FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

THE THERMODYNAMICS AND STATISTICAL MECHANICS OF PROTEIN FOLDING

A dissertation submitted in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

in

PHYSICS

by

Prem P. Chapagain

2005

To: Interim Dean Mark Szuchman
College of Arts and Sciences

This dissertation, written by Prem P. Chapagain, and entitled The Thermodynamics and Statistical Mechanics of Protein Folding, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

David C. Chatfield

Brian A. Raue

James R. Webb

Xuewen Wang

Bernard S. Gerstman, Major Professor

Date of Defense: June 2, 2005

The dissertation of Prem P. Chapagain is approved.

Interim Dean Mark Szuchman
College of Arts and Sciences

Dean Douglas Wartzok
University Graduate School

Florida International University, 2005

ACKNOWLEDGMENT

I would like to thank my advisor Prof. Bernard Gerstman for his guidance and support throughout the years of my Ph. D. program. In addition to the time spent together in front of SGI screens, hours of interesting discussions on history, politics, social issues comparing east and west, and science in general, have greatly expanded my knowledge.

I would also like to thank my committee members for their support and helpful comments. Thanks are also due to Eshel, Jose, and Shijun for their friendship and encouragement and for making the biophysics lab a very enjoyable place with their jokes and stimulating discussions. Thanks also to everybody at the department for making it such a nice place to work and be around.

I would like to thank the University Graduate School at Florida International University for the financial support through a Dissertation Year Fellowship allowing me to concentrate on my research and writing.

Finally, I want to thank my lovely wife Sushma for her cooperation and understanding during the writing of this dissertation, and my parents in Nepal for their faith, constant support, and encouragement.

ABSTRACT OF THE DISSERTATION

THE THERMODYNAMICS AND STATISTICAL MECHANICS OF PROTEIN FOLDING

by

Prem P. Chapagain

Florida International University, 2005

Miami, Florida

Professor Bernard S. Gerstman, Major Professor

The physics of self-organization and complexity is manifested on a variety of biological scales, from large ecosystems to the molecular level. Protein molecules exhibit characteristics of complex systems in terms of their structure, dynamics, and function. Proteins have the extraordinary ability to fold to a specific functional three-dimensional shape, starting from a random coil, in a biologically relevant time. How they accomplish this is one of the secrets of life. In this work, theoretical research into understanding this remarkable behavior is discussed. Thermodynamic and statistical mechanical tools are used in order to investigate the protein folding dynamics and stability. Theoretical analyses of the results from computer simulation of the dynamics of a four-helix bundle show that the excluded volume entropic effects are very important in protein dynamics and crucial for protein stability. The dramatic effects of changing the size of sidechains imply that a strategic placement of amino acid residues with a particular size may be an important consideration in protein engineering. Another investigation deals with modeling protein structural transitions as a phase transition. Using finite size scaling theory, the nature of unfolding transition of a four-helix bundle protein was investigated

and critical exponents for the transition were calculated for various hydrophobic strengths in the core. It is found that the order of the transition changes from first to higher order as the strength of the hydrophobic interaction in the core region is significantly increased. Finally, a detailed kinetic and thermodynamic analysis was carried out in a model two-helix bundle. The connection between the structural free-energy landscape and folding kinetics was quantified. I show how simple protein engineering, by changing the hydropathy of a small number of amino acids, can enhance protein folding by significantly changing the free energy landscape so that kinetic traps are removed. The results have general applicability in protein engineering as well as understanding the underlying physical mechanisms of protein folding.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

THE COMPLEX PHYSICS OF LIVING SYSTEMS

Sir Isaac Newton formulated the foundations of physics in the late 17th Century, but it was not until the latter half of the 20th Century that physicists began to understand biological systems in the manner that they had understood a variety of non-living systems over hundreds of years. Living systems have an extraordinary ability for a high level of self-organization. This maintenance of low entropy requires a level of complexity that is far above any other system. Complexity is a characteristic that physicists had little experience with. The self-organizing behavior of living systems is new physics.

The physics of self-organization and complexity is manifested on a variety of biological scales, from large ecosystems, down through individual organisms, to the molecular level. Protein molecules exhibit characteristics of complex systems in terms of their structure, dynamics, and function. They carry out a variety of functions, are flexible, and are self-organizing. They are also one of the simpler systems that display such complex behavior. Starting with Kendrew's first determination of the three dimensional structure of a protein in 1958 [1], much experimental work has been done on the structure and dynamics of protein molecules. In this dissertation, I report on theoretical research that investigates and explains the complex behavior of proteins.

A fundamental understanding of a system requires knowledge of the constituents of the system, knowledge of all the relevant forces, and an understanding of how the constituents respond to the forces. Macroscopic objects are affected by just the gravitational force and electromagnetic force, which may manifest itself in a variety of

ways such as the normal force and tension. The response of the objects to these forces is described by Newton's Laws of Classical Mechanics. Microscopic particles also feel the effects of nuclear forces, and the fundamental forces are recast as gravity, the strong force, and the electro-weak interaction. The response of microscopic particles to these forces is described by quantum mechanics. Relativity is required at high speeds or strong gravitational fields. Great progress has been made in the last 100 years in understanding the behavior of microscopic particles in a variety of systems. Electrons in conductors was an early triumph and our understanding of superconductivity has increased dramatically. Not so for living systems.

Biological physics was originally intended to apply the techniques of theoretical and experimental physics towards the understanding of living systems. However, the flow of ideas has also occurred in the other direction. Biology has opened up new fields of physics for the study of self-organization and complexity. The difficulty for physicists in making progress with biological systems arises from the inherent complexity of living systems. In order to self-organize and maintain low entropy, living systems must contain and process a tremendous amount of information. This requires a large number of different types of "particles" and a large number of distinct arrangements. This was discussed by Schrodinger in 1944 in his book *"What is life?"* [2]. The number of different types of constituents in living systems is larger than physicists are accustomed to working with. Electronic conduction in metals is determined by two types of constituents: electrons and positive ions, whereas proteins are composed of 20 different amino acids. The conduction electrons move in a periodic lattice arrangement of positive ions. The periodicity means that little information is needed to describe the position of all

the atoms of the system; you need to make measurements of the positions of only a few atoms. Living systems are aperiodic. The measurement of the position of a few amino acids in a protein will not allow you to predict the location of the positions of the other amino acids.

Physicists hope to find fundamental laws of life that will allow living systems to be understood more clearly. Quantitative descriptions of the underlying forces that control the behavior of living systems are not difficult since living systems are predominantly controlled by biochemical, inter-atomic electronic forces and the fundamental description of these forces is well known. However, the effects of these forces are difficult to predict in living systems because of the large number of different constituents and the aperiodic manner in which they are arranged. The information content required to describe living systems is not easily understood in terms of fundamental particles and fundamental forces. It is the arrangements of these particles that are complex.

The disciplines of physics that are most concerned with these matters are statistical mechanics and thermodynamics. These disciplines were created to understand multi-particle systems where all particles were identical, or only a few different species of particles were present. The number of different types of particles in living systems is an added complication. Fortunately, the increasing speed of computers allows for the study of more complex systems. Though the full complexity of even the simplest living system, or even a small protein, is still not possible to model, it is now possible to mathematically describe systems with many degrees of freedom and explore increasing levels of complexity. Along with theoretical and computational improvements,

experimental techniques are supplying more detailed information about the structure and dynamics of living systems.

Investigating the complex behavior of proteins requires extensive use of statistical mechanics and thermodynamics. The behavior of all systems is constrained by energy conservation which tells what behavior is possible. For complex systems, this is of little guidance because there are a huge number of possibilities. In this situation, the Second Law of Thermodynamics is much more valuable because it tells us what behavior is most probable. Physically and mathematically, this is explored through entropy. The Second Law states that the entropy of the Universe will be maximized, within the constraints imposed by the conservation laws. Determining the entropy of the Universe by counting its available states is a daunting task. Fortunately, the Universe can be viewed as a heat reservoir and changes of its entropy can be calculated thermodynamically without counting states. This allows us to focus on the statistical mechanics of counting the number of states of just our small system of interest. The statistical mechanical details of the entropy of our system of interest (protein molecules) and the entropy of the rest of the Universe can be combined through the concept of the free energy of our system of interest, which will be used extensively in this research.

Proteins are the fundamental building blocks of all living cells. They are the biological nano-machines found in all biological cells that carry out complex vital functions such as digesting food, fighting infections, transfer of electrons and energy, and the catalysis of crucial biochemical reactions. There are hundreds of thousands of different proteins found in living organisms. Each type has unique functions that maintain life. In order for a protein to carry out its unique functions it must take on a unique

configuration, called the native state. The large number of degrees of freedom available to a protein offers an astronomical number of possible configurations. In spite of this, proteins almost always fold to their native state. How they accomplish this is one of the secrets of life, and is known as the "Protein Folding Problem". In this work, I discuss my theoretical research into understanding this remarkable behavior of a complex system.

The dissertation is organized as follows. Chapter I contains a brief review of protein structure and dynamics. I briefly describe the concepts of free energy and energy landscape and the application of such concepts in understanding protein folding dynamics. Methods for calculating various thermodynamic quantities are described in Chapter II. In Chapter III, the computer lattice model and the simulation of the protein chain are described. My research on the effects of entropic excluded volume on protein unfolding is extensively discussed in Chapter IV. A model four-helix bundle is used and the effects of varying the size of amino acid side chains on the native state stability are investigated. In Chapter V, phase transition studies carried out in a model four-helix bundle are presented. Critical exponents and the Binder Cumulants are calculated in order to characterize the nature of the unfolding transition. Chapter VI discusses various aspects of protein engineering. Kinetic and thermodynamic analyses are carried out on a model two-helix bundle. A sequence is designed by strategic replacement of amino acid residues that changes the free-energy landscape so that the kinetic traps are removed and the folding is enhanced. Finally, in Chapter VII, the main results and findings from this research are summarized and directions for future research are discussed.

CHAPTER I

PROTEINS: STRUCTURE, DYNAMICS, AND ENERGY LANDSCAPE

1.1 Protein Structure

All types of proteins are composed from a set of only 20 different building blocks called amino acids. All amino acids consist of a central carbon atom (C_α) that has four bonds to other atoms. Three of the four bonds are the same for all amino acids: a hydrogen atom (H), an amino group ($-NH_2$), and a carboxyl group (-COOH). The only difference that distinguishes the 20 different types of amino acids is the group of atoms in the fourth position, called the side chain. All proteins are linear, unbranched chains composed of amino acid residues linked end to end by peptide bonds. A peptide bond connects the carbon atom of a carbonyl group (=CO) on one amino acid to the nitrogen atom of an amine group on the next amino acid. In forming the peptide bond, the carbonyl group loses an OH and the amine group loses an H. The OH and H join to form a water molecule. A peptide chain has a free amino group at the front end (N-terminus) and a free carboxyl group at the other end (C-terminus).

Protein chains can be tens or hundreds of amino acids long. The one-dimensional sequence of amino acids is labeled the primary structure of a protein. In the functional native state, the chain assumes a compact three-dimensional structure. The functional properties of proteins depend on their three-dimensional structures. These diverse and seemingly irregular but very specific folds have evolved through selection pressure to perform desired functions. However, there are regular features present in the protein structures. Protein structures can be classified in a hierarchical scheme in which a protein

molecule has a primary, secondary, and tertiary structure [3]. Multimeric proteins have yet another level of organization, referred to as the quaternary structure.

Segments of the peptide chain can form local regular structures, which are called secondary structures. Secondary structures provide local stability among groups of amino acids as well as promote global stability by allowing hydrophobic parts of the protein to be buried while hydrophilic parts, including the backbone chain itself, are exposed to the aqueous solvent. The two most common types of secondary structures are the α-helix and the β-strand. The α-helix consists of approximately 3.6 amino acid residues per turn with hydrogen bonds between the C=O of residue *i* and the NH of residue *i*+4 so that the structure is stabilized by hydrogen bonding and dipole interactions. α-helices vary considerably in length in globular proteins, ranging from a single turn to over a dozen turns. An α-helix can, in principle, be either right-handed or left-handed depending on the screw direction of the chain. However, individual amino acids are chiral and all protein molecules are composed of amino acids of the same chirality, L stereoisomers, so that the α-helices are almost always right-handed. Some amino acids are found more often in α-helices than other amino acids and this implies that these amino acids have a structural preference. For example, alanine, glutamic acid, leucine, and methionine are good α-helix formers, while proline, glycine, tyrosine, and serine are very poor. The helix forming propensity of these amino acids also depends on their position along the chain [4].

The other major secondary protein structure is the β-sheet. In contrast to the α-helix, this structure is built up from a combination of several regions, known as β-strands, of the polypeptide chain. β-strands are aligned adjacent to each other such that hydrogen

bonds can form between C=O groups of one β-strand and NH groups of adjacent β-strand and vice versa. The β-sheets thus formed can be in either parallel or antiparallel fashion. In parallel β-sheets, either the amino acids in the adjacent β-strands all run in same direction (amino to carboxy terminal) whereas in antiparallel β-sheets, the direction is alternated with amino to carboxy and carboxy to amino terminal.

Simple combinations of a few secondary structure elements, termed supersecondary structure or motifs, are frequently observed in protein molecules. Several motifs usually combine to form compact globular tertiary structures called domains. Domains are the simplest protein units that can carry out biological functions. For longer polypeptide chains these domains are the fundamental units of tertiary structure. The precise nature of the assembly of domains into larger proteins is crucial for the biomolecules to function. Large proteins comprise a number of relatively small domains, which are usually 100–250 residues long [5, 6]. Even though the combination of individual domains generates an enormous variety of proteins, the number of distinct domain folds is apparently limited to a few thousand [6, 7]. This has led to development of systematic structural genomics projects so that at least one example of each kind of domain folds can be found in the Protein Data Bank [8, 9].

Finally, some functioning biomolecules are complexes of multi-peptide chains and are called multimeric proteins. The peptide chains are held together by van der Waals interactions or salt bridges, as in the case of four myoglobin units comprising hemoglobin. This level of organization is usually referred to as quaternary structure. If the different peptide chains are identical, the subunits will have the same function and may act either independently of each other or cooperatively. For example, the allosteric

function of hemoglobin results from cooperative effects on oxygen-binding affinity when the subunits are arranged in their quaternary structure [10, 11]. This effect allows oxygen to be released from the molecule at venous oxygen pressure.

The first x-ray crystallographic structure of a globular protein, myoglobin, was reveled by Kendrew et al. [1] in 1958. Myoglobin (PDB entry 1mbn) is a very compact protein consisting of mainly α-helices as displayed in Fig. 1.1. This was a hallmark effort that laid the foundation for an understanding of the connection between the structure and function of a protein. Computers, synchrotron radiation, NMR, and improved detectors have changed both the quality and quantity of the protein structure determination. To date, more than 30,000 structures have been deposited in the Protein Data Bank.

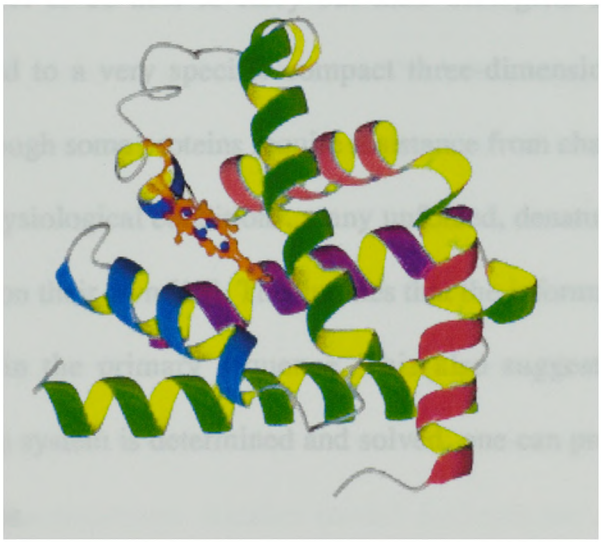


Fig. 1.1. Ribbon diagram of structure of myoglobin.

1.2 Protein Folding Dynamics

The question of how a given protein sequence efficiently and reliably folds to its specific native structure, out of an enormously large number of conformations, in a biologically relevant time, is known as the "Protein Folding Problem". Understanding protein folding has been one of the most fascinating and challenging problems in science for decades and, despite the intense research in fields as diverse as physics, chemistry, computer science, and biology, the search for the general principles that govern the folding process still goes on. However, significant progress has been achieved both theoretically and experimentally [12-14], thanks to improved experimental techniques, advances in the theoretical understanding of complex systems, and the continual increase in computational power.

All proteins start as a one-dimensional chain when they are synthesized in ribosomes. In order to be able to carry out their biological function, however, each protein has to fold to a very specific compact three-dimensional structure called the "native" state. Though some proteins require assistance from chaperone molecules, under the appropriate physiological conditions, many unfolded, denatured proteins fold back to their native states on their own [15]. This implies that the information to fold to the native state is encoded in the primary sequence. This also suggests that once the correct Hamiltonian of the system is determined and solved, one can predict the native structure from first principles.

A critical aspect of The Protein Folding Problem involves time scales. How does folding occur in biologically relevant time scales of milliseconds? If we make an underestimation that each amino acid has only eight possible states arising from two

possible configurations for each of three degrees of freedom (rotational degrees of freedom ψ, ϕ, and sidechain), then a typical protein made up of 100 amino acids would have $8^{100} \sim 10^{90}$ possible conformations. Even with the assumption of an overly fast sampling rate of 10^{-12}s per conformation, it would still take $\sim 10^{70}$ years to randomly find the native state, whereas the actual protein folding time is milliseconds. This is known as the Levinthal paradox, which suggests that there exist specific pathways for folding, rather than a random search for the native state throughout the conformation space [16]. An efficient path through the configuration space implies feedback amongst the various interactions, which act in a non-linear fashion in order to prevent misdirected wanderings in the multidimensional configuration space.

Different models have been proposed to reconcile the Levinthal paradox and explain the mechanisms that guide protein folding. In the *nucleation-growth model* [17], one or more critical kinetic structural nuclei are formed in the rate-limiting step and the rest of the structure grows around these nuclei. This model suggests that tertiary structure forms as an immediate consequence of the formation of secondary structure. In the *diffusion-collision* or *framework model* [18, 19], local secondary structure elements are first formed, independent of tertiary structure, followed by the docking of those elements to form tertiary interactions in the rate-limiting step. The model is supported by experimental findings [20] that secondary structure elements can be partially folded in the absence of tertiary interactions. Another model, *hydrophobic collapse model* [21, 22], considers the hydrophobic effect to be the driving force for folding. First, hydrophobic residues push out water in a nonspecific manner to form a more compact collapsed structure, and native protein conformation forms by rearrangement of the partially

collapsed structure. The collapsed state is the early step in the folding and is called a *molten globule*. Experiments have shown [23] that many large proteins adopt a molten globule state under mild denaturing conditions such as varying pH, supporting the hydrophobic collapse model. The framework and hydrophobic collapse models suggest the formation of kinetic intermediates that the nucleation model does not. The nucleation-condensation model [24, 25] incorporates the transition state and suggests that an extended folding nucleus is formed and consolidated through the transition state and the elements of secondary structures are formed concomitantly with tertiary structure. The nucleation-condensation model has been supported by experimental evidence from several small proteins. However, some proteins such as SH3, fold in an even higher hierarchical way where part of the structure forms early on, whereas other parts remain unstructured until the last steps of the folding reaction [26, 27]. Barnase folds with yet another mechanism, first to an intermediate, where two folding modules are independently formed according to a nucleation condensation mechanism [28, 29], and subsequently the two modules coalesce according to the framework mechanism to give the final native structure.

The nucleation–condensation might be understood as a unifying mechanism applicable in many proteins [30], where its extreme manifestation is the framework model when secondary structure becomes overstabilized or the hydrophobic collapse model when tertiary structure is overstabilized [31]. As can be seen from the number of different proposals just described, protein folding is not easily explained in terms of a single mechanism.

A more general theoretical framework of protein folding pathways has emerged more recently [32]. This describes the protein folding pathways in terms of funnel shaped free energy landscapes [33, 34] where many paths exist from the unfolded to the native state. Each of the above mechanisms may be possible. The free energy landscape has slopes with varying degrees of roughness and the protein molecule may follow the steepest path at each point. Alternatively, a protein may follow a slower path passing through several local minima and transition states in order to avoid a trap at the bottom of a steep pathway. The roughness of the slopes of the energy landscapes may present local minima that represent misfolded configurations, which act as kinetic traps and make the folding process less efficient. This phenomenon is known as *frustration*. The concept of free energy and energy landscape will be discussed in the next section.

Despite steadily increasing computational power, proteins are still too complex to simulate folding by following the behavior of all the atoms. However, simpler, reduced lattice models and Monte Carlo methods require less computer power and have been extremely useful for exploring protein folding conformational space, making it possible to investigate phenomena that are intractable experimentally. In this dissertation, a lattice computer model, discussed in a later chapter, is used to investigate various aspects of protein folding dynamics such as thermodynamics, phase transitions, and protein design.

1.3 Free Energy and Energy Landscape

Energy conservation determines the possible changes in a system. For a multi-particle system, the constraint of energy conservation will still allow a huge number of possible outcomes. The maximization of the entropy of the universe as required by the

Second Law of Thermodynamics must be used to determine which outcome is most likely. If we can enumerate the detailed nature of every energy state of the universe, we could then use the Second Law of Thermodynamics to predict the probable evolution of the universe at each instant of time. Enumerating the details of the universe is a daunting task. Fortunately, we are interested only in the time evolution of the structure of a protein molecule. We will try to enumerate the detailed structure of the states of a protein molecule and predict their probable evolution. In order to make use of the Second Law of Thermodynamics, we must take into account the entropy of the molecule as well as the entropy of the rest of the universe. The concept of free energy allows us to do both, while examining the details only of the molecule.

Free energy is one of the most important thermodynamic quantities. The term "free" comes from the fact that it is the amount of energy of a system that is available for work. Most of the physical properties of biochemical reactions depend directly or indirectly on the free energy of the system. For example, rate constants, binding constants, dissociation constants, and conformational preferences are all directly related to the difference in free energy between alternate states. Free energy is a statistical property, and like entropy, can be seen as a measure of the probability of finding a system in a given state, and depends on the extent of the phase space accessible to the system.

Consider a system, in our case a protein molecule, at constant temperature and pressure. In any spontaneous process, the combined entropy of the system and rest of the universe (reservoir) is given as

$$\Delta S_{total} = \Delta S + \Delta S' \tag{1.1}$$

where ΔS and $\Delta S'$ are the entropy changes of the protein molecule and the surrounding

heat reservoir, respectively. If a small amount of heat, H, is absorbed by the protein molecule from the surrounding, then the entropy change of the surrounding heat reservoir is given simply as $\Delta S' = -H/T$. Equation (1.1) becomes

$$\Delta S_{total} = \Delta S - \frac{H}{T} \tag{1.2}$$

Conservation of energy is equally satisfied whether the energy H resides in the molecule or in the heat reservoir, and does not favor either one.

The Second Law of Thermodynamics states that the most probable (equilibrium) distribution is the one in which ΔS_{total} is maximized. This allows us to determine how much H is likely to be transferred from the reservoir to the molecule. We rewrite Eq. (1.2) as

$$-T\Delta S_{total} = H - T\Delta S \tag{1.3}$$

The Gibb's free energy of a system such as a protein molecule is defined as

$$G \equiv -TS_{total} = H - TS \tag{1.4}$$

where H is the enthalpy and S is the entropy of the system, and T is the temperature of the heat reservoir. When the enthalpic term consists of only the internal energy E instead of H ($=E+PV$), the free energy is given as $F = E - TS$, where F is called the Helmohltz's free energy. Since $G=-TS_{Total}$, maximizing S_{Total} means minimizing G of the protein chain.

If a system whose external parameters except its volume are kept fixed is in thermal contact with a heat reservoir, the stable equilibrium situation is characterized by the condition of minimum free energy. Statistical mechanically, this can be quantified in terms of the number of ways that the Universe can arrange itself. The statistical physics definition of entropy is

$$S_{total} = k \ln \Omega_{total} \tag{1.5}$$

where k=1.38x10^{-23} J/K is Boltzmann's constant and Ω_{total} is the number of ways that the Universe can arrange itself. $\Omega_{total}=\Omega\Omega'$ where Ω is the number of accessible states of the protein and Ω' is the number of accessible states of the heat reservoir representing the rest of the Universe. The probability, $P(H)$, for the protein molecule to absorb a specific value of H from the reservoir is proportional to $\Omega_{total} = \Omega_{total}(H)$ for that transfer of H. From Eq. (1.4) and Eq. (1.5), we see that

$$\Omega_{total} = e^{S_{total}/k} = e^{-G/kT} \tag{1.6}$$

Equation (6) then gives

$$P(H) \propto e^{-G(H)/kT} \tag{1.7}$$

An expression similar to Eq. (1.7) can be written for the probability of finding our system with any specific value of any parameter y, i.e. $P(y) \propto e^{-G(y)/KT}$. This type of probability expression is extremely useful if the functional dependence of G on y is known.

Equation (1.7) shows that the most probable situation for the Universe is when G is a minimum. However, $G(H)$ may not be a monotonic function of H. The state of the system can be changed by transferring H from the reservoir to the molecule. This will lower the entropy of the reservoir, but raise the entropy of the molecule. The amount by which the entropy of the reservoir is lowered is given by $-H/T$. The amount by which the entropy of the molecule is raised when you transfer H depends on the details of the molecular structure, $\Delta S=\Delta S(H)=K\ln\Omega_f - K\ln\Omega_i$. Thus, precisely how much H should be transferred to maximize S_{total} (minimize G), depends on the detailed microscopic states of the molecule before and after H is transferred and T of the reservoir. This balancing

determines how much H will flow into the molecule at a given T.

The free energy G is defined so that the entropy of the Universe is maximized and equilibrium occurs when G is a minimum. This has the advantage of appearing analogous to the simple physical equilibrium situation of a ball rolling downhill to a potential energy minimum as a result of the gravitational force. The additional advantage of G is that it contains parameters H and S that depend only on the details of our small system of interest. We would like to understand the details of our protein molecule well enough to be able to use Eq. (1.5) to calculate the entropy of the molecule. In contrast, the heat reservoir has far too many degrees of freedom to possibly use Eq. (1.5). Fortunately, we are not interested in the detailed structure of the heat reservoir. The use of G requires that the only information needed about the reservoir is its temperature T, which is easy to measure. Thus, G allows us to determine how to maximize the entropy of the Universe by concentrating on details only of our small system of interest.

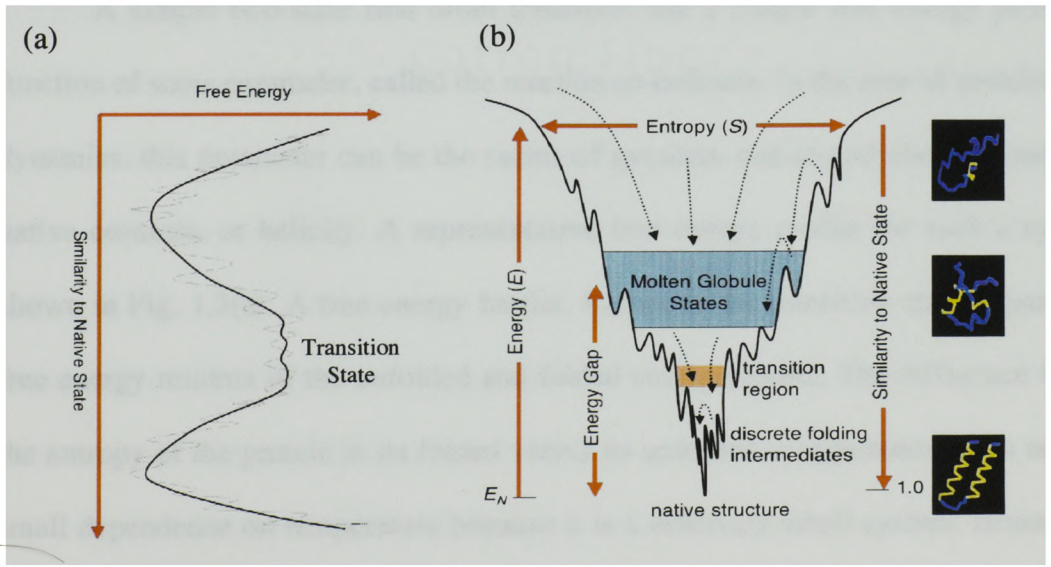


Fig. 1.2. (a) Free energy profile of a protein folding. The landscape is rugged as shown by the dashed curve. (b) Funnel shaped energy landscape.

Biological systems have a microscopic arrangement of states, $\Omega(H)$, that allow them to self-organize and maintain low entropy. Under proper biochemical conditions, protein folding that lowers S can occur spontaneously, which means that the more organized, low entropy folded state of the molecule is more probable. However, the lowering of the protein's entropy can only occur if the entropy of the surrounding environment (water) is increased by as much, or more. The entropy of the surrounding water can be increased through two processes. As the protein organizes and forms bonds, it can give off heat (enthalpy) to the water, or, as the protein folds, water molecules that were organized in clathrates around amino acids can become free to enter the bulk water, as occurs when hydrophobic amino acid sidechains come into contact with each other. Changes in free energy of the molecule compare the change in entropy of the molecule to the change in entropy of the surroundings. Protein structural transitions, such as from a random coil to a globular state, occur spontaneously if the free energy decreases.

A simple two-state first order transition has a simple free energy profile as a function of some parameter, called the reaction co-ordinate. In the case of protein folding dynamics, this parameter can be the radius of gyration, end-to-end distance, number of native contacts, or helicity. A representative free energy profile for such a system is shown in Fig. 1.2(a). A free energy barrier, known as the transition state, separates the free energy minima of the unfolded and folded configurations. The difference between the entropy of the protein in its folded versus its unfolded configuration has a relatively small dependence on temperature because it is a relatively small system. However, the change in entropy of the heat reservoir upon folding of the protein depends strongly on T because of the H involved, which can be spread over the large number of degrees of

freedom of the reservoir. Therefore, the determination of whether the unfolded or folded protein structure is the global minimum depends on temperature.

At low temperatures, the combined entropy is maximized and the free energy minimized when the protein is in its native folded state. Though the protein molecule's native state has low entropy, the combined entropy is maximized because the reservoir's entropy is high. This trade-off at low temperatures can be understood in several ways. Equations (1.1) and (1.2) show that at low T, any H given to the reservoir will cause a huge increase in the reservoir's entropy since $\Delta S' = H/T$. The folded state of the protein involves many bonds that release H to the reservoir and greatly increase its S'. Analogously, Eq. (1.4) shows that at low T, the S of the protein loses importance in determining G. The H of the protein is then the dominating factor, and smaller H results in a deeper minimum of G. The two viewpoints are equivalent. The more fundamental understanding comes from Eq. (1.2), but the free energy explanation is more useful for calculating probabilities and thermodynamic parameters.

