This is to certify that the

dissertation entitled

An Analysis of Protein Folding by Decoding the
Hierarchy of Native-State Structural
Interactions

presented by

Brandon Michael Hespenheide

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Biochemistry & Molecular
Biology and Physics & Astronomy

Leslie A Kuhn
Major professor

Date Aug. 9, 2002

An Analysis of Protein Folding by Decoding the Hierarchy of Native-State Structural Interactions

By

Brandon Michael Hespenheide

A Dissertation

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

Department of Biochemistry and Molecular Biology and

Department of Physics and Astronomy

2002

ABSTRACT

AN ANALYSIS OF PROTEIN FOLDING BY DECODING THE HIERARCHY OF NATIVE-STATE STRUCTURAL INTERACTIONS

By

Brandon Michael Hespenheide

Understanding the mechanism by which proteins fold is one of the most intensely studies problems in science. Here, an analysis of the native-state structures of proteins is presented as a means to study protein folding. The hypothesis is formed as follows. As a protein folds, hydrophobic collapse results in a compact, fluid structure with few, if any, specific contacts. As the protein begins to fold, hydrogen bonds and salt bridges begin to form, stabilizing the structure. These noncovalent bonds continue to form until the native state is reached. Assuming that these noncovalent bonds are maintained throughout the folding reaction, any stable substructure formed during folding should be visible as a subset of the interactions found in the native state of a protein.

An analysis of the observed packing geometry between helices and sheets in a set of nonhomologous proteins is presented in Chapter 2. The role of possible dipole interactions is evaluated by explicitly taking into account the N– to C–terminal orientations of the secondary structures. A reduced representation is used in which the structures are defined by 3-dimensional vectors fit to the C_α positions. Helix–sheet interactions are defined such that the geometry can be expressed by a single angle, Ω, which represents the dihedral angle formed by the helix, the strand in the sheet closest to the helix, and the line of closest approach between the helix and the strand. The results show that for helix–strand interactions in which no β-sheet dipole is present, no preferred Ω packing angle is observed. How-

ever, for β-sheets with a net dipole due to a partially or entirely parallel β-sheet topology, a strong preference for helices to pack at $\sim$180° relative to the strand is observed. This is expected, if dipoles play a key role in defining the packing geometry.

Chapters 3 and 4 present a novel means for measuring the flexibility in protein structures by using the program FIRST. Results of native-state flexibility analysis correlate well with experimentally observed native–state flexibility in several proteins, prompting the assumption that rigid regions represent folded structure and flexible regions represent unfolded structure. A means of simulating thermal denaturation is presented. We view the thermal unfolding of a protein as a process in which the hydrogen bonds and salt bridges break in an energy–dependent manner. This process is mimicked by breaking hydrogen bonds and salt bridges one by one, from weakest to strongest, and observing how the flexibility of a protein structure increases after each step. As the protein unfolds in response to this increased flexibility, an increasing number of residues in the protein become flexible, while others remain rigid. This proceeds until the entire protein becomes flexible when all hydrogen bonds have been removed. The mean coordination, $\langle r \rangle$, computed as the average number of bonds per atom in the structure, is determined at each step in the simulated denaturation, and is shown to be a relevant structural variable for tracking the unfolding reaction. Specifically, the number of bond–rotational degrees of freedom in the system, a free energy like quantity, can be monitored as a function of $\langle r \rangle$, and used to identify the rigid to flexible phase transition during the unfolding simulation. Finally, the folding cores for ten proteins are predicted by identifying the the last set of two or more secondary structures to remain mutually rigid, or stable, during simulated unfolding. The predicted folding cores are compared to those observed in hydrogen–deuterium exchange/NMR experiments, and the results for 8 out of the 10 proteins indicate a close correlation.

For my mother and father,

and Judy and Beth

ACKNOWLEDGMENTS

I would like to start by thanking my advisors, Dr. Leslie A. Kuhn and Dr. M. F. Thorpe. I began my graduate studies in the lab of Dr. Kuhn, whose encouragement and endless enthusiasm provided the driving force behind my application for a dual degree program in the Department of Biochemistry and Molecular Biology and the Department of Physics and Astronomy. Over the years I have had a chance to work closely with Dr. Kuhn and Dr. Thorpe, both in the lab and in the classroom. Observing and learning from the unique ways in which both of these professors approach problems has been the most rewarding experience of my graduate studies. I am extremely grateful for their mentoring, and I am looking forward to a future of exciting research using the skills they have taught me.

I extend an enthusiastic thank you to the members of my committee, Dr. Shelagh Ferguson-Miller, Dr. Robert Hausinger, Dr. Jack Watson and Dr. Phil Duxbury. They have all been key in shaping this interdisciplinary thesis. Their encouragement and critical assessment of this thesis has been greatly appreciated.

The research I have completed over the years would not have been possible without support and advice of many graduate students, post docs, and faculty that I have had the pleasure to work with over the years. Of particular note are Dr. Paul Sanchagrin, Dr. Michael Raymer and Dr. Volker Schnecke, all former members of Dr. Kuhn's lab. These three guys taught me how to write computer program properly, which has saved me hours

and sometimes days of frustration while performing experiments. I would also like to thank Maria Zavodszky, Rajesh Korde and Ming Lei for helping create an enjoyable working environment and providing useful feedback. Most importantly, I send a big thank you to A. J. Rader, a graduate student of Dr. Thorpe's, who is also doing a dual degree with Dr. Thorpe and Dr. Kuhn. Much of the research I've done over the years was completed after long discussions with A. J. on how to get the job done. We have both contributed to, and sometimes struggled with, the collaborative project between our respective labs, and I am thankful for his support, advice, and friendship over the years.

Finally, I would like to thank all of my friends and family who have made it all worth while. I would like to thank my whole family, from parents to fourth cousins, the best family in the world. Mom and dad, Judy, grandma and grandpa, uncle Keno, Ryan (Chachi), Alex, uncle Mike, aunt Dee, aunt Margaret and uncle Bob, aunt Mary, and Mr. and Misses Strunk, thank you for all your support. To all my friends from back in the day, and those I've made while a graduate student, Teri, Dr. Brad Mballs, Chachi again, Dave, Volker, Ina, Annika, Tim, John Moehn, John Centner, Frans, Josh, Bryan and Kirsten. Thanks for all the advice (beer) and good times (getting me in trouble). Lastly, I want to send all my love and thanks to Bethany Strunk. When things seemed impossible to do, she was my inspiration to keep going. Nakupenda, Beti.

TABLE OF CONTENTS

LIST OF TABLES

List of Figures

List of Abbreviations

$\langle r \rangle$	mean coordination
1D	1-dimensional
3D	3-dimensional
AFU	autonomous folding unit
BPTI	bovine pancreatic trypsin inhibitor
C_α	alpha carbon
CD	circular dichroism
CI2	chymotrypsin inhibitor 2
DC	diffusion–collision
DOF	degrees of freedom
DSSP	Dictionary of Secondary Structures of Protein
FIRST	Floppy Inclusions and Rigid Substructure Topography
FLOPS	floating point operations per second
H-D exchange	hydrogen–deuterium exchange NMR
HIV	human immunodeficiency virus
HZ	hydrophobic zipper
MD	molecular dynamics
NC	nucleation–condensation

NMA	normal modes analysis
NMR	nuclear magnetic resonance
PDB	Protein Data Bank
PE	protein engineering
RCD	rigid cluster decomposition
RCSB	Rutgers Collaboratory for Structural Bioinformatics
TS	transition state
TSE	transition state ensemble

Chapter 1

Protein Folding: A Transition from Flexible to Rigid

1.1 Computers and Biology

Perhaps the best known of the first fully electronic computers built was ENIAC (Electronic Numerical Integrator and Computer), which was constructed in 1946 (Adams et al., 1995). The physical size of the machine nearly filled a 7 by 13 meter room and required 18,000 vacuum tubes to run. Despite the impressive size, ENIAC could only perform 340 floating point operations per second (FLOPS). Technological advancements in processor design brought computers to their current position of dominance in modern society. The average workstation today occupies a physical space no larger than a shoe box, and boasts gigaFLOPS processing power. Recent advances in networking and computer architecture have allowed for many computers to be connected in parallel, acting as a single computational processor capable of solving large problems. In 2001, IBM announced that in collaboration with Lawrence Livermore National Laboratory, it would build a networked computer system, named Blue Gene/L, capable of $\sim$200 teraFLOPS, specifically designed for applications in the life sciences, particularity in the area of protein folding. An equivalent version of ENIAC in 1946 would require more surface area than is available on the planet.

Concurrent with advances in computer technology have come advances in all of the natural sciences. It is impossible to fathom the extent to which computers have added to our understanding of nature. A general case can demonstrate the point. Computers have allowed for quicker and more reliable analysis of data, allowing subsequent experiments to be performed more often and more accurately. As a specific example, all of structural biology has benefited from the speed at which protein crystal structure data is made available. Advances in computer technology allow for the design of better and better experimental equipment. Advances in processor speed allow for faster analysis and refinement of the diffraction data. And perhaps most important of all, the availability of the World Wide Web has allowed for easy public access to protein structures via the Protein Data Bank (PDB), hosted by Rutgers Collaboratory for Structural Bioinformatics (RCSB).

Of particular importance in regard to this thesis is the application of computers and computer science to problems in physics and biology that would otherwise be practically impossible to solve. The general field to which this thesis belongs could best be described as structural biology, and the following hierarchical category can be applied: *natural science* → *biophysics* → *protein folding* → *computational* → *native-state structure analysis*. This hierarchy can be extended one more step, where Chapter 2 would be → *geomet-*

ric analysis of secondary structure packing, and Chapters 3 and 4 would be → *graph-theoretical analysis of native-state bond networks*. All of the work presented here addresses one of the most challenging unsolved problems in structural biology, the protein folding problem.

1.2 The Protein Folding Problem

In the 1950s and early 1960s, Anfinsen published several experiments on the denaturation and renaturation of ribonuclease A (Anfinsen et al., 1954; Anfinsen and Haber, 1961). His main conclusion from this work of relevance to protein folding was that proteins can spontaneously refold to their native conformation after unfolding (Anfinsen, 1973). This experiment provided solid evidence for what has become a tenet of structural biology: all of the information required for a protein to fold is encoded in the primary structure, or sequence, of the protein. In Anfinsen's own words from his 1972 Nobel Prize speech, "The native conformation is determined by the totality of interatomic interactions and hence by the amino acid sequence, in a given environment."

The first clear indication of the complexity of protein folding emerged in 1968 with what has become know as "Levinthal's paradox" (Levinthal, 1968). The paradox arises from a simple estimation of the number of possible conformations a protein can adopt. Given the crude estimate that each amino acid in a protein can adopt 10 unique conformations that produce nonoverlapping structures, then the number of possible unique structures is 10^N, where N is the number of amino acids in the protein. For a protein 104 amino acids in length such as horse heart cytochrome *c*, this number is 10^{104}, or $\sim 10^{100}$. If a protein could convert between unique conformations on the order of molecular vibrations, say every femtosecond, it would still require on the order of 10^{80} years to sample every conformation, many times longer than the age of the universe. Given that most proteins fold in between microseconds and seconds, it is clear that proteins do not reach the native state by random sampling of conformations. Nonrandom searching implies the concept of a pathway, and leads to the question, by what mechanism do proteins fold?

1.3 The "Old" and "New" Views of Protein Folding

The protein folding pathway became an established concept as a direct consequence of Levinthal's paradox. Generally referred to the "old view" of protein folding in recent literature, early protein folding mechanisms envisioned the formation of a series of discrete intermediates along the free energy surface from a denatured state to a native state (Kim and Baldwin, 1990; Englander and Mayne, 1992). This view of protein folding has since been replaced by the "new view" of protein folding, perhaps best represented by the energy landscape theory of protein folding, in which folding occurs by the diffusion of *ensembles* of structures, rather than discrete intermediates, across the multidimensional free energy surface of a protein structure. A qualitative description of the energy landscape view of protein folding is given here. Detailed explanations can be found in a number of references (Onuchic et al., 2000; Nymeyer et al., 1998; Onuchic et al., 1997; Dill and Chan, 1997; Bryngelson et al., 1995; Bryngelson and Wolynes, 1987).

The energy landscape theory of protein folding employs a statistical mechanical description of the folding reaction in which the kinetics and thermodynamics are dictated by the self-organization of ensembles of structures. A funnel-shaped picture is often used to describe the energy landscape, as shown in Figure 1.1 (Leopold et al., 1992). Key for the interpretation of the folding funnel is the idea of a reaction coordinate or order parameter. The free energy of a protein will depend upon its structure, and therefore, the dimensionality of the free energy surface will depend on how many variables it takes to describe the structure of a protein. A common means of describing a protein structure is to use the x, y, and z positions of each atom in Cartesian space, which yields 3N variables to describe the structure of the protein, where N is the number of atoms. Visualizing high-dimensionality space is impossible, so the goal of the reaction coordinate is to provide a single variable that describes the structural features common to any conformation with a given free energy during folding. The most often cited reaction coordinate is the percentage of native contacts, Q, present at any point along the folding reaction, although other parameters such as surface area or radius of gyration have been used (Socci et al., 1996; Brooks III et al., 1998). The rough funnel shape of the energy landscape arises because evolution has selected for protein sequences that exhibit minimal frustration, as compared to random sequences. Frustration can manifest itself energetically or topologically, and examples of both are given here. Energetic frustration can arise when a main-chain amide group that is hydrogen bonded to solvent in the denatured state does not hydrogen bond in the native state. Topological frustration implies that certain native state conformations are easier to reach than others. For example, imagine a protein whose native state consists of a knot. It is very possible that a sequence could be mapped onto this conformation such that all bond lengths and angles are unstressed and every potential hydrogen bond is formed. This would be energetically unfrustrated. However, the energy barriers present in forming a knot would result in a folding funnel that resembles a golf course, with a tiny, extremely deep hole in the middle of the green. The roughness of the folding funnel shown in Figure 1.1 is indicative of energetic and topological frustration present in the protein during folding. The degree of roughness will depend on how well the protein was designed by nature. Fast-folding proteins will have smoother energy landscapes and less frustration relative to slow-folding proteins.

The two most important features of the folding funnel, which provide the clearest separation from the old view of protein folding, are the possibility for many folding paths from the denatured ensemble to the native state, and the idea of ensembles or macrostates, rather than discrete intermediates. Although a single pathway down the funnel representation may dominate, it is critical to understand that each point along that pathway represents an ensemble of structures. Identifying the key features of these ensembles, and hence the key features of the protein structure as it folds, has been the subject of many theoretical and experimental studies over the past decade.

