases for all hydrophobic strengths. The small magnitude of the fluctuation at high temperatures is because the six structural overlap measurements are uniformly around $\chi \approx 0.95$, the structural overlap of the native states with a random coil. There is little fluctuation from this value at high temperature. As the temperature decreases and the conformation collapses the six structural overlap measurements all decrease and the fluctuation stays relatively low. Eventually as temperature decreases the native states appear and one of the six structural overlap measurements sharply decreases. Now there are five high structural overlaps, $\chi \approx 0.5 - 0.7$, and one low structural overlap, $\chi \leq 0.1$. In Metropolis Monte Carlo on one processor and therefore one temperature the structural overlaps at low temperature would have little to no fluctuation since once the native state was folded it would remain in that state forever. In parallel tempering different native states can fold at different temperatures and then exchange positions in the temperature spectrum. This means that the configuration swapping can suddenly change the structural overlap values and greatly increase the fluctuation.

There is a small bump in each fluctuation curve in the medium and high strength regimes which shifts to higher temperatures at higher hydrophobic strength. The bumps also get wider at higher temperature. The bumps indicate a very large change in the structural overlap. The temperatures at the centre of each bump are listed in Table 4.5. The bumps are not visible in the low strength regime. These temperatures are all found to be similar to the collapse temperatures discussed above. The larger differences seen in the high strength regime are a result of the sparsity of temperatures in this part of the temperature spectrum making an accurate determination of the peak location difficult.

T_f can also be estimated from the density of native states in temperature space. A native state density can be estimated by counting the number of structures in one of the six native states for each temperature and hydrophobic strength, Figure 4.8. From this it is seen that native states, typified by a structural overlap < 0.36, are not found above a temperature of $T \approx 0.30\epsilon/k_B$. Further the maximum temperature at which native states are found decreases as the hydrophobic strength differs from an apparent ideal range of $C_h = 0.7 - 1.0$. This indicate that the folding transition

Table 4.5: Characteristic temperatures measured from specific heat, radius of gyration, structural overlap, and the native state density map are listed. Temperatures are reported in units of ϵ/k_B.

C_h	From C_v	From R_g^2	From $\Delta\chi$	From The Density Map
0.0	0.435	NA	NA	NA
0.1	0.44	NA	NA	NA
0.2	0.445	NA	NA	NA
0.3	0.45	NA	NA	NA
0.4	0.455	NA	NA	NA
0.5	0.45	NA	NA	NA
0.6	0.46	0.431	0.432	0.154
0.7	0.475	0.45	0.468	0.154
0.8	0.495	0.495	0.495	0.222
0.9	0.518	0.52	0.525	0.222
1.0	0.527	0.53	0.58	0.305
1.1	0.564	0.58	0.63	0.305
1.2	0.6	0.64	0.66	0.305
1.3	0.63	0.66	0.73	0.055
1.4	0.722	0.66	0.76	0.028
1.5	0.754	0.76	0.77	NA

temperatures are in fact lower than the collapse transition temperatures measured above and that the $\Delta\chi$ plots used to measure them were measuring the collapse transition instead.

Potential Energy Fluctuations

Large or sudden changes in the different energies which drive folding are indicative of different folding phases. For example, sudden changes in the forces driving secondary structure formation can indicate the temperature of a secondary structure formation phase transition. A careful analysis of the energy fluctuations reveals the folding path in temperature space for the four-bundle sequence studied.

The fluctuation of the bond angle potential energy is the same for all values of C_h at all temperatures, Figure 4.9. The fluctuation is small at low temperatures and increases as temperature increases with no noticeable peaks or other discernible features. This is because the harmonic bond angle potential changes a great deal for small changes in the bond angle and thus at high temperature the fluctuation is large because the allowed angular shifts are large. At high temperature, $T = 12.8\epsilon/k_B$, the average bond angle energy is $E_{BA} \approx 170\epsilon$ which puts the shifting angle at a high slope section of the energy parabola further increasing the fluctuation.