At high temperatures, the reservoir has such large entropy that the flow of H available from the protein has a very small effect on S'. In this case, the S of the protein plays a dominant role in determining S_{Total} in Eq. (1.2), or G_{min} from Eq. (1.4), and the higher S, unfolded protein configuration is more probable and is the global free energy minimum. At an in-between temperature, T_c, the effect of H in changing S' and the difference in S between the protein's folded and unfolded configurations are equal, i.e. $\Delta S'(\text{folded,unfolded}) = H/T_c = \Delta S(\text{folded,unfolded})$. The entropy of the Universe, or equivalently, the G of the protein, is the same whether H is in the protein (unfolded) or in

the heat reservoir (folded protein); S_{Total}(folded)= S_{Total}(unfolded) → G(folded)= G(unfolded). This T_c is known as the transition temperature, or the critical temperature.

The complex structure and dynamics of the protein molecule are a result of the energy landscape of attractors and basins. Simple systems have a simple low-dimensional energy landscape having a unique ground state minimum and few excited state local minima, whereas complex systems are characterized by complicated multidimensional and rugged energy landscapes with many humps and bumps. The folding scenario of a protein molecule is often understood in terms of a funnel-like energy landscape [33-36] as shown in Fig. 1.2(b). The higher-energy unfolded state has high entropy with many higher-energy configurations. As the folding proceeds, the conformational space is narrowed due to collapse of the chain into a molten globule and/or towards the native state. The low energy conformations are more and more native-like.

CHAPTER II

CALCULATION OF STATISTICAL MECHANICAL AND THERMODYNAMIC QUANTITIES

Deep insights into the dynamics of a system, such as a protein molecule, can be obtained from various thermodynamic and statistical mechanical quantities. The dynamics of the system are determined by various interaction energies along with entropic considerations. In this chapter, I give a brief introduction to some fundamental statistical mechanical and thermodynamic quantities and explain how such quantities are calculated from a time series of data about a system. In the following chapter, the computer model and the interaction Hamiltonian that is used to simulate protein dynamics and generate realistic simulated data will be discussed.

2.1 Free Energy

The free energy profile discussed in Chapter 1.3 can be calculated as a function of one or more protein structural parameters such as the radius of gyration (R_g), fraction of helical content (q), fraction of native tertiary contacts (Q), or end-to-end distance of the peptide (d_{ee}). At equilibrium, the probability that the system has a value y for a parameter is proportional to $e^{-F(y)/kT}$ as discussed in the previous chapter, i.e. $P(y)=C(T)e^{-F(y)/kT}$. Then the Helmholtz's free energy of the system as a function of y can be expressed as

$$F(y;T) = -kT \ln P(y;T) + c(T) \qquad (2.1)$$

where $P(y;T)$ is the probability for the configuration to have the parameter y at temperature T, and $c(T)=kT\ lnC(T)$ is a normalization factor that depends on T but is

independent of y. Since the dynamics of the system depend on relative values of F, we set $c(T)\equiv 0$ when comparing free energies for different y at the same T, giving $F(y;T) = -kT\ln P(y;T)$.

Because of its complexity, it is usually important to describe protein structure in terms of two or more structural parameters. The quantities in Eq. (2.1) would then be expressed as $F(y_1,y_2;T)$ and $P(y_1,y_2;T)$. The free energy profiles are then described as contour maps or landscapes, with the height of the landscape representing F. For example, if the probability of forming a peptide configuration with specific values for the fraction of native tertiary contacts Q and end-to-end distance d_{ee} is $P(Q,d_{ee};T)$, then the Helmholtz's free energy is calculated as

$$F(Q, d_{ee};T) = -kT \ln P(Q, d_{ee};T) \tag{2.2}$$

Free energy profiles at different temperatures can provide information on the nature of the protein folding kinetics and the stability of the protein. In many places in this dissertation, the T dependence will be implicit and quantities will be written without explicit T dependence; e.g. $F(Q,d_{ee};T)\rightarrow F(Q,d_{ee})$.

2.2 Heat Capacity

The heat capacity of a system is defined as

$$C_x = \left(\frac{dQ}{dT}\right)_x \tag{2.3}$$

where x is a thermodynamic parameter such as P or V, that is kept constant during the process. This heat capacity is a valuable quantity because it can give important statistical

mechanical information about the microscopic states of the system, and is straightforward to measure experimentally using techniques such as differential scanning calorimetry.

Under conditions of constant volume, the heat dQ added to the system goes entirely to increase the internal energy of the system i.e. $dQ = d\overline{E} + \overline{p}dV$ reduces to $dQ = d\overline{E}$, so that

$$C_v = \left(\frac{d\overline{E}}{dT}\right)_v \tag{2.4}$$

Therefore, measurements of C_v at several T supply information on the microscopic dependence $E=E(T)$. The heat capacity at constant pressure is not as straightforward to interpret. At constant pressure, the addition of heat causes the volume to expand and the system performs mechanical work. Thus, for the same dQ, there is a smaller increase in internal energy $d\overline{E} = dQ - \overline{p}dV$. Since dT depends on dE, the same dQ will cause a smaller dT at constant P than at constant V. Therefore, from Eq. (2.4), the heat capacity at constant pressure (C_p) is greater than the heat capacity at constant volume. The difference between C_p and C_v depends on the fraction of energy that is used to perform work, and at a given temperature can be expressed in terms of the isothermal compressibility, κ, and volume expansion coefficient, α,

$$C_p - C_v = VT\frac{\alpha^2}{\kappa} \tag{2.5}$$

where V and T are the volume and temperature of the system respectively.

If a system is compact and undergoes very small changes in volume, the difference in Eq. (2.5) is small and the ratio ($C_p/C_v = \gamma$) approaches 1. Experimental measurements of heat capacity in protein unfolding are carried out at constant pressure.

However, the folded state of a protein and the globular cluster unfolded configuration are both compact with similar volumes. The unfolded random coil configuration is not compact, but it is hard to define a volume for this configuration since it is effectively one-dimensional. The spacing between the amino acids along the chain is the same as in the compact configurations, but the amino acids undergo more "sideways" fluctuations in the random coil. Though the approximation of constant volume is not great, the heat capacity at constant volume is used frequently in theoretical studies of protein folding because it can be readily calculated from energy fluctuations, which allows comparison with experimental results.

In this work, the fluctuation in the total energy of the protein at temperature T is used to calculate the heat capacity at constant volume. The fluctuations in the total energy is defined by the dispersion relation [37]:

$$\overline{(\Delta E)^2} \equiv \overline{E^2} - \overline{E}^2 \tag{2.6}$$

Using the definition of the partition function

$$Z = \sum_i e^{-\beta E_i} \tag{2.7}$$

where i is the sum over all the microscopic states of the system, we have

$$-\frac{\partial \ln Z}{\partial \beta} = -\frac{1}{Z}\frac{\partial Z}{\partial \beta} = -\frac{1}{Z}\frac{\partial\left(\sum_i e^{-\beta E_i}\right)}{\partial \beta} = -\frac{1}{Z}\sum_i \frac{\partial\left(e^{-\beta E_i}\right)}{\partial \beta} = \frac{1}{Z}\sum_i \left(E_i + \beta \frac{\partial E_i}{\partial \beta}\right) e^{-\beta E_i} \tag{2.8}$$

The last term in Eq. (2.8) requires especial attention. The energy levels of the system, E_i, depend directly on V but do not depend directly on β ($=1/kT$). The last term

can be expanded to $\frac{\partial E_i}{\partial \beta} = \frac{\partial E_i}{\partial V}\frac{\partial V}{\partial \beta}$. If the energy fluctuation occur at constant volume, ∂V=0, then $\partial E_i/\partial \beta$ =0. In this case, Eq. (2.7) reduces to

$$-\frac{\partial \ln Z}{\partial \beta} = \frac{1}{Z}\sum_i E_i e^{-\beta E_i} = \overline{E} \tag{2.9}$$

A similar derivation gives

$$\overline{E^2} = \frac{1}{Z}\frac{\partial^2 Z}{\partial \beta^2} = \frac{\partial}{\partial \beta}\left(\frac{1}{Z}\frac{\partial Z}{\partial \beta}\right) + \frac{1}{Z^2}\left(\frac{\partial Z}{\partial \beta}\right)^2 = -\frac{\partial \overline{E}}{\partial \beta} + \overline{E}^2 \tag{2.10}$$

Therefore,

$$\overline{(\Delta E)^2} = \overline{E^2} - \overline{E}^2 = -\frac{\partial \overline{E}}{\partial \beta} \tag{2.11}$$

where the volume V is kept constant in taking the derivative. Substituting $T=1/k\beta$, and using Eq. (2.5), this becomes

$$\overline{(\Delta E)^2} = -\left(\frac{\partial \overline{E}}{\partial T}\right)_v \frac{\partial T}{\partial \beta} = kT^2\left(\frac{\partial \overline{E}}{\partial T}\right)_v = kT^2 C_v \tag{2.12}$$

Thus, the heat capacity is given as

$$C_v = \frac{\overline{E^2} - \overline{E}^2}{kT^2} \tag{2.13}$$

For convenience, the following expression is used to calculate the heat capacity at constant volume when T is given in the same units as E (e.g. $T \rightarrow RT$, Kcal/mol)

$$C_v = \frac{\overline{E^2} - \overline{E}^2}{T^2} \tag{2.14}$$

The averages of E and E^2 can be directly calculated from the energy time series in thermal equilibrium at temperature T. The reliability of the result depends on how well the energy data represents the actual phase space of the system. To ensure ergodicity and

generate time series of data that are canonically distributed, a computer model is used that has structural moves that allows the chain to reach any configuration from any other configuration, thus spanning all of structure space. To ensure a canonical distribution, the computer model used in this work incorporates a Metropolis test within a Monte Carlo algorithm. The details of the computer will be discussed in a later chapter.

The temperature dependence of the heat capacity can offer important insight into the microscopic behavior of the system, such as the extent of cooperativity (communication) throughout the peptide chain while folding or unfolding. A first-order-like protein folding transition (all-or-nothing in terms of the global structure) is characterized by a sharp peak in the heat capacity curve as a function of temperature. The thermal averages of Eq. (2.14) at a given simulation temperature can be directly computed by averaging a time series of energies. Plotting the temperature dependence of C_v requires performing simulations at different temperatures. This can be a very time consuming process because many independent simulations are required at each temperature to get reliable thermal averages. Since a single simulation of a long chain can take hours, 50 simulations at each of 10 different temperatures might take weeks to perform. At certain temperatures, the simulations may require frustratingly longer times to reach thermal equilibrium. This happens near phase transition points and the phenomenon is known as critical slowing down. Thus, it is computationally difficult to get a smooth temperature dependence curve, especially near the transition temperature. Fortunately, the canonical distribution of the conformations generated at a single temperature contains implicit information of the system's behavior at other temperatures. The histogram technique [38, 39] takes advantage of this fact and allows us to obtain the

temperature dependence of thermal averages for a range of temperatures from the trajectory simulated at a single temperature. The technique is very useful in lattice model simulations [40]. A brief description of the technique is given below.

2.3 The Histogram Technique

If the energy of each possible configuration of a system is known, Eq. (2.14) can be used to calculate C_v at any temperature. The averages in Eq. (2.14) are calculated as

$$\overline{E} = \frac{1}{Z}\sum_i E_i e^{-\beta E_i} \quad \text{and} \quad \overline{E^2} = \frac{1}{Z}\sum_i E_i^2 e^{-\beta E_i} \tag{2.15}$$

where the summations i is over every possible configuration of the system. Equivalently, these averages can be performed over the possible energy levels of the system, E_r, if the degeneracy n_r of each energy level is known, i.e. Eqs. (2.15) can be rewritten as

$$\overline{E} = \frac{1}{Z}\sum_r n_r E_r e^{-\beta E_r} \quad \text{and} \quad \overline{E^2} = \frac{1}{Z}\sum_r n_r E_r^2 e^{-\beta E_r} \tag{2.16}$$

These averages require knowledge of the energy of every possible configuration of the system. We investigate a peptide chain that is dozens of amino acids long, moving on a three dimensional lattice. The number of possible configurations is much too large to enumerate directly. Fortunately, the averages needed for C_v can be calculated at many different temperatures from sufficient data at a single temperature using the histogram technique. The histogram technique allows us to use a time series of data at one temperature to determine all n_r.

Consider a simulation performed at simulation temperature $T=T_s$, which generates a time series of configurations which has a canonical probability distribution in which

each configuration i appears with a frequency proportional to the Boltzmann factor, e^{-E_i/kT_s}, and each energy E_r appears with a frequency proportional to the product of the Boltzmann factor times the number of configurations n_r with that energy, $n_r e^{-E_{ri}/kT_s}$. In order for the data from the simulation to reflect the equilibrium canonical distribution, the simulation model must allow any configuration to be reached from any other configuration, the probability for making structural transitions must be given by the Boltzmann factor, and the simulation must be run long enough for the system to sample all of its possible structures. As we explain in the later chapter describing the computer model we used, all of these conditions are fulfilled. If the simulation generates configurations according to the Boltzmann equilibrium probability distribution, a histogram $h(E_r;T_s)$ of the frequency of appearance of a given energy at a single temperature T_s provides an estimate for the probability of appearance of that energy when the system is in equilibrium at T_s

$$P_r(E_r;T_s) = \frac{h(E_r;T_s)}{N} = \frac{n_r(E_r)e^{-E_r/kT_s}}{Z(T_s)} \tag{2.17}$$

where $h(E_r;T_s)$ is the number of times in the time series that a configuration with $E=E_r$ appears, N is the total number of configurations sampled in the time series, $n_r(E_r)$ is the degeneracy (density of states) of configurations for E_r, and $Z(T_s) = \sum_r n_r(E_r)e^{-E_r/kT_s}$ is the partition function at T_s.

Using Eq. (2.17), from the histogram $h(E_r;T_s)$, we can calculate the density of states $n_r(E_r)$:

$$n_r(E_r) = \frac{Z(T_s)}{N} h(E_r;T_s)e^{E_r/kT_s} \tag{2.18}$$

At any other temperature, the equilibrium canonical probability for the system to be found with energy E_r is given by the same expression

$$P_r(E_r;T) = \frac{h(E_r;T)}{N} = \frac{n_r(E_r)}{Z(T)} e^{-E_r/kT} \tag{2.19}$$

In comparing Eq. (2.17) at T_s with the general expression of Eq. (2.19), valid at any T, the Boltzmann factor and the partition functions will differ at different temperatures. However, if the volume of the system remains the same, then $n_r(E_r)$ will remain the same at all temperatures. This constancy of $n_r(E_r)$ allows us to get $P_r(E_r;T)$ for any E_r at any T from $P_r(E_r;T_s)$ obtained from the histogram $h(E_r;T_s)$ at a single temperature T_s. This is shown as follows.

From Eq. (2.18), we have

$$P_r(E_r;T) = \frac{h(E_r;T)}{N} = \frac{n_r(E_r)}{Z(T)} e^{-E_r/kT} = \frac{n_r(E_r)e^{-E_r/kT}}{\sum_r n_r(E_r)e^{-E_r/kT}}$$

Substituting Eq. (2.18) for n_r, we get

$$P_r(E_r;T) = \frac{\frac{Z(T_s)}{N} h(E_r;T_s)e^{E_r/kT_s}e^{-E_r/kT}}{\sum_r \frac{Z(T_s)}{N} h(E_r;T_s)e^{E_r/kT_s}e^{-E_r/kT}}$$

$$= \frac{\frac{Z(T_s)}{N} h(E_r;T_s)e^{E_r/kT_s}e^{-E_r/kT}}{\frac{Z(T_s)}{N}\sum_r h(E_r;T_s)e^{E_r/kT_s}e^{-E_r/kT}}$$

Thus,

$$P_r(E_r;T) = \frac{h(E_r;T_s)e^{\frac{E_r}{k}\left(\frac{1}{T_s}-\frac{1}{T}\right)}}{\sum_r h(E_r;T_s)e^{\frac{E_r}{k}\left(\frac{1}{T_s}-\frac{1}{T}\right)}} \tag{2.20}$$

Therefore, for any T, $P_r(E_r;T)$ can be calculated from the histogram generated at a single simulation temperature T_s. With Eq. (2.20) for $P_r(E_r;T)$, the histogram obtained at a single temperature T_s allows the calculation of the thermal average of any quantity Q at any temperature: $\overline{Q(T)} = \sum_r Q(E_r)P(E_r,T)$. In terms of $h(E_r;T_s)$, the thermal average can be written as

$$\overline{Q(T)} = \frac{\sum_r Q(E_r)h(E_r;T_s)e^{\frac{E_r}{k}\left(\frac{1}{T_s}-\frac{1}{T}\right)}}{\sum_r h(E_r;T_s)e^{\frac{E_r}{k}\left(\frac{1}{T_s}-\frac{1}{T}\right)}} \tag{2.21}$$

Thus, knowledge of the energy of every configuration in order to calculate Z is not required for calculating $\overline{Q}$. The histogram method greatly reduces the computational time needed to determine the temperature dependence of quantities such as C_v and makes it easier to locate peaks. Also, simulations run very close to a transition temperature can take an extremely long time due to critical slowing, making it impractical to map out the transition peak with simulations at many closely spaced temperatures. The histogram technique allows careful, high resolution mapping of the transition temperature region by running simulations at a single, nearby temperature and allows an opportunity for studying critical behavior.

This technique can be used for calculating the heat capacity as a function of temperature from a long Monte Carlo run at a single temperature using Eq. (2.16) and Eq. (2.21). This technique can also be used to calculate the fourth order cumulants that are useful in finite size scaling and phase transition studies (see Chapter V), which require the temperature dependence of $\overline{E^2}$ and $\overline{E^4}$.

Ideally, it would be nice to have knowledge of the partition function $Z(T_s)$. The definition in Eq. (2.7) requires knowledge of the energy of every possible configuration, an impossible task for our model. Fortunately, Eq. (2.17) shows that Z can be calculated if the degeneracy $n_o(E_o)$ of at least one energy level E_o, such as the native state energy level, is known:

$$Z(T_s) = \frac{N n_o(E_o) e^{-E_o/kT_s}}{h(E_o;T_s)} \tag{2.22}$$

With knowledge of the partition function, a complete density of states $n_r(E_r)$ for all energy levels r can be computed from Eq. (2.17). The model used in this research has enough structural degrees of freedom (3-*d*, sidechains) that the precise degeneracy of the native state energy level cannot be easily counted and this prevents the calculation of the partition function.

A serious limitation of the histogram technique is that it is valid only over a temperature range where the phase space can be properly sampled; i.e. fully covered, and with probabilities proportional to the Boltzmann factor. Getting the correct probability ratios requires multiple visits to many points in structural phase space, which requires long simulations. The simulation at any given temperature samples only a part of phase space and thus the density of states will be incorrect for regions not sampled properly. In order to get an appropriate sampling, the temperature should be chosen so that the trajectory visits all significant regions of the phase space. In this work, a simulation temperature is chosen that is close to the transition temperature so that it properly samples both the folded and unfolded regions.

CHAPTER III

COMPUTER MODEL FOR SIMULATING PROTEIN DYNAMICS

As explained in the previous chapter, a variety of important thermodynamic and statistical mechanical parameters can be calculated from a time series of data reflecting the dynamics of a protein. This data can be from experimental measurements or computer simulations, known as numerical experiments. The advantage of computer simulations is that numerical data can be obtained with high precision for a wide range of physical quantities. This allows the calculation of a variety of parameters that provide understanding of the underlying dynamics. However, in order for these calculated values to provide insight into the physics of actual proteins, the simulated data must be obtained from a computer model that realistically simulates protein behavior. There are fundamental considerations that must be fulfilled in simulating any system. The equilibrium results of the computer model must generate canonical ensembles in which the probability of finding the system in a specific configuration is given by the Boltzmann factor. I now describe the computer model used in the present work and explain the features that assure that his condition is fulfilled in a way that realistically simulates the folding dynamics of proteins.

3.1 Lattice Model and Interaction Hamiltonian

The large number of degrees of freedom and the competing free energy terms in a protein produce a complicated energy landscape. In order to investigate the effects of the large number of different peptide motions and interactions, I use a computer lattice model

to simulate the dynamics of proteins. The model used in the present study is based upon a lattice model developed by Skolnick and Kolinski [41]. The model has been shown to be effective at representing protein secondary and tertiary structure and associated structural transitions [41-45].

The model is a compromise between detailed treatments that include most atoms in a short chain [46-50] and models used to study longer chains, but with amino acid residues represented with very little detail [22, 51-54]. The models that have full representation of atoms are computationally limited and are useful only for the shorter chain length scales. Full atom simulation for longer peptide dynamics is still not feasible with present day computing power. Even though various interatomic forces are reasonably well known, the complicated dependencies on distance and details of multi-body effects are not fully known. Small uncertainties in these details can build-up for long chains and long simulation times. On the other hand, the minimalist models represent residues as just a single lattice site with no internal structure. They afford the advantage of allowing the investigation of long chains and for long times. The disadvantage is that without any internal structure or degrees of freedom for amino acids, important physics may be missed, especially entropic considerations involving sidechains. The model that I use allows chains long enough to form the tertiary structure of a four-helix bundle but also includes some structure and degrees of freedom for the side chains.

3.1.1 Lattice Modeling of Protein Chain Structure

The underlying lattice is a simple cubic lattice. An amino acid residue occupies an active lattice site for its backbone (C-C_α–N) and another active lattice site for the sidechain that is diagonally across a cube (±1, ±1, ±1) from its backbone. The orientation of the sidechain with respect to its backbone can vary, but is always constrained so that it gives the C_α left handed chirality, as is true with real amino acids. Backbones that are adjacent in the protein's primary sequence and connected by peptide bonds that are constrained to be connected on the lattice by a vector that is a permutation of (±2, ±1, 0). This constraint in the adjacent backbone-to-backbone distance of $\sqrt{5}$ represents a characteristic α-carbon-to-α-carbon distance 3.785 Å [55]. This connectivity has been found to allow actual secondary structures like α-helices and β-strands and tertiary structures such as a four-helix bundle and an α-β bundle to be effectively represented on the lattice [41]. By clever choices of interactions, these two degrees of freedom (backbone and sidechain constraints) may be sufficient to simulate protein dynamics that are realistic enough for statistical dynamical investigations.

In addition to a central point as the active lattice site, the backbone of a residue "occupies" the six nearest neighbor lattice sites, for a total of seven lattice points. Interaction energies between backbones are determined by the distance between the central, active site of each backbone. The other six occupied lattice points for each backbone are non-active, "virtual" backbone sites that act only to give thickness to the backbone and define a minimal excluded volume for the backbone. Two different amino acids cannot both occupy the same lattice point with either active or virtual points. The side chain of a residue occupies a sidechain active site and an additional three virtual

lattice points, filling out a quadrant of the lattice with respect to the backbone. The 11 points occupied by a residue are shown in Fig. 3.1(a), and the backbone active site and the sidechain active site are highlighted. Interactions between sidechains are determined by the distance between the active points of each sidechain. The other three sidechain points serve to exclude volume. If desired in the model, a residue can be chosen to be labeled as ''inert,'' in which case it has no sidechain and occupies only the seven lattice points of the backbone. This topic will be discussed in more detail later in Chapter IV.

The shape (configuration) of the chain is determined by the way it bends (conforms) at each residue as shown in Fig. 3.1(b). The backbone conformation of the i^{th} residue depends on both the distance between the $i-1^{th}$ and $i+1^{th}$ active sites, and their orientation with respect to i. In real proteins, the electronic orbitals of the C and N atoms result in bonds that do not permit the angle made by three consecutive residues to be either too sharp or completely straight. In order to simulate these constraints, we do not permit configurations that place residues $i-1$ and $i+1$ with a separation distance r closer than $r=\sqrt{6}$ or further apart than $r=\sqrt{18}$.

The cubic lattice discretizes the allowed values of r. The (210) constraint along with the constraint of $\sqrt{6} \le r \le \sqrt{18}$ restricts the possible squares of distance between $i-1$ and $i+1$ to seven values: r^2 = 6, 8, 10, 12, 14, 16, and 18. This gives a one-to-one mapping between squares of distances and $i-1$, i, $i+1$ bond angles as shown in Fig. 3.2. We represent the various bond angles as R-states of the residues. For example, if residues $i-1$ and $i+1$ are separated by a distance of $r=\sqrt{6}$, residue i is considered to be in R-state 6, or more specifically, $R6$. As shown in Fig 3.2, for some values of r, there may be multiple ways (degeneracy) of arranging $i-1$ and $i+1$. Though there are only seven possible r

values, the degeneracy of some of them along with the (210) connection constraint and cubic lattice allow a total of 18 possible configurations connecting residue i to i–1 and i+1. These 18 vectors are: R6a, R6b, R6c, R6d, R8a, R8b, R10a, R10b, R12a, R12b, R14a, R14b, R14c, R14d, R16a, R18a, R18b, R18c. With this scheme, a right-handed α-helix can be nicely formed with all the helical residues in R12b states, whereas a left-handed helix, not naturally occurring, would be composed of residues in R12a states. The discreteness of the lattice requires each turn of a helix to consist of 4 (integer) residues as opposed to 3.6 per turn in real α-helices.

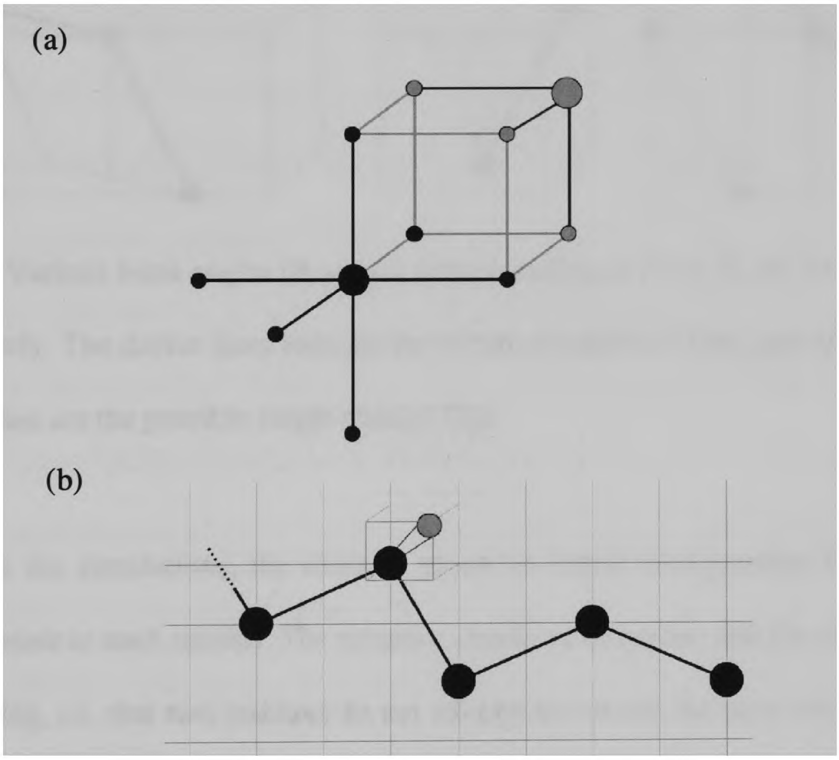


Fig. 3.1. (a) The lattice sites occupied by a residue. The dark circles represent the backbone occupied sites and the light circles represent the sidechain occupied sites. Circles for the active sites (interaction centers) are drawn bigger. (b) A peptide chain is laid out by combining backbone residues constrained with the (210) vector connectivity.

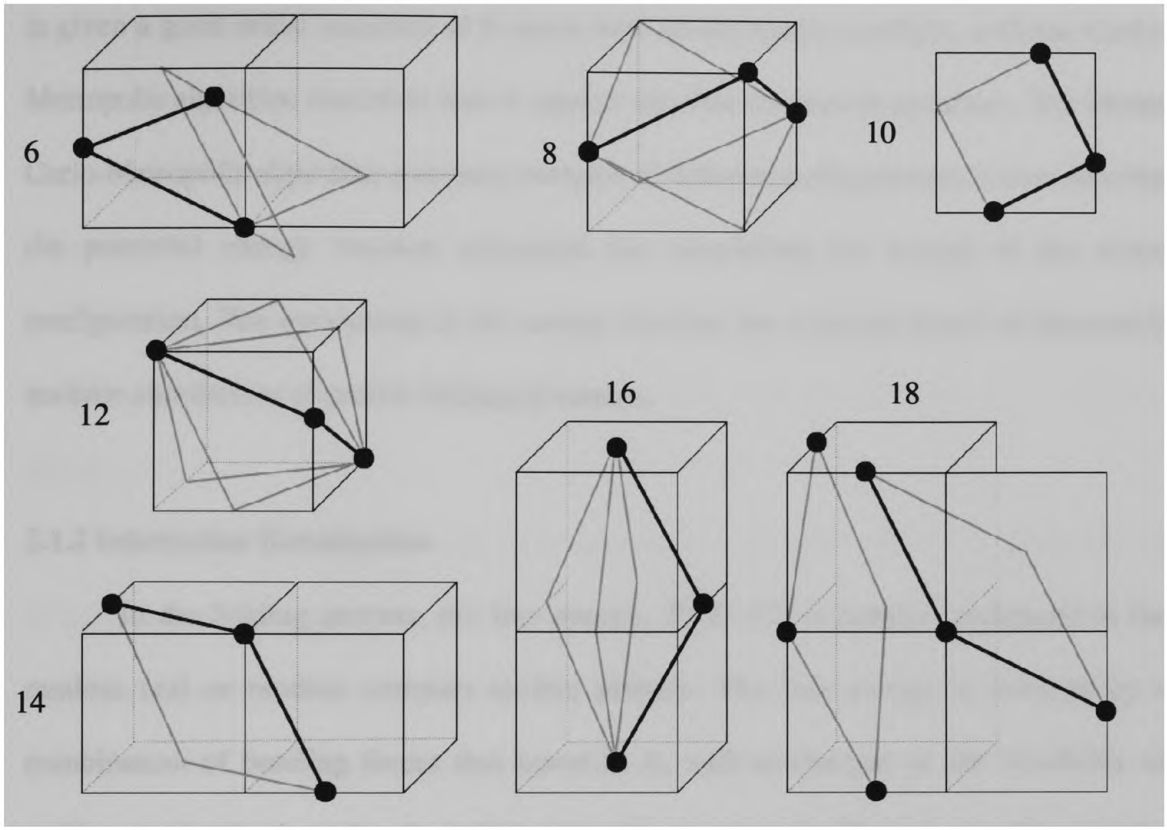


Fig. 3.2. Various bond angles (R-states) corresponding to r^2= 6, 8, 10, 12, 14, 16, and 18 respectively. The darker lines indicate the initial orientation of the pair of bonds, and the lighter lines are the possible single residue flips.

In the simulations, the chain is given an initial configuration by assigning an initial *R*-state to each residue. The program checks to be certain that the chain is non self-intersecting, i.e. that two residues do not attempt to occupy the same lattice site. If such an occupancy conflict occurs, the program rejects the initial layout and will not run, notifies the user to the two or more residues that are in conflict, and requires that a new initial layout be used as input. (This rejection of occupancy-conflicts can be disabled in the program in order to generate interesting, but non-physical shapes). When the program

is given a good initial sequence of R-states with no occupancy conflicts, a Monte Carlo-Metropolis algorithm described later is used to simulate the protein dynamics. The Monte Carlo-Metropolis algorithm compares energies of different configurations. I now describe the potential energy function employed for calculating the energy of the chain configuration. The interactions in the energy function are a crucial aspect of adequately realistic simulations of protein folding dynamics.