1.4 Overview of Protein Folding Models

Concurrent with the development of the energy landscape picture of folding has been the refinement of several phenomenological models describing the mechanism of protein folding. A common theme in all of these models is the formation of a small substructure(s) as

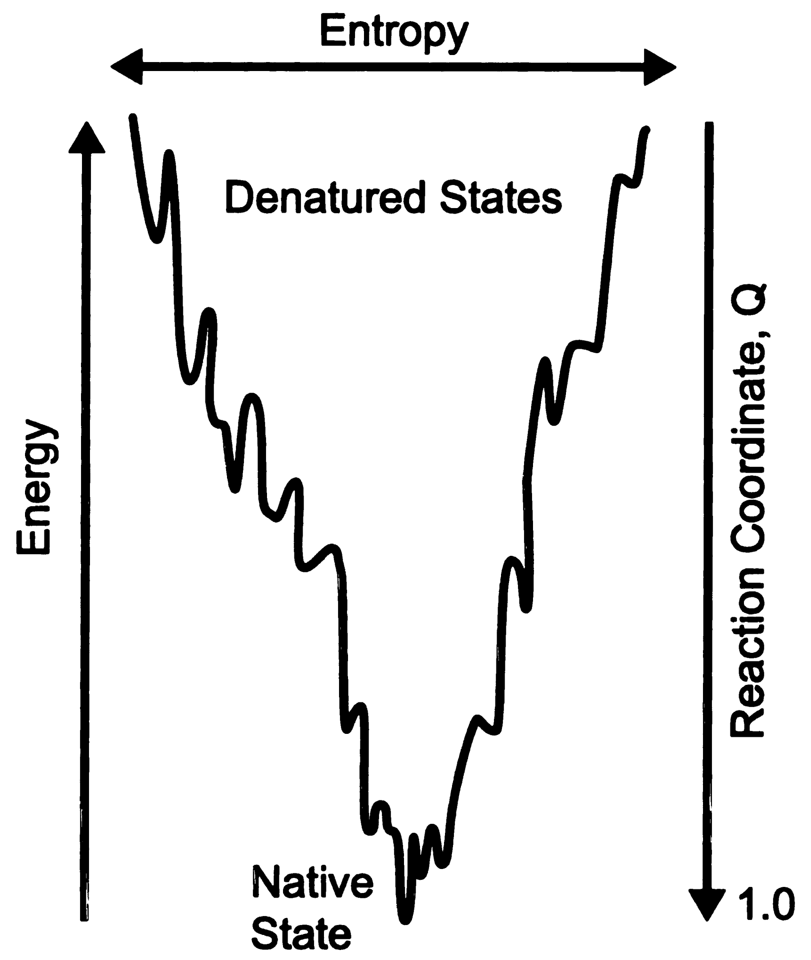

Figure 1.1: Generic energy landscape or folding funnel for a protein. The funnel shape represents the energetic bias towards the native state at the bottom of the funnel. The width of the tunnel roughly corresponds to the conformational entropy present in a protein as it folds. The top of the funnel, representing the denatured state, is wide indicating a large amount of conformational entropy. As the protein folds, it loses entropy and the width of the funnel shrinks. The many local small groves in the funnel represent local energy minima where the protein can get trapped for various amounts of time depending on the depth of the minima. The reaction coordinate, or order parameter, Q, is a number of native contacts present in any structure along the folding funnel and is a measure of similarity to the native state.

the first step of folding, but they differ in their emphasis on sequence local versus nonlocal interactions. An overview of these models is given here. For clarity, the terms sequence local and sequence nonlocal refer to the number of amino acids intervening between two interacting residues, and not the spatial distance between two atoms or residues. A typical sequence local interaction would be the $i \leftrightarrow i + 4$ hydrogen bond in an α-helix. The exact cutoff for defining a sequence nonlocal interaction can vary between experiments, but generally any interaction between residues i and j where $|i - j| \geq 8$ would be considered nonlocal.

A caveat of the following models is that they best describe the folding mechanism of single domain proteins that exhibit 2-state folding kinetics. A protein that folds by 2-state kinetics is believed to have only the denatured and native states populated at equilibrium, no intermediate steps are observed. Extension of these models to larger proteins appears feasible, as most large proteins split into domains that are usually capable of proper folding in the absence of the rest of the structure. Subregions of a whole protein that can fold independently of the rest of the structure are sometimes termed autonomous folding units (AFUs) (Fischer and Marqusee, 2000; Peng and Yu, 2000). One proposed model of folding in larger proteins is that it occurs via independent folding of multiple AFUs, as has been proposed for T4 lysozyme (Llinas and Marqusee, 1988), although this hypothesis is still being debated.

1.4.1 Nucleation–Condensation

The nucleation–condensation (NC) model describes a mechanism in which folding is initiated by formation of a stable nucleus consisting of both sequence local and nonlocal interactions (Mirny and Shakhnovich, 2001; Fersht, 2000; Thirumalai and Klimov, 1998; Fersht, 1997; Abkevich et al., 1994). These interactions result in the formation of native-like structure which collapses the protein into a phase that is more condensed than the denatured state. The specific interactions need not be unique (Guo and Thirumulai, 1997), although both theory and experiment suggest that several key residues are usually involved (Shakhnovich, 1998). This condensed phase is believed to be quite structurally similar to the native state, except that all of the non-nucleus forming interactions are weakened. The formation of the sequence local and nonlocal interactions together with the condensing of the structure represent the rate-limiting step in the folding kinetics.

The NC model provides a general description for how single domain, 2-state proteins may fold. The model fits within the theoretical framework of the energy landscape, and is fairly well accepted. The key experimental evidence supporting the NC mechanism has come from mutational experiments known as the protein engineering (PE) approach. Developed by Fersht and coworkers (Fersht et al., 1992), the PE approach attempts to identify whether a residue is important for nucleation by observing the affects of mutation on the height of the energy barrier for folding. The height of the barrier is expressed as the free energy of going from the denatured state to the transition state, $\Delta G_{D-\ddagger}$, and the difference in barrier height between the mutant and the wild-type protein is $\Delta G^{\mathrm{WT}}_{D-\ddagger} - \Delta G^{\mathrm{mut}}_{D-\ddagger} = \Delta\Delta G_{D-\ddagger}$. This number is normalized by dividing by change in the total free energy of folding, $\Delta\Delta G_{N-D}$, between the mutant and the wild-type, and this ratio a called the Φ value for the residue. The structural interpretation of Φ values near 1.0 is that they are as structured in the transition state as they are in the native state, and contribute to stabilizing the structure during folding. Likewise, if a mutation does not affect the free energy of the transition state ($\Delta G^{\mathrm{mut}}_{D-\ddagger} \cong \Delta G^{\mathrm{WT}}_{D-\ddagger}$), the residue will have a Φ value near 0.0. It is suggested that residues with Φ values near 0.0 are disordered in the transition state, as much as in the denatured state, and therefore are not important for nucleation. An important assumption in the PE method is that the mutations do not significantly alter the folding mechanism or the structure of the native state. This assumption assures that any observed changes in the folding kinetics can be attributed to stabilization/destabilization of the wild-type folding mechanism. For this reason, residues are usually mutated to alanine.

Extensive Φ value analyses has been performed on several proteins to date, and the results do seem to support the NC model (Clarke and Itzhaki, 1998; Nölting et al., 1997; Itzhaki et al., 1995). Typically, most residues have fractional Φ values, which can be interpreted in several ways. Because the transition state is represented by an ensemble of structures, a Φ value of 0.5 could mean that the given residue is structured in half of the structures, and disordered in the other half. It could also mean that the given residue is in the core, but the mutation caused a weakening of the interactions it makes. A third explanation could be the existence of parallel pathways, with parallel transition states. For example, if two pathways existed, the nucleus of pathway 1 could involve the given residue in the TSE, but the TSE of pathway 2 could use a different nucleus. Distinguishing between these interpretations experimentally is not trivial, and the issue is still being debated

(Myers and Oas, 2002; Ozkan et al., 2001)

In addition to results of PE experiments, there have been many computational studies using lattice models (Abkevich et al., 1994), off-lattice models (Li and Shakhnovich, 2001) and MD simulations (Kazmirski et al., 2001; Daggett et al., 1996) which have supported the NC model and provided a theoretical framework to describe the experimental observations. A good review of these computational techniques can be found in Mirny and Shakhnovich (2001).

1.4.2 The Diffusion–Collision Model

The diffusion–collision model bears some resemblance to the NC model described above, the key difference being that interactions local in sequence are much more strongly emphasized in the DC model (Karplus and Weaver, 1994, 1979). These nearby contacts lead to the formation of microdomains, which generally correspond to packed secondary structures such as α-helices or β-hairpins. These microdomains, which are marginally stable, diffuse through the solvent and collide with each other. Collisions that result in a stable tertiary interaction will produce larger substructures, and can occur in a unique order (single pathway) or in near random order (many parallel pathways). The rate-limiting step is dictated by how stable the individual structural units are, the probability of collisions forming native state conformation, and how quickly the units can diffuse through the media.

The DC model draws heavily from statistical mechanics and helix-coil transition theory (McCammon et al., 1980; Flory, 1969), in which the local $i \leftrightarrow i+1$ contacts of the α-helix provide the cooperativity required to cross the free energy barrier between random coil and structured helix. It is perhaps for this reason that the DC collision model has been successful for describing the folding of all-helical proteins such as apomyoglobin (Pappu and Weaver, 1998), as helix formation is so strongly driven by local interactions. Application of the DC model to proteins with significant β-sheet structure, or proteins with little or no secondary structure appears not to be a viable option, especially in light of experimental evidence for chymotrypsin inhibitor 2 (CI2) in which the folding nucleus specifically involves nonlocal interactions. In summary, it appears that the DC model is a special case of the NC model in which the protein folds via several nuclei that are localized to secondary structures.

1.4.3 The Hydrophobic Zipper Model

The hydrophobic–zipper (HZ) model (Dill et al., 1993) describes a mechanism that would initiate folding immediately after hydrophobic collapse through the interaction between nonpolar amino acids. Initially, a contact (HH) is made between a pair of hydrophobic residues i and j that are near each other spatially. Most likely this interaction will involve residues that are local in sequence, as this requires a smaller conformational search and will result in less loss of entropy upon folding, but nonlocal interactions are not excluded. Once an interaction is established, hydrophobic residues proximal to i and j will now be near each other spatially, and will have a higher chance of interacting than if the interaction between i and j had not been established. This scenario can be repeated indefinitely until all hydrophobic residues are in contact. Because each subsequent HH interaction gets easier to form due to a smaller loss of conformational entropy, this model implicitly describes a cooperative folding event.

It is important to note that the hydrophobic zipper model is not simply saying that the hydrophobic effect drives the entropic collapse of protein structures, resulting in burial of nonpolar residues (Tanford, 1980). The hydrophobic effect is a well accepted part of every modern theory of protein folding and implicitly occurs *a priori* in most mechanisms. It is for this reason that any qualitative description of protein folding begins with a compact denatured state, not a fully extended polypeptide, simply because the fully extended structure is not observed in nature.

While the HZ model appears to be a generalization of the hydrophobic effect to protein folding, its key difference is the formation of contacts between hydrophobic residues buried within the protein. The buried core of any denatured state is believed to be quite fluid because the hydrophobic effect results from nonspecific interactions between nonpolor groups. The HZ model suggests that as folding begins, some pairs of hydrophobic groups do form specific contacts, providing a nucleation site for the propagation of further hydrophobic contacts in a cooperative manner. The nonpolar amino acids can be ordered according to their relative "hydrophobicity values" (Bull and Breese, 1974), which implies that specific interactions may be possible. The predictions of the HZ model have been tested exhaustively using lattice simulations, and recent experimental evidence on the folding of proteins with coiled coil structure lends support (Hicks et al., 2002). Also, the HZ model favors the presence of multiple folding pathways.

It seems clear, from theory and experiment, that any mechanism of folding is going to require the formation of substructure or microdomains involving several key sequence local and most likely sequence nonlocal interactions that provide the cooperativity needed to scale the free energy barrier separating the denatured and native states. Furthermore, since it is implied that folding continues to the native state after the initial substructure is formed, the TSE along any pathway will have these substructures in common. Taken one step further, let us assume that the folding nucleus or microdomain is maintained all the way to the native state. In this case, the native state structure, as realized by experiment (X-ray crystallography, nuclear magnetic resonance (NMR), etc) should contain within its network of bonds a subnetwork corresponding to the folding nucleus or microdomains. It is this line of reasoning that has led to the development of many experiments, including those presented in this thesis, designed to answer one question: does the native state structure of a protein encode information about the folding mechanism?

1.5 Computational Analyses of Native-State Structure as a Tool To Study Protein Folding

The availability of high-resolution structural data for many proteins, with corresponding experimental data about the mechanism of folding, has facilitated the development of computational techniques to study protein folding based on protein structure. Experiments have made the important contribution that the interactions forming the folding nucleus or microdomains are generally conserved in the native state. The goal of the computational techniques discussed here is the *de novo* prediction of the folding initiation structure(s) and/or the TSE for a given protein by using only the native state structure.

The general scheme for most native-state analysis techniques is as follows. The topology of the native state network is derived from a high-resolution source of atomic coordinates, such as X-ray crystallography or NMR. In general, the topology of a system defines how things are connected, in this case we want to know how the atoms in a protein are connected. Depending on the method, the topology can be a reduced representation, describing only connections between C_α's within a certain radius of each other, or it can be extremely descriptive and include connections for each covalent bond, salt bridge, hydrogen bond and hydrophobic interaction. Once the topology is defined, an algorithm will proceed to dissect the native topology into subtopologies (or subgraphs) and measure a given quantity for each subgraph. The goal is to identify a subgraph that corresponds to a key structure along the folding reaction for a protein. Depending on how well the predicted structure compares to the corresponding structure observed experimentally, will dictate the viability of the method.

Galzitskaya and Finkelstein (1999) have a developed an algorithm for computing a free energy-like quantity for subtopologies of a given native state. They use a highly reduced model in which two (or four for larger proteins) consecutive residues are assigned a single site on the native state graph, and this site is referred to as a "link". For a protein with N links, any given subgraph is defined as having S links in native conformation, and $N - S$ links disordered. For each subgraph, they compute a free energy, where the enthalpy term is derived solely from S native state links, and the entropy is calculated based on the number and length of the $N - S$ disordered links. The free energy is then computed for every possible subgraph of S ordered and $N - S$ disordered links, subject to certain restrictions (for example, they only allow a fixed number of disordered loops). Their hypothesis is that the ensemble of subgraphs of covalent and noncovalent bonds with the highest computed free energy correspond to the TSE, as by definition the transition state is the most unstable species on any reaction path. In the case of proteins, structures in the TSE have exactly a 50.0% chance of proceeding to the native state and 50.0% chance of unfolding. Their computed TSE generally consists of thousands of structures that can be used to compute how often, on average, link i is found in a native conformation. This average is the computational analog of a Φ value. For example, if link 1 (corresponding to residues 1 and 2 in the protein) is part of an ordered region in exactly half of the TSE subgraphs, the computed Φ value is 0.5. Φ value predictions are made for each link in the protein and compared to experimental values.