The fluctuation of the Ramachandran angle potential energy shows a single peak for each hydrophobic strength, Figure 4.10. In the low and medium strength

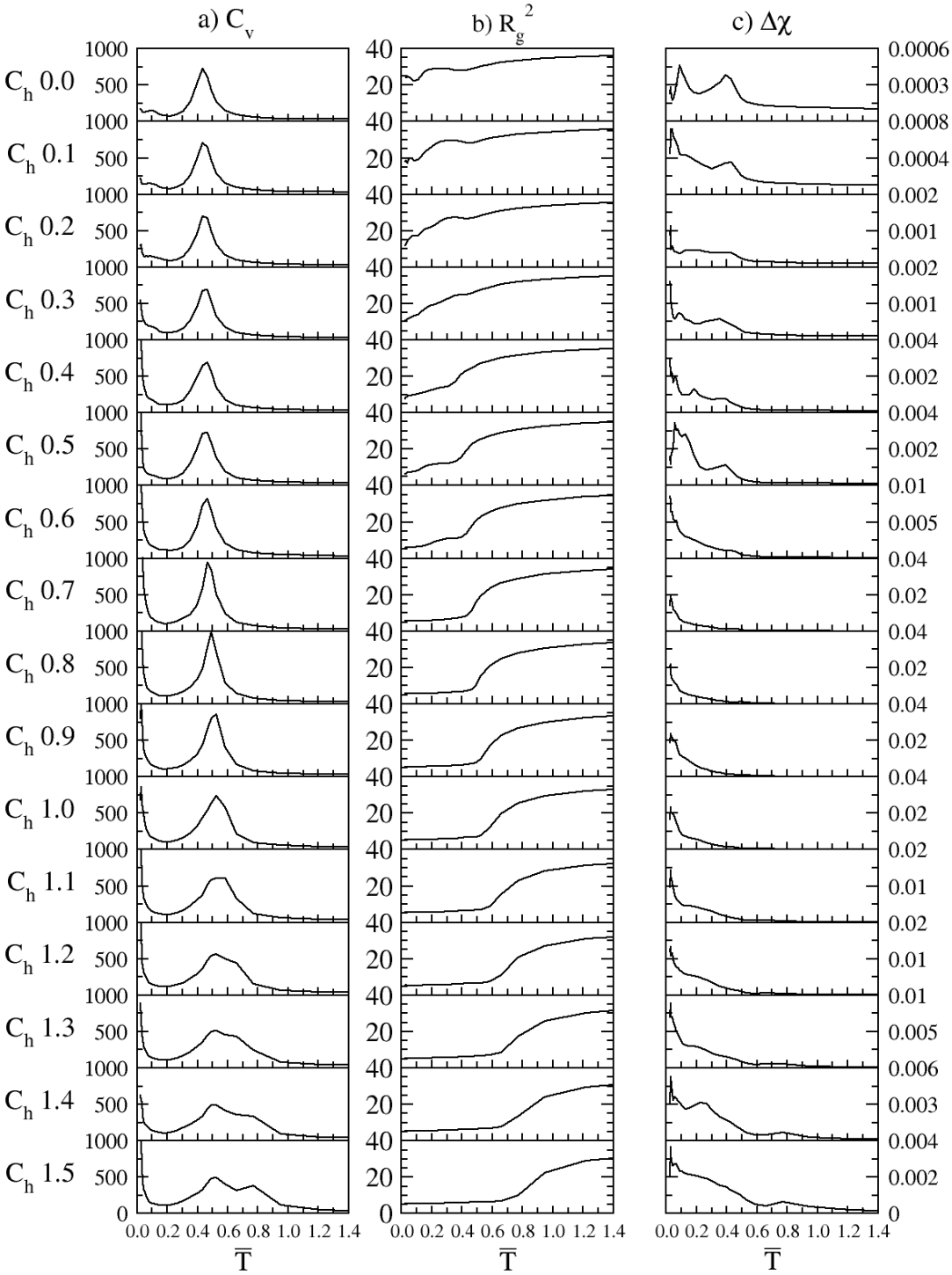


Figure 4.7: Shown is a) the specific heat versus temperature b) the radius of gyration versus temperature and c) the fluctuation in the structural overlap function vs temperature.