3.1.2 Interaction Hamiltonian

In the folding process, the free energy, $F=E-TS$, is reduced compared to the random coil or random compact molten globule. The free energy is lowered by a combination of bonding forces that lower E, as well as changes in the flexibility of residues and solvent molecules which can raise or lower S. These interactions can be specific in that they prefer that approaching residues orient themselves in a specific fashion, or non-specific in which two residues need merely to be close in distance to interact. These interactions guide the folding process and act in a non-linear fashion to produce a complicated potential energy landscape. Typically, folding is initiated by short-range interactions forming independent secondary structures such as α-helices. While these secondary structures are forming, they may also approach each other as the protein collapses inward into a condensed configuration. Large numbers of non-covalent, weak tertiary bonds (e.g. hydrophobic interactions) between helices act together to add up to a large energy contribution that promotes protein folding. In order to produce a stable folded protein conformation, the contribution of various interactions must compensate for the loss of entropy of the chain as it folds. A careful interplay must occur so that the

specific interactions can form organized structure while the non-specific interactions stabilize the collapsed state. If the collapse occurs too quickly, the specific interactions will not have sufficient time to act and will result in a collapsed state with no organized structure. The specific and non-specific interactions involved in the protein folding process that have been incorporated in our model are discussed below. The strengths of these interactions can be varied so that the simulations can be run to investigate various effects.

Hydrogen bonding is a specific interaction and represents a combination of covalent and electrostatic interactions. The strength of a hydrogen bond comes mainly from the electrostatic attraction between hydrogen donor and acceptor and is orientation dependent. The secondary structures are locally stabilized by hydrogen bonds. In α-helices, a hydrogen bond forms between NH of the backbone of residue i and the C=O of the backbone of residue i+4. In order to form the maximum number of hydrogen bonds along the length of the helix, the helix axis remains relatively straight. In both parallel and antiparallel β-sheets, all possible main chain hydrogen bonds are formed and stabilize the structure. In addition to hydrogen bonds, dipole moment interaction is also important for stabilizing secondary structure. The different polarity of NH and CO groups gives rise to a dipole moment (μ) on each residue. The dipole interaction energy has a dependence that is between r^{-4} and r^{-6}, depending on alignment and local inductive effects. In an α-helix, these dipole moments orient along the helical axis and the length of the helix produces a stronger dipole moment and stabilizes the helix. In our model, for simplicity, the backbone hydrogen bonding and dipole interaction are combined into one term, HBDIP, and it takes effect only when two backbones have a lattice separation

smaller than or equal to 4, and are not connected by a peptide bond (not sequential in the primary sequence). When the interaction takes effect, it constitutes an energy contribution that is weak and attractive.

Another interaction that we incorporate in our model is the steric repulsion that is part of the van der Waals interaction. The van der Waals interaction consists of a repulsive term that falls off approximately as r^{-12} and an attractive term that falls off approximately as r^{-6}. We do not explicitly model the attraction, due to induced dipoles, because it is weaker than other attractions that we do model. More important is the repulsive part of the van der Waals interaction. It can become especially strong if two residues approach too closely and the electron orbitals of the adjacent atoms begin to overlap and repel, owing to the quantum mechanical Pauli's exclusion principle. Since the interior of proteins is packed densely, a large number of close contacts are made in the folded form of the proteins. Although attractive van der Waals interactions are the weakest of the non-covalent forces, the dense packing of the atoms in proteins causes the repulsive part to sum up to a significant effect. Steric repulsion between the atoms provides an excluded volume to the protein and is extremely important in the folding and stability of the protein. To simulate the van der Waals repulsion, a soft-core repulsion and a hard-core repulsion are included for the backbone-backbone interaction. This gives thickness to backbones in addition to the infinitely hard-core virtual occupation of nearest neighbor lattice points. The soft-core repulsion, SCREP, occurs if backbone active sites come to with in a distance of $\sqrt{5}$ or less. If this distance becomes $\sqrt{3}$, the smallest possible, then the hard-core repulsion takes effect and is 3 times as strong than SCREP. My work on this topic will be discussed in detail later in Chapter IV.

Another very important non-specific interaction in protein folding dynamics and native state stability is the hydrophobic interaction. It is believed to be the main driving force in the initial collapse of the random coil chain into a more compact molten globule. It is a non-specific interaction in that it only depends on the distance between two non-polar sidechains and no special orientation is required. The strength of a hydrophobic interaction is not due to an intrinsic attraction between non-polar groups such as electrostatic or dipole. Instead, it lowers the free energy of the system by increasing the entropy of water molecules. A clathrate of highly ordered water molecules is formed around a non-polar molecule exposed in an aqueous solution. When two hydrophobic side chains come close, the solvent-exposed surface of each sidechain is reduced and some water molecules from each of the ordered clathrates are released to the bulk solvent. This is displayed in Fig. 3.3. The increase in entropy of the water is thermodynamically favorable because the free energy, $F=E-TS$, of the chain-water system is decreased. Thus, burial of hydrophobic sidechains, such as leucine and valine, in the folding process is free energy favorable. In contrast, sidechains of hydrophilic residues, such as arginine, do not cause formation of clathrates and will not lower the free energy of the system if they are buried. The polarity of hydrophilic sidechains allows them to make electrostatic bonds with water molecules and these sidechains prefer an aqueous environment and usually stick out from the surface into water. With larger sidechains, the hydrophilic interaction is slightly specific as opposed to the hydrophobic interaction. Though the hydrophilic interaction is explicitly included in our model, its small specificity is not included. In our model, a sidechain can be assigned a property to be either hydrophobic or hydrophilic, or neutral. During the simulation run, an interaction

between sidechains is triggered if the active sites of two side chains approach within a distance of $\sqrt{3}$. The strength of the interaction is given relative to the free energy the side chain would have if it were sticking out into water. If the two side chains are both hydrophobic, they would much prefer to be near each other compared to being surrounded by water and this interaction is thought to be the driving force behind that large-scale collapse of an extended chain. Therefore, the hydrophobic-hydrophobic interaction is strongly attractive and lowers the free energy of the system. If both side chains are hydrophilic, the interaction is weakly repulsive (compared to each being surrounded by water). Finally, the hydrophobic-hydrophilic interaction between sidechain is more strongly repulsive. For simplicity, we do not explicitly model the charged sidechain interactions, such as in a salt bridge, in out model.

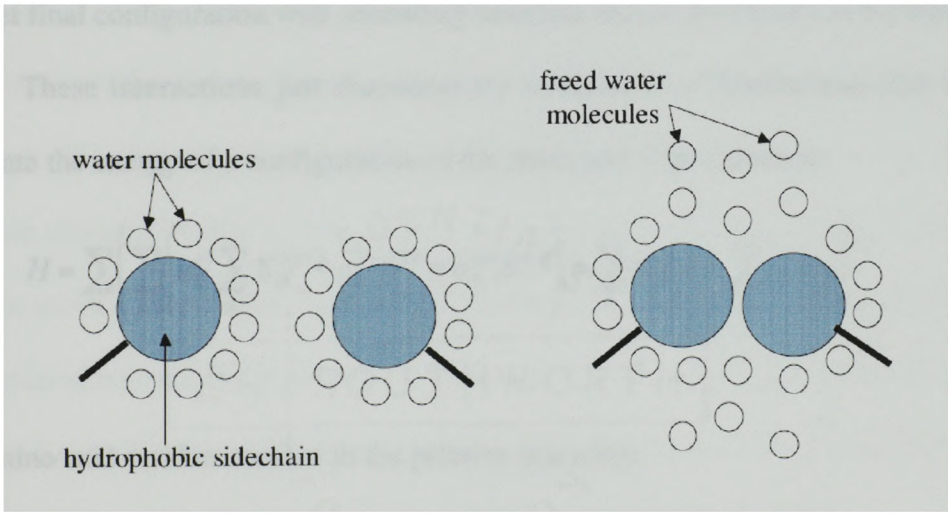


Fig. 3.3. Schematic representation of clathrate formation around hydrophobic (non-polar) sidechains in aqueous solution. When two hydrophobic sidechains approach each other, some water molecules forming the cage are freed. This increase in the entropy lowers the free energy.

Analysis of known protein structures in protein data banks, reveal that some amino acids have a propensity to appear in α-helices, and others in β-strands [3]. This propensity for an individual amino acid to form a type of secondary structure is called a local propensity, and its underlying cause in terms of electronic orbitals is not well understood. In our model, each amino acid can be assigned a local propensity for forming an α-helix, or β-strand or neither. Besides the local propensity, a medium range, or cooperative propensity is included to represent the possibility that two or more adjacent amino acids that are in an α-helix configuration may act to encourage further growth of the helix by stabilizing its structure. If adjacent residues having the same local preference are both in the preferred state, then the medium range propensity is activated and lowers the free energy of the chain. The local and medium propensities can be assigned to create a target final configuration with secondary structure that is pre-chosen in the input file.

These interactions just discussed are included in a Hamiltonian that is used to calculate the energy of a configuration of the chain and is given below:

$$H = \sum_i \left(\sum_{j>i} \left(a_{ij}^{sc} \sum_p E_{ij}^{scp} + a_{ij}^{bb} E^{bb} + a_{ij}^{rep} E^{rep} \right) + \sum_l a_i^L E_L + \sum_m a_i^m E_m \right) \qquad (3.1)$$

where

i,j : amino acid residue number in the primary sequence.

a_{ij}^{sc} : tests if sidechains i and j are close enough to interact; no=0, yes=1.

E_{ij}^{scp} : sidechain-sidechain interaction energy (p=1, 2, or 3 for hydrophobic-hydrophobic, hydrophilic-hydrophilic, or hydrophobic-hydrophilic, respectively).

a_{ij}^{bb} : tests if backbones i and j are close enough to interact; no=0, yes=1.

E^{bb} : backbone-backbone interaction energy (hydrogen bond, dipole).

a_{ij}^{rep} : tests if backbones i and j are too close so that steric repulsion takes effect; no=0, yes=1.

E^{rep} : Steric repulsion energy due to soft-core and/or hard-core repulsion.

a_i^l : tests if residues i–1, i, and i+1 are arranged so that i is in the preferred local configuration.

E_l : local propensity energy.

a_i^m : tests if residues i–1, i, i+1, and i+2 are arranged so that i and i+1 are in the same preferred local configurations.

E_m : medium range (cooperative) propensity energy.

All information for the initial configuration of the peptide chain as well as the strengths of various interactions appearing in the Hamiltonian and other parameters are fed as an input file to the computer program PROFOLD. Unlike real amino acids with 20 different side chains, in this model all side chains occupy the same volume. However, they are distinguishable in that the user chooses for each side chain whether it will interact as if it were either hydrophilic, hydrophobic, or neither. As explained earlier, each individual residue can be assigned a propensity for forming α-helices, β-strands, or neither. These three possibilities for each of these two independent characteristics gives nine possible amino acids, rather than the 20 found in nature.

After initialization of the protein's conformation, the Monte Carlo-Metropolis simulation is started and different moves are attempted according to an instruction set discussed below.

3.2 Monte Carlo Metropolis Method For Simulating Dynamics

The most fundamental understanding of the dynamics of any system occurs when every force acting on every component of the system is known. Newton's Second Law can then be used to predict the future behavior of the system. This is usually possible only for systems with few degrees of freedom and few forces contributing to the Hamiltonian. Multi-particle systems may have too many degrees of freedom and too many forces to be able to make deterministic predictions of the behavior of the various components. Even if all possible forces are known, the forces may act at random times, such as thermal fluctuations in the interactions with a solvent that acts as a heat reservoir. In this case, it is impossible to know precisely what forces act or what energy is transferred at each instant of time, and a completely deterministic prediction of the behavior is not possible. In these situations, a statistical mechanical and thermodynamic description of the behavior is the most insightful approach. This can be obtained by use of Monte Carlo techniques.

3.2.1 Monte Carlo Method

The Monte Carlo (MC) method is a technique that employs a sequence of random numbers to simulate the dynamics of a system and provide ensemble averages by performing statistical sampling. The name "Monte Carlo" was coined by Metropolis after the city in Monaco known for its gambling casinos. Although the use of random number sequences for statistical sampling dates back hundreds of years, it was not until the second half of the 20th Century that the Monte Carlo method gained widespread use in physics. This occurred as a result of the implementation of an additional constraint, called

the Metropolis test, which assures that the physical system will be distributed over its accessible states with a probability given by the canonical Boltzmann distribution. This method is particularly useful in multi-particle systems with a large number of degrees of freedom where exact computation of quantities of interest is impossible. The Monte Carlo method is applicable in simulating not only physical systems that are stochastic, or random, in nature, but also in highly deterministic systems with no apparent stochastic content, but with differential equations that are too complicated to solve analytically. There are a myriad of systems that have been investigated using the Monte Carlo Method.

As mentioned earlier, the deepest understanding of a system occurs when all the important interactions that control the dynamics of a system are known. It is usually difficult to measure these forces directly. Instead, measurements are made of the energies of the various states of the system and these energies give insight into the underlying forces. In complex systems with many degrees of freedom and many important interactions, there may be a huge number of conformational states of the system. These conformations may be too large in number and too closely spaced in energy for them to be measured individually. In this case, the best information available about the system may be thermal averages. If thermal averages can be obtained under a variety of different conditions, it may be possible to obtain information about the distribution of conformations and their energy, and ultimately the physical interactions and degrees of freedom that create the conformational space.

The statistical mechanical relationships between the individual states and the thermal averages are given as follows. The thermal average of a physical quantity <Q> is given as

$$<Q>=\frac{\sum_i Q_i e^{-E_i/kT}}{Z} \tag{3.1}$$

where E is the energy of a state of the system and $Z = \sum_i e^{-E_i/kT}$ is the partition function.

For complex systems, enumerating all the states can be a daunting task. The complete enumeration of the conformation space of proteins is limited to smaller chains as the computing time increases exponentially with the system size in general. For larger system size, this problem can be overcome by employing an importance-sampling MC method. In such a method, only a tractable set of M conformations is sampled as representative of the full conformation space so that an estimate of <Q> in Eq. (3.2) can be written as:

$$\langle Q\rangle_{est} = \frac{\sum_i^M Q_i e^{-E_i/kT}}{\sum_i^M e^{-E_i/kT}} \tag{3.3}$$

In order to get accurate averages, the sampling must be done properly so that small number of conformations accurately represents large conformational space. In particular, the sampling is biased towards the conformations that are significantly populated at equilibrium. If the conformations are chosen with a probability $P_i \propto e^{-E_i/kT}$, the estimate for the thermal average becomes:

$$\langle Q\rangle_{est} = \frac{\sum_i^M Q_i}{M} \tag{3.4}$$

The conformations with such a property can be sampled using a Metropolis algorithm [54, 55] which ensures the Boltzmann distribution. The Metropolis algorithm will be discussed later in this Chapter.

In this research, the dynamics of the protein chain is simulated using a Monte Carlo method consisting of a set of moves that realistically reflect the actual protein dynamics. The different types of moves employed in the model are discussed in the following section. With such a move set, random numbers are used to select a new configuration of the chain, and a Metropolis test, described later, is performed to decide if the chain is updated to the new configuration.

3.2.2 The Move Set

A Monte Carlo step consists of N attempts of moving a single residue (singleton move), where N is the number of residues in the chain, and two attempts at multi-residue moves. Each singleton starts with a random number that picks an amino acid i and attempts to move it to a new lattice site without moving the backbones of i–1 or i+1. To maintain the correct chirality of the chain, the sidechains of i–1 and i+1 may be reoriented as part of the move. Since i–1 and i+1 remain in the same location, the R-value of i remains the same. Based upon the lattice restrictions, there are a limited number of new possible positions. The precise number and location of new positions depend on the present R-state and are listed in a table. The computer program checks the R-state of the chosen residue, and uses a file to determine how many positions of i are possible. Another random number determines which of the new positions the active sites of the backbone and sidechain of residue i will attempt to move to. If any of the 11 new

positions are already occupied by any of the 11 points associated with another residue, there is an occupancy conflict and the move is rejected. Another move is then attempted. Occupancy conflicts are rare when the peptide chain is in a stretched out random coil configuration. As the chain collapses, residues are closer together and fewer moves are available that avoid occupancy conflicts. If an attempted move satisfies the occupancy test, a Metropolis test, which is described later in this Chapter, is performed. Once the Metropolis test is passed, the chain is updated to a new configuration due to the change in position of residue i and a possible change in orientation of the sidechains of i–1 and i+1. Fig. 3.4(a) shows a representative single residue move (corresponding to R-state of $R10$). All the possible singleton flips are displayed in Fig. 3.2.

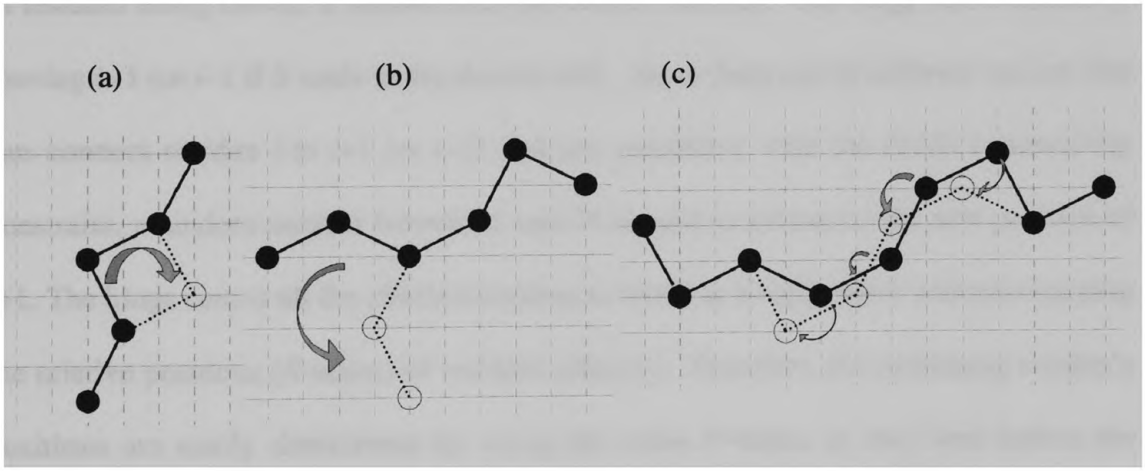


Fig. 3.4. Three different types of moves employed in the computer model (a) single residue move (singleton flip) corresponding to R-state $R10$. (b) hinge move with i as the anchor residue. (c) wave-like move.

In real proteins there is the possibility that several residues move simultaneously as a result of a concerted push from the solvent or a local concentration of enough internal energy in the chain. To emulate such large-scale fluctuations, we include two types of multi-residue attempts in each MC step: hinge and wave-like. For hinge moves, one end of the peptide chain moves as a rigid hinge as shown in Fig. 3.4(b). An anchor residue i is first selected randomly. Since moves are less likely to occur as the number of residues to be moved increases, the probability for attempting a hinge move is weighted against moving a large number of residues. The residues that will be moved are on the shorter end from i so that the number involved is always $< N/2$. This is done by selecting another random number from 2 to $N/2$ and comparing it with the number of residues effectively being moved in the hinge attempt. The attempt is accepted only if the number of residues being moved is smaller than the random number. The hinge move occurs by moving $i+1$ (or $i-1$ if it leads to the shorter end). Since there are 24 different vectors that can connect residue i to $i+1$ (or $i-1$) and are consistent with the (210) connectivity constraint, a random number between 1 and 24 is used to determine the new position of $i+1$. The hinge causes all the affected residues to move as a rigid object without changing the relative positions (R-states) of residues after $i+1$. Therefore, the remaining residue's positions are easily determined by using the same R-states as they had before the attempted move. The attempt is successful only if it passes the occupancy test for all residues that are moved, and the Metropolis test, and then the chain configuration is updated.

In addition to the hinge move, another multi-residue wave-like move is incorporated in the simulation. Unlike the hinge move, the wave move usually involves a

segment of residues in the middle of the chain. A wave involving the entire chain is possible, but weighted to be highly unlikely. The wave move allows the propagation along the chain of multi-residue structural elements. Two residues, i and j, are chosen to act as the ends of the segment of residues that will move. As in the hinge move, a random number weighs the probability to attempt to move a segment of the chain. If the length of the segment is smaller than a random number chosen between 2 and N–1, the move is attempted. In a wave-like move the R-states are permuted in a cyclic fashion inside the segment as shown in Fig. 3.4(c). The (210) vector connecting residues i–1 and i is used to join i and i+1, the old vector between i and i+1 becomes the new vector connecting i+1 and i+2, and so on until old vector j–1 and j becomes new vector j and j+1. To complete the cyclic permutation, the old vector connecting i–1 and i is replaced by the old vector that connected j and j+1. This wave propagation temporarily places the residues in new positions and if the move passes both the occupancy and Metropolis tests, the chain conformation is updated.

The equivalent real time for a MC step can be estimated as follows. Each singleton move involves approximately 10 atoms. Atomic vibrations in solids have periods of approximately 10^{-12} seconds. Therefore, a singleton move requires approximately between 10^{-11} and 10^{-10} seconds of real time, and a multi-residue move takes about 10 times as long. Therefore, in a chain of approximately 100 residues, a single Monte Carlo step is equivalent to approximately 10^{-9} seconds [56].

3.2.3 Metropolis Test

The central ingredient in the MC method is the idea of equilibrium, which is expressed as the condition of detailed balance. The condition of detailed balance states that for a system in equilibrium, the rates at which the system makes a transition into and out of any state must be equal. If the probability of these transitions between states is same, the idea of equilibrium is satisfied. If we write p_μ as the probability to be in a state μ, and $P(\mu \rightarrow \nu)$ for the probability of a transition from μ to ν, the restrictive form of detailed balance can be written as

$$p_\mu P(\mu \rightarrow \nu) = p_\nu P(\nu \rightarrow \mu) \tag{3.5}$$

In thermal equilibrium, p_μ is the Boltzmann probability $e^{-E_\mu / kT} / Z$ and we can write

$$\frac{P(\mu \rightarrow \nu)}{P(\nu \rightarrow \mu)} = \frac{p_\nu}{p_\mu} = e^{-(E_\nu - E_\mu)/kT}, \tag{3.6}$$

where $(E_\mu - E_\nu)$ is the energy difference between the new and the old states. To assure thermal equilibrium, the move from one conformation to another is possible only if the inverse move is also possible. With appropriate choice of a move set, Metropolis algorithm is enough to ensure that any conformation can be arrived at from any other state for the lattice and thus ergodicity is satisfied. In our model, Metropolis algorithm in simulating the protein folding dynamics selects conformations with a rule of move set described earlier and such a choice of moves have been shown to assure ergodicity [41]. As long as the condition of detailed balance is satisfied, the order of moves carried out does not have an effect on the equilibrium canonical distribution reached after sufficiently long time.

A Metropolis algorithm that meets the above-mentioned criteria is used in simulating the protein folding dynamics in this study and is described as follows. When a move is attempted, if the new lattice sites are vacant, the program calculates the new energy of the system using the interaction Hamiltonian described above. If the new energy is lower than the current energy, the move is accepted and the configuration of the chain is updated. If the new energy is higher, the move is not automatically rejected but undergoes a Metropolis test to determine if the move will be accepted. There is a non-zero probability that random interactions with the solvent will allow the protein to move to a state of higher energy. The relative probability of finding the system in the higher energy state ν compared to the lower energy state μ is given by the Boltzmann factor

$$\frac{p_\nu}{p_\mu} = \frac{e^{-E_\nu / kT} / Z}{e^{-E_\mu / kT} / Z} = e^{-(E_\nu - E_\mu)/kT} = e^{-\Delta E / kT} \tag{3.7}$$

where ΔE is the energy difference. The relative probability of Eq. (3.7), which has the same exponential form as a Boltzmann factor, is compared to a random number. The move will be accepted and the conformation is updated only if the relative Boltzmann factor $e^{-\Delta E / kT}$ is greater than the random number. Otherwise the move is rejected and the current conformation is kept until a move attempt is successful. In this method, attempted moves with larger increases in energy, ΔE, have exponentially decreasing probability for success.

The criteria of acceptance of an attempted move can be summarized as

$$P_{accept} = \begin{cases} e^{-\Delta E / kT} & \text{if } \Delta E \text{ is positive} \\ 1 & \text{otherwise} \end{cases} \tag{3.8}$$

The combination of the Monte Carlo method for choosing moves and the Metropolis test for determining if they are accepted have been shown to assure that the simulations satisfy the important statistical mechanical and thermodynamic behavior of the real system.

To summarize, the computer model simulates the dynamics of a protein by following these steps:

1. The user supplies strengths for the various interaction terms used to calculate the free energy of a chain configuration.
2. The initial conformation of a chain is supplied by the user with a sequence of R-states for each amino acid along the primary sequence.
3. The computer program checks to see if this initial sequence is non self-intersecting. If it is self intersecting, the simulation does not start, and a new initial sequence of R-states must be supplied. If it is not self-intersecting, the program continues.
4. The computer chooses random numbers in a Monte Carlo approach to attempt moves that change the position of single, or groups, of amino acids.
5. Random numbers are used in a Metropolis test to decide if a move will be accepted or rejected.
6. When a move is accepted, the program examines the new configuration and calculates a variety of quantities of interest such as energy of the chain (E), radius of gyration (R_g), end-to-end distance (d_{ee}), fraction of helical content (q), fraction of native tertiary contacts (Q) etc. Along with the position of the active site of every backbone and sidechain, these quantities are written to output files that can

be used in statistical mechanical and thermodynamic analysis of the behavior of the program. The output file containing information of the locations of the residues can be fed to a graphics program, PROVIEW, that displays the chain configuration with a ball and stick model. This allows the configuration to be inspected frame by frame for easy identification of structural elements and changes of conformation during the simulation, and the time evolution of the quantities of interest.

7. The program continues running for a user-defined number of Monte Carlo steps.

In the following chapters, I discuss the statistical mechanical and thermodynamic analysis of the numerical data generated by the computer model and explain the insight into protein folding dynamics obtained from this analysis.

CHAPTER IV

EXCLUDED VOLUME ENTROPIC EFFECTS ON PROTEIN UNFOLDING AND STABILITY

Various intrachain interactions, as well as the interaction of the amino acid residues with the solvent, guide the folding to the native state structure. In addition to various other interactions such as hydrophobic and hydrogen bonding, another very fundamental non-specific interaction called steric repulsion contributes to the free energy of the peptide chain as discussed earlier. This entropic excluded volume effect is a very important factor in protein dynamics, as it is always present regardless of any other attractive or repulsive interaction. Recently, there has been a growing interest in the effect of macromolecular crowding and confinement on biochemical processes such as protein folding [59-63]. Such studies have shown that the volume exclusion due to the presence of macromolecules in the media has a significant effect on protein denaturation. The folding rates of many proteins may be different in the intracellular environment than in an ideal solvent [61, 64]. The effect of molecular shape and the size of the solute as well as the solvent have also been of considerable interest in polymer dynamics [65]. In this dissertation, I focus on volume exclusion effects on the protein stability and unfolding times due to the protein chain itself rather than macromolecules present in the medium. I will show that the excluded volume due to presence of sidechains has significant effects on stability and the unfolding dynamics of the native state of a model four-helix bundle protein.

4.1 Excluded Volume in Lattice Model

The only difference among the 20 different kinds of amino acids is in their sidechains. The 20 sidechains differ in terms of enthalpic interaction properties such as hydrophobicity and size, which causes different entropic excluded volume effects. I am interested here in how different amino acid volumes give rise to different folding and unfolding dynamics, regardless of similarities in other properties such as hydrophobicity or hydrophilicity. The strategic placement of an amino acid residue of a particular size at a particular place in the primary sequence might be crucial for the chain to correctly fold and for the stability of its native structure. To determine the importance of excluded volume, the effects of changing the size of sidechains on the unfolding dynamics of a model four-helix bundle protein have been investigated. In addition, the effects of changing the thickness of the chain's backbone are investigated. These investigations have applicability for understanding the effects of side chain size differences, and also have relevance to the behavior of synthetic polymers where the size of the constituent units can be varied.

As discussed in Chapter III, in our model amino acid residues move on a cubic lattice. Each amino acid residue occupies lattice points representing the backbone and additional lattice points representing the sidechain. As shown in Fig. 4.1, the backbone of a residue occupies seven lattice points; a central point and the six nearest neighbors. The six occupied lattice points around the active lattice site for each backbone give thickness to the backbone and define a minimal excluded volume for the backbone.

The model has the capability to increase the excluded volume in addition to these seven lattice points. The sidechain of a residue occupies four additional lattice points- an

active site and three lattice points for an excluded volume of the sidechain. In the model, a residue can be chosen as either hydrophobic or hydrophilic or inert. A residue with hydrophobic or hydrophilic sidechain occupies a total of eleven lattice points. An inert residue, however, has no sidechain and occupies only the seven lattice points of the backbone. The lattice points occupied by a residue with both a backbone and a sidechain are displayed in Fig. 4.1. The interaction points, or active sites, for the backbone and the sidechain are displayed as darker circles. To reflect the excluded volume of the backbone and side chains, in the lattice model backbones are not allowed any closer than within $\sqrt{3}$ lattice spacings of other backbones or any closer than within $\sqrt{2}$ lattice spacings of other residues' sidechains.

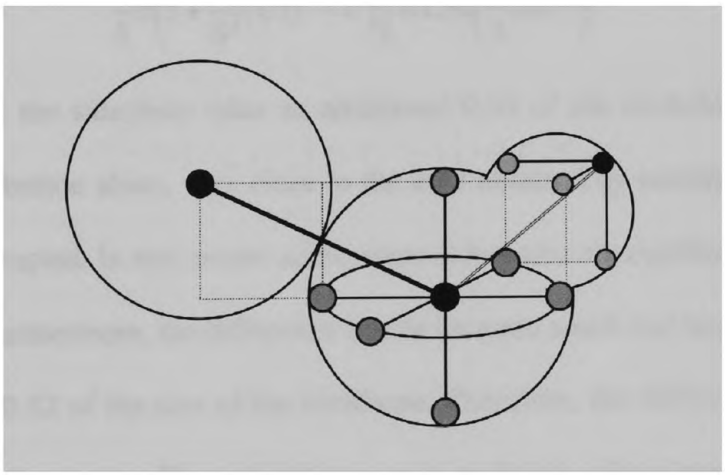


Fig. 4.1. Lattice points and excluded volume occupied by amino acid residues and their sidechains. The large sphere on the left shows the excluded volume occupied by a residue with no sidechain.