Comparison between predicted and experimental Φ values was performed for five proteins: CI2, barnase, CheY, SRC SH3 domain and α-spectrin SH3 domain. The average correlation coefficient between prediction and experiment for all five proteins was 0.46, with a highest value of 0.56 for CI2 and a low value of -0.02 (no correlation) for SRC SH3 domain. The poor correlation can be attributed to deficiencies in the free energy calculation, such as the absence of a term describing potentially stabilizing interactions that occur within disordered loops. (The authors made the assumption that disordered links could not adopt stable nonnative conformations.) Despite the unimpressive correlation coefficients, the predictions are better than random, and imply that the topology of a protein can encode information about folding. Extensions of this work have been performed in which an ensemble of dynamically generated structures, such as from an off-lattice Monte Carlo simulation, was analyzed instead of just the native state topology. Predicted folding nuclei and features of the TSE in these experiments have shown much better correlation to experiment (Dokholyan et al., 2002; Vendruscolo et al., 2001).

Nussinov and coworkers have developed a different type of native-state analysis algorithm that is designed to dissect a protein into hydrophobic folding units or building blocks, and the folding of a protein is described as the hierarchic assembly of these building blocks (Tsai et al., 2000, 1998). The details of how the building blocks are identified can be found in (Tsai and Nussinov, 1997). Briefly, the protein structure is exhaustively dissected into fragments of contiguous sequence called building blocks, ranging from the entire structure down to a minimal block size of seven residues. For each building block, an empirical score is computed based on its solvent accessible surface area, compactness, hydrophobicity and isolation. The score is designed to represent how stable each building block would be if it were isolated from the rest of the protein structure. Low scoring building blocks are discarded. The remaining set of building blocks can be assembled in various ways to form the complete protein such that the sequences represented by the building blocks don't overlap by more than a few residues. Because many building blocks will generally be found for a protein (78 were found for actin, which has 373 residues), it is possible to build the whole protein from different assemblies of building blocks.

The pathways identified by the above algorithm can most readily be associated with both the HZ model of folding initiation (formation of the building blocks), followed by a mechanism of hierarchic folding in which the building blocks are assembled, similar to the DC model. Taken together, Nussinov refers to the predicted folding assemblies as the "building block" model of folding. The model is consistent with energy landscape theory in that it allows for single or parallel folding pathways, depending on the order in which the building blocks are assembled. Despite their making available the anatomy trees for every protein in the PDB (via the following web site: http://protein3d.ncifcrf.gov/tsai/anatomy.html), very few experimental correlations have been published. It appears that the method is quite good at dissecting a protein into domains, supersecondary structures, and subsequently individual secondary structures. However, the specific interactions forming the folding nucleus cannot be distinguished, and *de novo* prediction of which building blocks compose the TSE would be difficult. Furthermore, the anatomy tree for barnase does not include the N-terminal helix in the first level, corresponding to an early forming structure along the folding reaction. This result does not correlate with mutational experiments that suggest an early interaction between the N-terminal α-helix with several strands of the C-terminal β-sheet.

A third method, developed by Wallqvist et al., (1997) is described as, "a computational method useful for identifying the existence of stable structural components of a protein and rank ordering their stability". The details of their algorithm are quite complicated and outside the scope of this introduction. Briefly, they compute an "unfolding penalty" for each residue in a protein based on an empirically derived free energy-like equation. The free energy equation was parameterized through the analysis of a large number of nonhomologous protein structures, not unlike the parameterization of force fields used in MD simulations. The free energy unfolding penalty can be thought of as the degree to which a given residue will resist unfolding, and they depend on the geometry and the amino acid composition in the vicinity of the given residue.

The authors compare their unfolding penalties to protection factors determined by hydrogen-deuterium exchange NMR (H-D exchange) experiments (H-D exchange is described more thoroughly in the next section). Under native conditions and at equilibrium, H-D exchange experiments measure the rate at which the main-chain amide groups of specific residues exchange their protons with solvent. It is believed that the protein must unfold to a certain extent for exchange to occur, and that the distribution of observed exchange rates indicates the degree to which each amide must unfold. Residues that exchange quickly easily unfold, whereas residues that exchange slowly resist unfolding. From these exchange rates a protection factor can be computed, which is the experimental analog of the computed unfolding penalties.

Correlation coefficients greater than 0.5 between unfolding penalties (predicted) and protection factors (observed) were reported for several proteins; plastocyanin, staphylococcal nuclease and three different cytochrome *c*'s (Wallqvist et al., 1997). For horse heart cytochrome *c*, the correlation coefficient between the predicted and experimental data was 0.71, and the qualitative overlap was very good for all proteins studied. The authors defined the subset of structure with the highest unfolding penalties as the folding core of the protein. In relation to the folding mechanisms described above, this folding core represents a substructure that forms after nucleation (in the NC model) or after a favorable collision (in the DC model). The authors make no suggestion that the residues with the highest unfolding penalties should be involved in the nucleus or the microdomains. Overall, these data, together with the two methods described above, provide encouraging results that the native state structure does indeed encode information about how the protein folded.

It should be noted that the application of H-D exchange data to the study of protein folding pathways has come under some scrutiny lately, particularly in light of results from PE experiments. These issues are addressed in the next section, in which H-D exchange methodology is described. The concerns raised by PE experiments are addressed, and a possible alternative to the interpretation of Φ values is discussed.

1.6 H-D Exchange, Φ values, and the Protein Folding Core

To verify any computational or theoretical prediction of protein folding it is necessary to have reliable experimental data for comparison. NMR measurements of hydrogen–deuterium exchange of protein backbone amide groups provides a powerful tool for the study of structural fluctuations in proteins. An outline of the method is given here. A protein is expressed and isolated under environmental conditions favoring the native state. An NMR spectrum of the protein is recorded in ^1H_2O, and the observed chemical shifts are assigned to specific backbone amides in the protein. The protein is then transferred into a buffer composed of deuterated water, ^2H_2O. Under native conditions, a protein will experience dynamic fluctuations that can range from localized unfolding events to complete denaturation via a global unfolding pathway. These fluctuations have the effect of exposing amides to solvent, allowing hydrogen–deuterium exchange to occur. Under conditions favoring the native state, local unfolding, or breathing, can arise as a result of protein function, or simply be due to the absence of sufficient bond forces in a given area. (According to thermodynamics, even global unfolding to high energy conformations is expected to occur in a small number of protein molecules at equilibrium based on the Boltzmann distribution.) Several experiments are performed, and in each the protein is allowed to exchange in ^2H_2O for a different time period. At the end of each time period, a new NMR spectrum of the protein is recorded. Because deuterium will not produce a signal in these experiments, exchange can be observed as a decay in the signal intensity for each amide proton. By observing exchange over many different time periods, an exponential can be fit to the signal intensity decay, and an exchange rate computed.

The mechanism of hydrogen–deuterium exchange in proteins is believed to occur according to an unfolding reaction (local or global) according to equation 1.1, as initially proposed by Linderstrøm-Lang (1958). In this equation, C represents a closed form of the amide group. Exchange cannot occur from this state. Likewise, O represents an open or exchange competent form of the amide proton. Equilibrium between these two forms is defined by the rate constants for opening, k_{op}, and closing, k_{cl}. Once in an exchange competent form the amide can exchange its hydrogen with solvent. Because the apparent rate of exchange depends on both the rate of opening, k_{op}, and the rate of exchange, k_{int}, it is nearly impossible to determine these rates individually in the context of whole protein studies. Therefore, k_{int} is typically determined from the rate of exchange observed, for each amino acid type, within the structure of small model peptides (Bai et al., 1993; Molday et al., 1972), for which no "opening" reaction is required.

$$\mathrm{C^{H}} \underset{k_{cl}}{\overset{k_{op}}{\rightleftarrows}} \mathrm{O^{H}} \xrightarrow{k_{int}} \mathrm{O^{D}} \rightleftarrows \mathrm{C^{D}} \qquad (1.1)$$

Under conditions that favor folding, $k_{cl} \gg k_{op}$, and the observed rate of exchange, k_{ex}, can be expressed using equation 1.2.

$$k_{ex} = \frac{k_{op} k_{int}}{k_{cl} + k_{int}} \tag{1.2}$$

Based on equation 1.2, two limiting scenarios of exchange arise. The first case occurs if $k_{int} \gg k_{cl}$, in which the observed rate of exchange from 1.2 can be reduced to $k_{ex} = k_{op}$. This scenario is named the EX1 limit for exchange. The EX1 limit for exchange is rarely observed in proteins under native conditions. The fact that exchange occurs more quickly than reprotection of the amide suggests a significant structural instability for the protein. This observation is valid, as experiments have shown that most amides favor the EX1 mechanism at increasing concentrations of denaturant.

The alternative scenario occurs when $k_{cl} \gg k_{int}$, referred to as the EX2 limit. In this case, equation 1.2 can be reduced to equation 1.3. Because the term $k_{op}/k_{cl} = K_{op}$ represents the equilibrium constant for opening and closing the amide, and this represents the rate-limiting unfolding required for exchange, an apparent free energy of exchange can be computing from the observed exchange rate, k_{ex}, and the intrinsic exchange rate, k_{int}, by using equation 1.4. EX2 exchange has been shown to be the dominant mechanism of exchange under native conditions, allowing the apparent free energies of exchange to computed.

$$k_{ex} = \frac{k_{op}}{k_{cl}} \cdot k_{int} = K_{op} \cdot k_{int} \tag{1.3}$$

$$\Delta G_{ex}^{\text{app}} = -RT \ln K_{op} \tag{1.4}$$

The usefulness of H-D exchange as a means to study protein folding is based on the thermodynamic premise that a protein can sample all of its higher energy conformations along the folding pathway according to a Boltzmann distribution. This means that even under native conditions, at any given time a small population of protein molecules will be in an unfolded state. The protein will rapidly refold, but during the time it is denatured, H-D exchange can occur, and the highly sensitive NMR technique can observe the exchange.

The exchange rates observed in proteins can vary by several orders of magnitude. Fast exchange rates are generally associated with local fluctuations or breathing. The slowest exchanging residues correspond to global unfolding events; that is, these residues only exchange if the protein completely unfolds. It is assumed that the slowest exchanging amides require global unfolding for exchange to occur. This is verified by comparing the apparent free energy of exchange to the free energy of folding, ΔG_{D-N}. If $\Delta G_{ex}^{\text{app}} \sim \Delta G_{D-N}$, then exchange requires global unfolding. Assigning a global unfolding mechanism to slow-exchanging residues can also be accomplished by comparing the change in $\Delta G_{ex}^{\text{app}}$ that occurs upon mutation. If $\Delta\Delta G_{ex}^{\text{app}}$ for a given residue is approximately equal to the change in the stability between the wild type and mutant proteins, $\Delta\Delta G_{D-N}$, then that residue exchanges by a global unfolding pathway.

Based on the results of H-D exchange experiments, Woodward and coworkers proposed the idea of a slow-exchange core, defined as the minimal collection of residues that include the slowest-exchanging amides observed under native-state conditions. This slow exchange core can be further expanded to define a "folding core", using the following definition. If a slow-exchange amide is found in a secondary structure, then that structure is part of the folding core. Occasionally, slow exchange core residues are found within turns or loops, and these are excluded from the definition of the folding core. Therefore, the folding core is the set of secondary structures that encompass the slowest exchanging amides. The folding core provides a low-resolution picture of the earliest stable substructure formed on a folding pathway.

Beginning in the early 1990's, around the same time the folding core concept was introduced, Fersht and coworkers began applying the protein engineering (PE) method to probe the structure of the TSE in barnase and other small proteins. Since that time, PE results, presented as Φ values, have become available for a number of proteins. In particular, barnase, barstar, CI2 and SRC SH3 domain have Φ value data corresponding to mutations in over50% of the amino acids in each protein. Determining Φ values is a considerable task, given that each requires a site-specific mutant protein be made, verified, and thermodynamically characterized. Φ values provide an additional experimental means to identify residues important for folding. In the mid-1990's, Fersht and coworkers began to notice discrepancies between Φ value results and H-D exchange exchange rates. In particular, Φ values indicate the N-terminal helix of barnase to be in native-state conformation in the transition state, even though almost all of the residues in this helix have fast exchange mechanisms.

In 1999, Li and Woodward presented a review article of H-D exchange results for many proteins, together with their identification of the folding core in each protein based on the published exchange rates. In this paper, they defend the concept of the folding core, and show that the correlation between structures with high Φ values and structures in the folding core is quite high. In the case of the N-terminal helix of barnase, the following explanation has been given. The only slow-exchanging residue in the helix is L14, in the middle of the secondary structure. This residue also has a high Φ value ($\sim$1.0) and so it is reasonable to assume that the main-chain hydrogen bond of L14 is formed early in the folding pathway. Once this bond is formed, local interactions will become favored (relative to their random association), according to any of the folding models described above (NC, DC, HZ models). Especially since this interaction is in a helix, we would expect the formation of adjacent $i \leftrightarrow i + 1$ hydrogen bonds to contribute to the cooperativity required for folding. The fact that the amides adjacent to L14 exchange by local fluctuations does not exclude the possibility that they formed early, as Fersht asserts. However, the slow exchange rate of L14 does indicate that it will only exchange when the entire helix is disrupted as a result of global unfolding. It is this line of reasoning that allows a single slow exchange residue to impart an "early folding" label to a whole secondary structure. While it would seem more reasonable that the folding core definition should only be extended to the small section of a secondary structure local to the slowly exchanging residues, Woodward points out that the rate constants cannot be resolved well enough to allow for a clear delineation of which part of the structure is involved. It is for this reason that the folding core provides a low-resolution picture of the early folding structure.

Given that most of the controversy surrounding the folding core definition has come from PE studies, it is necessary to discuss the structural interpretation of Φ values. It is generally accepted that a Φ value near 0.0 indicates that the side chain of the given residue does not contribute to stabilizing the TSE. In fact, a Φ value of 0.0 can also occur if a given residue is significantly structured in the denatured state. The denatured state is most often envisioned as an ensemble of random coil states, random implying no specific interactions. However, some studies have indicated that native interactions may exist in a protein under conditions that are often considered denaturing. If this is true, mutating a residue that is structured in the denatured ensemble will affect the free energy of the denatured state as much as the free energy of the transition state, resulting in a $\Delta\Delta G_{D-\ddagger} \sim 0.0$, and subsequently a Φ value near 0.0. The explicit definition of a Φ value near 0.0 is that the residue is as structured in the TSE as it is in the denatured state. But if it's structured (native-like) in the denatured state, and therefore in the TSE, it should have a Φ value of 1.0, hence the misinterpretation of the data. Also, many Φ values have values less than 0.0 or greater than 1.0, and the structural interpretation of these residues is unclear, although several interpretations have been suggested (Ozkan et al., 2001; Myers and Oas, 2002).