Temperature	Ch 0.0	Ch 0.1	Ch 0.2	Ch 0.3	Ch 0.4	Ch 0.5	Ch 0.6	Ch 0.7	Ch 0.8	Ch 0.9	Ch 1.0	Ch 1.1	Ch 1.2	Ch 1.3	Ch 1.4	Ch 1.5
0.025	0	0	0	0	0	0	7	9	10	10	9	8	8	5	1	0
0.028	0	0	0	0	0	0	7	9	9	10	9	7	8	4	2	0
0.033	0	0	0	0	1	1	5	9	9	10	9	6	7	5	0	0
0.039	0	0	0	0	0	0	4	7	10	10	7	5	4	4	0	0
0.047	0	0	0	0	0	1	3	8	7	8	8	6	6	2	0	0
0.055	0	0	0	0	0	1	0	6	6	7	9	3	4	1	0	0
0.066	0	0	0	0	0	0	0	5	5	7	7	2	3	0	0	0
0.077	0	0	0	0	0	0	1	4	3	6	6	2	3	1	0	0
0.090	0	0	0	0	0	0	2	0	4	5	6	2	2	0	0	0
0.107	0	0	0	0	0	1	1	1	2	4	5	0	2	0	0	0
0.128	0	0	0	0	0	0	0	1	2	3	2	0	2	1	0	0
0.154	0	0	0	0	0	1	1	2	1	2	3	0	2	0	0	0
0.184	0	0	0	0	0	0	0	0	2	2	3	0	1	0	0	0
0.222	0	0	0	0	0	0	0	0	1	1	3	1	1	0	0	0
0.265	0	0	0	0	0	0	0	0	0	0	1	0	3	0	0	0
0.305	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0
0.354	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.395	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.432	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.468	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.495	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.527	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.584	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.660	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.770	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.950	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.220	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.700	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2.600	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4.200	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7.000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12.800	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 4.8: The density of native states is shown for each temperature and hydrophobic strength. The numbers represent a count of the number of native state configurations found out of a possible ten.

regimes the centre and width of the peaks increases slightly while the amplitude decreases slightly with increasing temperature. In the high strength regime the peaks are essentially identical. The temperatures corresponding to the centre of each peak are summarized in Table 4.6. This indicates that around a temperature of $T \approx 0.5\epsilon/k_B$ there is a sudden increase in the number of Ramachandran angle pairs within the energy well. It should be noted that even though many of the Ramachandran angles lie within the well this does not necessarily indicate a sudden increase in secondary structure since the Ramachandran energy well is constructed to have a large area in ψ and ϕ space.

The fluctuation of the hydrogen bonding potential energy shows a single peak for each hydrophobic strength, Figure 4.10. The fluctuation curves in the low strength regime are almost identical across different hydrophobic strengths. The centre of the peaks shifts slightly higher in the temperature spectrum as hydrophobic strength increases. In the medium strength regime the centre and width of the peaks increases while the amplitude decreases with increasing temperature. In the high strength regime the peaks continue the behaviour observed in the medium strength regime with one exception. There is a small plateau centred around $T \approx 0.5\epsilon/k_B$ which was not observed in the other regions. This plateau is essentially constant for all values in the high strength regime. These plateaus are indicative of the formation and breaking of hydrogen bonds independent of the hydrophobic effects in this model. This is a strong indicator that secondary structure formation primarily begins happening as the temperature decreases below $T \approx 0.5\epsilon/k_B$. The

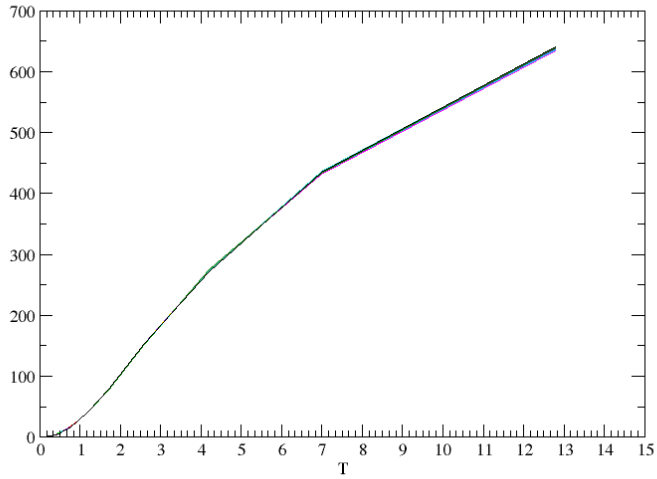


Figure 4.9: The fluctuation in the bond angle potential for each hydrophobic strength regime is shown. All of the curves are nearly identical.

temperatures corresponding to the centre of each peak are summarized in Table 4.6.