In terms of excluded volume effects, the sidechain adds 4/7 (0.57) of the lattice sites of the backbone. This ratio can be further solidified if we use hemispherical volume sections to fill out the excluded volume that is occupied by lattice points, as displayed in Fig. 4.1. The seven lattice sites of the backbone alone can be thought of as excluding a sphere with a radius equal to one lattice unit and a volume of $\frac{4}{3}\pi(1)^3$. The distance from the center of the backbone to the active site of the sidechain is $\sqrt{3}$. The sidechain can be thought of as excluding a volume of $\frac{1}{8}\cdot\frac{4}{3}\pi\left(\sqrt{3}\right)^3$. Therefore, the sidechain excludes a volume that is $\frac{1}{8}(\sqrt{3})^3 = 0.65$ of that of a backbone alone. However, the seven lattice points occupied by just the backbone on the cubic lattice results in the backbone alone excluding a volume of $\frac{1}{8}\cdot\frac{4}{3}\pi(1)^3$ in the same region even if the sidechain is not present. Therefore, an amino acid residue with backbone and sidechain excludes a volume of

$$\frac{4}{3}\pi\left(1+\frac{1}{8}\left((\sqrt{3})^3-1^3\right)\right)=1.52\left(\frac{4}{3}\pi(1)^3\right).$$

This means that the sidechain adds an additional 0.52 of the excluded volume of the backbone contribution alone, very close to the 0.57 obtained by merely counting lattice sites that are occupied. In real amino acids, some sidechains are significantly bigger than the backbone. Furthermore, the difference in size between small and large sidechains can be greater than 0.52 of the size of the backbone. Therefore, the difference of 52% may actually underestimate the effects of differences in excluded volume due to differences in side chain sizes.

There is an energy term, E(SCREP), as part of the E^{rep} in the interaction Hamiltonian of Eq. (3.1), representing the possibility that the backbone may have a

thickness greater than that determined by the excluded volume of the seven backbone lattice points. These seven lattice points are equivalent to an overlap energy of ∞, and overlap will always be prevented by failure to pass the Metropolis test. With this absolute excluded volume, the active sites of two backbones can get as close as a distance of $\sqrt{3}$, but no closer. If an energy penalty is imposed when two backbones try to approach very closely, the effective excluded volume can be expanded. This energy penalty is equivalent to a soft-core repulsion, E(SCREP). In simulations with a larger value of E(SCREP), the soft-core repulsion becomes more important and the excluded volume of each residue is increased. The simulation is designed so that if two backbones try to approach to a separation distance of either $\sqrt{5}$ or 2, the soft-core repulsion is activated and the configuration energy is increased by E(SCREP). If two backbones try to approach to the minimum possible separation of $\sqrt{3}$, a much stronger physical repulsion is simulated by increasing the configuration energy by 3 times E(SCREP). The overall effect of the steric repulsion due to soft and hard-core repulsions is displayed schematically in Fig. 4.2. The inclusion of soft and hard-core repulsion reduces the overall accessible space (water), which ultimately reduces the total number of possible configurations in the phase space. In the present research work, the size of the sidechains, the hydropathy of the sidechains, the thickness of the backbone, or a combination of all three are changed systematically, which makes it possible to distinguish entropic excluded volume effects from enthalpic effects.

(a) (b)

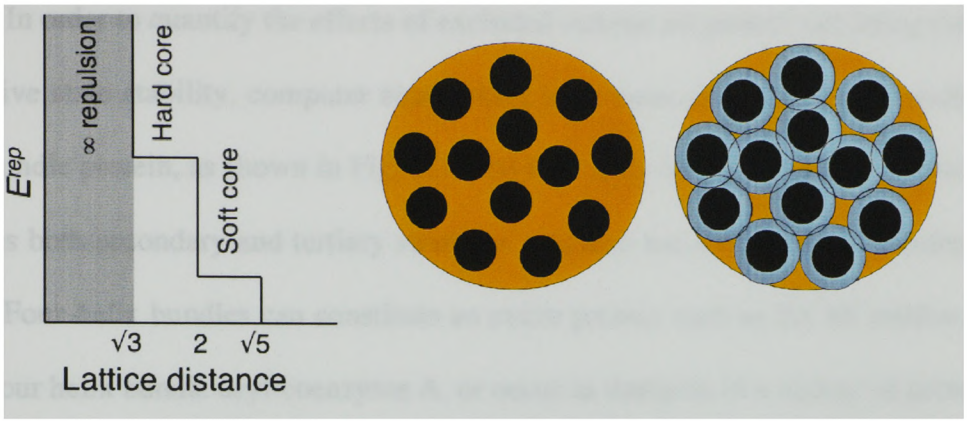


Fig. 4.2. (a) Schematic representation of the soft and hard-core repulsion in the lattice model. (b) With the inclusion of soft-core repulsion, the effective volume available for the particles (yellow area) is reduced. The dark circles represent the space occupied by particles and the envelopes represent the area where the soft and hard-core repulsions take effect.

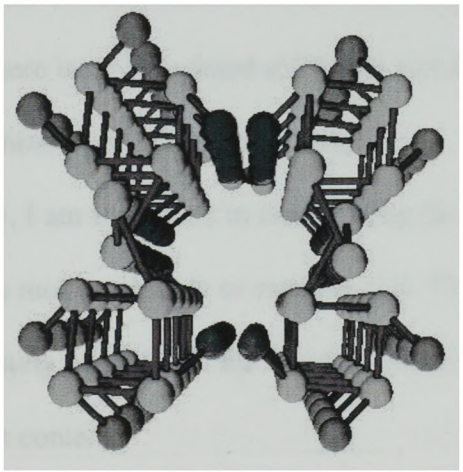


Fig. 4.3. Model four-helix bundle. Backbones are shown as lightly colored spheres, hydrophobic sidechains are dark spheres located between helices, and hydrophilic sidechains are gray spheres extending outwards from helices.

4.2 Analysis of Native State Stability

In order to quantify the effects of excluded volume on protein unfolding times and the native state stability, computer simulations have been performed on a model four-helix bundle protein, as shown in Fig. 4.3. The four-helix bundle is used as a model as it contains both secondary and tertiary structure, and also has a well-defined hydrophobic core. Four-helix bundles can constitute an entire protein such as the 86 residue, single chain four helix bundle acyl-coenzyme A, or occur as domains in a variety of proteins [3] such as myohemerythrin, cytochrome c', cytochrome b562, ferritin and the coat protein of tobacco mosaic virus. The initial configuration is set as a four-helix bundle of 83 residues with 18 residues in each helix, and additional residues in the loops connecting the helices. The chain is constructed with amino acid residues that are either hydrophobic (*H*), hydrophilic (*P*) or inert (*I*). For each helix, we choose a repeating sequence of either *–HHII–* or *–HHPP–* for the helical elements, and *–IIII–* for the loops connecting the helices. For an *I* residue, there is no associated sidechain and therefore only the backbone interactions in the Hamiltonian of Eq. (3.1) are possible.

In the present study, I am interested in determining the relative stability of a four-helix bundle compared to a molten globule or random coil. The four-helix bundle protein is considered to be completely unfolded if the fraction of the helical residues is reduced to 15% of the original helix content.

To get a good statistical sample of the dynamics, for each set of conditions, 50 simulations of length 1×10^6 MC steps each are run. The 50 different simulations have identical initial conditions but differ in their random number seed and this gives an ensemble of simulations. From these simulations, a wealth of statistical mechanical and

thermodynamic information is available for analysis. As explained in Chapter II, a free-energy function can be calculated as $F(q)= -kT \ln P(q)$, where $P(q)$ is the probability of finding the chain with the particular value of helical content q. The free-energy function can quantify the relative stabilities of the native and unfolded states. The increased stability of any state is characterized by increased depth in its corresponding free energy minimum.

The relative stabilities of the native state and unfolded state can be investigated further by calculating the heat capacity, C_v, as a function of temperature T, using the fluctuations in the energy. The thermodynamic averages in the above expression were calculated using a Monte Carlo histogram technique as described earlier in Chapter II. The histogram technique can be used to calculate the ensemble average of thermodynamic quantities, such as energy, at various temperatures from a run performed at a single simulation temperature. Since the most accurate histogram for sampling the energy space is obtained when the simulation is run at the transition temperature, simulations were run at different temperatures in search of the transition temperature. At each simulation temperature, the heat capacity curve was obtained as a function of temperature using Eq. (2.14). Simulations were run at different temperatures until a simulation temperature was found at which the temperature of the peak position in the heat capacity curve matched closely to the simulation temperature that was used to generate the curve. This temperature is used as the simulation temperature T_s.

The histogram $h(E)$ in Eq. (2.21) is the number of times a particular energy state appears in the canonical ensemble of 50 simulations. The histogram obtained by combining multiple runs has approximately 50 times as much data and therefore gives a

more reliable picture of the phase space than use of any single run. There are however, advantages to using 50 separate runs and then combining them, rather than running a single simulation that is as long as all 50 combined. A single long run may get trapped into a specific structure and never leave. This structure may in actuality, be improbable in the overall dynamics, yet it will appear to be the dominant structure because it occupies a large fraction of the MC steps. To avoid this mis-representation of any single structure, we carry out 50 different simulations. In addition, multiple runs allow the error in the histogram to be estimated by regrouping the 50 simulation runs. For example, several groups with 40 simulation runs can be obtained by excluding 10 runs each time. Then the average and standard deviation for various quantities can be directly calculated. The error estimates obtained this way in the present study were insignificant as compared to the average histogram.

4.3 Results of Varying the Excluded Volume

Different types of interactions help stabilize the four-helix bundle. Backbone interactions help stabilize individual helices, and interactions between hydrophobic sidechains on different helices help stabilize the bundle. The effects of excluded volume are investigated by changing the occupancy of sidechains and modifying other interactions.

4.3.1 Enhanced stability due to sidechain excluded volume

Two different sizes for the inert sidechains are used. In one case, the inert sidechains of the helices are the same size as the hydrophobic sidechains. As described in

Sec. 4.2 and displayed in Fig. 4.1, these sidechains occupy a volume of four lattice sites. In the other case, the inert residues do not have sidechains and only the backbone is represented so that the residues are effectively glycine-like. For the loop segments connecting helices, the inert residues always had no sidechains. Fig. 4.4 displays the crucial difference in the volume excluded by the native states of four-helix bundles with these two schemes of sidechain sizes of the inert residues. Later, the size of the large sidechains are maintained, but changed them from inert to hydrophilic. As will be shown, it is found that the enthalpic effect of the exposed hydrophilic sidechains is much less important than the entropic excluded volume effect due to the change in size.

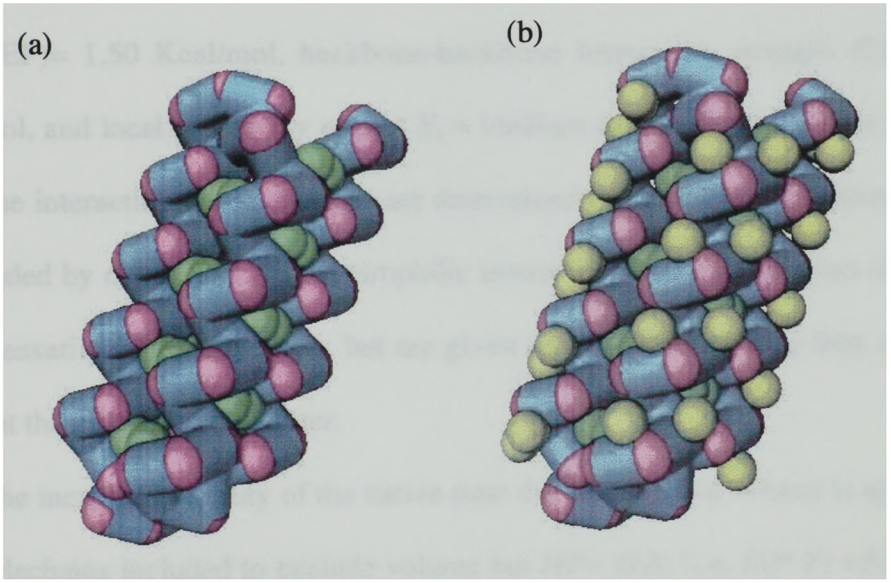


Fig. 4.4. (a) Four-helix bundle with no sidechains for the inert residues. The backbones are shown in purple-blue shading and the hydrophobic sidechains in the buried core are green. (b) The exposed inert (or hydrophilic) residues have sidechains shown in yellow. The buried hydrophobic sidechains are still present in the core. (Sizes of sidechains and backbones are not drawn to scale).

A dramatic increase in the stability of the four-helix bundle is observed when the small, glycine-like inert residues are replaced by residues with the larger inert side chains that exclude more volume. In Fig. 4.5, I display the heat capacity as a function of temperature for three different sidechain conditions. The solid curve, with a heat capacity peak at 313 K, is for glycine-like residues with no sidechains. The dotted line, with a peak at 328 K, has full size sidechains that are inert. The dashed line, with a peak at 331 K, also has full size sidechains, but the sidechains have been made hydrophilic, with a hydrophilic-hydrophilic interaction strength $E(P\text{-}P)$= 0.10 Kcal/mol, and hydrophilic-hydrophobic interaction strength $E(P\text{-}H)$ = 0.50 Kcal/mol. For all three conditions, the hydrophobic-hydrophobic interaction $E(H\text{-}H)$ strength is set at –1.10 Kcal/mol, E(SCREP)= 1.50 Kcal/mol, backbone-backbone interaction strength $E(B\text{-}B)$= –0.25 Kcal/mol, and local propensity energy E_L = medium range energy E_M = –0.50 Kcal/mol. All these interactions' free energies are determined relative to the situation of a residue surrounded by only water. The hydrophilic interaction strengths between sidechains are not necessarily inherently weak, but are given small values because they are not much different than the effects of water.

The increased stability of the native state due to excluded volume is apparent. With inert sidechains included to exclude volume but HP= (0,0) [i.e. $E(P\text{-}P)$ =$E(P\text{-}H)$= 0, the dotted curve], the transition temperature is increased by 15 K compared to no sidechain-excluded volume (solid curve). In order to compare the entropic effects of sidechain excluded volume to enthalpic effects, the dotted curve and the dashed curve are compared. Both simulations have full size sidechains that exclude volume. With the inclusion of hydrophilic sidechain interactions, the HP= (0.1,0.5) curve [$E(P\text{-}P)$= 0.1

Kcal/mol, $E(P\text{-}H)$= 0.5 Kcal/mol], the transition temperature increases by only 3 K, a much smaller effect than the excluded volume.

The greater relative importance of sidechain excluded volume compared to sidechain hydrophilicty is also demonstrated when we examine the probability for the four-helix bundle to unfold at T=312 K. With no sidechains, 74% of the 50 runs unfolded. With inert sidechains, this dropped dramatically and only 8% of the runs unfolded. When we made the sidechains hydrophilic, 10% of the runs unfolded, which is very similar to the large but inert sidechains.

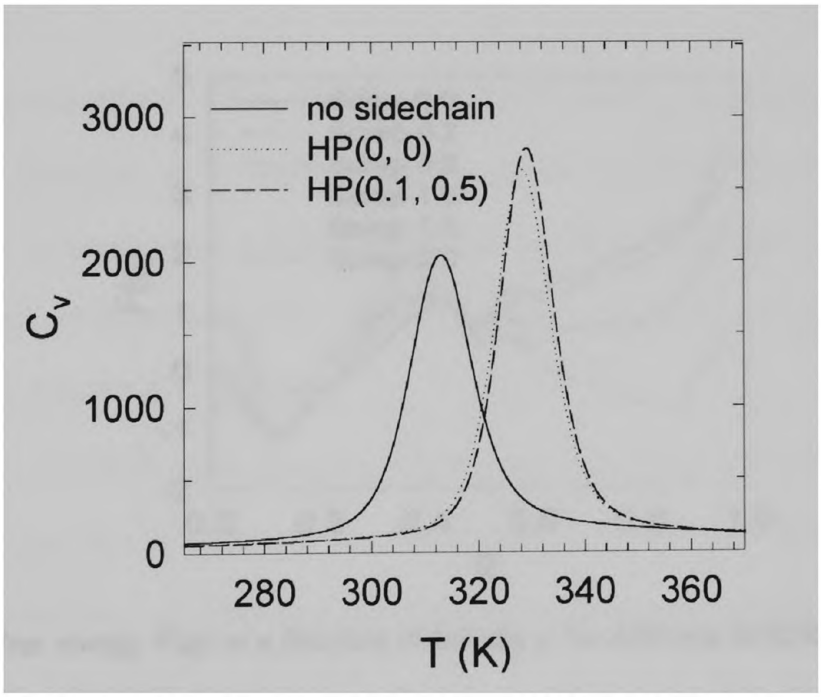


Fig. 4.5. Heat capacity curves for different types of sidechains. The presence or absence of the sidechains affects the transition temperature much greater than the hydrophilicty of the exposed sidechains.

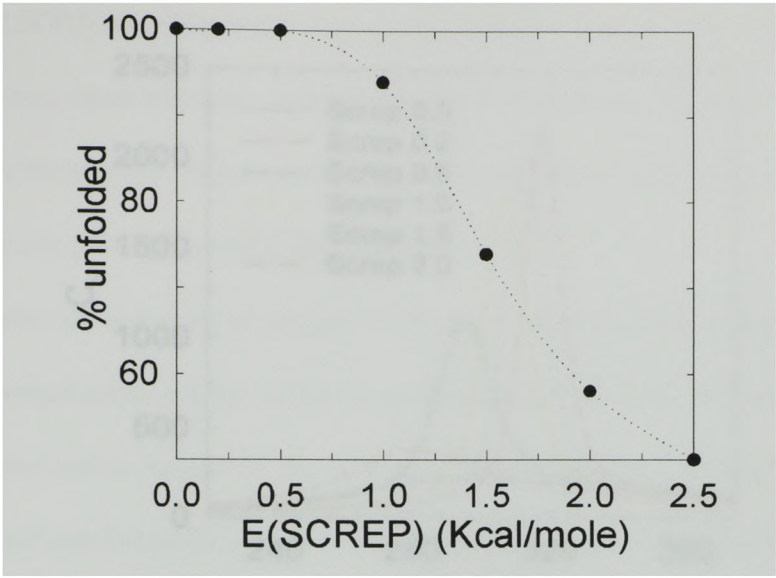


Fig. 4.6. Decreasing probability for a four helix bundle to unfold as the excluded volume is increased by increasing *E*(SCREP).

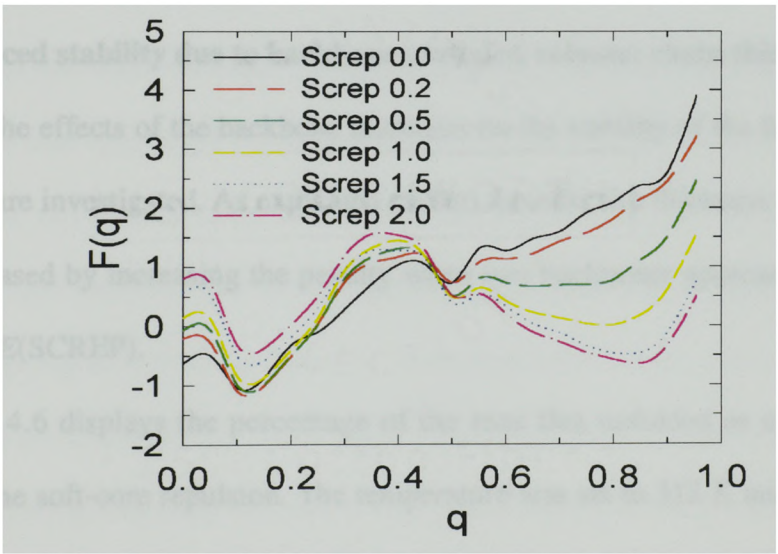


Fig. 4.7. Free energy $F(q)$ as a function of helicity q for different *E*(SCREP). All curves are plotted at the same temperature. Lower values of the soft-core repulsion (less backbone excluded volume) favor the unfolded state (low q) and higher values of *E*(SCREP) (greater backbone excluded volume) give rise to free energy landscapes with absolute minim near the native state configuration (high q).

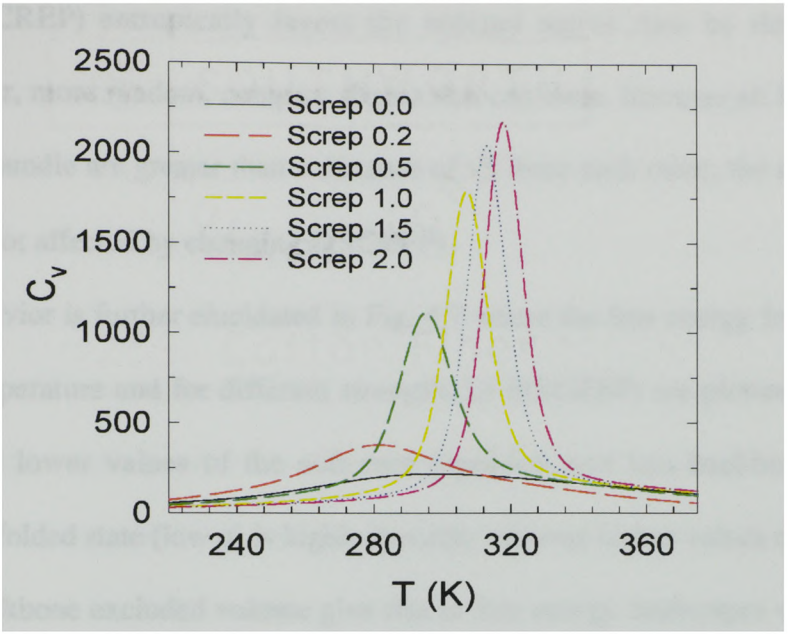


Fig. 4.8. Heat capacity curves as functions of temperature for various *E*(SCREP).

4.3.2 Enhanced stability due to backbone excluded volume: chain thickness

Next, the effects of the backbone thickness on the stability of the four-helix bundle native state are investigated. As explained earlier, the effective thickness of the backbone can be increased by increasing the penalty when two backbones approach by increasing the value of *E*(SCREP).

Figure 4.6 displays the percentage of the runs that unfolded as a function of the strength of the soft-core repulsion. The temperature was set to 312 K and the interaction energies are the same as listed above. The only sidechains present were the *H* ones in the helices; the *I* residues carried no sidechains. Again, we defined the protein to be completely unfolded if the helical content is 15% or less of the original helical content. From the figure it is evident that as *E*(SCREP) increases, the four-helix bundle becomes relatively more stable against thermal fluctuations to the denatured molten globule state.

Increased E(SCREP) entropically favors the ordered native state by decreasing the number of other, more random, compact shapes that can exist. Because all backbones in the four-helix bundle are greater than a distance of $\sqrt{5}$ from each other, the energy of the native state is not affected by changing E(SCREP).

This behavior is further elucidated in Fig. 4.7 where the free energy functions $F(q)$ at a single temperature and for different strengths of E(SCREP) are plotted. The figure shows that for lower values of the soft-core repulsion and less backbone excluded volume, the unfolded state (low q) is highly favored, whereas higher values of E(SCREP) and greater backbone excluded volume give rise to free energy landscapes with absolute minima near the native state configuration (high q).

In order to further investigate the enhanced stability of the four-helix bundle due to the increase in soft-core repulsion excluded volume, the heat capacity, C_v, as a function of temperature T was calculated. As described above, the histogram technique was used, but first the best simulation temperature was determined for use with each value of E(SCREP). Heat capacity curves as functions of temperature are plotted in Fig. 4.8. While undergoing an unfolding transition from four-helix bundle to molten globule, the transition temperature shifts towards higher temperature as the soft-core repulsion is increased, showing increased stability of the native state four-helix bundle. In addition, the sharpness of the heat capacity curves increase as the soft-core repulsion increases. This implies that as the excluded volume of the backbone increases, the structural transition between the native state and the molten globule becomes more like a phase transition.

4.4 Effect on Protein Unfolding Times

It was shown in Sec. 4.3 that the four-helix bundle is thermodynamically more stable when the amino acid residues of the heteropolymer chain have a greater excluded volume, through either a larger backbone or sidechain. In addition to the thermal stability, the kinetics of the unfolding process is investigated and it is found that larger excluded volume also slows down the transition from the four-helix bundle to the molten globule.

Fig. 4.9 shows the median first passage time, MFPT defined as the number of Monte Carlo steps required for q to drop below 0.15. These results are all for T=327 K. At lower T, most of the *HP* runs did not unfold. At 327 K, a majority of simulations unfolded for each condition and the choice of which simulation to use as the median was unambiguous. As E(SCREP) and therefore the excluded volume of the backbone increases, MFPT increases. The same trend of increasing MFPT occurs when the excluded volume is increased by changing the size of the sidechain. When the inert sidechains are not present, MFPT=1.0×10^5 MC steps; when the inert sidechains are present, *HP*(0,0), there is an increase to MFPT=6.0×10^5 MC steps, and with hydrophilic sidechains, *HP*(0.1,0.5), MFPT changes slightly to 6.4×10^5 MC steps.

The kinetics is further examined by calculating the autocorrelation function [66] of the helicity q defined as

$$C(\tau) = \frac{\overline{\Delta q(t)\Delta q(t+\tau)}}{\overline{q(t)^2} - \overline{q(t)}^2} \tag{4.5}$$

where $\Delta q(t) = q(t) - \overline{q(t)}$. In Fig. 4.10(a) shows the decay of $C(\tau)$ for increasing values of E(SCREP). By fitting the average autocorrelation curve to a decaying exponential, we are able to quantify the trend by determining the autocorrelation

relaxation time τ_{cor}. In Fig. 4.10(b), τ_{cor} vs. E(SCREP) is plotted. It is clear that as the backbone excludes greater volume, τ_{cor} increases, except for the largest value of E(SCREP)= 2.0 Kcal/mol. For this large value of E(SCREP) almost half of the simulation runs did not unfold. In averaging 50 runs in which approximately half are dominated by large scale motion, but the other half are not, the interplay between small scale and large scale structural motions may complicate the dynamics, making the autocorrelation function harder to interpret. The overall trend of increasing τ_{cor} with increasing E(SCREP) also implies a longer time for the unfolding transition to occur [67].

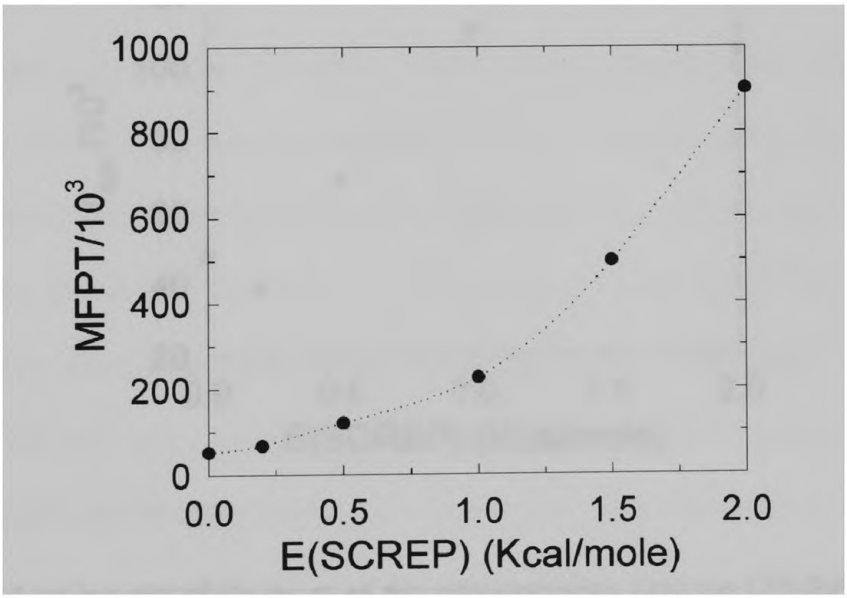


Fig. 4.9. Median first passage time (MFPT) as a function of E(SCREP). All points were from simulations at T=327 K

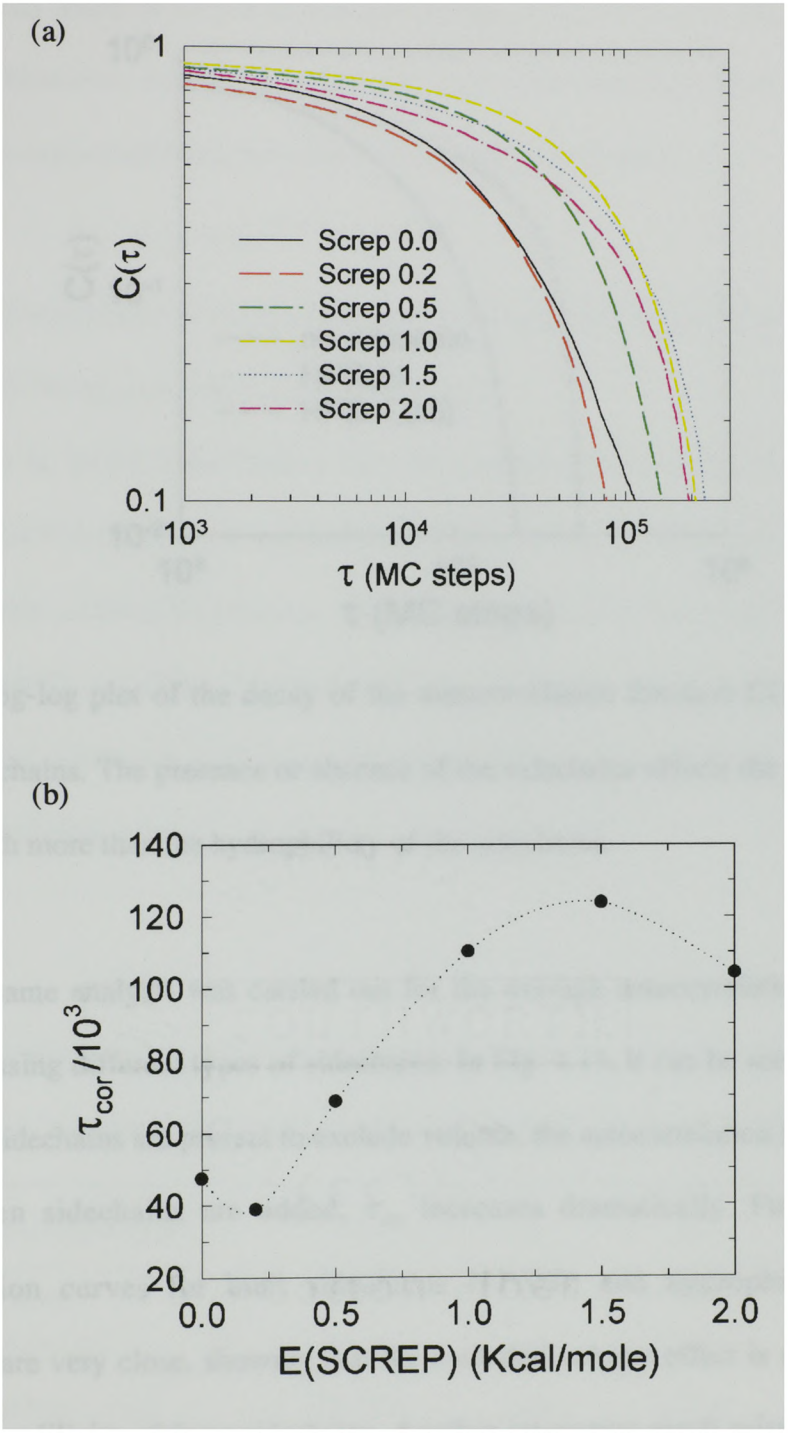


Fig. 4.10. (a) Log-log plot of the decay of the autocorrelation function $C(\tau)$ for increasing values of E(SCREP). (b) Autocorrelation relaxation time τ_{cor} as a function of E(SCREP).