Despite the controversy surrounding H-D exchange as a method to study folding *pathways* (Clarke et al., 1997), the assignment of a folding core based on slow-exchanging residues remains a low-resolution way of identifying structures that form early in the folding pathway. Thus, folding core data provide a useful dataset for validating computational techniques designed to probe early forming substructures (Torshin and Harrison, 2001), as presented in this thesis. H-D exchange, particularly in conjunction with other techniques such as mutagenesis or mass spectrometry, continues to be widely used experimental probe of protein folding (Perrett et al., 1995).

1.7 Protein Flexibility

Conformational flexibility is an intrinsic and necessary property of protein structures (Jacobs et al., 2001). The very concept of "folding" a protein implies that a structural deformation is required to change from a denatured state to a native conformation. The importance of native-state flexibility has been discussed for over 20 years (Huber, 1979; Brooks et al., 1988), especially in the context of enzymes (Gavish, 1986). Multiple experimental techniques, such as fluorescence quenching of tryptophan residues, circular dichroism (CD) spectroscopy, NMR, and hydrogen-deuterium exchange (H-D exchange) have been applied to probe native state fluctuations. Catalytic mechanisms can require a broad range of flexibility from individual side-chain rotations (Cobessi et al., 2000), to small loop movement as in the flaps of HIV protease (Venable et al., 1993), up to the concerted motion of multiple domains, as in ATP synthase (Sabbert et al., 1997). Regulation of proteins via allosteric mechanisms has also been shown to require structural flexibility (Bustos-Jaimes et al., 2002).

Theoretical approaches to predicting flexibility in the native states of proteins arose as the number of high-resolution crystal structures increased. Molecular dynamics (MD) simulation is perhaps the most straightforward method to probe the structural flexibility observed in proteins, using techniques such as essential dynamics (Amadei et al., 1999, 1993). However, these methods are computationally expensive. Running a simulation long enough, even for small proteins, is prohibitive. Alternative methods have been developed in which flexibility can predicted empirically based on a combination of structural and chemical features, such as atomic density and the distribution of polar and nonpolar residue type (Ragone et al., 1989), or based on sequence alone (Bhaskaran and Ponnuswamy, 1988). These methods have met with limited success (Vihinen et al., 1994), suggesting that sequence and/or chemistry alone are not the sole discriminants of protein flexibility. Perhaps the explicit interactions between residues needs to be taken into account.

Methods that include structural information, specifically the chain topology, when identifying flexible regions in proteins fall into two broad categories. The first rely on comparison between structures of the same protein in different conformations. These conformations can arise from several sources, such as alternative crystal packings or ligand-free versus ligand-bound states. Comparison of the structures is accomplished using various geometric parameters, such as the difference between inter-C_α distances (Nichols et al., 1995) or differences in dihedral angles (Korn and Rose, 1994). These methods provide solid evidence for the location of flexible regions in protein, but are severely limited in that multiple structures must be available. Thus possible alternative conformations are not probed. The second class of algorithms designed to predict flexibility in proteins using structural information is based on physical forces. MD falls into this category, as does normal-mode analysis (NMA). NMA was first applied to proteins in the 1980's (Go et al., 1983; Brooks and Karplus, 1983). It is believed that the lowest frequency vibrational modes, or "soft modes", represent the largest fluctuations in the structure, and therefore are associated with functionally relevant motion. Interestingly, much of the verification that normal mode analysis gives reasonable results came through comparison of NMA results to those of crystal structures comparisons mentioned above (Thomas et al., 1996; Ma and Karplus, 1997). As in MD, NMA is restricted by the size of the protein being analyzed, due to a computationally intensive step (diagonalization of the Hessian matrix), although clever alternatives have arisen to help overcome this limitation (Go et al., 1983; Brooks et al., 1995). Although faster than MD, NMA is subject to the same criticisms: lengthy computation time, and reliance on empirical force fields.

Rigidity theory provides an alternative technique to measure flexibility. The mathematics of rigidity theory allows one to describe the deformability of any structure, given internal constraints on the structure. The key to applying rigidity theory to real life problems lies in properly representing a physical structure in mathematical terms, such that conditions required by the theory hold true. For molecular structures, a proper representation lies in accurately representing the bond forces that hold the atoms together. For glass networks, which have been extensively studied by such techniques (), the dominant bond forces are the covalent bonds. Flexibility analysis of these networks has been successful using rigidity theory, and led to accurate prediction of their material properties. In particular, the mean coordination of the networks has been identified as the relevant structural reaction coordinate or order parameter. The variation in the flexibility of glass networks, as a function of the mean coordination, can accurately define the phase transition between rigidity and flexibility in these structures.

Recently, an approximate representation of proteins has been developed such that protein flexibility can be analyzed using rigidity theory (Jacobs et al., 1999, 2001; Rader et al., 2001). These advances have been embodied in the computer program FIRST (Floppy Inclusions and Rigid Substructure Topography) which identifies each bond in a protein as being rotatable (flexible) or nonrotatable (rigid). Furthermore, coupling between flexible and rigid bonds allows decomposition of a protein into rigid regions and flexible regions, and these flexible regions have been shown to correlate well with structurally significant motion in many proteins (Jacobs et al., 2001). The bulk of the work presented in this thesis (chapters 3 and 4) builds on FIRST analysis, and a description of the program is given in chapter 3.

1.8 Work Presented in This Thesis

The motivation for this thesis has been to address the hypothesis that native-state topology encodes information about protein folding. Chapter 2 presents an analysis of the geometry of secondary structure packing in a set of nonhomologous protein structures, specifically α-helices interacting with β-sheets. The results can be divided into two categories, those interactions in which a dipole is present in the sheet, and those interactions in which no dipole is present in the sheet. For the latter case, no preferred packing geometry is observed. However, for helix–sheet interactions in which a dipole is present in the sheet, a strong preference is observed for the helix to align its dipole in the opposite direction relative to the sheet dipole.

Chapter 3 presents an introduction to the flexibility analysis of proteins using the program FIRST. Comparison of native-state flexibility results to experimentally observed flexible regions show good correlation, validating that the bond forces in a protein can be accurately modeled. A simple model of protein folding is assumed in which hydrophobic collapse leads to a compact structure, which is then stabilized by specific hydrogen bonds as the protein folds to the native state. The key to this simple view of folding is that hydrogen bonds form during folding, and therefore break during unfolding. Protein unfolding is simulated by breaking hydrogen bonds, in an energy-dependent manner, in a method called hydrogen bond dilution. The changes in protein flexibility that occur as hydrogen bonds are diluted from the structure are tracked and related to corresponding experimental observables of protein unfolding. The mean coordination of a protein at any given point during hydrogen bond dilution is shown to be a useful reaction coordinate for the unfolding of a protein. Chapter 4 presents a method for predicting protein folding cores from these hydrogen bond dilution results, and the correlation with experimentally observed folding cores from H-D exchange experiments is shown to be very good. A summary and perspective of the results is given in chapter 5, with a qualitative interpretation of the hydrogen bond dilution results. Also, potential future directions of FIRST analysis are discussed, including methods to predict Φ values for a protein.

Chapter 2

An Analysis of Helix-Sheet Packing Geometry in a Set of Nonhomologous Protein Structures

2.1 Abstract

Here I present an analysis of the packing geometry observed between α-helix and β-sheet secondary structures. The structures are represented as finite size vectors fit to the C_α coordinates. A packing interaction is defined by any helix-strand pair within 13.0Å of each other, and whose line of closest approach intersects both finite vector representations of the secondary structures. These criteria ensure that the packing geometry can be described by a single dihedral angle, Ω. A strand that is interacting with a helix can be in one of five orientations, depending on a parallel or antiparallel hydrogen bonding pattern with respect to its neighbors, and whether it is the terminal strand in a sheet. α-helices packing against β-sheets were searched for in a set of 1316 proteins non-homologous protein crystal structures determined at better than 2.2Å resolution. From this set, helix-sheet packing

interactions were found in 391 (29.7%) proteins. Bias in the distribution of Ω angles is accounted for by dividing the observed distribution by the expected uniform random distribution of packing angles that exhibits a $\sin\Omega$ dependence. For most helix-strand pairs no preferred Ω packing angle is observed. However, for helix-strand interactions in which the strand is parallel to both of its neighboring strands, we see a strong preference for the helix to align antiparallel to the strand, with a packing Ω angle near 180°.

2.2 Introduction

The mechanism by which a protein folds from a denatured state to a folded conformation is an intensely studied, unsolved problem in the natural sciences. Many models describing the reaction have been proposed and supported by experimental evidence, and one single model may not hold for all known proteins. In one particular folding model, known as the framework or diffusion–collision model (Karplus and Weaver, 1994), a subset of secondary elements form partial or complete structures early in the folding reaction. These substructures then interact forming a super-secondary structure that is representative of the transition state ensemble, and folding then continues to the native state. Both mutagenesis (Kippen et al., 1994) and hydrogen-deuterium out-exchange (H-D exchange) experiments (Perrett et al., 1995) have shown the framework model to be a valid scenario for the folding of barnase, in which the N-terminal α-helix packs against several strands of the C-terminal β-sheet to form the folding core.

Assuming the framework model is a valid scenario for protein folding, it is an interesting question to ask whether secondary structures prefer to adopt specific geometries when they coalesce. Research on observed packing geometry for secondary structures extends back 20 years. In one of the earliest studies by Chothia et al., (1981), they analyzed 50 helix-helix packing interactions from 10 protein structures. The results led them to propose the "ridges into grooves" model for helix-helix interactions, in which helix pairs prefer to adopt specific geometries so as to avoid steric overlap between the side chains. Since that time, advances in computer technology have allowed for not only an invaluable increase in the number of protein crystal structures available, but also the development of algorithms to parse out proteins with homologous sequences whose structures may bias the data. More recent studies have expanded the analysis of helix-helix packing interactions to a dataset of 687 interactions from 220 protein structures with less than 35% sequence identity and better than 2.2Å resolution (Walther et al., 1996).

In these studies, and similar experiments, the secondary structures are most often represented by best-fit lines though the C_α coordinates of the residues in each structure. The geometry of the interaction can then be uniquely expressed by a distance and two angles. The key angle, named Ω, is defined as the dihedral angle formed by two interacting structures and the line of closest approach between them (Figure 2.1). Initially, observed distributions of Ω packing angles for helix-helix interactions exhibited distinct peaks (Walther et al., 1996). However, Bowie (1997), with further developments by Walther et al. (1998), demonstrated that the expected uniform random distribution of Ω is biased towards angles near 90°. As described in (Walther et al., 1998), there are simply more ways to pack two helices at 90° than there are to pack them at 10°. When this bias was taken into account, the observed peaks in the helix-helix Ω angle distribution were significantly attenuated. However, parallel studies in which specific details of the helix-helix interface were measured as a function of Ω angle yielded new correlations. These analyses continue to provide useful information in the field of protein design.

Measuring the packing geometry for helix-sheet packing interactions has proven a more difficult task than helix-helix interactions due to the non-symmetric structure of the β-sheet. Early work by Janin and Chothia (1980) stated that the Ω angle for a helix packing against a sheet should be near 0°, indicating that only small angles allowed for complementary packing of the helix side chains within the groove created by a twisting β-sheet. This observation of near parallel helix-sheet packing was further supported by work published by Cohen et al. (1982) a few years later. A theoretical study by Chou et al. (1985), in which low energy helix-sheet conformations were predicted, further agreed that a helix-strand packing Ω angle near 0° was a favorable interaction. An analysis of 163 helix-sheet packing interactions observed from proteins of known structure showed a predominate peak near 0°. In all of these studies the packing angles were measured by approximating inherently twisted β-sheets as a plane. Also, the Ω angle was measured on the range, $-90° \leq \Omega \leq 90°$, therefore the N–terminal to C–terminal direction of the structures was not taken into account.

In this chapter the analysis of helix–strand packing interactions is extended. Five possible strand orientations within a sheet are defined depending on the strand direction relative to its neighbors (parallel versus antiparallel) and present the observed distributions of Ω packing angles, with geometric bias taken into account, for each of the five cases. The Ω

packing angle is measured over the range, $-180° \leq \Omega \leq 180°$, to observe any correlation between parallel/antiparallel packing and Ω angle. We use a coordinate transformation to measure the Ω packing angle that does not require fitting a plane to the β-sheet. The results indicate a strong preference for an helix to pack antiparallel to a sheet composed of parallel strands, indicating that the dipole-dipole interaction may be important for this type of supersecondary structure.

2.3 Methods

2.3.1 Protein Dataset

The culled Protein Data Bank (PDB) list (Hobohm et al., 1993) from March 8th, 2002 was used to create a dataset of protein crystal structures that had less than 20% sequence identity, better than 2.2Å resolution, and R-factors below 0.2. Only proteins whose PDB files contained HELIX and SHEET records were included. The final dataset consisted of 1316 proteins.

2.3.2 Representing Secondary Structures as Vectors

The residues forming regular secondary structure in each protein structure were identified according to the HELIX and SHEET records in each PDB file. Helices were required to have at least seven residues, corresponding to two complete turns of a regular α-helix. Strands were required to have at least 3 residues for proper fitting of a vector to the C_α

coordinates. Occasionally, the ordering of strands in a PDB file is not consistent with the order they are observed in. For each sheet identified, the closest distance between neighboring strands was measured, and any sheet that had a closest interstrand distance greater than 5.0Å was visually checked to see that the strands were in the proper order. Errors in strand order within a PDB file were fixed manually.

The α-carbon positions of each residue in a helix and strand were used to compute the best fit line through a given structure using a parametric least squares algorithm (Christopher et al., 1996). Because an individual strand can severely deviate from linearity, the degree to which each strand bowed was also computed. Strand bow was calculated using the following equation:

$$Bow = \frac{||\mathbf{m}||}{||\mathbf{d}||} \tag{2.1}$$

Where **d** is the distance between the first C_α and the last C_α in the strand and **m** is the distance from the C_α in the middle of the strand projected onto **d**. If the strand contained an even number of residues, the average position of the middle two C_α's was used to compute **d**.