The fluctuation of the hydrophobic potential energy shows a single peak for each hydrophobic strength, Figure 4.10. In all hydrophobic strength regimes the centre, amplitude, and width of the peaks increases with increasing hydrophobic strength. At low hydrophobic strength the amplitude of the fluctuation is very small since the hydrophobic well depth is tiny. In this regime the hydrophobic energy has little effect on the folding path. The few hydrophobic interactions that contribute to the energy are typically between C_α atoms within the same helix. The amplitude of the peak is related to the the size of the energy decrease as the random coil collapses and so as hydrophobic strength increases so does the peak amplitude. The peaks shift to higher temperature as hydrophobic strength increases since the energy well is deeper and the tendency to collapse is stronger. From this it is deduced that the driving force behind the bundle collapse in the medium and high strength regimes is the hydrophobic energy. The collapse of the bundle to create hydrophobic bonds will bring the peptide units close together and result in the creation of several hydrogen bonds. This is the shifting peak of the hydrogen bonding fluctuation. The temperatures corresponding to each peak are summarized in Table 4.6.

Conclusions on the Folding Pathway

From the energy analysis the relation of the folding process to temperature in this model is deduced. First, the random coil collapses into a compact structure mostly due to hydrophobic contributions as evidenced by the large hydrophobic

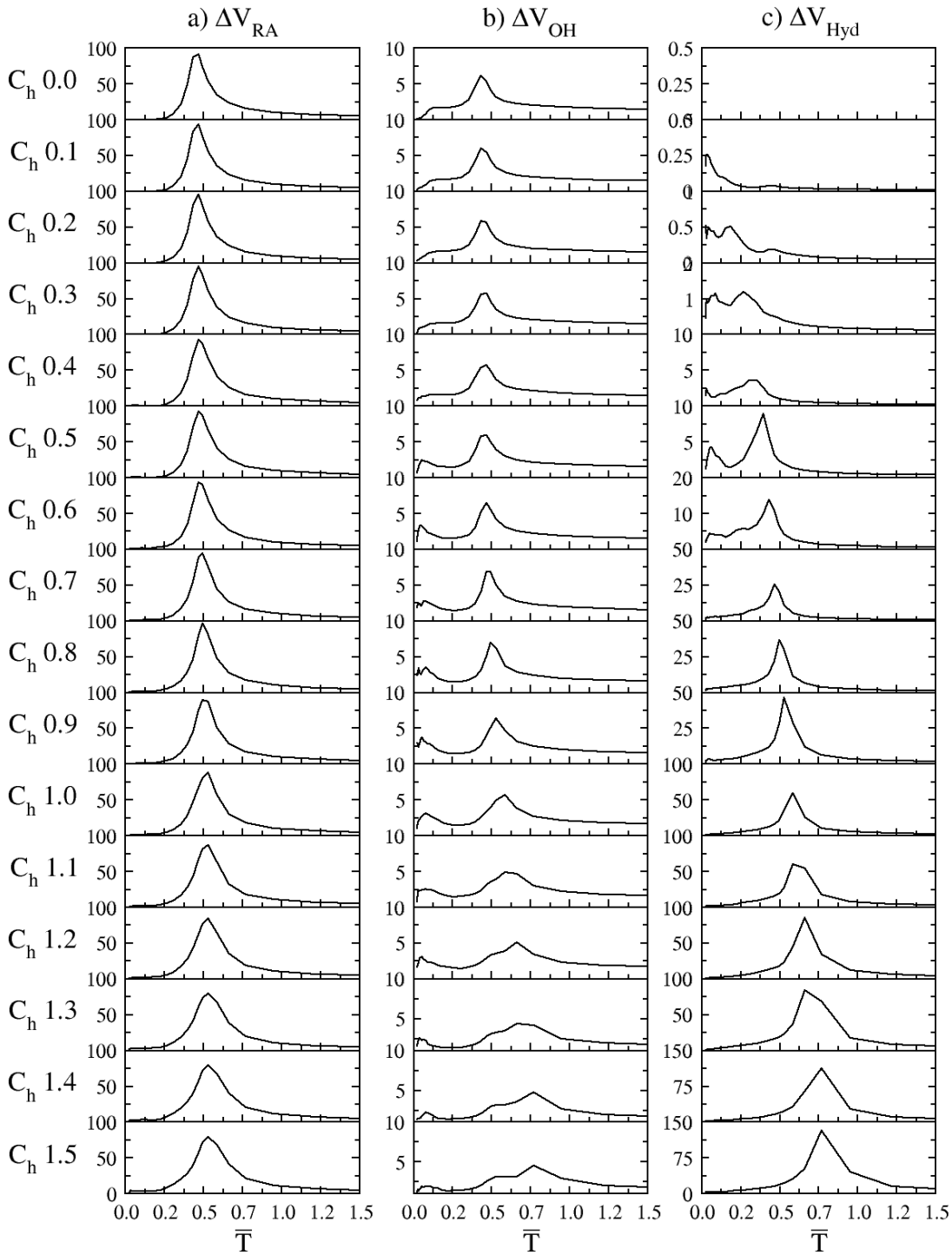


Figure 4.10: Fluctuation in the potential energy is shown for the Ramachandran angle energy (left), the hydrogen bonding energy (centre), and the hydrophobic energy (right).

Table 4.6: The temperatures at the centre of the peaks in the energy fluctuation plots are listed by hydrophobic strength. Temperatures are reported in units of ϵ/k_B.

C_h	From ΔE_{RA}	From ΔE_{OH}	From ΔE_{Hyd}
0.0	0.46	0.435	NA
0.1	0.47	0.44	0.03
0.2	0.47	0.445	0.175
0.3	0.47	0.45	0.265
0.4	0.475	0.45	0.335
0.5	0.475	0.45	0.395
0.6	0.475	0.47	0.435
0.7	0.49	0.48	0.471
0.8	0.495	0.50	0.50
0.9	0.51	0.528	0.528
1.0	0.52	0.575	0.58
1.1	0.52	0.61	0.60
1.2	0.52	0.655	0.66
1.3	0.52	0.70	0.68
1.4	0.52	0.76	0.76
1.5	0.52	0.77	0.77

fluctuation peaks. Then the compact state experiences the formation of secondary structure as indicated by the constant peak in both the Ramachandran angle and hydrogen bonding potentials around $T \approx 0.52\epsilon/k_B$. At low hydrophobic strength secondary structure forms but no compact tertiary structure is folded. The collapse is driven by hydrogen bonding and hydrophobic interactions play little to no role. In the medium strength regime the two folding stages happen simultaneously in temperature space and there is only one stage to the folding. In the high strength regime these stages are increasingly separated in the temperature spectrum as the initial collapse shifts to higher temperature. This supports the hydrophobic collapse hypothesis.