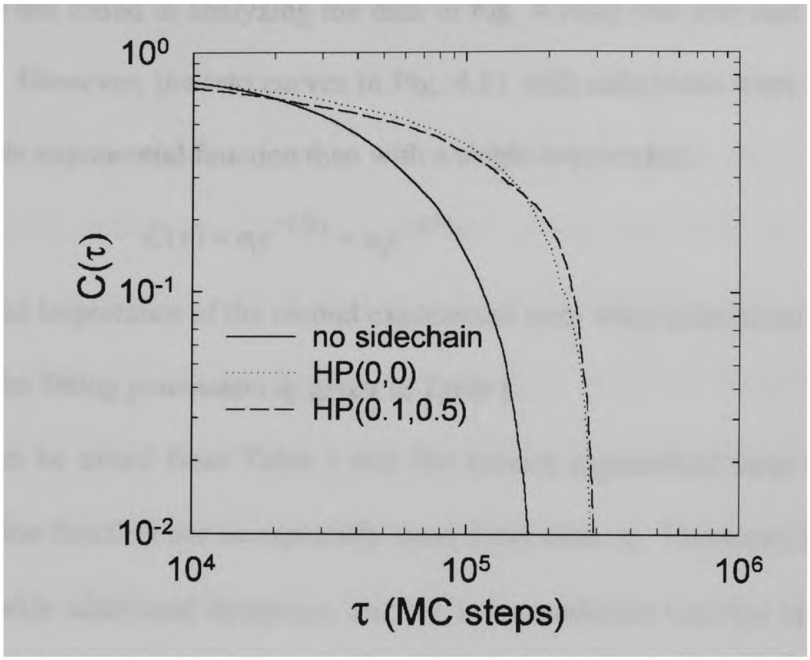


Fig. 4.11. Log-log plot of the decay of the autocorrelation function $C(\tau)$ for different types of sidechains. The presence or absence of the sidechains affects the autocorrelation function much more than the hydrophilicty of the sidechains.

The same analysis was carried out for the average autocorrelation function for simulations using different types of sidechains. In Fig. 4.11, it can be seen that when no hydrophilic sidechains are present to exclude volume, the autocorrelation function decays fastest. When sidechains are added, τ_{cor} increases dramatically. Furthermore, the autocorrelation curves for inert sidechains $HP(0,0)$ and hydrophilic sidechains $HP(0.1,0.5)$ are very close, showing that the excluded volume effect is more important than the hydroplilicity of these sidechains. Another interesting result relates to the fitting of the autocorrelation curves when sidechain are added. For the curve in Fig. 4.11 with no sidechains, a single exponential gave an excellent fit, as expected from the good single

exponential fits found in analyzing the data in Fig. 4.10(a) that also had no hydrophilic sidechains. However, the two curves in Fig. 4.11 with sidechains were fit much better with a double exponential function than with a single exponential.

$$C(\tau) = a_1 e^{-t/\tau_1} + a_2 e^{-t/\tau_2} \tag{4.6}$$

The increased importance of the second exponential term when sidechains are present can be seen by the fitting parameters a_2 given in Table I.

It can be noted from Table I that the second exponential term used to fit the autocorrelation function has an especially short delay time, τ_2. Therefore, the presence of sidechains adds additional dynamics, and the autocorrelation function is exhibiting the ability to quantify this effect. Again, the excluded volume entropic effect, this time from the presence of sidechains in the water-exposed surface, appears more important than the hydrophilic enthalpic effect.

TABLE I. Double exponential fitting parameters from Eq. (4.6) for different types of sidechains. The autocorrelation times, τ_1 and τ_2, are given in units of 100 MC steps.

Sidechain type	a_1	τ_1	a_2	τ_2
None	0.932	527	0.075	2
HP(0,0)	0.796	1416	0.186	23
HP(0.1, 0.5)	0.729	1514	0.230	38

4.5 Discussion

It is found that excluded volume effects are crucially important for the stability of the ordered native state structure of the four-helix bundle. Increasing the volume excluded by sidechains or backbones of a protein significantly favor the native state. In addition, the heat capacity peak at the transition temperature becomes sharper implying that the unfolding structural transition becomes more like a first order phase transition. The increased favoring of the native state may have important significance for protein folding. Different amino acids with similar biochemical properties (hydropathy) have side chains of different sizes. (e.g. small valine vs. large methionine or phenylalanine). However, in some locations along the primary sequence it may be critical to have large sidechains to provide the excluded volume necessary to prevent the stabilization of the molten globule. If smaller sidechains with the same hydropathy are incorrectly located at these locations, folding may never occur because the peptide chain becomes trapped in a molten globule intermediary. If this strategic placement of large amino acids is not sterically possible, or has other evolutionarily unfavorable effects, then chaperones might be necessary to allow folding to occur by supplying the excluded volume to block the molten globule. Therefore, proteins that require chaperones for folding may be able to fold without chaperones if there are strategic replacements in the primary sequence of amino acids with smaller sidechains by replacing them with larger sidechains with similar hydropathy.

CHAPTER V

PHASE TRANSITION STUDIES IN PROTEIN STRUCTURAL TRANSITIONS

The denaturation (unfolding) of globular proteins is a structural transition from the highly organized biologically active native state of a protein to a less organized molten globule, which is not biologically active. The unfolding may also take the molten globule to a more extended random coil. Structural transitions between collapsed states and extended states have been studied in great detail using computer simulations for homopolymers [68, 69] as well as protein-like heteropolymers [70]. Several studies [71, 72] have examined phase diagrams in simulations of proteins containing secondary and tertiary structure. These studies provide a great deal of insight on the thermodynamics of protein folding. Detailed statistical mechanical investigations have been carried out that have resulted in the calculation of critical exponents [68, 69] in protein simulations. These important calculations were done for a single α-helix, which is a common element of protein secondary structure. In this chapter, I briefly review phase transition and finite size scaling theory and present the results of their application in protein unfolding transition by examining thermodynamic properties of a four-helix bundle.

5.1 Phase Transition and Critical Phenomena

A phase transition is characterized by a discontinuity in derivatives of the free energy. In a phase transition, a significant change in a property of the system is observed. For example, transitions from liquid to gas or from paramagnet to ferromagnet show such behavior. If there is a finite discontinuity in one or more of the first derivatives of the free

energy, the transition is called a first order transition [72, 73]. For example, the Gibb's free energy as a function of temperature in a liquid-gas transition at constant pressure has a discontinuity in the slope at the transition temperature as displayed in Fig. 5.1(a). In the diagram, the tangent is the first derivative of the free energy function with respect to temperature at constant pressure, which is the entropy of the system, $S = -(\partial G/\partial T)_P$. As displayed in the T-S isobar in Fig 5.1(b), the entropy changes abruptly at the temperature corresponding to the discontinuous tangent in the G-T curve, a typical characteristic of a first order transition. Such a discontinuity corresponds to the absorption or release of heat, and is called the latent heat.

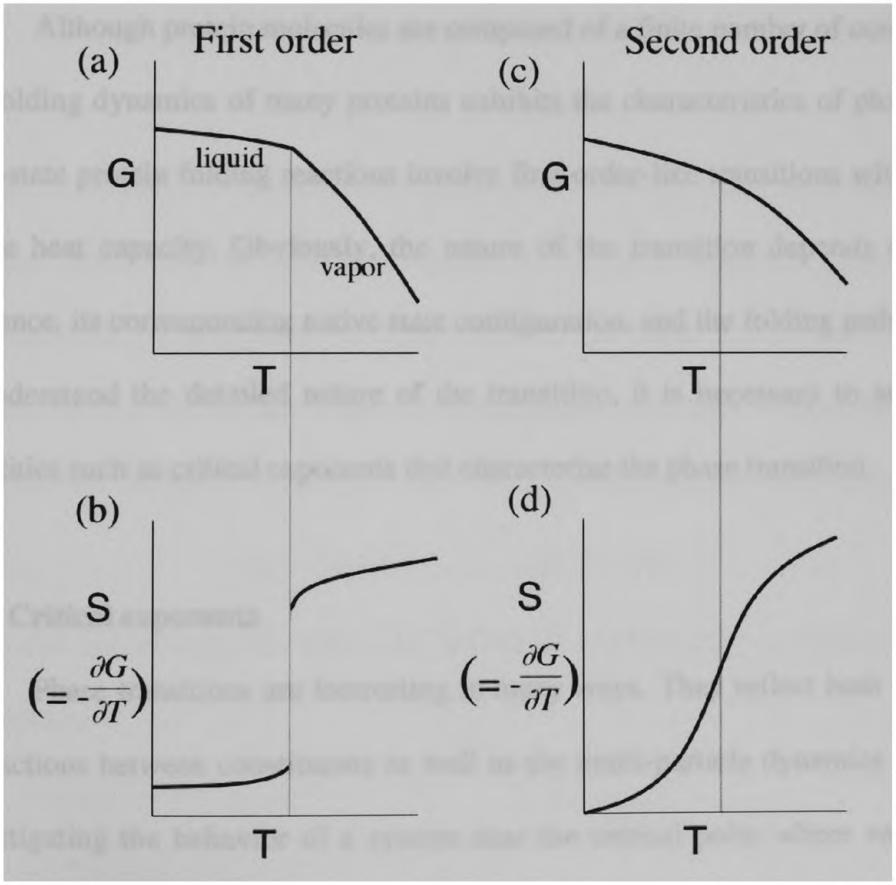


Fig. 5.1. Schematic display showing first and second order phase transitions.

If the first derivatives are continuous but second (or higher) derivatives are discontinuous, the transition is called a second (or higher) order transition [72, 73]. Such transitions are also called continuous or critical transitions and are characterized by divergent susceptibility and a power law decay of correlations. Figures 5.1(c) and 5.1(d) display typical free energy and entropy curves of a system undergoing a continuous phase transition. Some systems display quite unusual behavior in the vicinity of a continuous phase transition and such phenomena are known as critical phenomena. For example, an optical effect known as critical opalescence causes large fluctuations in the refractive index of critical mixtures of liquids such as methanol and hexane and causes normally transparent liquids to appear milky.

Although protein molecules are composed of a finite number of constituent atoms, the folding dynamics of many proteins exhibits the characteristics of phase transitions. Two-state protein folding reactions involve first-order-like transitions with a sharp peak in the heat capacity. Obviously, the nature of the transition depends on the protein sequence, its corresponding native state configuration, and the folding pathways. In order to understand the detailed nature of the transition, it is necessary to analyze various quantities such as critical exponents that characterize the phase transition.

5.1.1 Critical exponents

Phase transitions are interesting in many ways. They reflect both the underlying interactions between constituents as well as the multi-particle dynamics of the system. Investigating the behavior of a system near the critical point where various physical quantities possess singularities can be especially enlightening. These singularities are

customarily expressed in terms of power laws characterized by a set of quantities called critical exponents, which describe the behavior of various quantities of interest near the transition point. A complete review of critical exponents is beyond the scope of the present work and the details can be found elsewhere [73-75]. A brief introduction to some of the critical exponents relevant to the present work is given below.

Consider an order parameter m whose value is determined by the corresponding ordering field h. In the limit $h \to 0$, $m \to m_0$ with the property that $m_0 = 0$ for $T \geq T_c$ and $m_0 \neq 0$ for $T < T_c$. For magnetic systems, the order parameter can be the magnetization M and the magnetic field H is the ordering field h. For a liquid-gas system, the density differential $|1-\rho/\rho_c|$ can serve as m and the pressure differential $(1-P/P_c)$ as h.

The critical exponent α is defined by the divergence of the specific heat:

$$C \sim t^{-\alpha} \qquad (t>0) \tag{5.1}$$

where t is the reduced temperature defined as $t = (T - T_c)/T_c$ and measures the difference in temperature from the critical temperature, T_c. In a liquid-gas transition, the heat capacity in Eq. (5.1) is taken at constant volume. The same type of relation holds true for the specific heat at constant magnetization in phase transitions in magnetic systems. The exponent β refers to the behavior of the order parameter in zero ordering field:

$$m_0 \sim (-t)^{-\beta} \qquad (t<0,\ h \to 0) \tag{5.2}$$

Similarly, the manner in which the low-field susceptibility χ diverges, defines the exponent γ:

$$\chi \sim t^{-\gamma} \qquad (t>0,\ h \to 0) \tag{5.3}$$

The divergence of the correlation length ξ defines the critical exponent ν in the following way:

$$\xi \sim t^{-\nu} \qquad (t>0, h\to 0) \tag{5.4}$$

Corresponding to each of the above exponents, there exist separate exponents, namely α', β', γ', and ν', when the critical point is approached from temperatures below T_c, i.e. $t<0$. There are more critical exponents from similar power law behaviors of the system near the critical point. However, some critical exponents are not independent of others. For example, the critical exponents α, β, and γ are related by the relation $\alpha + 2\beta + \gamma = 2$. The values of the exponents differ very little as we go from one system to another, implying universality. The values of the critical exponents as well as the equality and inequality relations provide insights into the singularities in thermodynamics functions near the transition and broader understanding, and characterization the nature of the phase transition in a system.

5.1.2 Finite Size Scaling

Phase transitions are well defined only in the thermodynamic limit of an infinite number of constituent particles. Protein molecules are finite in size. A typical protein of 100 amino acids has on the order of 10^3 atoms, and $\sqrt{N}/N$ is not close to zero, showing that it is not near the thermodynamic limit. Fortunately, it is possible to calculate critical exponents and extrapolate to critical behavior in the limit of infinite size by analyzing phase-transition-like behavior in finite systems using finite size scaling. Finite size scaling investigations have been very valuable in investigations of Ising systems,

homopolymer transitions, and other systems composed of identical units. A brief discussion on finite size scaling pertinent to the dissertation is presented here. More detailed treatments can be found elsewhere [74].

Consider a d-dimensional system in the thermodynamic limit that undergoes a phase transition at a finite critical temperature $T_c(\infty)$. The critical exponents are d dependent and obey the hyperscaling relation $\nu d = 2 - \alpha = 2\beta + \gamma$. In the thermodynamic limit, a lattice of size $L_1 \times ... \times L_d$ means that each $L_j \to \infty$. In the case of finite system size, some L_j are allowed to stay finite. The main aim of finite size scaling is to determine the L dependence of various physical quantities when the system undergoes a phase transition. Near the critical point, it can be shown [74] that the susceptibility, heat capacity and the correlation length scale with system size L as follows:

$$\chi \sim L^{\gamma/\nu}, \quad C \sim L^{\alpha/\nu}, \quad \xi \sim L \qquad \text{(5.5a, b, c)}$$

For infinite systems, these parameters display singularities at $T=T_c$ whereas for finite size systems, these parameters exhibit peaks at T_c. The dependencies of these parameters on t exhibit universality if t is scaled as

$$t \sim L^{-1/\nu} \qquad \text{(5.6)}$$

This scaling will be used later in analyzing heat capacity calculations.

Thus, the critical exponents can be directly calculated from the size scaling of these thermodynamic quantities. In the present work, the dimension of a protein chain undergoing a structural transition is difficult to associate to a certain integer (or non-integer) number and therefore the dimension is kept as d without assigning a value.

5.2 Effect of Hydrophobic Interaction in Protein Structural Transitions

Since the hydrophobic interaction plays a major role in the dynamics of the protein chain and the stability of the native state, it is important to understand how the nature of the transition depends on the strength of the hydrophobic interaction. Simulations have been performed for a model four-helix bundle protein [Fig. 5.2] in order to investigate the nature of the unfolding transition by calculating critical exponents. This model protein contains the secondary structural elements of the individual helices and the tertiary structure of the bundle. The helices are formed with repeating sequences of *–HHPPHHPP–* and the loops connecting the helices are *–PPPP–*. The free energy of the system is lowered and the structure is stabilized when hydrophobic sidechains are buried in the middle of the bundle so that they are in contact with each other in a solvent-excluded volume. Each helical turn is composed of four amino acids with the sequence *–HHPP–*. When the protein is in its native state four-helix bundle, each *H* amino acid has its side chain pointing into the middle of the bundle where it makes a hydrophobic interaction with one hydrophobic sidechain from a neighboring helix [Fig. 5.2(a)].

5.2.1 Finite Size Scaling in a Protein Heteropolymer

Recent studies using finite size scaling techniques to investigate structural transitions in a single helix [68, 69] used a homopolymer polypeptide in which the repeating unit was alanine and four different chains up to length 30 were investigated. In the present work, I use a heteropolymer chain composed of two distinct types of amino acids; hydrophobic and hydrophilic. The hydrophobic interactions between buried

sidechains stabilize the tertiary structure of the four-helix bundle. The identical repeating unit that is varied is not the individual monomers, but the number of helical turns, N. First, simulations are run for a four-helix bundle in which each helix had four turns plus an additional half turn that leads into a loop segment connecting two helices, giving N=18. Then, one turn from each helix is removed, giving N=14, and the simulations are re-run. This procedure of removing one turn from each helix is repeated. In this study I have used N=10, 14, or 18, corresponding to chains of amino acid length of 51, 67, or 83 as shown in Fig. 5.2(b-d). If the four-helix bundle was composed only of helical segments, then the ratio of the number of helical turns to the number of amino acids would be the same for all chain lengths, since each helical turn contains four amino acids. This is not true, due to the connecting loops that join the four helices into a bundle.

To study the effect of the hydrophobic interaction strength on the nature of the structural transition using finite size scaling, the number of turns in each of the four helices is progressively reduced. As explained below, critical exponents are determined from slopes of graphs that plot different properties as a function of N. If finite size scaling is applicable, then each plot will have a unique critical exponent and therefore each plot should be a straight line. If finite size scaling is not applicable, then there will not be a well-defined critical exponent and a plot will not be a straight line. It is found that when calculating slopes, the three points for N=18, 14, 10 always fell nicely on a straight line for all of the plots. This consistency supports the usefulness of finite scaling for this system.

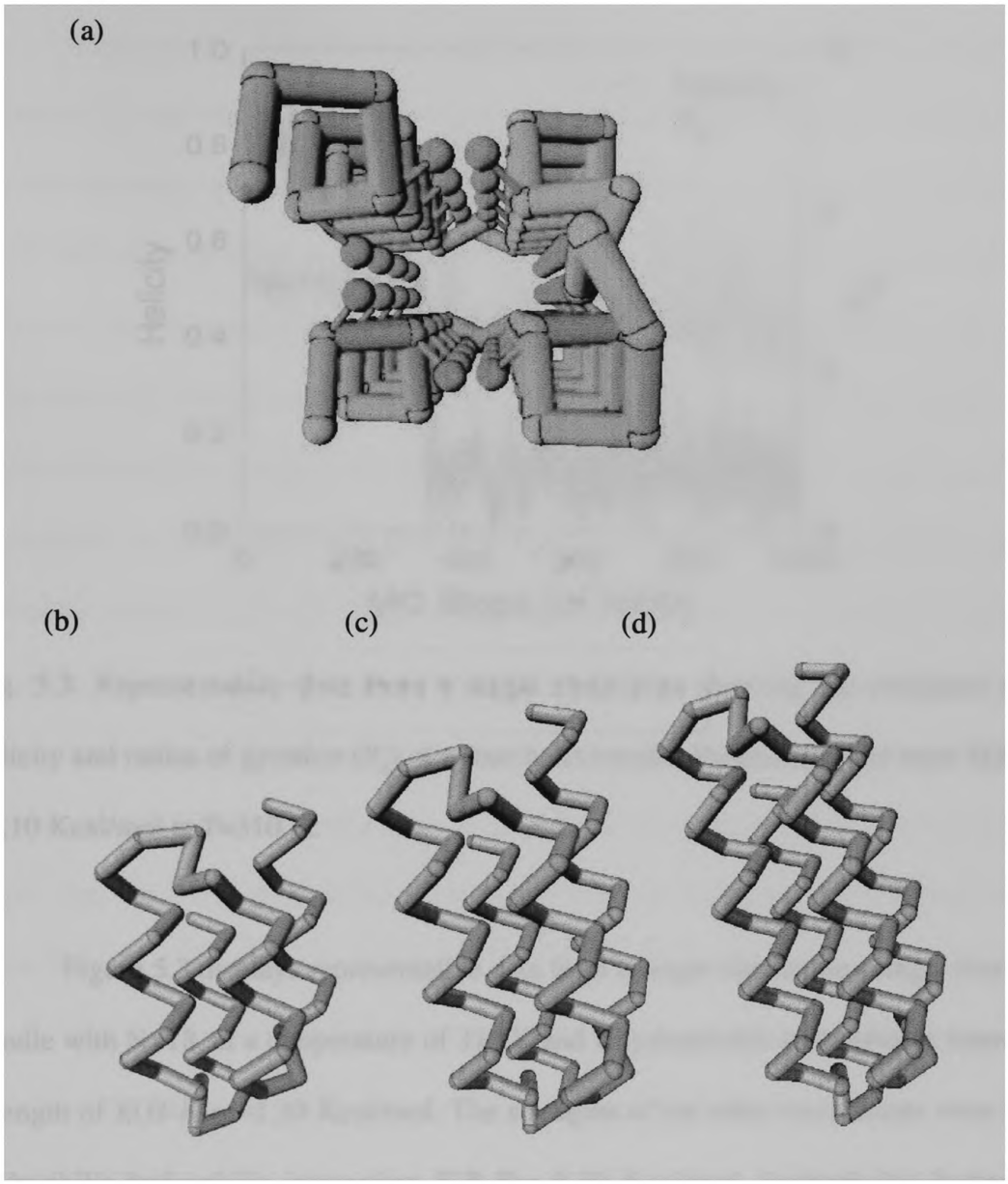


Fig. 5.2. (a) Ball and stick display of a four-helix bundle with the number of helical turns N=18, corresponding to a chain length of 83 amino acid residues (top view). Side view of the four-helix bundle (b) N=10, chain length of 51 amino acid residues. (c) N=14, chain length of 67 amino acid resides (d) N=18, chain length of 83 amino acid residues.

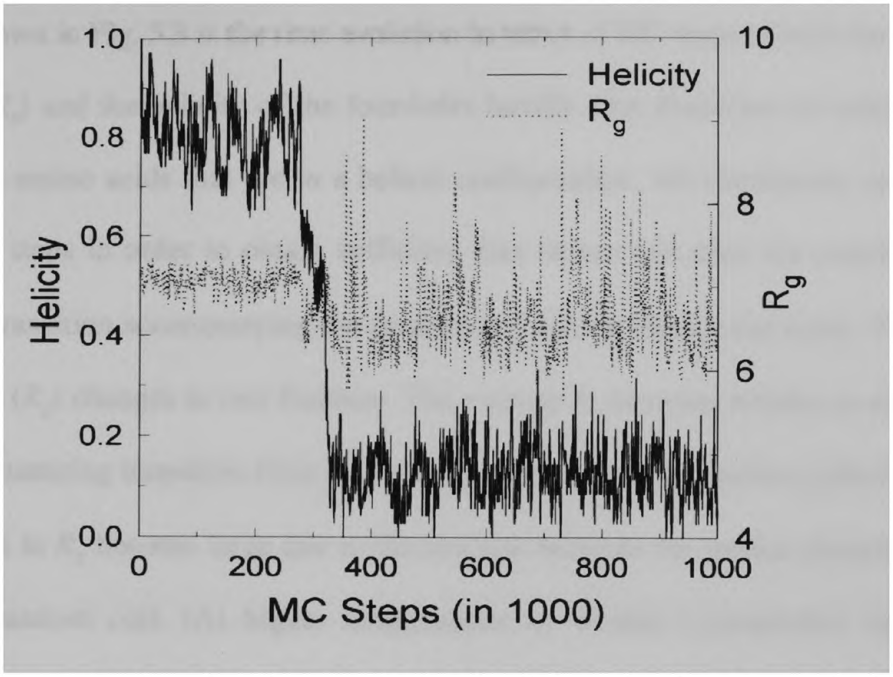


Fig. 5.3. Representative data from a single simulation showing the evolution of the helicity and radius of gyration (R_g) of a four-helix bundle. Parameters used were $E(H\text{-}H)$= –1.10 Kcal/mol at T=310 K.

Figure 5.3 displays representative data from a single simulation using a four-helix bundle with N=18, at a temperature of 310 K and a hydrophobic-hydrophobic interaction strength of $E(H\text{-}H)$= –1.10 Kcal/mol. The strengths of the other interactions were set at: hydrophilic-hydrophilic interaction $E(P\text{-}P)$= 0.10 Kcal/mol, hydrophobic-hydrophilic interaction $E(H\text{-}P)$= 0.50 Kcal/mol, local propensity E_L= –0.50 Kcal/mol, cooperative interaction E_M= –0.50 Kcal/mol, hydrogen bond and dipole interaction E(HBDIP)= –0.25 Kcal/mol, E(SCREP)= 1.50 Kcal/mol. The relative strengths of the interaction parameters are chosen to be qualitatively similar to the experimental results and scaled so that the corresponding transition temperature is close to body temperature.

Shown in Fig. 5.3 is the time evolution in terms of MC steps of both the radius of gyration (R_g) and the helicity of the four-helix bundle. Our definition of helicity is the fraction of amino acids that are in a helical configuration. All simulations are run for 1×10^6 MC steps in order to obtain sufficient data before and after the transition. The structural transition accompanying the denaturation can be clearly discerned. The radius of gyration (R_g) changes in two fashions. The average R_g becomes smaller as the system makes a denaturing transition from the native state to a compact molten globule, but the fluctuations in R_g become large due to fluctuations between the molten globule and the extended random coil. (At higher temperatures or weaker hydrophobic interaction strengths, the denatured state not only has larger R_g fluctuations than the native state, but spends more time in the extended random coil configuration and the average R_g of the denatured state becomes larger than the R_g of the four-helix bundle.) Figure 5.3 also shows that the fraction of amino acids in a helical configuration drops almost to zero when the protein denatures.

Using a Monte Carlo histogram method, the specific heat capacity per turn, C_N, is calculated for different system sizes and different hydrophobic strengths as a function of temperature [68]:

$$C_N = (\overline{E^2} - \overline{E}^2)/NT^2 \tag{5.7}$$

In Eq. (5.7), N is the total number of turns in the four-helix bundle, T represents RT, where R is the universal gas constant and T is the temperature in Kelvin. The averaged quantities in the above expression are calculated with the histogram technique.

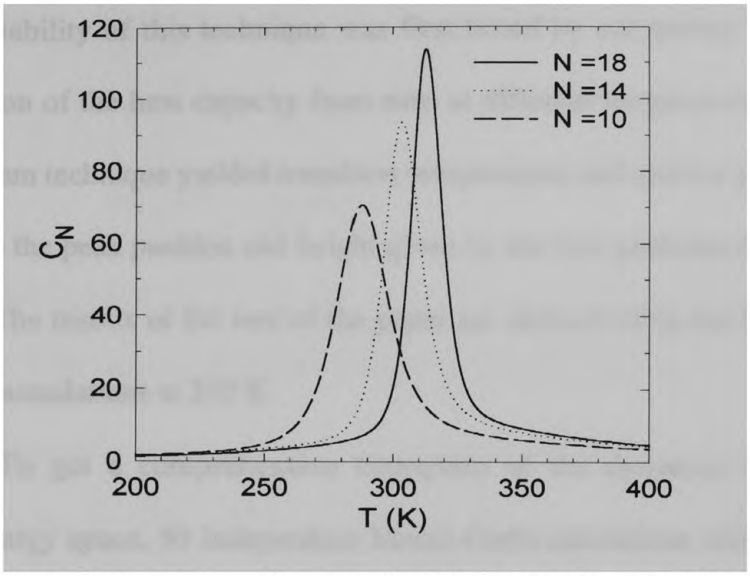


Fig. 5.4. Heat capacity per turn, C_N, as a function of temperature at $E(H\text{-}H)$= –1.10 Kcal/mol for various chain lengths. As the system size increases, the transition temperature increases, the peak height increases, and the relative width decreases.

TABLE I. The transition temperature T_c from the peak of C_N, the height $C_N(\max)$, and Γ, the full width at half height of the peak of C_N, as a function of both N and the strength of the hydrophobic interaction, $E(H\text{-}H)$, given in Kcal/mol.

$E(H\text{-}H)$	N	T_c	$C_N(\max)$	Γ
	18	311	137	12.3
–1.00	14	301	120	14.3
	10	286	95.2	17.6
	18	312	113	14.3
–1.10	14	303	93.3	18.1
	10	288	70.4	24.9
	18	317	82.7	19.0
–1.20	14	307	69.5	23.9
	10	289	43.2	47.6
	18	321	72.7	21.6
–1.30	14	308	52.4	32.5
	10	309	31.2	75.0

The reliability of this technique was first tested by comparing the results with direct calculation of the heat capacity from runs at different temperatures. It was found that the histogram technique yielded transition temperatures and relative peak heights that agree well with the peak position and height given by the runs performed over a range of temperatures. The results of the rest of the paper are derived using the faster histogram technique with simulations at 310 K.

To get a comprehensive histogram of the dynamics that thoroughly samples the energy space, 50 independent Monte-Carlo simulations were used for each hydrophobic interaction strength. Each simulation started with the same four-helix bundle configuration, but with a different random number seed. Figure 5.4, with $E(H\text{-}H) = -1.10$ Kcal/mol, shows the temperature dependence of the heat capacity for four-helix bundles with N=10, 14, and 18, corresponding to 2.5, 3.5, and 4.5 turns per helix, respectively. It is clear from the figure that the transition temperature increases as the system size increases, i.e. $T_c = T_c(N)$. This is likely due to increased stability of the four-helix bundle due to an increase in the number of hydrophobic sidechains in the core. Also, the peak is higher and sharper for the largest and most stable four-helix bundle. The results are summarized in Table I for several different hydrophobic strengths.