2.3.3 Identifying a Pair of Interacting Secondary Structures

Each helix in a protein is represented in 3D by a vector **h**, and each strand is represented by a vector **s**. These vectors, and their corresponding secondary structures, are shown graphically in Figure 2.1. The distance between the midpoints of **h** and **s** is defined as *MD*. The point of closest approach between **h** and **s** was computed using equations described by

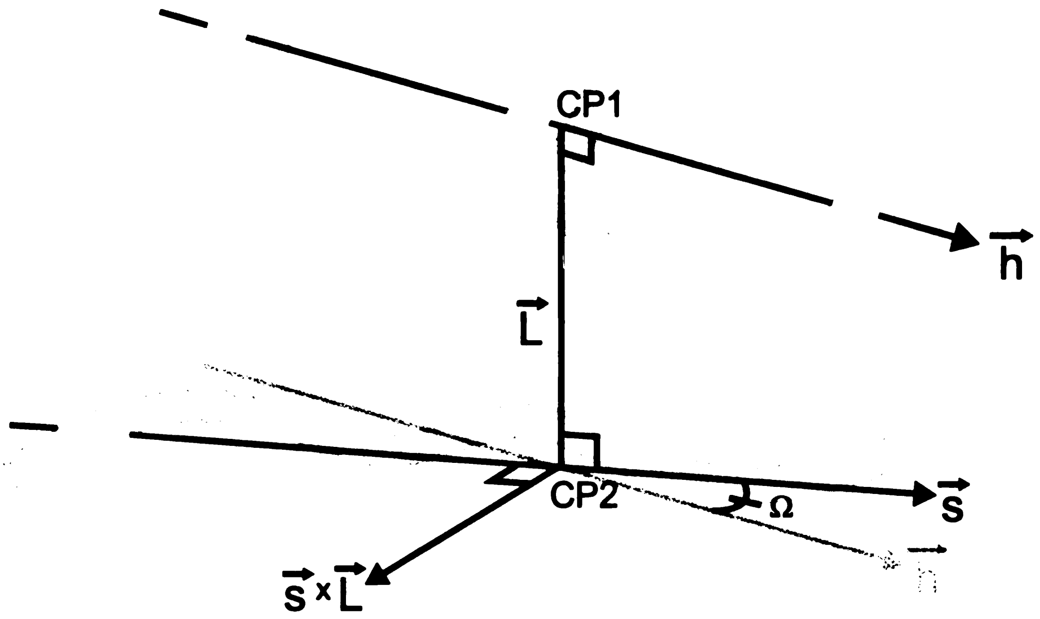

Figure 2.1: Graphical representation of a helix-sheet packing geometry. The helix and the strand are shown as light gray ribbons. The vector representations of the helix, $\vec{h}$, and the strand, $\vec{s}$, are shown as black arrows. The line of closest approach is labeled $\vec{L}$, and intersects the helix at a point labeled CP1, and the strand at the point CP2. Because $\vec{L}$ is perpendicular to $\vec{s}$, their cross product, $\vec{L} \times \vec{s}$ is perpendicular to both. The Ω packing angle is measured as the angle between $\vec{s}$ and the projection of $\vec{h}$, shown as a light gray arrow, onto the plane defined by $\vec{s}$ and $\vec{L} \times \vec{s}$

Chothia et. al. (1981). These equations compute two scalar quantities, cp1 and cp2. The point on the helix vector **h** that is closest to strand **s** is defined as CP1 and the point on the strand vector **s** that is closest to helix **h** is defined as CP2. Examples of CP1 and CP2 are shown graphically in Figure 2.1.

The scalar quantities cp1 and cp2 can be less than 0.0 or greater than 1.0, in which case the line of closest approach does not intersect with one or both of the secondary structures. Likewise, if both cp1 and cp2 are between 0.0 and 1.0, then the line of closest approach intersects both secondary structures. The vector **L** is defined as the line of closest approach between **h** and **s** and is computed as, **L** = CP1 - CP2. The length of the closest distance between **h** and **s** is defined as *CD*, and is computed as the magnitude of the vector **L**. As seen in Figure 2.1, **L** can be viewed as both the projection of CP1 onto **s**, and the projection of CP2 onto **h**, and consequently **L** is orthogonal to both **h** and **s**. Using the measured quantities *MD*, *CD*, CP1, and CP2, a helix is defined as interacting with a strand if the following criteria are met:

1. $MD \leq 20.0$Å.
2. $CD \leq 13.0$Å.
3. $CD_j \leq 13.0$Å; $CD_k \leq 13.0$Å, where j and k are the two closest strands to **s**.
4. $0.01 \leq$ CP1, CP2 ≤ 0.99
5. Bow ≤ 0.25. Also the neighboring strands (or neighboring strand if **s** is the last strand in a sheet) must have Bow ≤ 0.25.

Criteria 1 and 2 are designed to limit the search of helix-strand pairs within a given structure to those that are near each other in 3D. Initially larger cutoff values for *MD* were chosen, however all of the additional helix-strand pairs identified failed to meet any of the subsequent criteria, and were discarded. Criterion 3 states that if a helix is interacting with strand *i*, then it should also be within 13.0Å of strands *j* and *k*, which are the two nearest strands to strand *i* within the sheet. This criterion was implemented to discard cases where the helix is interacting with the hydrogen bonding edge of one of the strands, and not the side-chain face of a sheet. Criterion 4 ensures that surface of the helix is packing against surface of the sheet to which the interaction strand belongs. If CP1 and CP2 are allowed to be less than 0.0 and/or greater than 1.0, it is possible that the helix and the strand are oriented perpendicular to each other (that is, the helix vector is normal to an approximate sheet plane). Criterion 4 also ensures that the line of closest approach, **L**, will be perpendicular to both **h** and **s**, and thus **h** and **s** will be coplanar when projected onto a common plane normal to **L**. Criterion 5 will discard interactions in which the interaction strand or its neighbor(s) are excessively bowed. Excessive bow can result from the occurrence of a β-bulge or the presence of a residue in the strand that has Φ,Ψ angles that are outside the β region of the Ramachandran plot (Salemme, 1983).

2.3.4 Assigning Local Strand Orientation

The orientation of a strand relative to its hydrogen bonded neighbor(s) can be assigned using the "sense" field assigned to columns 39–40 of the SHEET record in a PDB file. This field gives the N-terminal to C-terminal direction of a strand with respect to the previous strand in the sheet. The first strand in the sheet is assigned a sense of 0. If the second strand is parallel to the first strand, it is assigned a sense of 1, if it is antiparallel, it is assigned a

Table 2.1: Assigning a unique orientation value for each strand in a sheet. The left-hand column shows the strand we are computing an orientation value for, depicted as a double-lined arrow, and its neighbors. Orientations −1 and 1 correspond to strands at the end of a sheet.

Strand	Orientation
Order	Value
↓ ⇑ ↓	−2
⇑ ↓	−1
↑ ⇑ ↓	0
⇑ ↑	1
↑ ⇑ ↑	2

sense of −1. The orientation value of strand i is computed as the sense of strand i plus the sense of strand $i+1$. For example, if the second strand in a sheet is parallel with respect to the first one, it will have a sense value of 1. If the third strand in the sheet is parallel with the second strand, it will also have a sense value of 1. The orientation value of the second strand is the sum of these two values, $1 + 1 = 2$. Table 2.1 lists the five possible orientation values that can occur for a strand in a sheet. The left hand column shows a cartoon representation of a portion of a sheet. The strand we are computing an orientation value for is shown as a double-lined arrow. The right hand column lists the orientation value assigned to each case. Orientations −1 and 1 correspond to stands at the end of a sheet.

2.3.5 Measuring the Packing Geometry of a Helix–Strand Interaction

Because the line of closest approach, **L**, between the helix and the strand is perpendicular to both **h** and **s**, the packing geometry between the two structures can be defined by a single dihedral angle, Ω. The angle is computed by orienting the vector **s** along the positive x-axis, with the N-terminal end positioned at the origin. The system is then rotated about the x-axis such that **L** lies along the positive z-axis. The result of the final transformation is shown in Figure 2.1. In this orientation both **h** and **s** are coplanar with the **L-s** plane (which can also be viewed as the x-y plane), and Ω can be computed as the angle between the transformed coordinates of **h** and **s** using equation 2.2.

$$\Omega = cos^{-1}\left(\frac{\mathbf{h}\cdot\mathbf{s}}{||\mathbf{h}||\,||\mathbf{s}||}\right) \tag{2.2}$$

2.3.6 Measuring Local Sheet Twist

The local sheet twist was measured to observe the degree to which sheet twist effects the Ω packing angle for helix-strand interactions. For a given helix-strand interaction, the vectors **s** and **L** are orthogonal, and can be used as a basis for a 2-dimensional subspace, W = {**s**,**L**}. The two closest strands to **s**, **a** and **b**, are then projected onto W using equations 2.3 and 2.4.

$$proj_W\mathbf{a} = \langle \mathbf{a}, \hat{s}\rangle \hat{s} + \langle \mathbf{a}, \hat{L}\rangle \hat{L} \tag{2.3}$$

$$proj_W\mathbf{b} = \langle \mathbf{b}, \hat{s}\rangle \hat{s} + \langle \mathbf{b}, \hat{L}\rangle \hat{L} \tag{2.4}$$

where $\hat{L}$ and $\hat{s}$ are unit vectors in the direction of **L** and **s**, respectively. The twist angle, T, is then found by using equations 2.5, 2.6, and 2.7, which is the average of the angles the projected vectors, **a** and **b**, make with strand **s**, in the plane of W.

$$\tau_1 = cos^{-1}\left(\frac{proj_W \mathbf{a} \cdot \hat{s}}{||proj_W \mathbf{a}||}\right) \tag{2.5}$$

$$\tau_2 = cos^{-1}\left(\frac{proj_W \cdot \hat{s}}{||proj_W \mathbf{b}||}\right) \tag{2.6}$$

$$T = \frac{\tau_1 + \tau_2}{2} \tag{2.7}$$

2.3.7 Normalizing the Helix–Sheet Ω Angle

The Ω packing angle measured for packing secondary structures has an inherent bias towards angles near 90°. This bias was originally shown for helix-helix interactions by Bowie (1997), and further developed by Walther et al. (1998). The bias arises from non-uniform probability distribution in the 10° bin sizes used to tabulate the Ω angle results. For example, two vectors of unit length forming a 90° angle generate an area, $A_1 = \sin(90°) = 1.0$. Likewise, two unit vectors forming a 45° angle create an area, $A_2 = \sin(45°) = 0.7071$. In the case of A_1, if we keep one of the unit vectors fixed in space, the other vector can position its endpoint anywhere within the area A_1 and keep $\Omega = 90°$. It can be readily seen that $A_1 > A_2$, and therefore there are more ways in which two unit vectors can form a 90° angle than there are ways to form a 45° angle, and the observed bias is proportional to $\sin \Omega$. To eliminate this bias from our data, the number of observed occurrences for each 10° bin was divided by the number expected from the uniform random distribution. The number of occurrences expected for each 10° bin was computed using equation 2.8. In these unbiased data a value of 1.0 indicates that the given angle occurs just as often as we would expect it to if secondary structures packed together at random angles. Values less than 1.0 indicate unfavorable packing angles, and values greater than 1.0 indicate a preferred packing angle.

$$\int_{\Omega_1}^{\Omega_2} \sin \Omega \, d\Omega \tag{2.8}$$

2.4 Results

2.4.1 Helix-Strand Ω Packing Angle as a Function of Strand Orientation

For each helix-strand packing interaction the strand can be in one of five orientations depending on whether it is hydrogen bonded in parallel or antiparallel with respect to its neighbors, and whether or not it is the first or last strand in the sheet. The distributions of observed Ω packing angle, divided by the expected distribution, is presented in Figure 2.2 for each of the five possible strand orientations. A cartoon representation of the strand orientation within the sheet is shown in the upper left of each panel (see also Table 2.1). Orientations of 1 and -1 correspond to strands with only one neighbor, as they are at the end of a sheet. For each 10° bin in the plots, a value of 1 indicates that the number of observed Ω angles in that bin occurred just as often as would be expected randomly. Values less than 1.0 suggest the packing angle occurs less often than expected, and values greater than 1.0 indicate preferred packing angles. Ω angles in the range $90° \leq \Omega \leq -90°$ represent an interaction in which the N-terminal to C-terminal direction of the helix is parallel to the direction of the strand. If $\Omega \leq -90°$ or $\Omega \geq 90°$, the helix is packed antiparallel to the

Figure 2.2: Distribution of helix-sheet Ω packing angles for each of the five strand orientations. A cartoon representation of the strand orientation is shown in the upper left of each histogram. The number of observed occurrences for each 10° bin was divided by the number expected from a uniform random distribution. Here, a value of 1.0 indicates that the given Ω angle occurs just as often as would be expected by random. Values less than 1.0 are unfavored, and values greater than 1.0 indicate preferred packing angles. Helix–strand interactions in which the strand is in an orientation of 1 or 2 show a strong preference to pack antiparallel (Ω angles near 0.0° or 180°). Strands in orientations 0, −1 and −2 shown no strong angle preference when packing with a helix.