Chapter 5

Conclusions

A knowledge based minimal model for the folding of protein molecules is presented. The model improves upon previous work by Imamura [1] in several ways. The representation of the α-helix is improved by calculating the bond angle energy with the peptide unit atoms instead of the backbone C_α atoms. This allows the helix to adopt the correct number of residues per helical turn while maintaining the correct bonding pattern. The ambiguity of the first peptide unit orientation in the Imamura model was removed by freeing the peptide units to rotate about their associated C_α-C_α bonds. The Ramachandran angle potential is added to mimic steric collisions and counteract the large increase in possible structures from the freeing of the peptide units.

A four-helix bundle consisting of four α-helices and three loop regions is simulated with the parallel tempering Monte Carlo method. It is found that there are six native states of very similar conformation and energy. Four of these states are of the U-bundle type and the remaining two are of the Z-bundle type. They are all found to have scaled energies of $E \approx -218\epsilon$ and all seem equally likely to appear in simulation. There appears to be an ideal hydrophobic strength value near $C_h \approx 0.8$ where the tendency to fold to a native state is highest. As the hydrophobic strength is decreased from this value there is less and less drive for the helices to come together in a bundle and the native states are less and less likely. Conversely as the strength is increased from the ideal value the hydrophobic interactions tend to break the helices through bending thus destroying secondary structure. This value agrees with the $C_h = 0.7$ used by Imamura in his studies of α to β structural conversion.

A detailed analysis of the structures and energetics of the four-helix bundles shows that there are multiple folding stages in the temperature spectrum dependent on the strength of the hydrophobic interactions. The two stages observed as temperature decreases are 1) the hydrophobic energy causes the random coil to collapse into a compact globule resulting the the realization of several hydrophobic and hydrogen bonding interactions 2) the secondary structure is largely formed starting below a temperature of about $T \approx 0.52\epsilon/k_B$. At low hydrophobic strength

there is a small collapse driven by hydrogen bonding interactions. In the medium strength regime the two folding stages happen at the same temperature. As the hydrophobic strength increases the two stages separate and the initial collapse into a compact globule happens at higher and higher temperatures.