In order to determine the order of the transition between the native state and the unfolded state, critical exponents are calculated using finite size scaling theory. In this analysis I follow the work of Alves [69] and equate the relevant linear length L to $N^{1/d}$. As with Alves, no assumptions are made about relating d to a particular integer geometrical dimension and values for the combined quantity $d\nu$ are determined, where ν is the correlation critical exponent. Equation. (5.6) discussed scaling of temperature

differences, $t \sim L^{-1/\nu}$. An important choice of t is the width of the specific heat capacity curve Γ_c. Using $L \sim N^{1/d}$ and $t \sim L^{-1/\nu}$ gives $\Gamma_c(N) \sim N^{-1/d\nu}$ so that

$$\ln \Gamma_c(N) = -\frac{1}{d\nu} \ln N + \ln \Gamma_o \tag{5.8}$$

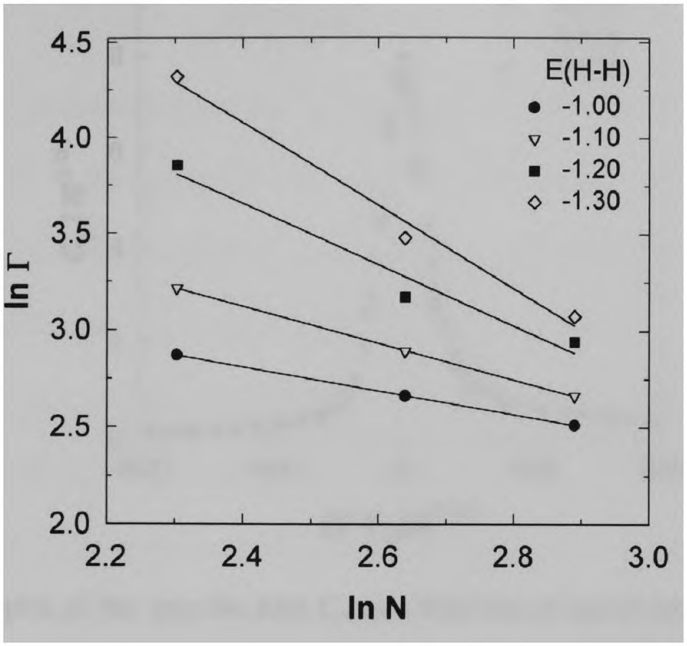
where Γ_o is independent of N. In Eq. (5.4), we use the customary definition of $\Gamma_c(N) = T_2(N) - T_1(N)$ such that $C_N(T_1) = C_N(T_2) = \frac{1}{2} C_N(T_c)$ and the critical temperature $T_c(N)$ is the temperature where $C_N(T)$ has its maximum. Similarly, the size of the heat capacity peak scales as $C_N^{\max} \sim N^{\alpha/d\nu}$ where α is the specific heat critical exponent, so that

$$\ln C_N^{\max} = \frac{\alpha}{d\nu} \ln N + \ln C_o \tag{5.9}$$

where C_o is independent of chain length N. Group theoretical arguments predict for a first order phase transition values of $d\nu = \alpha = 1$ [69, 76, 77].

I now study the behavior of the critical exponents as the strength of the hydrophobic interaction changes. From the slopes of the plots in Fig. 5.5(a) and 5.5(b), it can be seen that the critical exponents change significantly as the strength of the hydrophobic interaction is increased. Table II gives the values of the critical exponents as determined from Figs. 5.5(a) and 5.5(b) and Eqs. (5.8) and (5.9). It can be that at the weaker hydrophobic strength, the critical exponents are near 1.0, implying a first order phase transition. As the hydrophobic interaction strengthens, the critical exponents deviate more and more from 1.0, showing that the transition is no longer first order. Another analysis discussed below is used to show that as $E(H\text{-}H)$ increases, the transition shifts from first to second order.

(a)



(b)

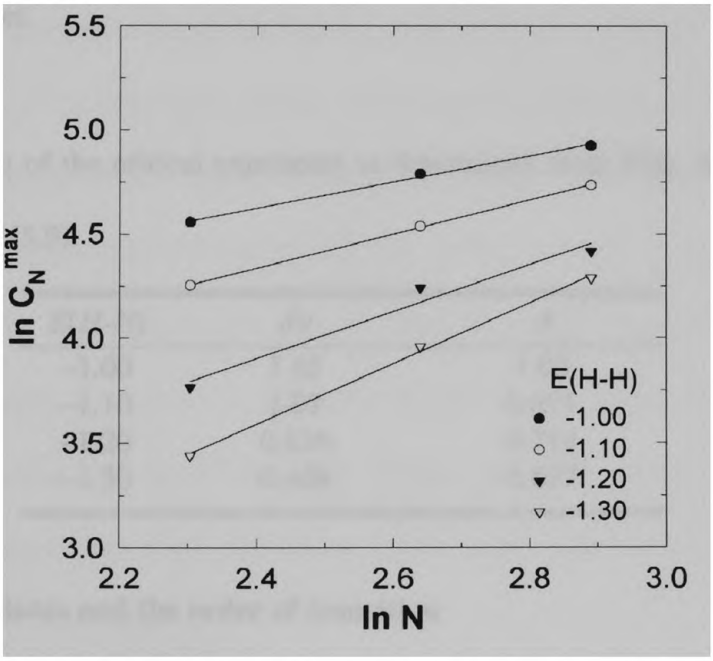


Fig. 5.5. (a) The width, Γ_c of the specific heat capacity curve versus the number of helical turns N (b) The size of the peak of the specific heat, C_N^{max}, versus the number of helical turns N.

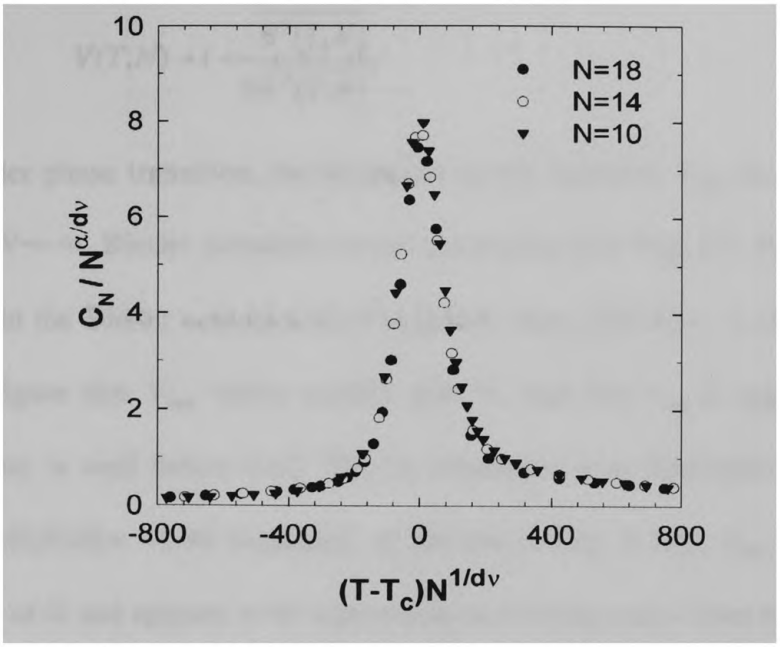


Fig. 5.6. A scaling plot of the specific heat C_N as a function of temperature for $E(H\text{-}H) = -1.10$ Kcal/mol showing that the different curves for N=10, 14, and 18 all superimpose on top of each other.

TABLE II. Values of the critical exponents as determined from Figs. 5.5(a) and 5.5(b) and Eqs. (5.8) and (5.9).

$E(H\text{-}H)$	$d\nu$	α
–1.00	1.65	1.03
–1.10	1.07	0.871
–1.20	0.636	0.714
–1.30	0.468	0.677

5.3 Binder Cumulants and the order of transition

The above results on the order of the transition are tested using a different analysis by calculating the Binder's reduced cumulant [75, 78], $V(T,N)$, which is sensitive to the nature of a phase transition,

$$V(T,N) = 1 - \frac{\overline{E^4(T,N)}}{3\overline{E^2(T,N)}^2} \qquad (10)$$

For a second order phase transition, the minimum of this quantity, V_{min}, is expected to approach 2/3 as $N \rightarrow \infty$. Binder cumulant curves are displayed in Fig. 5.7. Figure 5.7(a) shows the trend in the Binder cumulant as N increases when $E(H\text{-}H) = -1.10$ Kcal/mol. We see in this figure that V_{min} varies weakly with N, and that V_{min} is approaching a limiting value that is well below 0.67. This is consistent with first-order dynamics. However, when $E(H\text{-}H) = -1.40$ Kcal/mol, as shown in Fig. 5.7(b), V_{min} is a much stronger function of N and appears to be approaching a limiting value close to 2/3. Thus, as $E(H\text{-}H)$ increases, the transition shifts towards second-order behavior.

This trend towards second order behavior is also displayed in Fig. 5.8 which shows the Binder cumulant curves for N=18 as $E(H\text{-}H)$ is systematically changed. As the hydrophobic strength increases, the curves monotonically approach 2/3 and second order behavior. The validity of the various analyses is further strengthened if we note that $T_c(N=\infty)$ can be approximated from the Binder cumulant curves as the temperature at which the curves for different N cross. In Fig, 5.7(a), this temperature is at approximately 320 K, which is very close to the limiting value of the position T_c of the heat capacity peak in Fig. 5.4, which also has $E(H\text{-}H) = -1.10$ Kcal/mol. In Fig. 5.7(b), where $E(H\text{-}H) = -1.40$ Kcal/mol, the curves cross at a higher temperature, showing, as expected, that T_c increases as $E(H\text{-}H)$ increases.

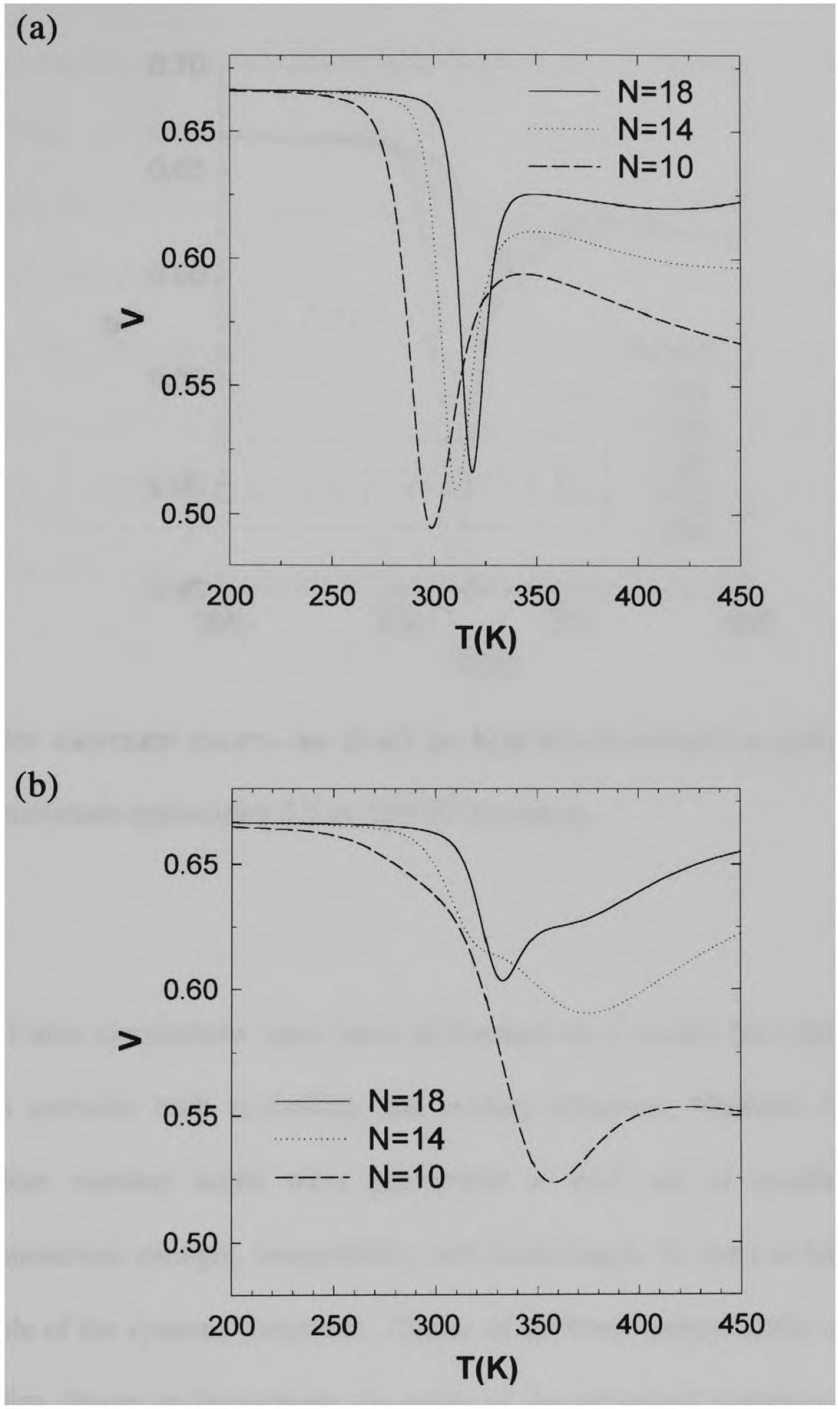


Fig. 5.7. Binder cumulant curves. (a) Binder cumulant as a function of N for $E(H\text{-}H)$= –1.10 Kcal/mol showing a weak N dependence. (b) Binder cumulant as a function of N for $E(H\text{-}H)$= –1.40 Kcal/mol, showing a much stronger dependence on N and approach to a limiting value close to 2/3.

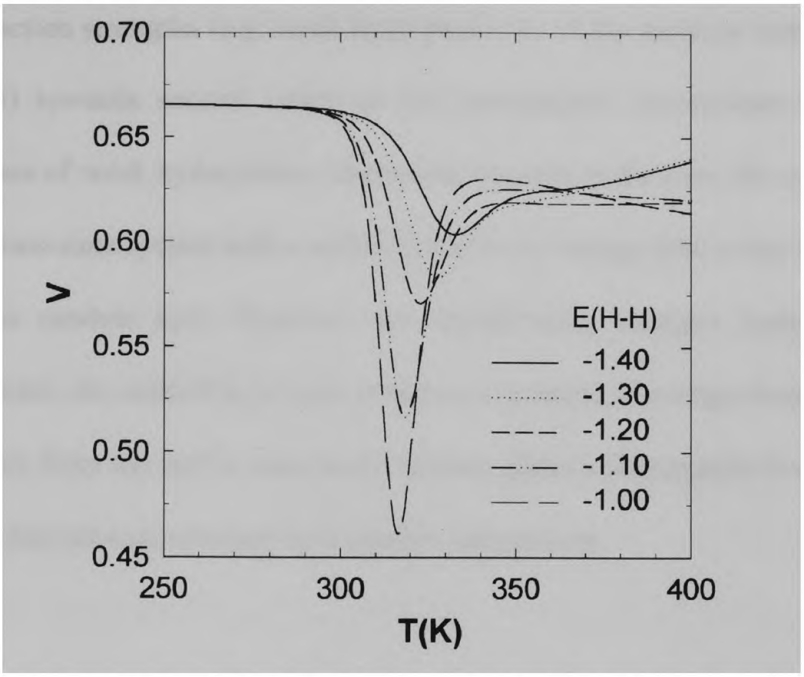


Fig. 5.8. Binder cumulant curves for N=18 as $E(H\text{-}H)$ (Kcal/mol) is systematically changed. The minimum approaches 2/3 as $E(H\text{-}H)$ increases.

5.4 Summary

Monte Carlo simulations have been performed on a model four-helix bundle protein, which contains both secondary and tertiary structure. Multiple runs using different random number seeds were performed at each set of conditions, i.e. hydrophobic interaction strength, temperature, and chain length, in order to have a good statistical sample of the systems' behavior. Chains of different lengths allow us to apply finite size scaling theory to investigate the order of the structural transition from the native four-helix bundle state to a denatured configuration. The results from finite size scaling and the calculation of Binder cumulants on this particular system of four-helix bundle reveal that the order of the structural transition shifts from first order at weak

hydrophobic interaction strengths (e.g. weak hydrophobicity of the residues forming the hydrophobic core) towards second order as the hydrophobic interaction strength increases. In the case of weak hydrophobic interaction strength in the core, the unfolding transition is like a two-state system with a sudden jump in the energy time series from the native state to the random coil. However, for significantly stronger hydrophobic interactions in the core, the unfolding is more continuous in terms of energy change since the transition occurs from the native state to the molten globule (disorganized unfolded but compact state) that consists of many hydrophobic interactions.

CHAPTER VI

FREE ENERGY LANDSCAPES AND KINETICS: PROTEIN ENGINEERING

A crucial aspect of protein folding and design is that the primary sequence of a protein must fold in a biologically relevant time of milliseconds or less [79, 80] to a stable native state [81]. The presence of non-native kinetic traps may significantly lengthen the overall protein folding time to the point of uselessness. I use a model α-turn-α helical hairpin peptide to show how changing only a few amino acid residues along the sequence can greatly increase the speed and reliability of the folding process, as seen experimentally [82]. These results support the arguments of Zhou and Karplus [83] that removal of kinetic traps, obligatory or non-obligatory, is crucial for fast folding. The trap-free folding follows single exponential kinetics. The folding kinetics for the sequence with traps does not follow single exponential kinetics and reflects multiple distinct pathways for folding. Not only does the sequence with traps fold slower, but it is also less reliable because the traps can be almost as stable as the native state and prevent the sequence from fully folding. Both the kinetics and thermodynamics are quantified to show how protein engineering can be used to enhance protein folding.

6.1 Protein Engineering

The interplay between secondary and tertiary structure formation is of crucial consideration in protein engineering and the understanding of protein folding. When engineering a protein, it is worthwhile to look for principles that will make a design more effective. Detailed studies [84-86] have been carried out on the role of secondary

structure preference in sequence design and the effect of amino acid substitution on helix formation. [87]. In the case of helical hairpin designs, an important consideration for tertiary structure formation is getting the correct inter-helical contacts [88]. In my research, the primary sequence is engineered with a small number of amino acid substitutions. These substitutions are made in a way that does not affect the formation of helical secondary structure but does encourage the formation of native tertiary contacts between helices, which is the rate limiting transition step. The hydrophobic core of a protein is crucial for the folding process as well as stabilizing the protein once it has attained the native state. In this study, it is shown that the strategic placement of hydrophilic residues in the core can increase both the speed and reliability of the folding process, and still maintain the stability of the native state. This strategy is used by nature in the leucine zipper, and may be an important consideration in future human engineering of synthetic proteins.

6.2 Sequence Design: Strategic substitution of amino acids to remove kinetic traps

An amino acid sequence is designed which is similar to the *de novo* design of Fezoui et al. [89, 90] that forms a two-helix bundle in the native state. It has already been shown that this simple structure can be engineered in the laboratory [89-91], and has been proposed as a useful system for more detailed investigations. The two-helix bundle considered here in the model consists of 41 residues. Each helix contains 18 residues and a turn segment composed of five residues connects them. The amino acids in the helical regions are given a preference for forming helical secondary structures. The residues in the turn segment are not given any preference for forming specific structure.

The thermodynamics and kinetics of peptide chains are investigated with two slightly different primary sequences at the interface between the two helices. The native state of Seq A has all hydrophobic sidechains, *H-H-H-H*, along the interface of each helix as shown in Fig 6.1(a). All four of the native state sidechain contacts in the core involve a hydrophobic sidechain from each helix. The folding dynamics gives rise to kinetic and thermodynamic structural traps which are configurations in which the protein spends significant time at an energy close to that of the native state. The deep traps occur when the helices are misaligned due to offset as shown in Fig. 6.1(b).

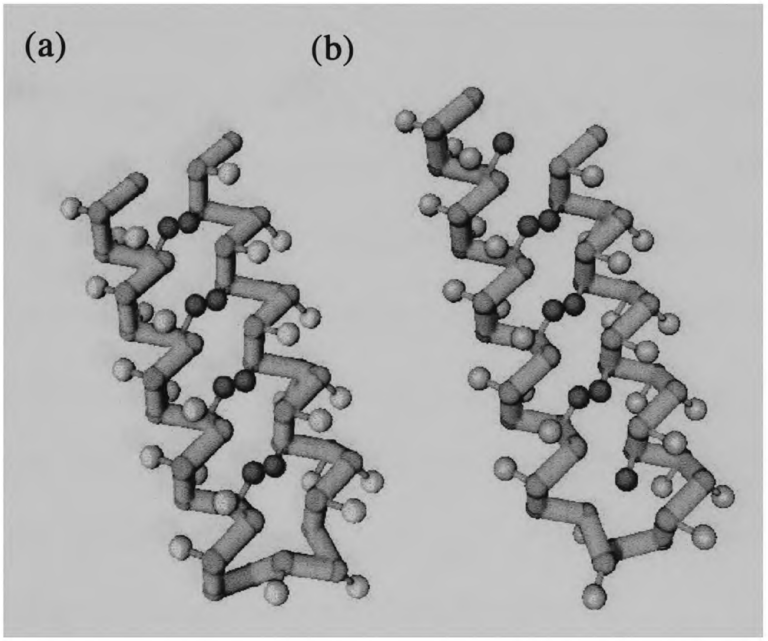


Fig. 6.1. (a) Native state of two-helix bundle. All the side chains in the interface between two helices make hydrophobic contacts. (b) Three hydrophobic contacts in the interhelical interface stabilize the misaligned trap conformation. Two hydrophobic sidechains that are not making any contacts may lead to protein aggregation.

This sequence is compared to Seq B in which we slightly redesign the primary sequence in a way that removes the tertiary structural traps by strategically replacing a single hydrophobic residue of each helix with a hydrophilic residue. The misaligned structure that is quite stable for Seq A [Fig 6.1(b)] is not stable for Seq B [Fig. 6.2(b)]. The misaligned structure with Seq B still has three contacts between the two helices, but two of these contacts are repulsive because they involve *H* from one helix and *P* from the other. Both sequences have the same turn segment composed of neutral residues that are neither hydrophobic (*H*) nor hydrophilic (*P*). In both sequences, the residues in the helices that are not at the interface are hydrophilic.

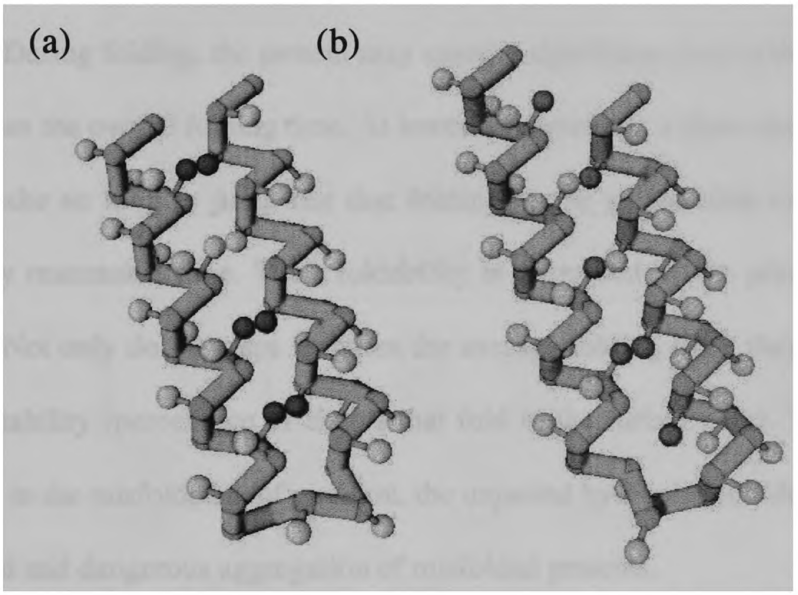


Fig 6.2. (a) Native state structure of two-helix bundle for Seq B. Sidechains in the interface between two helices make three strongly attractive hydrophobic contacts a weakly repulsive hydrophilic contact in the native state. (b) Only one attractive hydrophobic contact and two repulsive hydrophobic-hydrophilic contacts make this misaligned trap conformation very unstable.

For Seq A, all four of the native state sidechain contacts in the core involve a hydrophobic sidechain from each helix. However, the non-native structure of Fig. 6.1(b) is a kinetic and thermodynamic trap in which the chain spends significant time at an energy close to that of the native state. This can be seen in Fig. 6.3(a), which is a time series of the behavior of Seq A. The difference in tertiary structure can be quantified by the parameter Q, which is the fraction of tertiary native contacts between the helices that exist at a given time. In the native state of Fig. 6.1(a) when all four interface contacts are properly made, Q=1. The energy of the trap structure is almost as low as the native state because the misalignment still allows three hydrophobic interactions between the two helices. Due to the misalignment, none of these three hydrophobic interactions are native, and Q=0. During folding, the protein may spend a significant time in the deep traps and this increases the overall folding time. At lower temperatures, a chain that first falls into a trap may take so long to jump out that folding to the native state cannot occur in a biologically reasonable time. Thus, foldability is decreased by the presence of traps in two ways. Not only do the traps lengthen the average folding time, they also reduce the folding reliability (percentage of chains that fold to the native state). There is also the hazard that in the misfolded configuration, the unpaired hydrophobic sidechains may lead to unwanted and dangerous aggregation of misfolded proteins.

In spite of having one less hydrophobic interaction, the native state conformation with Q=1 for Seq B is long lived and has an energy almost as low as the native state for Seq A. Though the native state is quite stable for both Seq A and Seq B, the non-native misaligned structure of Fig. 6.1(b) loses almost all of its stability if Seq A is converted into Seq B. The misaligned Q=0 structure has three hydrophobic contacts for Seq A and

is almost as stable as the native state, and acts as a long-lived kinetic trap. In contrast, the three sidechain interactions in the misaligned Q=0 structure for Seq B, Fig. 6.2(b), has only one hydrophobic contact. The other two sidechain-interactions are repulsive interactions between a hydrophilic sidechain of one helix and a hydrophobic sidechain from the other helix. The strategic substitution of one amino acid on each sidechain removes the kinetic trap for Seq B by destabilizing the misaligned structure and making it short lived.

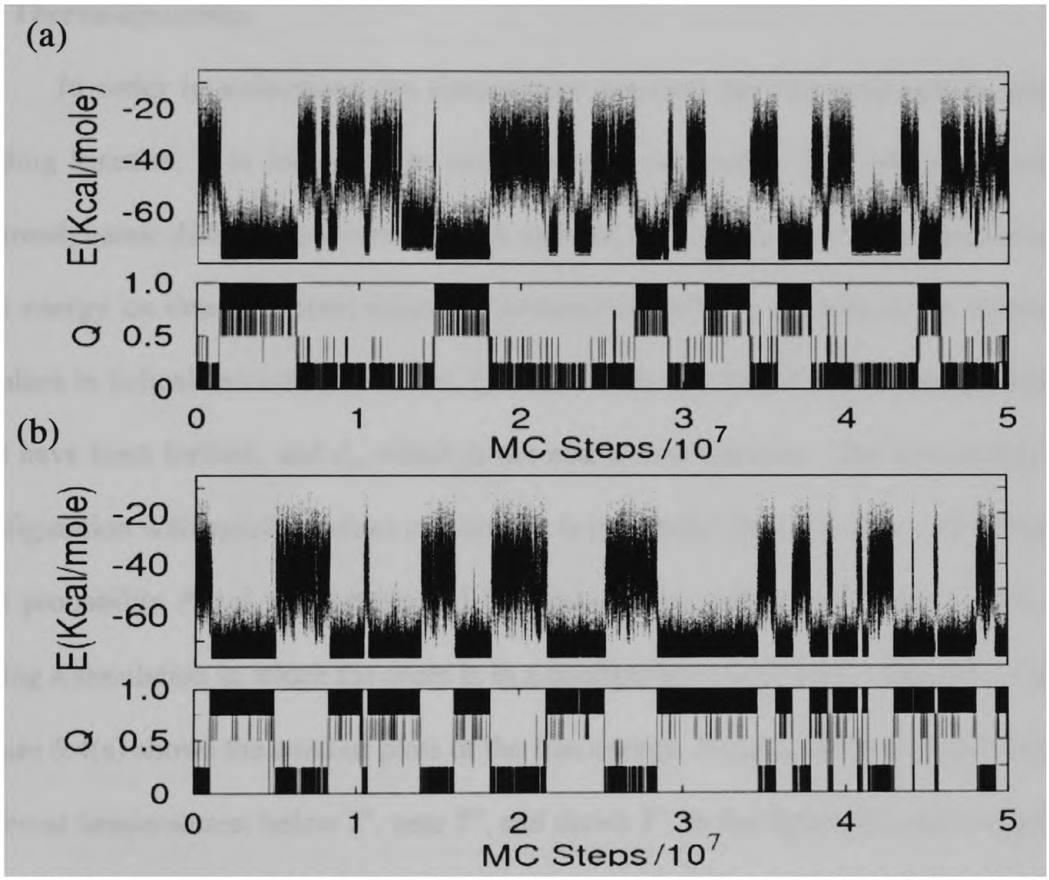


Fig. 6.3 Representative time series data for E and Q for (a) Seq A in which a significant time is spent in the low E conformations corresponding to misaligned structures where the value of Q is still close to zero. (b) Seq B where the removal of the low energy non-native structures is apparent.

Figure 6.3(a) contains a characteristic energy time series and Q time series for Seq A and Fig. 6.3(b) contains the same for Seq B. Figure 6.3(a) and Fig. 6.3(b) both display fluctuations in energy and Q due to folding and partial unfolding (Q>0). However, Fig. 6.3(a) has deep, long-lived traps where Q=0, which Fig. 6.3(b) does not have. The trap of Seq A slows down folding appreciably, as can be seen by the longer time scale of Fig. 6.3(a) compared to Fig. 6.3(b).

6.3 Thermodynamics

In order to understand the connections between the thermodynamics and the folding kinetics, it is important to use structural parameters that reflect the crucial thermodynamic differences between Seq A and Seq B. I calculate the dependencies of the free energy on three different structural parameters: helicity q which is the fraction of residues in helical secondary structure, Q which is the fraction of native tertiary contacts that have been formed, and d_{ee} which is the end to end distance. The free energy of a configuration with specific values of q and d_{ee} is calculated from $F(q,d_{ee})= -kT \ln P(q,d_{ee})$. The probability $P(q,d_{ee})$ is determined by counting the number of Monte Carlo steps during a simulation in which the chain is in a configuration with these values of q and d_{ee}. Figure 6.4(a) shows the contour plots of the free energy, $F(q,d_{ee})$, for Seq A and Seq B at different temperatures; below T', near T', and above T'. In the figure, the color represents the relative value of the free energy. The increasing darkness of the shades corresponds to the decreasing free energy. The native state free energy minimum is the basin around q~1 and d_{ee}~7. The non-native minimum has low q and high d_{ee}.