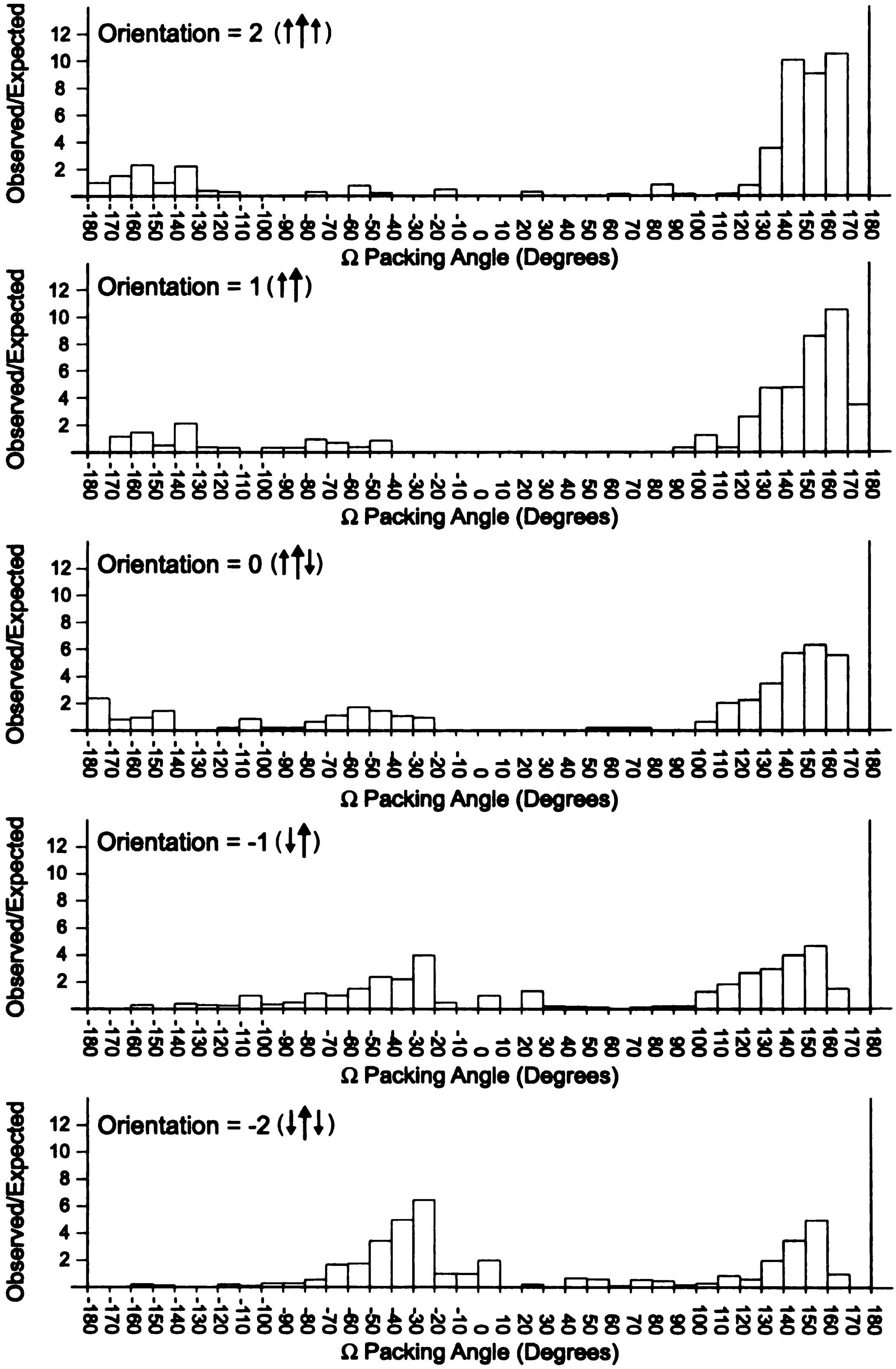
Orientation = 2 (↑↑↑)
Orientation = 1 (↑↑)
Orientation = 0 (↑↑↓)
Orientation = -1 (↓↑)
Orientation = -2 (↓↑↓)
Observed/Expected
2
4
6
8
10
12
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
Ω Packing Angle (Degrees)

Figure 2.2

strand. The top three panels in Figure 2.2 show that for type 0, 1 and 2 strand orientations, there is an increasing preference for the helix to pack antiparallel to the strand. Type 2 strand orientations exhibit the strongest preference, with almost no parallel packing interactions observed in real proteins. Figure 2.3 shows an ideal type 2 helix-strand interaction present in the protein IIB cellobiose from *E. coli* (PDB code: 1iib) (von Montfort et al., 1997). The arrows on the strands (colored yellow) point in the N- to C-terminal direction. The strand determined to be interacting with the helix is the second one from the left, and it can be seen that this strand is parallel to both its neighbors. The N- to C-terminal direction of the helix is from the upper-right to the lower-left. The Ω packing angle for this interaction is 118.04°.

Type -1 and -2 strand orientations show no preference for parallel versus antiparallel packing, however there is a preference to pack at angles near $-25°$ and $155°$, which represent the same angle if you disregard the N- to C-terminal direction of the structures.

2.4.2 Ω Packing Angle as a Function of Local Sheet Twist

For each helix-sheet interaction found, the geometry is measured relative to a single strand in the sheet that is closest to the helix. The local twist of the sheet is then measured by using the strand interacting with the helix, and its neighbors. A scatter plot of local sheet twist versus Ω packing angle is shown in Figure 2.4. The points are colored according to the orientation value of the interaction strand. Correlation coefficients between T and Ω were computed for each of the five strand orientations, and the results are shown in Table 2.2. No correlation between sheet twist, T, and helix-sheet Ω packing angle was observed

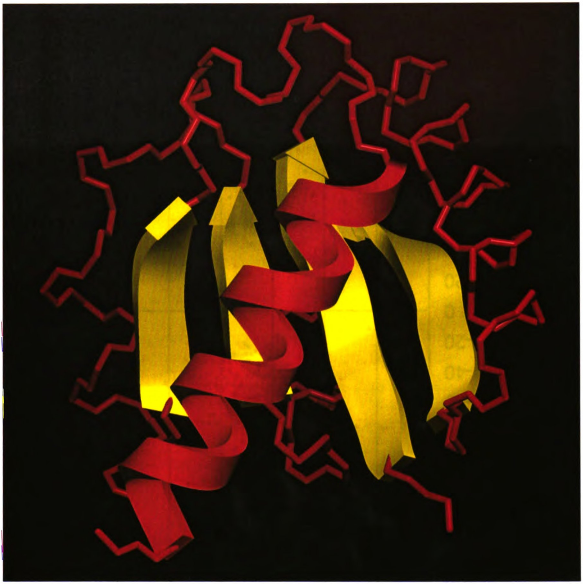

Figure 2.3: Example of a helix-sheet packing interaction found in the protein IIB cellobiose from *E. coli.* The geometry of the interaction was measured relative to the second strand from the left, which has an orientation value of 2. The N- to C-terminal direction of the strands is indicated by the arrow heads. The N- to C-terminal direction of the helix is from upper-right to lower-left. The measured Ω angle 118.04°

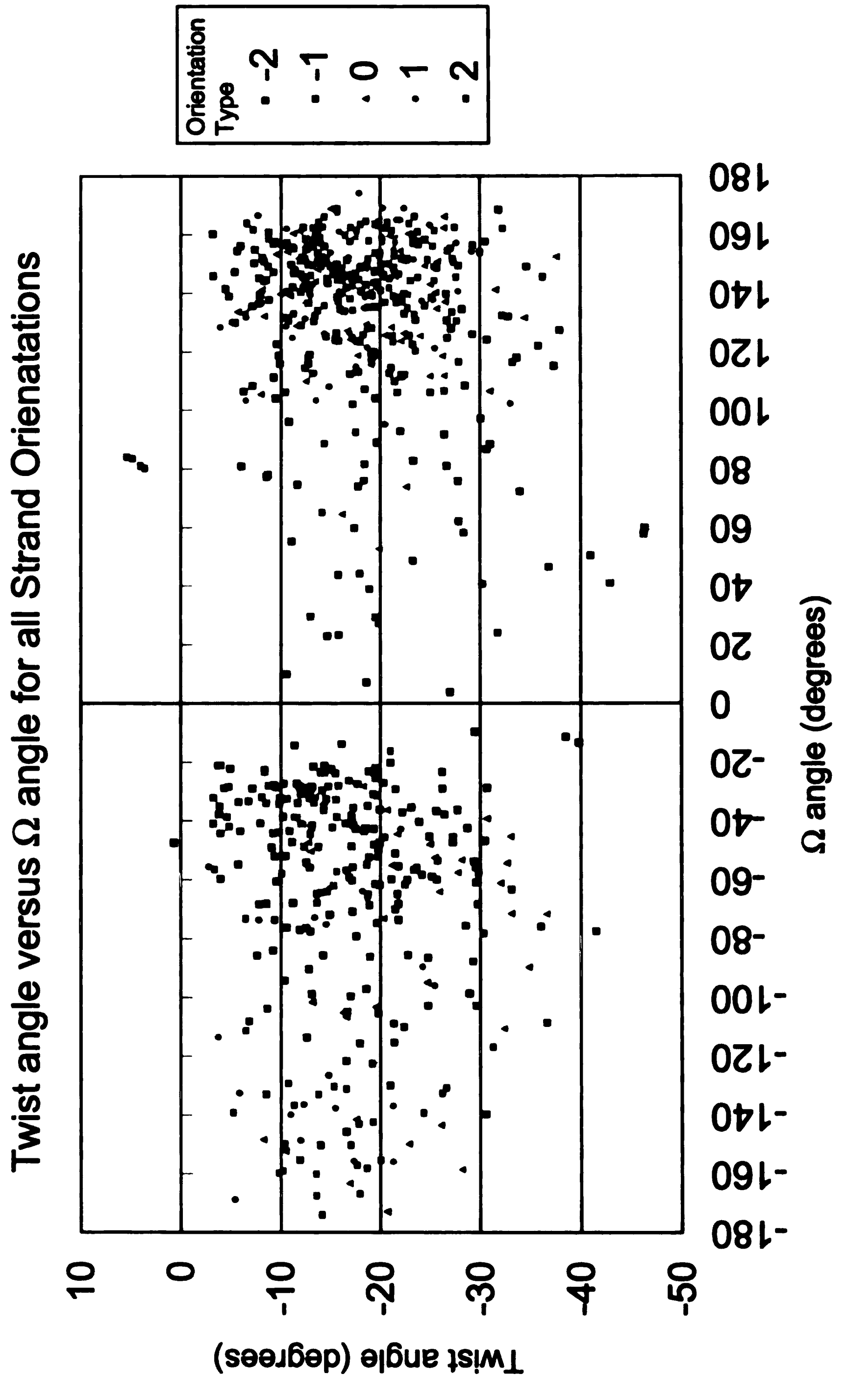

Figure 2.4: Scatter plot of local sheet twist versus Ω packing angle for each of the five strand orientations. The right-handed twist common to most β-sheets is indicated by the large number of negative twist angles observed. No clear correlation between twist angle and Ω angle was observed for any of the five strand orientations.

Table 2.2: Correlation coefficients between sheet twist, T, and Ω packing angle for all five possible strand orientations.

Strand Orientation	Correlation Coefficient
-2	0.0761
-1	-0.0289
0	0.2304
1	-0.2533
2	-0.2525

for any of the five possible strand orientations.

2.5 Conclusions

The coiling and right-handed twist associated with β-strands depends on the Φ/Ψ values of the individual residues (Chothia, 1983). These Φ/Ψ values in turn depend on the type of residue at any given position and the hydrogen bonding pattern between adjacent strands. Ideally, a flat (uncoiled) β-sheet, like the one proposed by Pauling and Corey (1951), would have optimal interchain hydrogen bonding geometry. However, this also required the residues in the sheet to adopt a perfect 2-fold helix symmetry, which is energetically unfavorable. To minimize the energetic frustration, residues within a strand adopt Φ/Ψ angles that lead to a right-handed twist, resulting in poor hydrogen bonding complementarity between strands. To realign the hydrogen bond donors and acceptor of adjacent strands, successive residues in a strand adopt different Φ/Ψ values producing twisted, coiled β-strands, and subsequently giving rise to a twisted β-sheet. This compromise between maximizing the number of hydrogen bonds formed and minimizing the conformational energy of each strand has been predicted theoretically and observed in proteins of known structure (Salemme, 1983).

An individual sheet can consist of all parallel, all antiparallel, or mixed parallel and antiparallel strands. This diversity in hydrogen bonding pattern, along with varying amino acid composition, can lead to sheets in which the twist and coil vary depending on where in the sheet you are looking. Here I presented a novel geometric definition for the measurement of the local twist of a β-sheet. The hypothesis was that as the twist of a sheet deviates farther from planarity, steric interactions would cause the helix to turn, creating a larger Ω packing angle to better fit in the groove formed by the strands of the sheet. A plot of local sheet twist versus Ω would then reveal a correlation between the degree to which a sheet twists, and the angle at which the helix will pack against the sheet. Table 2.2 and Figure 2.4 clearly indicate that there is no correlation between our measure of sheet twist and Ω packing angle. This can arise from several reasons, most likely, due to the side-chain conformations. Side chains can vary in size, and most exhibit conformational flexibility. By not taking into account the specific interactions occurring in the helix-sheet interface, we assume that there are specific side chains within the interface between all observed helix-sheet pairs. The hypothesis also assumes that the surface created by the side-chains, the surface to which a helix is actually interacting, can be approximated by the backbone atoms of the strands. This appears not the case, and an extended analysis of these interfaces, similar to what has been reported for helix-helix packing interfaces, is warranted.

In the distributions of Ω angle for each strand orientation shown in Figure 2.2, only those orientations where the interacting strand is parallel to its neighbors show a strong Ω angle preference. In these cases, orientations 1 and 2, the helix prefers to pack antiparallel to the strand, near 180°. One possible explanation for seeing a preference in these orientations and not the others is the presence of a net dipole arising from the hydrogen bonding pattern in parallel strands. The hydrogen bonds between parallel strands make a 20° angle with respect to the N to C direction of the protein backbone (Figure 2.5A). This leads to a net dipole moment of about 1.15 Debyes (Hol et al., 1981). If a helix-strand dipole interaction is occurring, we would expect the helix to orient its dipole in the opposite direction as the strands. This expected antiparallel packing interaction is indeed what is observed. For the remaining strands in orientations 0, -1, or -2, the interacting strand is antiparallel to one or both of its neighbors. The hydrogen bonds between antiparallel strands are nearly perpendicular to the protein backbone (Figure 2.5B), and a negligible net dipole moment is produced. In these helix-strand interactions, the dipole would not be expected to play a role, and we observe no strong preference for Ω angle.

Another possible explanation for the observed Ω angle packing preference in type 2 and 1 strand orientations is the structure of the sheet. Sheets composed entirely of parallel strands have been shown to be flatter and less flexible than purely antiparallel sheets (Salemme, 1983), increasing the net dipole moment relative to a highly twisted sheet. Also, purely parallel sheets are uncommon, and tend to be buried within protein of α/β architecture (Chothia, 1983). In these cases, optimizing the packing interaction between a helix and a sheet would be beneficial to maintaining a compact protein structure.