The characteristic collapse temperature, T_θ, is measured with several methods at each hydrophobic strength. Attempts to measure T_f from the structural overlap function proved to be difficult due mostly to the presence of six minima and the complications that arose in the parallel tempering Monte Carlo scheme. A very rough estimate of T_f is obtained at each hydrophobic strength by noting the highest temperature at which the first minimum states are typically found. T_f is found to be significantly lower than T_θ.

There are several ways this model could be improved in the future. A more accurate modelling of steric restrictions would further refine the conformational space and reduce the number of allowed non-realistic native states. This can be accomplished by implementing a unique bead size for each of the twenty possible residues and fine tuning the Ramachandran angle potential. Additionally it might be wise to separate the side chain interaction centres from the C_α backbone atoms. This can be implemented by adding virtual interaction centres to the centre of mass positions for each side chain. More than just α-helix secondary structures can be modelled by adding the β and left handed α-helix regions to the Ramachandran potential energy. The depth and widths of each well would need to be balanced to promote the correct structure formation in different cases.

References

[1] Hideo Imamura. *Minimal model for the secondary structures and conformational conversions in proteins.* PhD thesis, University of Waterloo, Waterloo, Ontario, Canada, 2005. iii, 1, 2, 4, 7, 11, 12, 13, 19, 43

[2] J.D. Honeycutt and D. Thirumalai. The nature of folded states of globular proteins. *Biopolymers*, 32:695–709, 1992. 1, 12, 13, 24

[3] Edgar Haber and Christian B. Anfinsen. Regeneration of enzyme activity by air oxidation of reduced subtilisin-modified ribonuclease. *J. Biol. Chem.*, 236(2):422–424, 1961. 1

[4] Cyrus Levinthal. Are there pathways for protein folding? *Journal de Chimie Physique*, 65(1):44–45, 1968. 2

[5] Joseph D. Bryngelson, Jose Nelson Onuchic, Nicholas D. Socci, and Peter G. Wolynes. Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins*, 21:167–195, 1995. 2

[6] H. Jane Dyson, Peter E. Wright, and Harold A. Scheraga. The role of hydrophobic interactions in initiation and propagation of protein folding. *Proc. Natl. Acad. Sci. USA*, 103(35):1305713061, 2006. 2

[7] George D. Rose and Siddhartha Roy. Hydrophobic basis of packing in globular proteins. *Proc. Natl. Acad. Sci. USA*, 77(8):4643–4647, 1980. 2

[8] David J. Hill, Matthew J. Mio, Ryan B. Prince, Thomas S. Hughes, , and Jeffrey S. Moore. A field guide to foldamers. *Chem. Rev.*, 101(12):3893–4012, 2001. 3

[9] D. Thirumalai and D.K. Klimov. Deciphering the timscales and mechanisms of protein folding using off-latice minimal models. *Curr Opin Struct Biol.*, 9(2):197–207, April 1999. 2, 12, 13, 24, 25

[10] Carl Branden and John Tooze. *Introduction to Protein Structure.* Garland Publishing Inc., New York: New York, 1991. 2, 3, 4, 5, 6, 7, 8, 9, 12

[11] Shun-Cheng Li, Natalie K. Goto, Karen A. Williams, and Charles M. Deber. alpha-helical, but not beta-sheet, propensity of proline is determined by peptide environment. *Proc. Natl. Acad. Sci. USA*, 93:6676–6681, 1996. 4

[12] Raghavendran Subramanian and Kazem Kazerounian. Improved molecular model of a peptide unit for proteins. *J. Mech. Des.*, 129(11):1130–1136, November 2007. 9, 19

[13] Thomas Creighton. *Protein Folding*. W. H. Freeman and Company, New York, 1992. 9

[14] G. N. Ramachandran and C. M. Venkatachalam. Stereochemical criteria for polypeptides and proteins. iv. standard dimensions for the cis-peptide and conformation of cis-polypeptides. *Biopolymers*, 6:1255–1262, 1968. 9