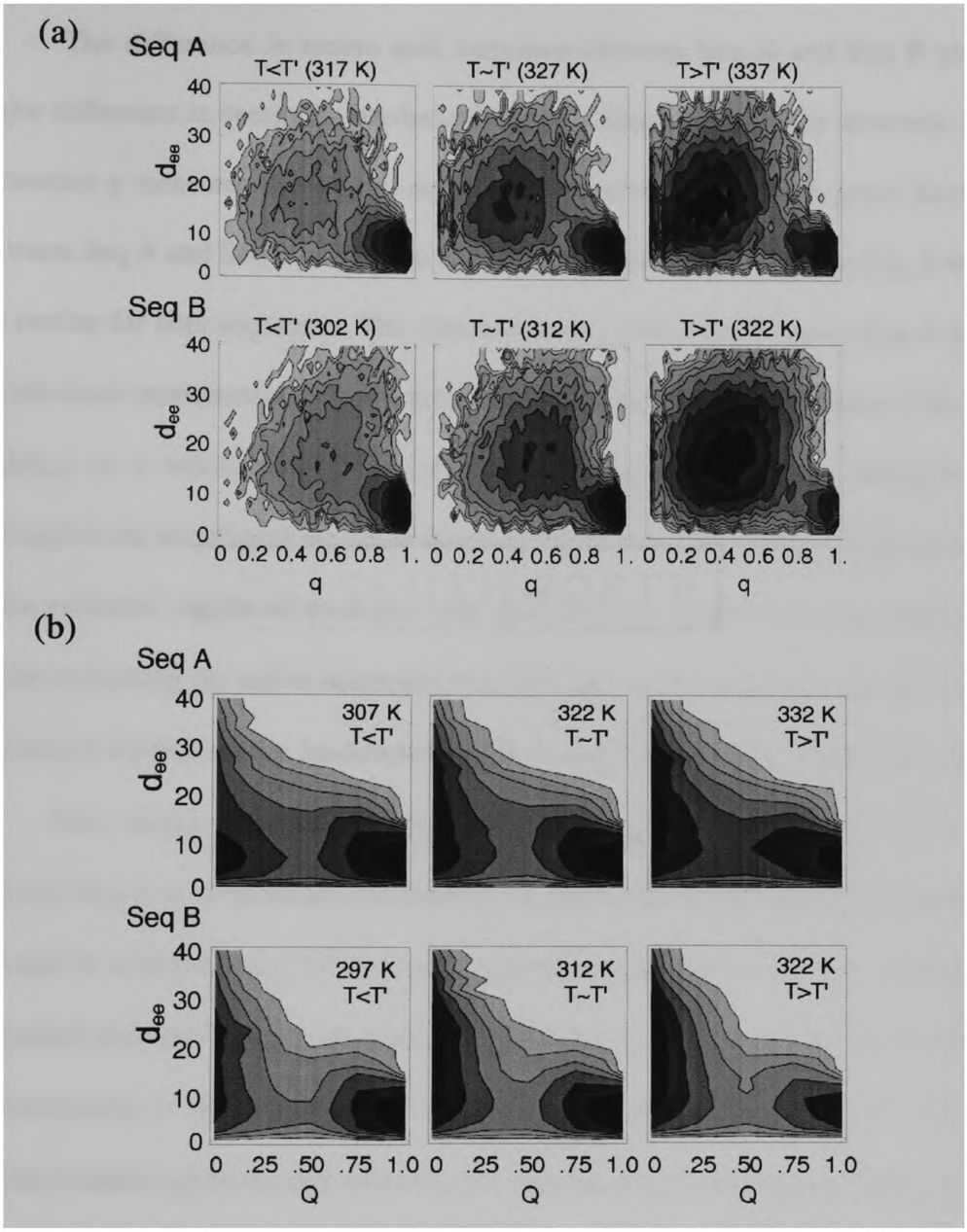


Fig. 6.4. Free energy contour plots for Seq A and Seq B. Darker shades correspond to lower free-energies (a) $F(q,d_{ee})$ as a function of q and d_{ee} (b) $F(Q,d_{ee})$ as a function of Q and d_{ee}. As the temperature is lowered, the non-native minimum (darkest shade) of Seq A remains as a deep trap and shrinks in extent but the non-native minimum of Seq B remains broad, but becomes shallower.

The difference in amino acid sequence between Seq A and Seq B produces a major difference in energy in interhelical contacts formed by tertiary structure. Since the parameter q measures secondary, not tertiary, structure, it is not a good discriminator between Seq A and Seq B and this is manifested in the contour plots in Fig. 6.4(a) which are similar for both sequences. The thermodynamic difference between Seq A and Seq B is seen more prominently in Fig. 6.4(b), where I display the contour plots of free energy, $F(Q,d_{ee})$, as a function of Q and d_{ee}. Q is a measure of tertiary structure and can distinguish the misaligned structure from the native structure. Here, the non-native basin is the extended region of contours with Q~0 that are darker than the contours of the barrier separating the native minimum from the non-native minimum. As the temperature is lowered, the free energy landscapes of Seq A and Seq B change in different ways.

Free energy is a combination of enthalpy and entropy, $F=E-TS$. The non-native basin for Seq A is more localized than the non-native basin for Seq B. This is because the non-native minimum F_{NN} of Seq A is enthalpic and due to the low entropy, well-organized, but misaligned trap conformation of Fig. 6.1(b). The low free energy is due predominantly to three non-native hydrophobic contacts which lower E. High entropy unfolded molten globules and random coils with large d_{ee} contribute at outer portions, but do not contribute to the deepest part of the non-native minimum. In contrast, the non-native minimum F_{NN} of Seq B is due mostly to high entropy unfolded molten globules and random coils with a large range of d_{ee}. The enthalpy is not low because there are fewer enthalpy lowering interactions. As T is lowered, contributions to lowering the free energy from high S, unfolded configurations become less important. This decreasing entropic importance is displayed in Fig. 6.4(b) for Seq B. As the temperature is lowered,

the non-native basin continues to cover as much phase-space, but becomes less deep (lighter shading): F_{NN}(322 K)–F_{NN}(297 K)= –1.35 Kcal/mol. Seq A displays the decreasing importance of high entropy unfolded states in a different fashion. As the temperature is lowered, the depth of the non-native minimum remains as deep because it is mostly due to enthalpic contributions from hydrophobic interaction energies in the misaligned trap state: F_{NN}(332 K)–F_{NN}(307 K)= –0.01 Kcal/mol. Unlike Seq B, as the temperature is lowered and high S states lose importance, the extent of the basin decreases in size because only the outer regions of the basin are due to high entropy, high d_{ee} states. At 307 K, the non-native basin has significantly shrunk as compared to its size at 332 K.

The heat capacity as a function of temperature was calculated using the Monte Carlo histogram technique. I used a very long run near the transition temperature that undergoes large energy fluctuations so that it samples a large region of configuration space. To confirm that this heat capacity curve is correct, multiple simulations at seven different temperatures, which bracket the transition temperature, were performed. This assures that the configuration space is accurately sampled at high and low energies. As shown in Fig. 6.5, both sequences exhibit a distinct peak, implying that both fold through a first-order-like phase transition. The transition temperature T' for Seq A (324 K) with traps is ~15K higher than the transition temperature of trap-free Seq B (310 K). Seq A with the traps has a higher T' and is more stable because its native state has four hydrophobic contacts between inter-helical residues whereas, in order to remove the kinetic traps, Seq B has only three.

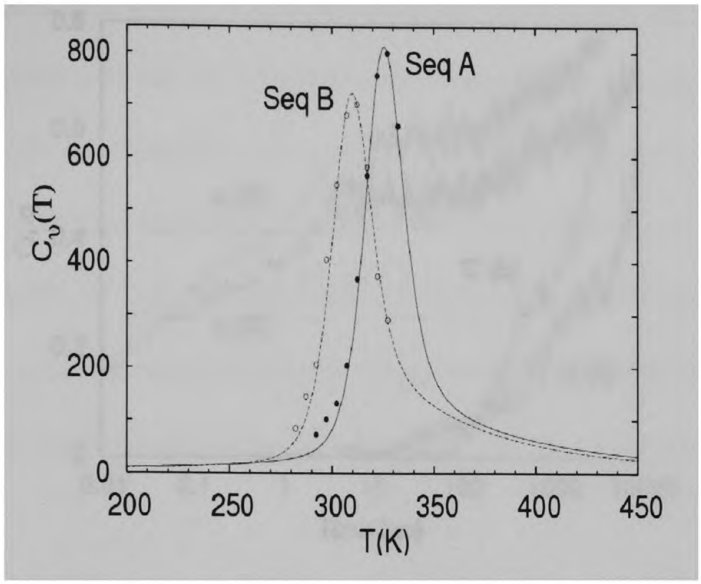


Fig. 6.5. Heat capacity curves as functions of temperature. The sharp peaks imply a first-order-like transition.

6.4 Protein Folding Kinetics

The free energy surfaces and their temperature dependencies have a direct effect on the folding kinetics and fast folding requires the avoidance of traps in the free energy landscape. The importance of traps is quantified by comparing folding times for trap-containing Seq A with Seq B which is designed to remove misfolded traps. Simulation time was converted to a reduced time with an estimate of each MC step as equivalent to 1 ns of real time. Figure 6.6 shows the time evolution of the parameters Q (tertiary) and q (secondary) for both sequences A and B at temperatures 322 K and 307 K respectively, slightly below their respective transition temperatures.

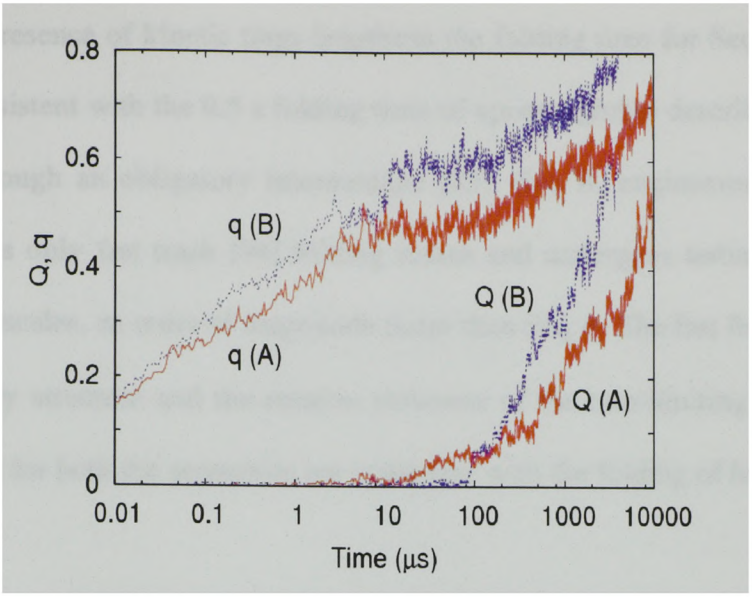


Fig. 6.6. Time evolution of the parameters q (helicity) and Q (fraction of native tertiary contacts) for Seq A and Seq B.

The parameter q measures the helical content of the chain but does not measure the relative positions of the two helices. Each of the trajectories are the ensemble averages of 50 independent runs that consisted of several folding and unfolding time sequences as in Fig. 6.3. The formation of secondary structure, q, occurs at the same rate for both Seq A and Seq B, and occurs at time scales hundreds of times faster than formations of tertiary structure, Q. The high q, low Q structure is an extended, high d_{ee} configuration consisting of native secondary helical segments that are not yet in tertiary contact. Quick formation of secondary structure is important for fast folding [92]. The secondary helical content begins to form extremely quickly, within a few nanoseconds, and reaches 50% for both sequences by 10 μs. This preorganized secondary structure must then wait for the folding process to be completed by the formation of the tertiary contacts between the helices. This waiting time differs significantly for Seq A compared

to Seq B. The presence of kinetic traps lengthens the folding time for Seq A to greater than 0.01 s, consistent with the 0.5 s folding time of apomyoglobin described in ref. 93, which folds through an obligatory intermediate [83]. Seq B, engineered not to have kinetic traps, has only fast track [94] folding routes and undergoes tertiary folding on millisecond timescales, an order of magnitude faster than Seq A. The fast formation of α-helical secondary structure and the relative slowness of the rate-limiting formation of tertiary structure for both the sequences are consistent with the folding of helical proteins [83, 95].

In Figs. 6.7, I show how the folding and unfolding times of Seq A and Seq B vary with temperature. The times given in Figs. 6.7 are median first passage times, MFPT. The folding MFPT is the time at which half of the simulations starting in the same random coil configuration (with different random number seeds) have folded. The error bars give a measure of the uncertainty in the MFPT calculated from the simulations. To calculate the error bars, the following analysis was performed. First, the 200 folding times were arranged in a random order. The first 40 folding times are removed and the MFPT of the remaining 160 was calculated. Then the first 40 times were re-inserted back into the list and the second group of 40 (41-80) was removed, and the MFPT for this second group of 160 times was calculated. The MFPT for each of the five groups of 160 runs were calculated. The average MFPT is then plotted with the standard deviation as the error bars.

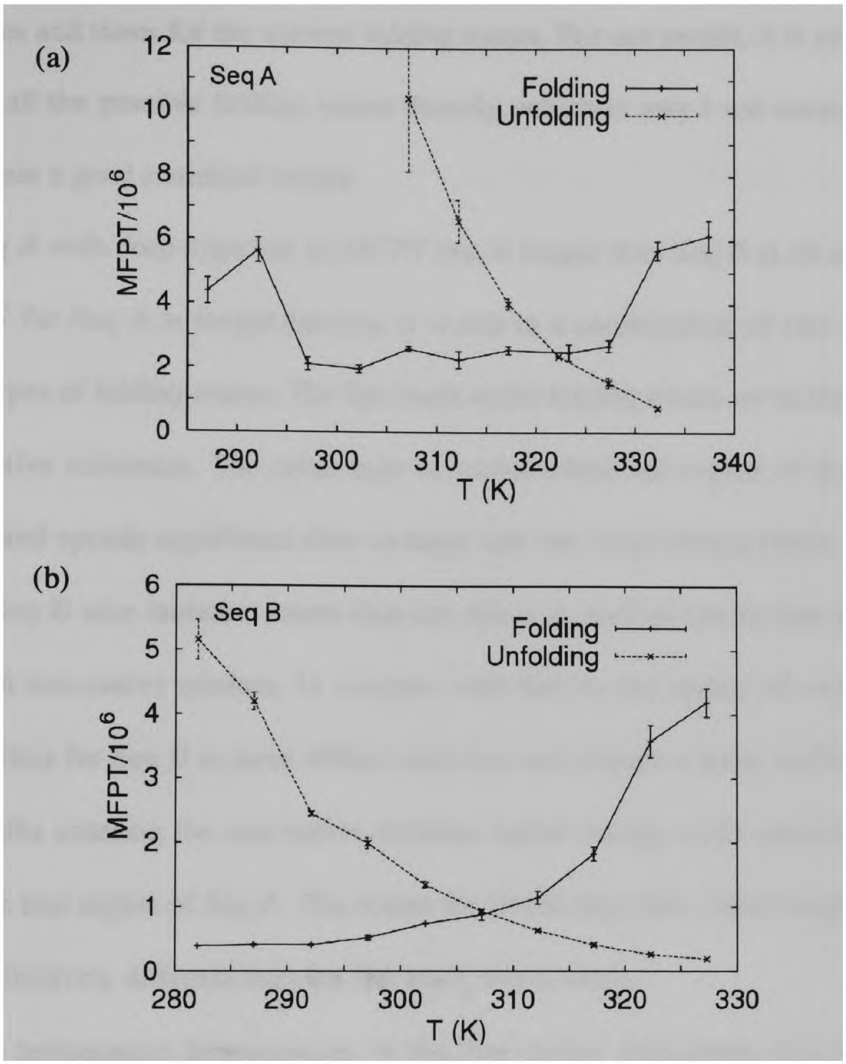


Fig. 6.7. Median first-passage times for folding and unfolding (a) Seq A and (b) Seq B.

Figure 6.7(a) for Seq A and Fig. 6.7(b) for Seq B display crucial differences in MFPT folding times that are the direct result of the differences in their free energy landscapes and their temperature dependencies discussed above. The extent of the non-native minimum affects the probability for a folding route to enter that region. The deeper the minimum, the longer the time that is required to leave the region and fold towards the native state. The detailed nature of the contours, such as gradients, affects both the

probabilities and times for the various folding routes. For our model, it is not possible to determine all the possible folding routes directly, which is why I use enough computer runs to obtain a good statistical sample.

Seq A with deep traps has an MFPT that is longer than Seq B at all temperatures. The MFPT for Seq A is longer because it is due to a combination of two qualitatively different types of folding routes. The fast-track quick folding routes avoid the region with the non-native minimum. The other type of routes enters the region of the non-native minimum and spends significant time in traps and has long folding times. The folding routes of Seq B also includes routes that are direct as well as routes that go through a region with non-native minima. In contrast with Seq A, the region of non-native free energy minima for Seq B is more diffuse and does not contain a deep, well-defined trap. Folding paths entering the non-native minima region escape more quickly than those entering the trap region of Seq A. The routes for Seq B may take a little longer to fold but are not qualitatively different than the fast-track direct routes.

The temperature dependencies of the free energy landscapes also have a direct effect on the temperature dependencies of the MFPT. As the temperature is lowered, the folding MFPT for Seq B decreases. That is because its non-native minimum becomes less deep and easier to leave. Table I shows that as the temperature is lowered from 327K down to 292 K, a temperature drop of approximately 10%, F_{NN} drops by 40%. Thus, the ease and rate of escape from the non-native minimum, which depends on $\exp(-F_{NN}/RT)$, increases which lowers the escape time and decreases MFPT. For Seq A, as the temperature is lowered, the non-native free energy minimum remains at approximately the same depth, as can be seen in Table I. This causes longer delays in the traps at low

temperatures. However, the region of the non-native minimum also shrinks in extent, which decreases the probability for a folding route to fall into the trap. At mid-temperatures, these two effects tend to cancel and produce an MFPT that is independent of T. As T is further lowered, the extent of the region of the minimum stops shrinking and the constant depth makes it harder to leave as T is lowered. This causes the MFPT to increase at the lower temperatures in Fig. 6.7(a).

TABLE I. Free energies calculated using $F(Q,d_{ee})= -kT \ln P(Q,d_{ee})$ for the non-native (*NN*) minimum, barrier (*B*), and native state minimum (*N*) for the contour plots for Seq A and Seq B.

T(K)	T(Kcal/mol)	F_{NN}	F_B	F_N
		Seq A		
302	0.60	–1.88	0	–3.11
307	0.61	–1.87	0	–2.32
312	0.62	–1.85	0	–2.03
317	0.63	–1.71	0	–2.48
322	0.64	–2.00	0	–2.39
327	0.65	–1.42	0	–1.08
332	0.66	–1.88	0	–1.20
		Seq B		
292	0.58	–2.28	0	–3.84
297	0.59	–2.42	0	–3.35
302	0.60	–2.33	0	–3.10
307	0.61	–2.76	0	–3.26
312	0.62	–2.95	0	–2.77
317	0.63	–3.27	0	–2.60
322	0.64	–3.77	0	–2.23
327	0.65	–3.89	0	–2.04

6.4.1 Fast and slow folding routes

In order to more precisely relate the free energy landscapes to kinetic pathways and quantify the difference between trap-free fast folding pathways versus pathways with traps, characteristic folding times (τ) were calculated by fitting survival probability curves S at each temperature

$$S(t,T) = 1 - n(t,T) \tag{6.1}$$

where $n(t,T)$ is the fraction of the runs that have "succeeded" by time step t at temperature T. For folding simulations that start in a random coil unfolded configuration, "succeeding" means that they have folded based upon the criteria that Q=0.75 or higher. For simulations investigating unfolding, the chain is initialized to be in the folded configuration and succeeding means unfolding to Q=0.25 or less. Fig. 6.8(a) and 6.8(b) show representative curves of survival probability of the folding runs (probability to remain in the unfolded state) for Seq A and Seq B respectively. The probability to remain in the unfolded state decreases from 1 when all the runs are in the unfolded state to zero when all the runs find their native states. For Seq A, the folding rate is fastest at T' and becomes slower for both $T>T'$ and $T<T'$. For Seq B, there is a consistent order of faster to slower as the temperature is increased.

Sine the kinetics are complicated due to the presence of traps in the case of Seq A, it was found that the curves could not be fit well with a single exponential function. For this sequence, the protein folding kinetics have two qualitatively different dominant pathways: one direct route which leads the unfolded chain into the native state without going through the traps (route 1) and the other that goes through the kinetic traps created by the relatively stable misaligned structures (route 2).

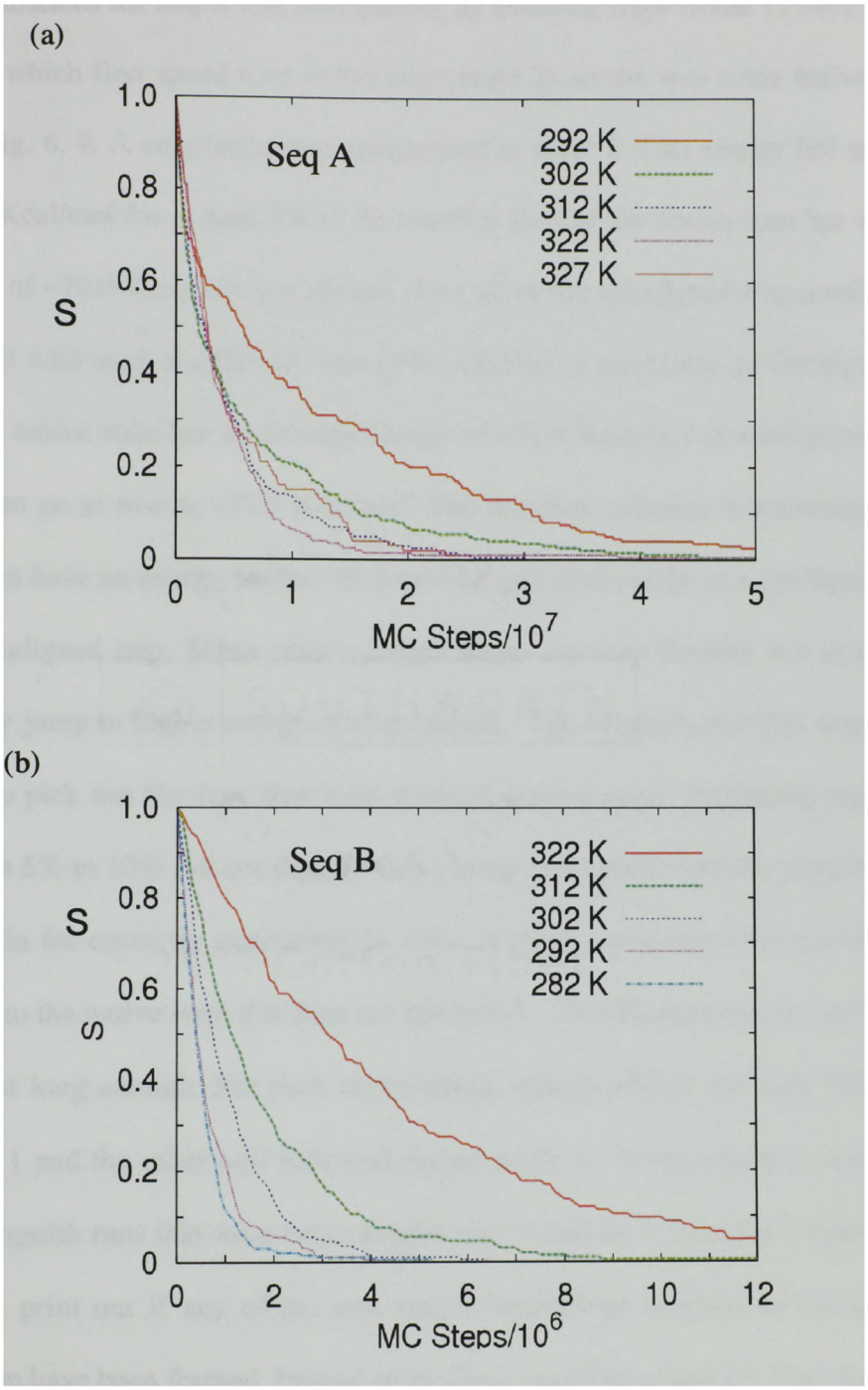


Fig. 6.8. Survival curves for folding runs (probability of the chain to remain in the unfolded state) for (a) Seq A and (b) Seq B.

Simulations for Seq A that fold quickly by avoiding traps (route 1) were separated from those which first spend time in the trap (route 2) on the way o the native state, as shown in Fig. 6. 9. A simulation was categorized as route 2 if its energy fell and stayed below –70 Kcal/mol for at least 5% of the stability time of the native state but with Q=0. The energy of –70.0 Kcal/mol was chosen since all of the misaligned trap conformations are captured with such a criterion. The Q=0 criterion is necessary to distinguish traps because the native state has an average energy of –71.4 Kcal/mol at a temperature close to T' and can go as low as –77.8 Kcal/mol. The duration criterion is necessary because the chain can have an energy below –70 Kcal/mol and Q=0 but be in a configuration that is not a misaligned trap. These other configurations are very flexible and unstable and very quickly jump to higher energy configurations. The duration criterion was robust in its ability to pick out the runs that went through kinetic traps. Increasing the duration criteria from 5% to 10% did not significantly change the results and this supports the use of the criteria for correctly separating the runs. A run is considered to have followed a direct route to the native state if it does not fall below –70.0 Kcal/mol or in case it does, it does not last long enough. For each temperature, almost half of the runs followed the faster route 1 and the other half followed slower route 2. In retrospect, a more precise way to distinguish runs that went into a kinetic trap would be to have the computer model periodically print out if any of the non-native interhelical contacts of the misaligned configuration have been formed. Instead of Q, these could be called Q'. The chain will be in the misaligned configuration if Q'=3, or possibly even if Q'=2.

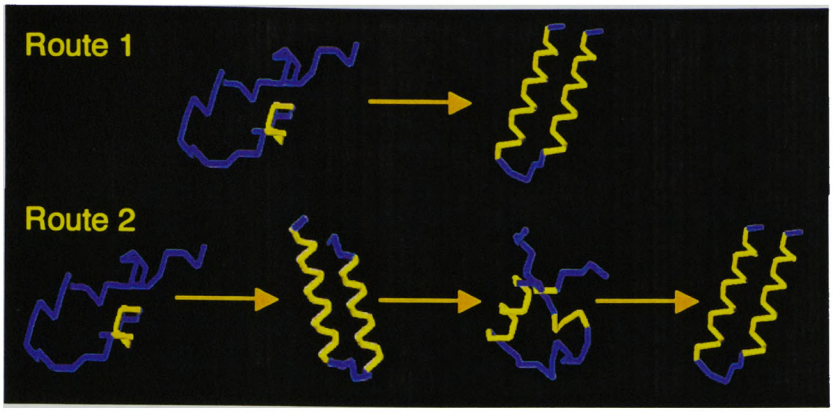


Fig. 6.9. Schematics of fast folding route 1 and the slow folding route 2. The formation of α-helical secondary structure is highlighted.

Once the folding routes for Seq A have been separated, the survival functions for route 1 and route 2 separately can each be well fit by single exponential functions (not shown). At a given temperature, the characteristic folding times, τ_1 and τ_2 determined from the separate fits to route 1 and route 2 were then used to fit the survival function from the set of all the runs, regardless of their paths, with a double exponential function, $S(t,T)= a_1 e^{-t/\tau_1} + a_2 e^{-t/\tau_2}$. The survival function at each temperature is always well fit by a double exponential function which is used to get the prefactors a_1 and a_2. As an example, the fit for the survival function at 307 K (below T') is displayed in Fig. 6.10. The survival functions from the unfolding runs fit well with single exponential functions. Table II summarizes the fit parameters. The shorter folding time scales are due to the runs following route 1 and the longer time scales are from the runs that followed route 2. The temperature dependence of both the folding and the unfolding characteristic times from the fits for Seq A are displayed in Fig. 6.11(a).

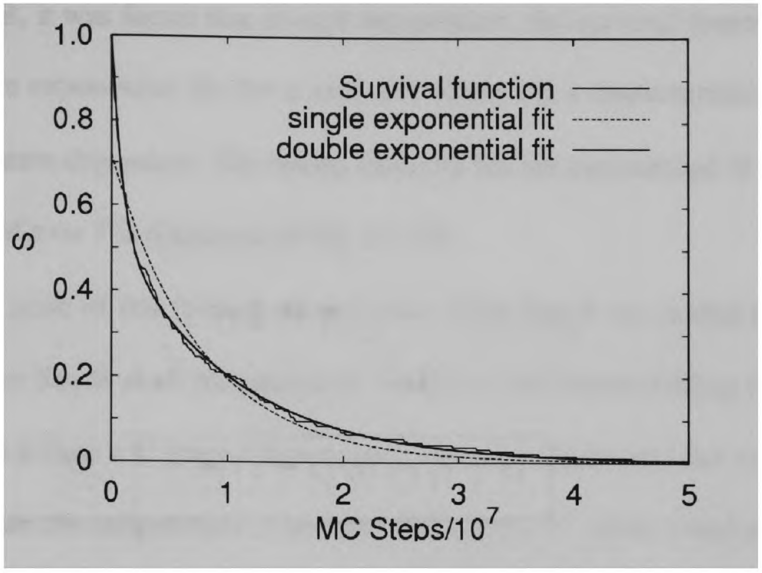


Fig. 6.10. Representative fits with single and double exponential functions to the survival function from the computer simulations for Seq A.

TABLE II. Fit parameters from the fitting of survival functions of Fig. 6.8(a) with a double exponential function $S = a_1 e^{-t/\tau_1} + a_2 e^{-t/\tau_2}$. τ_1 and τ_2 are first determined from the separate single exponential fits to route 1 and route 2, and then a_1 and a_2 are determined from the double exponential fit to S of all the combined runs of Fig. 6.8(a). N_1 is the number of routes, out of a total of 200 runs, following route 1and N_2 is the number of remaining runs following route 2. The characteristic times are given in MC steps.

T(K)	N_1	a_1	τ_1 (MC)	N_2	a_2	τ_2 (MC)
297	96	0.99	6.08×10^5	104	1.02	1.13×10^7
302	106	1.02	8.85×10^5	94	1.14	1.02×10^7
307	93	1.01	8.86×10^5	107	1.42	5.72×10^6
312	95	1.01	9.49×10^5	105	1.24	6.12×10^6
317	95	0.99	1.09×10^6	105	1.38	4.69×10^6
322	99	1.04	1.41×10^6	101	1.56	4.25×10^6
327	108	1.01	1.91×10^6	92	1.10	7.72×10^6

For Seq B, it was found that at each temperature, the survival function can be fit nicely by a single exponential $S(t,T)= a \exp(-t/\tau)$ where τ is a characteristic folding time which is temperature dependent. The results from the fits are summarized in Table III and the dependence of τ on T is displayed in Fig. 6.11(b).

The time scale of fast folding routes (route 1) for Seq A are similar to the folding times for trap-free Seq B at all temperatures. However, the slower folding time scales of route 2 for Seq A follow a U shaped dependence with the temperature due to the presence of the traps. When the temperature is lowered from above T', route 1 and route 2 of Seq A, and Seq B display a decrease in folding time. Unlike route 1 and Seq B, at lower temperatures, route 2 has an increase in folding time because of increased difficulty in climbing out of the kinetic trap.