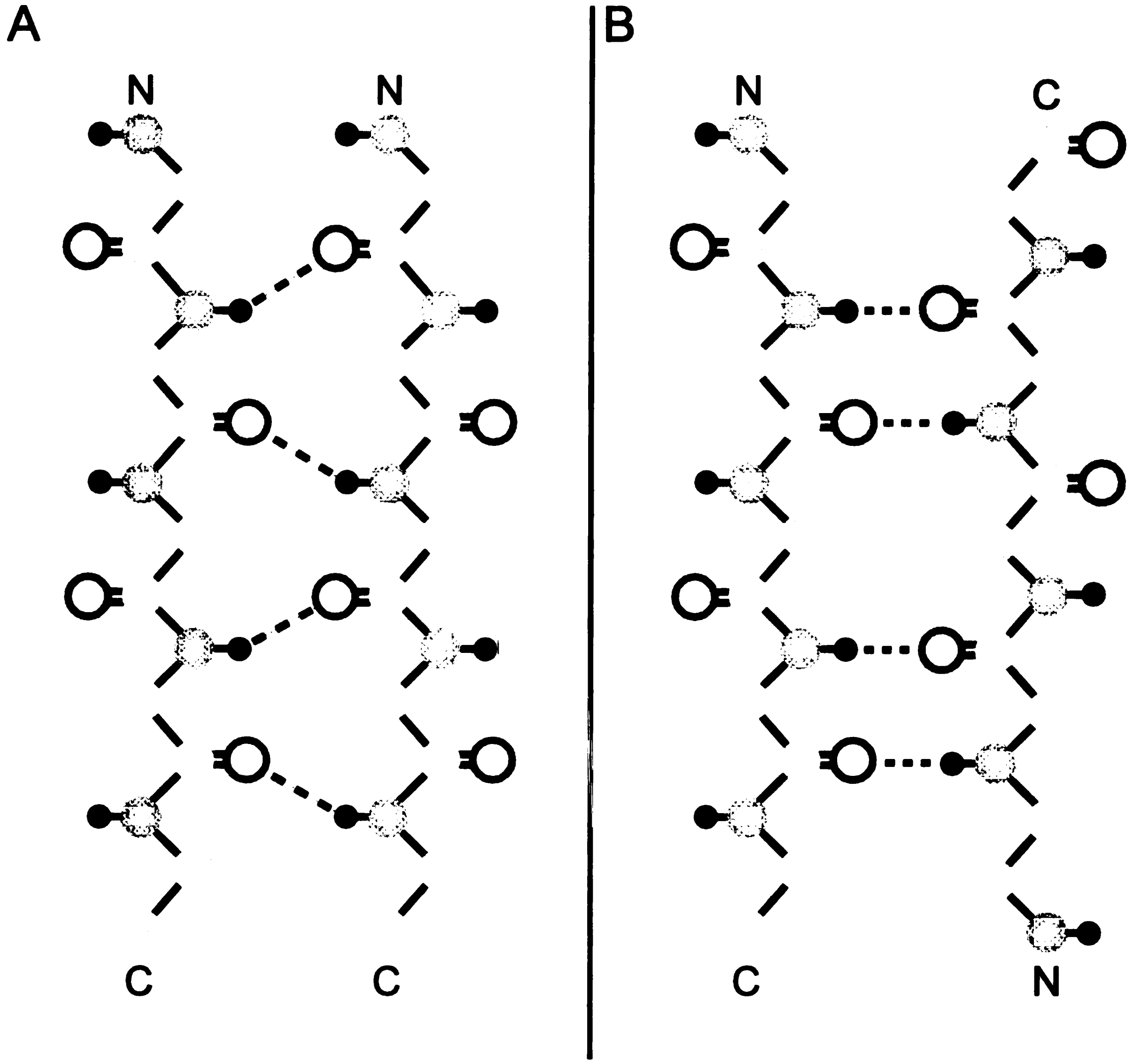

Figure 2.5: Hydrogen bonding pattern for parallel and anti-parallel β-strands. Carbons are depicted as light gray spheres, nitrogen as dark gray spheres, oxygen as open spheres, and hydrogen as small black spheres. Hydrogen bonds are shown as dashed lines between the main-chain oxygen and hydrogen atoms of adjacent strands. A. The hydrogen bonds between parallel strands form a 20° angle with respect to the protein backbone resulting in a net dipole in the C→N direction. B. The hydrogen bonds between antiparallel strands are nearly perpendicular to the protein backbone, and no net dipole is produced.

Chapter 3

FIRST Flexibility Analysis and Hydrogen Bond Dilution as a Method to Simulate Thermal Denaturation

Research presented in this chapter is based on work that has appeared in the following publications:

B. M. Hespenheide, A. J. Rader, M. F. Thorpe, and L. A. Kuhn. Identifying protein folding cores from the evolution of flexible regions during unfolding. *J. Mol. Graph. Model.*, In press.

A. J. Rader, B. M. Hespenheide, L. A. Kuhn, and M. F. Thorpe. Protein unfolding: Rigidity lost. *Proc. Natl. Acad. Sci.*, 99:3540-3545, 2002

3.1 Abstract

Here I present the application of a novel computational technique, FIRST is presented for measuring flexibility in protein structures. The flexibility present in a molecular structure is a property that depends upon the bond forces present in the structure. FIRST treats bond forces, such as covalent and hydrogen bonds, as distance constraints that put restrictions on the conformational space available to the atoms in a protein. Once all the bond forces have been identified and modeled, FIRST computes the resulting flexibility, and produces a rigid cluster decomposition (RCD) of the protein structure. The RCD reports for each bond in a protein whether it is free to rotate (flexible) or not free to rotate (rigid). The RCD for the native state of HIV protease, in both ligand bound and unbound forms, correlates well with experimentally identified rigid and flexible regions in this protein. Also, we present a method for mimicking thermal denaturation in a protein based on dilution of the hydrogen bonds and salt bridges within a protein. We show that the unfolding of a protein can be viewed as a rigid to flexible transition, and this transition can be tracked by observing how the flexibility of a protein changes at each step during hydrogen bond dilution (simulated thermal denaturation). A novel graphical representation is presented for displaying the data. Finally, the transition state is determined from the inflection point in the change in the number of independent bond-rotational degrees of freedom, or *floppy modes*, of the protein as its mean atomic coordination decreases. The first derivative of the fraction of floppy modes as a function of mean coordination is similar to the fraction-folded curve for a protein as a function of denaturant concentration or temperature. The second derivative, a specific heat-like quantity, shows a peak around a mean coordination of $\langle r \rangle = 2.41$ for 26

diverse proteins. As a protein denatures, it loses rigidity at the transition state, proceeds to a state where just the initial folding core remains stable, then becomes entirely denatured or flexible. This universal behavior for proteins of diverse architecture, including monomers and oligomers, is analogous to the rigid to floppy phase transition in network glasses. This approach provides a unifying view of the phase transitions of proteins and glasses, and identifies the mean coordination as the relevant structural variable, or reaction coordinate, along the unfolding pathway.

3.2 Introduction

Much interest is currently focused on the rapid and faithful folding of proteins from a one-dimensional (1D) sequence of amino acids in a random coil, to a three-dimensional (3D) biologically functional structure in the native state (Bryngelson et al., 1995; Honig, 1999; Baker, 2000). A general view of protein folding is that it begins with hydrophobic collapse, in which the random coil changes to a compact state, with the hydrophobic groups in the interior region and polar groups at the surface interacting with the surrounding water. The packing is not yet optimal, with hydrophobic groups somewhat free to slide about in the interior of the globule, until residues are locked in place by the formation of specific hydrogen bonds. These hydrogen bonds can be regarded as a sort of *velcro* that locks the various structural elements in the folded protein together. Once these interactions are optimized, the native state is predominantly rigid with flexible hinges or loops at the surface – the number and distribution of these depending on the particular protein.

There have been many significant theoretical advances in understanding protein folding in recent years – including the concept of a funnel-shaped free energy landscape (Bryngelson et al., 1995; Onuchic et al., 1997; Chan and Dill, 1998; Brooks III et al., 2001), simplified lattice models that are more tractable for simulations of folding (Chan and Dill, 1998; Klimov and Thirumalai, 1999; Mirny and Shakhnovich, 2001), and more detailed but computationally intensive off-lattice models and molecular dynamics (MD) simulations (Daggett et al., 1996; Duan et al., 1998; Shea and Brooks III, 2001). These approaches have increased our understanding considerably, but the actual steps along the folding pathway continue to remain elusive. Experimentally, chemical and thermal denaturation of proteins are standard techniques to determine protein folding and unfolding equilibria and kinetics (Jackson, 1998; Eaton et al., 2000). However, to probe the range of time scales involved in folding, from microseconds to seconds, a series of challenging experiments is required (Eaton et al., 2000; Gruebele, 1999), and detailed structural information is generally not available.

I have concentrated on a simpler problem – that of analyzing the unfolding mechanism by dilution of noncovalent contacts in the native structure. For proteins in which the unfolding process is reversible, this approach also provides information about the folding pathway. I postulate that information about the folding pathway is contained within the density, strength, and specific location of the hydrogen bonds in the native state. To simulate denaturation, the hydrogen bonds and salt bridges within the structure are ranked according to their relative energies and broken one by one, from weakest to strongest, similar to the way these bonds would break in response to slowly increasing temperature. The transition towards a flexible, denatured ensemble in the protein is observed as the hydrogen-bond and salt-bridge network is disrupted. In chapter 4, these results are found to be robust against the introduction of some noise, or stochastic character, into the order in which the hydrogen bonds are broken.

In this chapter the program FIRST (Floppy Inclusions and Rigid Substructure Topography) is introduced as a computational tool to study protein folding. FIRST can decompose a protein structure into rigid clusters and flexible regions. When hydrogen bonds are removed from a structure, as during simulated unfolding, a protein will become increasingly flexible. The results of FIRST analysis on native-state structures are shown to agree with known flexible and rigid regions of folded proteins. This leads to the conclusion that rigid regions of a protein represent folded structure, and flexible regions represent unfolded or non-native structure. Using this definition of rigid = folded, flexible = unfolded, we can track the unfolding of a protein by observing the evolution of flexible regions during a simulated unfolding experiment. Also, the ability of FIRST to present detailed information on the phase transition between native (rigid) and denatured (flexible) states of the protein is presented.

3.3 Methods

3.3.1 FIRST Flexibility Analysis

The program FIRST was developed as a computational tool to measure flexibility in protein structures. At the core of the program is a graph-theory algorithm named the *3D pebble* *game* which is a 3-dimensional extension and implementation of results in mathematical rigidity theory that have developed over the past few years (Jacobs and Hendrickson, 1997; Jacobs and Thorpe, 1995, 1998). The roots of this work go back to Lagrange's (1788) introduction of constraints on the motion of mechanical systems in the late eighteenth century, which Maxwell (1864) used in the mid-nineteenth century to determine whether structures were stable or deformable. The applications of this kind of work have traditionally been to solve problems in engineering, such as the structural stability of different truss configurations in bridges. A very significant advance occurred with Laman's theorem (Laman, 1970), which exactly determines the degrees of freedom (DOF) within 2-dimensional networks, and allows the rigid regions and flexible joints between them to be found. A rigorous application of Laman's theorem to 3D structures has not yet been proven, however, the molecular framework conjecture proposed by Tay and Whiteley suggests that Laman's theorem will hold for a specific class of 3D networks called bond-bending networks, in which vertices (atoms) are connected by edges (bonds) and every angle between edges is defined (each bond angle is fixed) (Tay and Whiteley, 1984). For 3D bond-bending networks, the flexibility in the system derives from dihedral or torsional rotations of the bonds that are not locked in by the network. A brief introduction into rigidity theory as applied to macromolecules, such as proteins, is presented here. More detailed accounts can be found in (Jacobs et al., 1999, 2001) and references therein.

The results of FIRST rely on accurately counting the DOF and distance constraints in a system. Each atom in the system is assigned 3 DOF associated with motion in any direction in 3 dimensions. When bonds form between atoms the motion of the atoms becomes

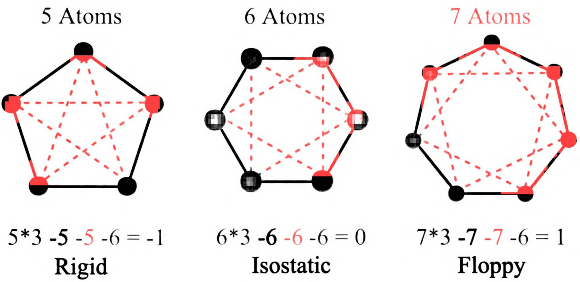

Figure 3.1: Determining the number of internal degrees of freedom in 3 small rings using constraint counting. Examples are shown for five, six, and seven-fold rings. The internal degrees of freedom (DOF) are counted by determining the total DOF, 3 for each atom (shown in green), and subtracting the number of distance constraints that arise from central-force bonds (shown in black), bond-bending constraints (shown in red), and the macroscopic rigid-body DOF (indicated by the light blue -6 in each equation). A negative value for the number of internal DOF (as in the five-fold ring) indicates that the structure is rigid, and overconstrained. It has more than enough constraints to be rigid. A value of 0 (as in the six-fold ring) indicates the structure is rigid and isostatic. This structure has just enough constraints to be rigid. A positive value for the number of internal DOF (as in the seven-fold ring) indicates that the structure is flexible or underconstrained.

restricted. Bond forces impose *distance constraints* on the atoms, that is, a pair of bonded atoms can no longer move independent of each other. The Euclidean distance between bonded atoms is held constant, and the net effect is the loss of DOF in the system. An example of how the internal DOF of three small rings can be computed by counting the distance constraints is shown in Figure 3.1. For the five-fold ring, which could represent the side-chain of a histidine residue, there are 5 atoms, so the system consists of $3 * 5 = 15$ DOF. There are 5 covalent bonds (thick black lines) and 5 bond-bending constraints (red dashed lines), resulting in $15 - 10 = 5$ DOF in the system. However, it is necessary to subtract off 6 trivial DOF, referred to the macroscopic or rigid body DOF in order to determine the internal DOF. Rigid body DOF refer the fact that you can take all 5 atoms in the five-fold ring and translate or rotate them together and it doesn't change any of the properties of the system. Because we are in 3 dimensions, there are 3 rigid-body translational DOF and 3 rigid-body rotational DOF for a total of 6 rigid body DOF. If we subtract off these rigid body DOF (indicated by the light blue -6 in the equations of Figure 3.1), then we see that the five-fold ring has -1 internal DOF. The physical interpretation of the negative value is that there are more constraints than are necessary to make the five-fold ring rigid. The common name for these type of structures is *overconstrained.* If another atom is added to the system, as in the six-fold ring, the final constraint count shows there are 0 DOF in the six-fold ring. This means that there are just enough constraints to make this structure rigid. Add a bond and it will become overconstrained. Remove a bond and it will become flexible. A structure with 0 DOF is rigid and is referred to as *isostatic*. For completeness, the constraint counting for a seven-fold ring is shown. Here, the final count yields 1 DOF. Positive values in the number of DOF indicate *flexible* or *floppy* structures.

For a protein, the total number of DOF will be the number of atoms observed in the crystal structure times three. Because the intricate bond network of a protein structure consists of many large and small rings, it is possible to have multiple overconstrained, isostatic and flexible regions in a protein at the same time. Determining the size and location of these regions, after all the DOF and distance constraints have been accounted for, is practically impossible to do by hand, and requires the program FIRST, specifically the 3D pebble game, to do the counting. At this point it becomes necessary to identify all of the distance constraints that can arise due to bond forces.

The bond forces in a molecular structure such as a protein will range from strong (i.e. covalent bonds) to weak (i.e. van der Waals interactions) (Figure 3.2). For the purposes of flexibility analysis all bond forces that are as strong or stronger than hydrogen bonds are included. By setting this cutoff, it is assumed that weaker bond forces, such as van der Waals interactions, are not strong enough to impose a distance constraint between a pair of atoms. The specific bond forces included the model are covalent bonds, salt bridges and hydrogen bonds. These bond forces are used to build the bond-bending network that FIRST requires for proper analysis of the flexibility in the structure. The connections between atoms generate the central-force distance constraints. The required angular constraints arise because each bond angle is treated as constant. To represent a constant angle in the bond-bending network, the distance between second-nearest neighbor atoms is fixed. An example of both of bond length and bond angle distance constraints for the main-chain atoms of an amino acid are shown in Figure 3.3. The bond length distance constraints are shown as thick black lines between N–C_α and C_α–C atoms. The bond-angle constraint, which results from a constant angle α, is shown a dashed, gray line between the N and C atoms.