[15] Sven Hovmöller, Tuping Zhou, and Tomas Ohlson. Conformations of amino acids in proteins. *Acta Cryst.*, 58:768–776, 2001. 10

[16] Martin J. Field. *a practical introduction to the simulation of molecular systems*. Cambridge University Press, Cambridge, 1999. 11, 15

[17] Andrew R. Leach. *Molecular Modelling Principles and Applications*. Prentice Hall, London, second edition, 2001. 11, 15

[18] Valentina Tozzini. Course-grained models for proteins. *Curr Opin Struct Biol*, 15:144–150, 2005. 12, 14

[19] MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, and Karplus M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B*, 102:3586–3616, 1998. 12

[20] Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, and Kollman PA. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J Am Chem Soc*, 117:5179–5197, 1995. 12

[21] David De Sancho and Antonio Rey. Evaluation of coarse grained models for hydrogen bonds in proteins. *J Comput Chem*, 28:11871199, 2007. 12

[22] D. Thirumalai, D.K. Klimov, and T. Veitshans. Protein folding kinetics: timescales, pathways and energy landscapes in terms of sequence-dependent properties. *Folding and Design*, 2(1):1–22, December 1996. 12, 13, 24, 25

[23] Kevin A. T. Silverstein, A. D. J. Haymet, and Ken A. Dill. A simple model of water and the hydrophobic effect. *J. Am. Chem. Soc.*, 120(13):3166–3175, 1998. 12

[24] Chris Thachuk, Alena Shmygelska, and Holger H Hoos. A replica exchange monte carlo algorithm for protein folding in the hp model. *BMC Bioinformatics*, 8:342–362, 2007. 13

[25] Ken A. Dill, Sarina Bromberg, Kaizhi Yue, Klaus M. Fiebig, David P. Yee, Paul D. Thomas, and Hue Sun Chan. Principles of protein folding - a perspective from simple exact models. *Protein Science*, 4:561–602, 1995. 13

[26] Yuzo Ueda, Hiroshi Taketomi, and Nobuhiro Go. Studies on protein folding, unfolding, and three-dimensional lattice model of lysozyme fluctuations by computer simulation ii. a three-dimensional lattice model of lysozyme. *Biopolymers*, 17:1531–1548, 1978. 14

[27] Kurt Binder. *Monte Carlo Methods in Statistical Physics*, volume 7 of *Topics in Current Physics*. Springer-Verlag, New York, second edition, 1986. 15, 16

[28] Antonio Rey and Jeffrey Skolnick. Computer modeling and folding of four-helix bundles. *Proteins*, 16:8–28, 1993. 15, 28

[29] Kurt Binder and Dieter W. Heermann. *Monte Carlo Simulation in Statistical Physics: An Introduction*, volume 80 of *Spring Series in Solid State Sciences*. Springer-Verlag, New York, 1988. 16

[30] Cristian Predescu, Mihaela Predescu, and Cristian V. Ciobanu. The incomplete beta function law for parallel tempering sampling of classical canonical systems. *J. Chem. Phys.*, 120(9):4119–4128, March 2004. 17

[31] Nitin Rathore, Manan Chopra, and Juan J. de Pablo. Optimal allocation of replicas in parallel tempering simulations. *J. Chem. Phys.*, 122, October 2005. 024111 article number? 17, 18

[32] Aminata Kone and David A. Kofke. Selection of temperature intervals for parallel-tempering simulations. *J. Chem. Phys.*, 122, 2005. 206101 article number? 18

[33] Daniel Sindhikara, Yilin Meng, and Adrian E. Roitberg. Exchange frequence in replica exchange molecular dynamics. *J. Chem. Phys.*, 128, 2008. 024103 article number? 18

[34] Herbert Goldstein, Charles Poole, and John Safko. *Classical Mechanics*. Addison Wesley, Toronto, third edition, 2002. 20, 21

[35] Valentina Tozzini, Walter Rocchia, and J. Andrew McCammon. Mapping all-atom models onto one-bead course-grained models: General properties and applications to a minimal polypetide model. *J. Chem. Theory Comput.*, 2(3):667–673, 2006. 21

[36] Z. Guo and D. Thirumalai. Kinetics and thermodynamics of folding of a de novo designed four-helix bundle protein. *J. Mol. Biol.*, 263:323–343, 1996. 23, 28