TABLE III Characteristic folding and unfolding times from single exponential fits for the survival curves in Fig. 6.8(b) (Seq B)

T (K)	τ-fold (MC)	τ-unfold (MC)
282	5.07×10^5	8.32×10^6
287	6.00×10^5	5.72×10^6
292	6.17×10^5	3.67×10^6
297	7.71×10^5	2.84×10^6
302	1.00×10^6	1.85×10^6
307	1.27×10^6	1.48×10^6
312	1.67×10^6	9.24×10^5
317	2.80×10^6	5.64×10^5
322	4.31×10^6	3.82×10^5
327	6.11×10^6	2.54×10^5

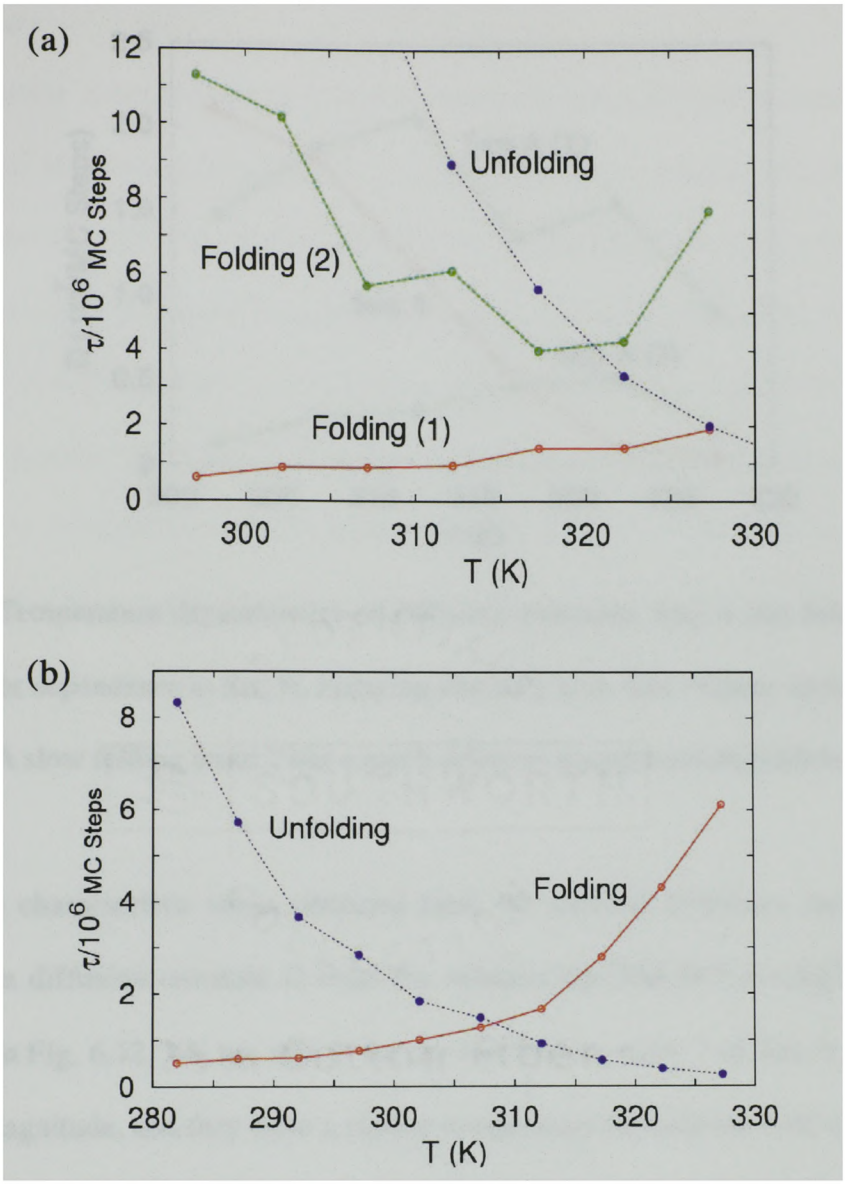


Fig. 6.11. Characteristic folding and unfolding times for (a) Seq A and (b) Seq B as a function of temperature determined by fits to the survival function $S(t,T)$. At each temperature Seq A folding is best described by the combination of two exponential processes whereas Seq B folding is best described by a single exponential function at each temperature. Unfolding is well described by a single exponential at all temperatures for both Seq A and Seq B.

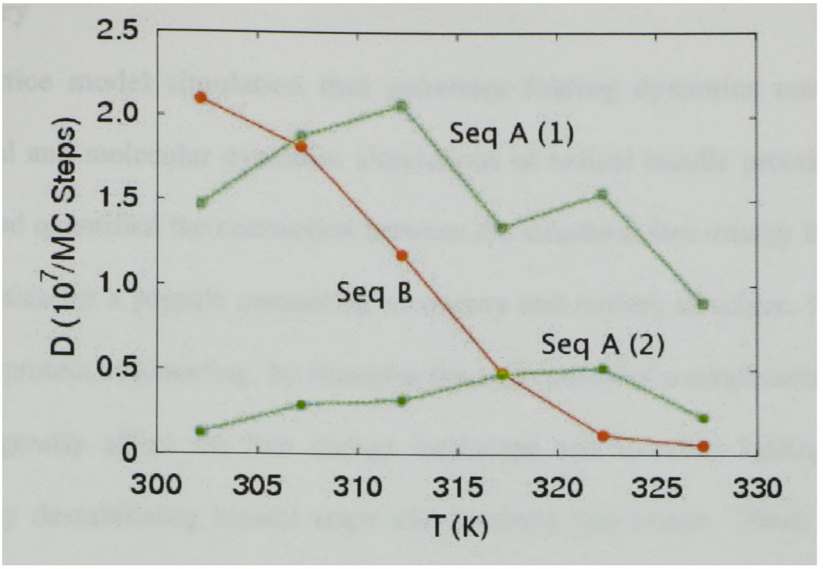


Fig. 6.12. Temperature dependencies of diffusion constants. Seq A fast folding route 1 has a similar dependence to Seq B, implying that they both fold without spending time in traps. Seq A slow folding route 2 has a much different temperature dependence.

The characteristic times obtained from the survival functions can be used to determine a diffusion constant D from the relationship [96] $D(T)=<\Delta Q^2>/\tau$ and are displayed in Fig. 6.12. We see that D for the fast folding route 1 of Seq A is similar to Seq B in magnitude, and they have a similar temperature dependence with a temperature offset due to the difference in their T'. D_{A1} and D_B both decrease as T increases because the unfolded free energy minimum becomes deeper because it is due to high entropy configurations. Route 2 of Seq A has a significantly smaller D at all temperatures. Unlike D_{A1} and D_B, D_{A2} increases with temperature because as the temperature increases, it becomes easier to get out of the kinetic trap whose free energy minimum is predominantly enthalpic and remains constant.

6.5 Summary

Using a lattice model simulation that generates folding dynamics consistent with experimental and molecular dynamics simulations of helical bundle proteins [83, 95] I explained and quantified the connection between the structural free-energy landscape and folding kinetics for a peptide containing secondary and tertiary structure. I have shown how simple protein engineering, by changing the hydropathy of a small number of amino acids, can greatly affect the free energy landscape and increase folding speed and reliability by destabilizing kinetic traps and favoring fast tracks. These results have general applicability in protein engineering as well as for understanding the mechanisms of protein folding.

CHAPTER VII

CONCLUSIONS AND FUTURE DIRECTIONS

The complexity of the Protein Folding Problem has attracted researchers from a variety of fields and it has been one of the most challenging problems in molecular biology for decades. Progress has been made during recent years, both experimentally and theoretically. Since the function of a protein is intimately related to its three dimensional structure, which is determined by its primary sequence, an insight into the sequence-structure relation will allow the design of novel proteins that are engineered to carry out specific molecular tasks. Thus, a deeper understanding of the Protein Folding Problem will have a tremendous impact on on biomedicine and nanotechnology. In this dissertation, I focused on quantifying how various changes in the primary sequence of amino acids can greatly affect the dynamics of folding. These substitutions changed either the size of the affected amino acid or its hydropathy. I also explored various statistical mechanical and thermodynamic theoretical techniques to quantify the effects of the amino acid substitutions on the folding process.

I investigated the excluded volume entropic effects on the native state stability of a simulated four-helix bundle protein. The lattice model I used allowed me to systematically change the chain thickness in two ways: by changing the size of the sidechain and by changing the soft-core repulsion part of the backbone interaction which is equivalent to changing the size of the backbone. The results show that increasing the size of the sidechains in the water-exposed surface significantly increases the relative stability of the native state compared to the biologically useless molten globule. In the

case of the four-helix bundle, the effect of the excluded volume is more important than the hydrophilicity of the water-exposed sidechains. This result suggests that a strategic placement of residues with sidechains of particular sizes may be an important consideration in protein engineering. Relevant to the synthetic polymers, the results also show that increasing backbone size increases the relative stability of the native state. Also, the unfolding transition changes its nature as a function of excluded volume. The greater the excluded volume, the sharper is the heat capacity peak at the transition temperature which implies that the structural unfolding transition becomes more like a first order thermodynamic phase transition in an infinite system.

Another investigation dealt with the effect of changing the strength of the hydrophobic interaction on the stability of the native state, and the nature of the structural transition. I showed that even in a protein, it is possible to choose a valid repeating unit that allows the finite size protein folding or unfolding structural transition to be analyzed as a phase transition. This allows the nature of the transition to be studied using finite size scaling theory. Finite size scaling on unfolding dynamics of a model four-helix bundle shows that the order of the phase transition depends on the hydrophobic interaction strength. For relatively weak (but strong enough to assure native state stability) hydrophobic strength in the core, unfolding is a first-order-like transition. Unfolding is shifted from first-order towards a higher-order, more continuous transition when the hydrophobic interaction strength is significantly increased. Once folded, the increase in the hydrophobic strength increases the stability of the native state and the unfolding transition temperature is raised. However, the relative stability of the molten globule compared to the native state is also increased due to non-native hydrophobic contacts.

During the folding process, a chain that collapses to the molten globule may not be able to leave to continue folding to the native state. This could make folding less reliable and therefore substituting amino acids in the core with other amino acids that are significantly more hydrophobic is not preferable. A small increase in the hydrophobicity of the core residues may be helpful to increase the stability of the native state.

The dissertation culminated in a chapter describing my research on the folding of a model α-helical hairpin peptide. The α-helical hairpin peptide forms a two-helix bundle in the native state and offers an important model system for studying the kinetics and thermodynamics of helical proteins. I showed how simple protein engineering, by selective substitution of a small number of amino acids, can enhance the folding process by making it faster and more stable. I investigated this on a fundamental level by showing how the enhanced folding can be explained in terms of the effect of the substitutions on the free energy landscape. I showed how the substitutions remove kinetic traps in the free energy landscape and how this increases the speed and reliability of the folding. The results from the analysis have general applicability in the design and engineering of helical proteins.

Having used a variety of statistical mechanical and thermodynamic techniques in the work presented in this dissertation, the direction for future work is clear. I showed that there are several ways that protein folding can be enhanced by strategic substitution of amino acids. I quantified these important effects in chapters IV and V by calculating heat capacities, correlation functions, critical exponents, and Binder Cumulants. In chapter VI, I investigated the dynamics on a deeper level by calculating the free energy landscape. At this level, I was able to explain and predict the changes in various

thermodynamic and statistical mechanical parameters that describe the dynamics of the folding process. The next step is to go to the deepest possible level of understanding. This is to enumerate the states of a protein as a function of energy and, equally important, to describe their structural connectivity, i.e. their relative ease of interchange. The structural connectivity of microscopic configurations is where the complexity of the system is manifested on the most fundamental level.

This research might be carried out as follows. The computer model can be used to determine how many different configurations (given by the model as *R*-state sequences) and their energies are visited at a given temperature. Changes of configuration in the computer model require specific structural moves that connect the configurations and prevent unphysical structural transitions. The allowed moves have attempt probabilities that are weighted by the number of amino acids that participate and therefore impose statistical mechanical constraints. The model also imposes thermodynamic constraints through the Metropolis test. These combine to provide the visitation probability with the ability to act as a measure of the connectivity, though other measures may be found that are more valuable. The visitation probability is the number of times a specific configuration is visited during a run. A configuration can be defined by its sequence of *R*-States. It is also necessary to categorize each structure in terms of its secondary and tertiary structure and how close it is to the native state. Categorizing a configuration by its energy is important but not sufficient because of degeneracy; i.e. several configurations may have the same energy. Keeping track of the energies of the different states is important because it will allow the calculation of the average energy as a function of temperature. The counting of the microscopic states by their configuration

and their visitation probabilities are important because they will allow the calculation of the entropy as a function of temperature and average energy. The calculation of entropy and average energy as a function of temperature and structural organization will allow the calculation of free energy, and all other statistical and thermodynamic properties.

The amount of information to describe a single configuration is large, and quantifying the connectivity in a multi-dimensional hyperspace will also be large. Therefore, investigations will be theoretically and computationally limited initially to small chains. If the states and their connectivity can be enumerated, all aspects of the dynamics can be calculated, such as entropy, free energy, kinetic rates of folding, order of phase transitions, heat capacities, etc. The calculation of entropy and free energy will allow the determination of whether there is a preferred route through configuration space that the chain will always follow to a well-defined free energy minimum. This is a characteristic of folding to a native state. This will lead to the ability to predict if a primary sequence of amino acids has a native state and what it looks like. This will also give the ability to predict how a change in the amino acids of the primary sequence will change the free energy landscape, and the effect that it may have on the shape and stability of the native state. These predictive abilities, based upon the underlying physics of the amino acid chain, will then solve the Protein Folding Problem.

REFERENCES

1. Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wycoff, H. W., and Phillips, D. C. A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature* **181**, 662-666 (1958).

2. Schrodinger, E. *What Is Life?* Cambridge University Press, Cambridge, England (1944).

3. Branden, C. and Tooze, J. *Introduction to protein structure.* Garland Publishing, New York and London (1991).

4. Engel, D. E. and DeGrado, W. F. Amino acid propensities are position-dependent throughout the length of a-helices. *J. Mol. Biol.* **337**, 1195–1205 (2004).

5. Hubbard, T. J. P., Ailey, B., Brenner, S. E., Murzin, A. G., and Chothia, C. SCOP: a structural classification of proteins database. *Nucleic Acids Res.* **27**, 254-256 (1999).

6. Holm, L. and Sander, C. Protein folds and families: sequence and structure alignments. *Nucleic Acids Res.* **27**, 244-247 (1999).

7. Orengo, C. A., Jones D. T. and Thornton, J. M. Protein superfamilies and domain superfolds. *Nature* **372**, 631-634 (1994).

8. Sali, A. 100,000 protein structures for the biologist. *Nat. Struct. Biol.* 5, 1029-1032 (1998).

9. Terwilliger, T. C. Waldo, G., Peat, T. S., Newman, J. M., Chu, K., and Berendzen, J. Class-directed structure determination: Foundation for a Protein Structure Initiative. *Protein Sci.* **7**, 1851-1856 (1998).

10. Monod, J., Wyman, J., and Changeux, J. -P. (1965). On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.* **12**, 88-118 (1965).

11. Perutz, M. F. Stereochemistry of cooperative effects of hemoglobin. *Nature* **228**, 726-739 (1970).

12. Dobson, C. M. and Karplus, M. The fundamentals of protein folding: bringing together theory and experiment. *Curr. Opin. Strcut. Biol.* **9**, 92-101 (1999).

13. Alm, E. and Baker D. Matching throry and experiment in protein folding. *Curr. Opini. Struct. Biol.* **9**, 189-196 (1999).

14. Eaton, W. A, Muñoz, V., Hagen, S. J., Jas, G.S., Lapidus, L. J., Henry, E. R., and Hofrichter, J. Fast kinetics and mechanism in protein folding. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 327–59 (2000).

15. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **181**, 223-230 (1973).

16. Levinthal, C. Are there pathways for protein folding? *J. Chem. Phys.* **65**, 44-45 (1968).

17. Wetlaufer, D. B. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl. Acad. Sci. U.S.A.* **70**, 697–701 (1973).

18. Ptitsyn, O. B. and Rashin, A. A. A model of myoglobin self-organization. *Biophys. Chem.* **3**, 1-20 (1975).

19. Kim, P. S. and Baldwin, R. L. Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. *Annu. Rev. Biochem.* **51**, 459-489 (1982).

20. Muñoz, V. and Serrano, L. Local versus nonlocal interactions in protein folding and stability—an experimentalist's point of view. *Fold. Des.* **1**, R71–77 (1996).

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

22. Dill, K. A., Bromberg, S., Yue, K. Z., Fiebig, K. M., Yee, D. P., Thomas, P. D., and Chan, H. S. Principles of protein folding - A perspective from simple exact models. *Protein Sci.* **4**, 561-602 (1995).

23. Bychkova, V. E. and Ptitsyn, O. B. The molten globule state of protein molecules is becoming a rule rather than exception. *Biophysics* **38**, 58-66 (1993).

24. Fersht, A. R. Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* **7**, 3–9 (1997).

25. Fersht, A. R. Optimization of rates of protein folding: the nucleation-condensation mechanism and its applications. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 10869–10873 (1995).

26. Martinez, J. C. and Serrano, L. The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nat. Struct. Biol.* **6**, 1010-1016 (1999).

27. Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Alm, E., Ruczinski, I., and Baker, D. Experiment and theory highlight role of native state topology in SH3 folding. *Nat. Struct. Biol.* **6**, 1016-1024 (1999).

28. Matouschek, A., Kellis, J. T., Serrano, L., and Fersht, A. R. (1989) Mapping the transition state and pathway of protein folding by protein engineering. *Nature* **340**, 122-126 (1989).

29. Serrano, L., Matouschek, A., and Fersht, A. R. The folding of an enzyme III. Structure of the transition state for the unfolding of barnase analysed by a protein engineering procedure *J. Mol. Biol.* **224**, 805–818 (1992).

30. Fersht, A. R., Itzhaki, L. S., ElMasry, N. F., and Matthews, J. M. Single versus parallel pathways of protein folding and fractional formation of structure in the transition state. *Proc. Natl. Acad. Sci. U.S.A.* **91**, 10426–10429 (1994).

31. Dagget, V. and Fersht, A. R. Is there a unifying mechanism for protein folding. *Trends Biochem. Sci.* **28**, 19-26 (2003).

32. Pande, V. S., Grosberg, A., Tanaka, T. and Rokhsar, D. S. Pathways for protein folding: Is a new view needed? *Curr. Opin. Struct. Biol.* **8**, 68–79 (1998).

33. Bryngelson, J. D., Onuchic, J. N., Socci, N. D., and Wolynes, P. G. Funnels, pathways, and the energy landscape of protein folding: a synthesis *Proteins Struct. Funct. Genet.* **21**, 167–195 (1995).

34. Wolynes, P. G., Onuchic, J. N., and Thirumalai, D. Navigating the folding routes. *Science* **267**, 1619–1620 (1995).

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

36. Onuchic, J. N., Wolynes, P. G., Luthey-Schulten, Z., and Socci, N. D. Toward an outline of the topology of a realistic protein folding funnel. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 3626 –3630 (1995).

37. Reif, F. *Fundamentals of Statistical and Thermal Physics.* McGraw-Hill Book Co. Singapore (1985).

38. Ferrenberg, A. M. and Swendsen, R. H. New Monte Carlo technique for studying phase transitions. *Phys. Rev. Lett.* **61**, 2635-2638 (1988).

39. Ferrenberg, A. M. and Swendsen, R. H. Optimized Monte Carlo data analysis. *Phys. Rev. Lett.* **63**, 1195-1998 (1989).

40. Socci, N. D. and Onuchic, J. N. Kinetic and thermodynamic analysis of proteinlike heteropolymers: Monte Carlo histogram technique. *J. Chem. Phys.* **103**, 4732-4744 (1995).

41. Skolnick, J. and Kolinski, A. Dynamic Monte Carlo simulations of a new lattice model of globular protein folding, structure and dynamics. *J. Mol. Biol.* **221**, 499-531 (1991).

42. Skolnick, J. and Kolinski, A. Simulations of the folding of a globular protein. *Science* **250**, 1121-1125 (1990).

43. Gerstman, B. S. and Garbourg, Y. Structural information content and Lyapunov exponent calculation in protein unfolding. *J. Polym. Sci.*, Part B: Polym. Phys. **36**, 2761-2769 (1997).

44. Chapagain, P. P. and Gerstman, B. S. Finite size scaling of structural transitions in a simulated protein with secondary and tertiary structure. *J. Chem. Phys.* **119**, 1174-1180 (2003).

45. Chapagain, P. P. and Gerstman, B. S. Excluded volume entropic effects on protein unfolding times and intermediary stability. *J. Chem. Phys.* **120**, 2475-2481 (2004).

46. Abagyan, R. and Totrov, M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.* **235**, 983-1002 (1994).

47. Kuhlman, B., Dantas, G, Ireton, G. C., Varani, G., Barry, L., Stoddard, B. L., and Baker, D. Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364-1368(2003).

48. Shimada, J. and Shakhnovich, E. I. The ensemble folding kinetics of protein G from an all-atom Monte Carlo simulation. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 11175-11180 (2002).

49. Sorin, E. J. and Pande, V. S. Exploring the helix-coil transition via all-atom equilibrium ensemble simulations. *Biophys. J.* **88**, 2472–2493 (2005).

50. Shen, M. and Freed, K. F. All-atom fast protein folding simulations: The villin headpiece. *Proteins* **49**, 439-445 (2002).

51. Abkevich, V. I., Gutin, A. M., and Shakhnovich, E. I. Improved design of stable and fast-folding model proteins. *Fold Des.* **1**, 221-230 (1996).

52. Dill, K. A. and Chan H. S. From Levinthal to pathways to funnels. *Nat. Struct. Biol.* **4**, 10-19 (1997).

53. Dinner, A. R., Sali, A., and Karplus, M. The folding mechanism of larger proteins: role of native structure. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 8356–8361 (1996).

54. Muñoz, V. and Eaton, W. A. A simple model for calculating the kinetics of protein folding from three-dimensional structures *Proc. Natl. Acad. Sci. U.S.A.* **96**, 11311–11316 (1999).

55. Kolinski, A., Milik, M., and Skolnick, J. Static and dynamic properties of a new lattice model of polypeptide chains. *J. Chem. Phys.* **94**, 3978-3985 (1991).

56. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087- 1092 (1953).

57. Newman, M. E. J. and Barkema, G. T. *Monte Carlo Methods in Statistical Physics*, Oxford University Press, New York (1999).

58. Thirumalai, D. From minimal models to proteins: time scales for protein folding kinetics. *J. Phys.* I France **5**, 1457-1467 (1995).

59. Minton, A. P. The influence of macromolecular crowding and macromolecular confinement on biochemical reactions in physiological media. *J. Biol. Chem.* **276**, 10577-10580 (2001).

60. Li, J., Zhang, S., and Wang, C. -C. Effects of macromolecular crowding on the refolding of glucose-6-phosphate dehydrogenase and protein disulfide isomerase. *J. Biol. Chem.* **276**, 34396-34401 (2001).

61. Minton, A. P. Effect of a concentrated "inert" macromolecular cosolute on the stability of a globular protein with respect to denaturation by heat and by Chaotropes: A statistical-thermodynamic Model. *Biophys. J.* **78**, 101-109 (2000).

62. Ellis, R. J. Macromolecular crowding: an important but neglected aspect of the intracellular environment. *Curr. Opin. Struct. Biol.* **11**, 114-119 (2001); R. J. Ellis, R. J. and A. P. Minton, A. P. Join the Crowd. *Nature* **425**, 27-28 (2003).

63. Ping, G., Yuan, J. M., Vallieres, M., Dong, H., Sun, Z., Wei, Y., Li, F. Y., and Lin, S. H. Effects of confinement on protein folding and protein stability. *J. Chem. Phys.* **118**, 8042-8048 (2003).

64. van den Berg, B., Wain, R., Dobson, C. M., and Ellis, R. J. Macromolecular

crowding perturbs protein refolding kinetics: implications for folding inside the cell. *EMBO J.* **19**, 3870-3875 (2000).

65. Chan, H. S. and Dill, K. A. Solvation: Effects of molecular size and shape. *J. Chem. Phys.* **101**, 7007-7026 (1994).

66. Box, G. E. P and Jenkins, G. M. *Time Series Analysis*, Holden-Day Series in Time Series Analysis, Holden-Day, San Francisco (1970).

67. Socci, N. D., Onuchic, J. N., and Wolynes, P. G. Diffusive dynamics of the reaction coordinate for protein folding funnels. *J. Chem. Phys.* **104**, 5860-5868 (1996).

68. Hansmann, U. H. E. and Okamato, Y. Finite-size scaling of helix-coil transitions in poly-alanine studied by multicanonical simulations. *J. Chem. Phys.* **110**, 1267-1276 (1999).

69. Alves, N. A. and Hansmann, U. H. E. Partition function zeros and finite size scaling of helix-coil transitions in a polypeptide. *Phys. Rev. Lett.* **84**, 1836-1839 (2000).

70. Kolinski, A. and Madziar, P. Collapse transitions in protein-like lattice polymers. The effect of sequence patterns. *Biopolymers* **42**, 537-548 (1997).

71. Pande, V. S., Grosberg, A. Yu., and Tanaka, T. thermodynamics of the coil to frozen globule transition in heteropolymers. *J. Chem. Phys.* **107**, 5118-5124 (1997).

72. Yeomans, *J.M. Statistical Mechanics of Phase Transitions*, Oxford University Press, New York (1992).

73. Stanley, H.E. *Introduction to Phase Transitions and Critical Phenomena*, Oxford University Press, New York (1971).

74. R. K. Pathria, *Statistical Mechanics,* Butterworth-Heinemann, Oxford (1996).

75. Challa, M. S. S., Landau, D. P., and Binder, K. Finite-size effects at temperature-driven first-order transitions. *Phys. Rev. B* **34**, 1841-1852 (1986).

76. Alves, N. A., Berg, B. A., and Villanova, R. Potts models: Density of states and mass gap from Monte Carlo calculations. *Phys. Rev. B* **43**, 5846-5856 (1991).

77. Imry, Y. Finite-size rounding of a first-order phase transition. *Phys. Rev. B* **21**, 2042-2043 (1980).

78. Landau, D. P. and Binder, K. A *Guide to Monte Carlo Simulations in Statistical Physics*, Cambridge University Press, Cambridge, UK (2000).

79. Mirny, L. A., Abkevich, V. I., and Shakhnovich, E. I. How evolution makes proteins fold quickly. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 4976-4981 (1998).

80. Shakhnovich, E. I. and Gutin, A. M. Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl. Acad. Sci. U.S.A.* **90**, 7195-7199 (1993).

81. Scalley-Kim, M. and Baker, D. Characterization of the folding energy landscapes of computer generated proteins suggests high folding free energy barriers and cooperativity may be consequences of natural selection. *J. Mol. Biol.* **338**, 573-583 (2004).

82. Vendruscolo, M., Paci, E., Dobson, C. M., and Karplus, M. Three key residues form a critical contact network in a protein folding transition state. *Nature* **409**, 641-645 (2001).

83. Zhou, Y. and Karplus, M. Interpreting the folding kinetics of helical proteins. *Nature* **401**, 400-403 (1999).

84. Structure-based conformation preferences of amino acids. *Proc. Natl Acad. Sci. U.S.A.* **96**, 12524–12529 (1999).

85. DeGrado, W. F., Summa, C. M., Pavone, V. Nastri, F., and Lombardi, A. *De novo* design and structural characterization of proteins and metalloproteins. *Annu. Rev. Biochem.* **68**, 779-819 (1999).

86. Martinek, T. A. and Fülöp, F. Side-chain control of β-peptide secondary structures. *Eur. J. Biochem.* **270**, 3657-3666 (2003).

87. Sung, S. and Wu, X. Molecular dynamics simulations of helix folding: the effects of amino acid substitution. *Biopolymers* **42**, 633-644 (1997).

88. Baltzer, L. and Broo, K. S. De novo Designed Polypeptide Catalysts with Adopted Folded Structures. *Biopolymers* **47**, 31-40 (1998).

89. Fezoui, Y., Weaver, D. L., and Osterhout, J. J. *De novo* design and structural characterization of an α-helical hairpin peptide: A model system for the study of protein folding intermediates. *Proc. Natl. Acad. Sci. USA* **91**, 3675-3679 (1994).

90. Fezoui, Y., Weaver, D. L., and Osterhout, J. J. Strategies and rationales for the *de novo* design of a helical hairpin peptide. *Protein Sci.* **4**, 286-295 (1995).

91. Ramagopal, U. A., Ramakumar, S., Sahal, D., and Chauhan, V. S. *De novo* design and characterization of an apolar helical hairpin peptide at atomic resolution: Compaction mediated by weak interactions, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 870-874 (2001).

92. Myers, J. K., and Oas, T. G. Preorganized secondary structure as an important determinant of fast protein folding. *Nat. Struct. Biol.* **8**, 552-558 (2001).

93. Cavagnero, S., Dyson, H. J., and Wright, P. E. Effect of H helix destabilizing mutations on the kinetic and equilibrium folding of apomyoglobin. *J. Mol. Biol.* **285**, 269-82 (1999).

94. Jackson, S. E. How do small single-domain proteins fold? *Fold. Des.* **3**, R81–R91 (1998).

95. Mayor, U., Guydosh, N. R., Johnson, C. M., Grossmann, J. G., Sato, S., Jas, G. S., Freund, S. M. V., Alonso, D. O. V., Daggett, V., and Fersht, A. R. The complete folding pathway of a protein from nanoseconds to microseconds. *Nature* **421**, 863-867 (2003).

96. Socci, N. D., Onuchic, J. N., and Wolynes, P. G. Diffusive dynamics of the reaction coordinate for protein folding funnels. *J. Chem. Phys.* **104**, 5860-5868 (1996).

VITA

PREM P. CHAPAGAIN

Born: 1973, Syangja, Nepal.

EDUCATION:

Ph. D., Physics, 2005
Florida International University, Miami, FL
Research Area: Theoretical Biophysics (Protein Folding)

M. Sc., Physics (Distinction), 1998
Tribhuvan University, Kathmandu, Nepal
Relevant courses: Condensed Matter Physics

B. Sc., Physics (First class), 1996
Tribhuvan University, Kathmandu, Nepal
Relevant courses: Physics, Chemistry, and Mathematics

RESEARCH INTERESTS:

Protein Folding, Non-linear Dynamics, Complex Systems, Self-Organized Criticality

ACADEMIC / PROFESSIONAL EXPERIENCE:

Teaching Assistant/ Research Assistant, Florida International University (2000-2004)
Medical Physicist, Bhaktapur Cancer Hospital, Bhaktapur, Nepal (1999-2000)
Assistant Lecturer, Department of Physics, Tribhuvan University, Nepal (1998-2000)
Science Teacher, Balmandir School, Lumle Agr. Research Centre, Nepal (1995-1996)

LIST OF PUBLICATIONS:

1. Prem P. Chapagain and Bernard S. Gerstman, "Enhanced protein folding by removal of kinetic traps: thermodynamics and kinetics of a model α-helical hairpin peptide," (In preparation).
2. Bernard S. Gerstman and Prem P. Chapagain, "Self-organization and the hydrophobic interaction in protein folding," *J. Chem. Phys.* 123, 054901 (2005).
3. Prem P. Chapagain and Bernard S. Gerstman, "Excluded volume entropic effects on protein unfolding times and intermediary stability," *J. Chem. Phys.* 120, 2475 (2004).
4. Prem P. Chapagain and Bernard S. Gerstman, "Finite size scaling of structural transitions in a simulated protein with secondary and tertiary structure," *J. Chem. Phys.* 119, 1174 (2003).
5. Bernard S. Gerstman, Prem Chapagain and Lu Lu, "Polymer dynamics and chaos in protein unfolding," *Current Trends in Polymer Science*, Vol. 7 (2002).