Representing a bond force, such as a covalent bond, as distance constraint assumes that the distance between the two atoms is constant. These constant distances are defined in a protein structure either explicitly as equilibrium bond lengths, or implicitly as equilibrium bond angles. By fixing the distance we neglect high-frequency motion (bond-stretching,

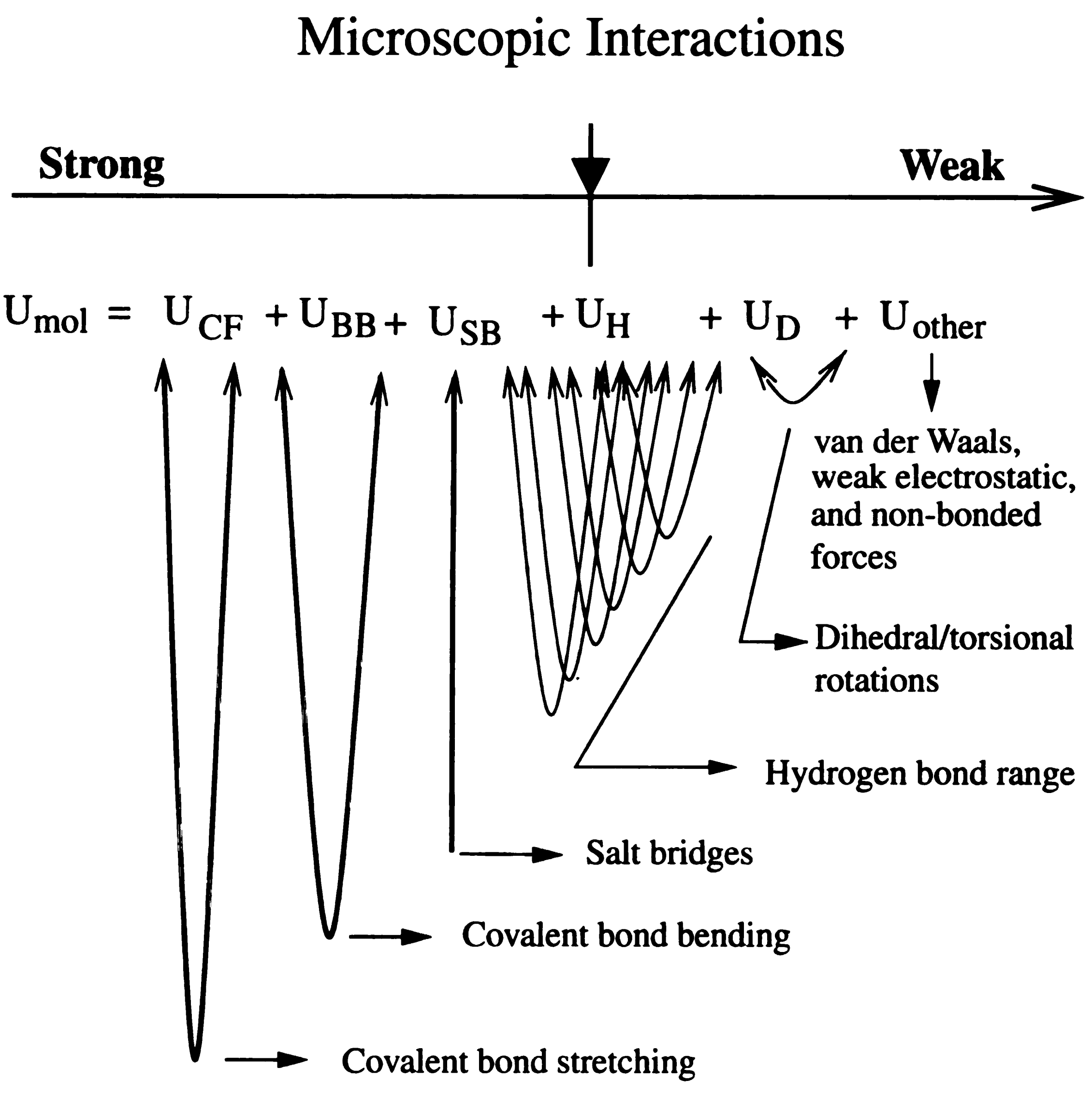
Microscopic Interactions
Strong
Weak
$U_{mol} = U_{CF} + U_{BB} + U_{SB} + U_{H} + U_{D} + U_{other}$
van der Waals, weak electrostatic, and non-bonded forces
Dihedral/torsional rotations
Hydrogen bond range
Salt bridges
Covalent bond bending
Covalent bond stretching

Figure 3.2: A schematic representation of microscopic bond forces ordered from strongest to weakest. U_{mol} represents the total potential energy of the bond forces in a protein. It is necessary to select which bond forces impose distance constraints by setting an appropriate energy cutoff. For the purposes of protein flexibility analysis, hydrogen bonds (with energies ≥ -0.1kcal/mol), salt bridges and covalent bonds are modeled as distance constraints. Weaker forces such as van der Waals interactions as not included.

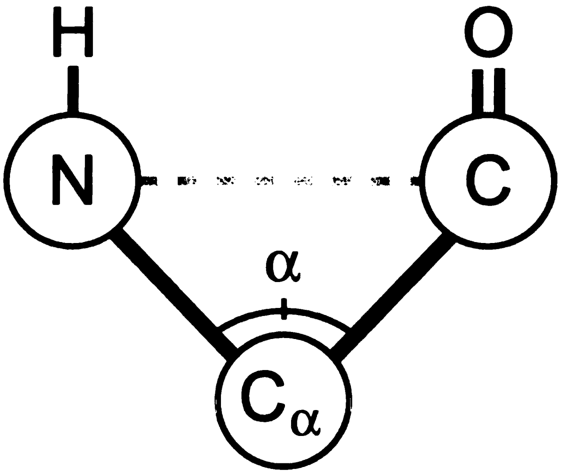

Figure 3.3: Example of bond-length and bond-angle distance constraints for the main-chain atoms of an amino acid. The positions of the N, C_α, C atoms are crystallographically defined, and the sp^3 hydridization of the C_α atom defines the bond angle α. Because the angle α is constant, the distance between the N and C atoms, shown as dashed, gray line, is also constant. The thick black lines between the N–C_α and C_α–C atoms represent bond-stretching distance constraints that arise from covalent bonds.

bond-bending) that would be expected due to thermal motion. This leads to the interpretation that FIRST results are meaningful only on time scales longer than those observed for bond bending and bond stretching frequencies, which generally occur in the range of 4000 – 200 cm^{-1} (120.0 - 6.0 femtoseconds) (Fadini and Schnepel, 1989). The structural flexibility required for protein folding (Jackson, 1998) or domain motion (Epstein et al., 1995) occur as a result of dihedral rotations, which are low-frequency modes that occur on much longer time scales ($\geq$ microseconds). Therefore FIRST results can give us information about flexibility in these processes.

The peptide bond of a protein represents a special case of a bond force due to its partial-double bond character that arises from resonance with the main-chain carboxylate. All double and partial double bonds are viewed as non-rotatable dihedral angles, and special care is taken within the FIRST program to lock these bonds.

In addition to modeling the strong bond forces mentioned above, hydrophobic interactions are also included as distance constraints. However, in contrast to covalent bonds and hydrogen bonds, hydrophobic interactions are modeled such that they restrict the motion between two hydrophobic atoms, and do not fix it constant. This is accomplished by linking a pair of hydrophobic atoms via a series of artificial atoms and bonds. These *pseudoatoms* increase the number of DOF associated with a hydrophobic interaction, and the intervening *pseudobonds* create distance constraints that reduce the number of DOF. The net effect is the loss of 2 DOF/hydrophobic interaction. A further description of how hydrophobic tethers are modeled is given below in the Methods section: Identifying and Modeling Hydrophobic Interactions.

Once all of the bond forces and hydrophobic interactions have been identified, it is possible to create the bond-bending network of a protein. In this 3D network, each of the vertices represents the position of an atom from the protein structure. Each edge represents a distance constraint that arises from fixed bond lengths and angles. This generic bond-bending network is what FIRST analyzes using 3D constraint counting. The algorithm will identify which distance constraints in the network are adding stress to the network. These redundant bonds are associated with nonrotatable dihedral angles in the protein. A set of interconnected nonrotatable bonds form a rigid cluster. Also computed are the number of *floppy modes*, which is specifically the number of bond-rotational DOF that remain in the protein after the nonredundant distance constraints have been subtracted. Floppy modes are usually associated with a collective motion (a concerted motion of many bonds within a protein, such as a large domain motion). In general, because floppy modes are associated with a collective motion consisting of many bonds, the number of floppy modes will be less than the total number of flexible bonds in a protein.

It is worth mentioning that the algorithms encoded in FIRST are extremely efficient. Alternative methods to identifying rigid and flexible bonds in protein will generally scale with a computational complexity of order $O(N^7)$, where N is the number of atoms in the protein (Jacobs et al., 1999). Theoretically, FIRST scales as order $O(N^2)$, however, in practice is usually linear in the number of atoms (of order $O(N)$). The worst case that has been observed was of order $O(N^{1.2})$ (M. F. Thorpe, personal communication).

3.3.2 Preprocessing Protein Structures for Analysis

Given the absence of electron density for hydrogen atoms in most X-ray crystal structures, positions for polar hydrogen atoms (including those in bound water molecules) were assigned using the software WhatIf (Vriend, 1990). The WhatIf software uses a combination of heuristic criteria and hydrogen bond energy functions to optimize the placement of polar hydrogen atoms in a protein structure. Comparison of hydrogen positions determined by WhatIf to those observed in neutron diffraction structures for five proteins have been shown to overlap well (the worst case had 94.3% of the hydrogen positions in common between the computational and experimental results) (Jacobs et al., 2001). WhatIf was run on a protein in the presence of all crystallographic water molecules found in the structure. However, for all subsequent analyses only buried water molecules were included. Buried waters were identified using the PRO_ACT software (Williams et al., 1994).

The program WhatIf will not add hydrogens to atoms or molecules defined with the HETATM (heteroatom) field of a Protein Data Bank (PDB) file. HETATMs are typically small ligands such as metals, cofactors, inhibitors, and substrates or substrate analogs. To add hydrogen atoms to HETATM groups the Biopolymer programs of InsightII molecular graphics package (Biosym, Molecular Simulations) was used.

In the choice of protein structures to analyze, the stereochemical quality of the structure can have a significant influence on the definition of its network of hydrogen bonds, due to their angular dependence (described in the next section). The result is that FIRST analysis on a structure with poor stereochemistry is likely to indicate the protein as being more flexible than it actually is, due to missing hydrogen bond distance constraints. It is advisable to assess the main-chain stereochemistry through a Φ, Ψ plot, as well as focus on high-resolution, well-refined structures for FIRST analysis.

3.3.3 Identifying and Modeling Hydrogen Bonds

Hydrogen bonds were identified between donor and acceptor groups according to the following geometric criteria (Stickle et al., 1992; McDonald and Thornton, 1994), shown graphically in Figure 3.4:

1. Donor-Acceptor distance, d $\leq$ 3.6Å.
2. Hydrogen-Acceptor distance, r $\leq$ 2.6Å.
3. Donor-Hydrogen-Acceptor angle, $90° \leq \theta \leq 180°$.

The energy of each hydrogen bond was measured using a modified Mayo potential (Dahiyat et al., 1997). The function evaluates the favorability of the observed hydrogen-bond length relative to the optimal, equilibrium length for that pair of atoms based on their electron orbital hybridization, as well as the favorability of the angles between the donor and acceptor groups. The modification avoids non-physical H-bonds with angles near 90° (e.g., between C=O(i) and NH(i+3), rather than the important C=O(i)$\leftrightarrow$NH(i+4) interactions in the middle of α-helices). Salt bridges were identified between the negatively charged groups of aspartate, glutamate, or the carboxy-terminus of the protein, with the positively charged groups of histidine, lysine, arginine, or the amino-terminus. The energies of hydrogen bonds, E_{HB}, and salt bridges, E_{SB}, were calculated using equations

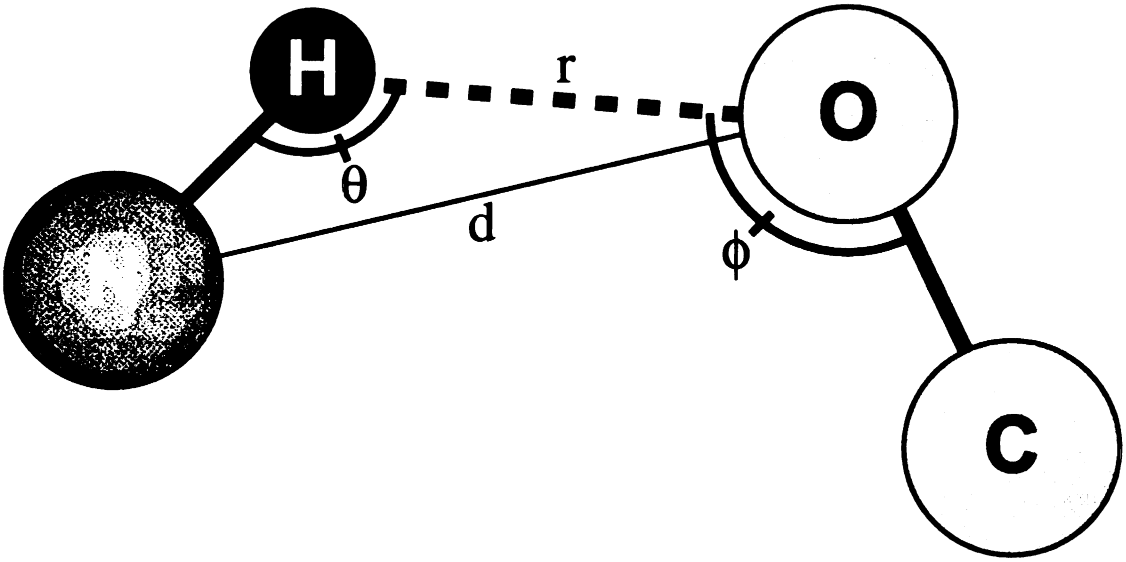

Figure 3.4: Geometric parameters used to identify hydrogen bonds and measure their energy. The hydrogen bond is depicted as a dashed line between the hydrogen and the acceptor oxygen. r is the hydrogen-acceptor distance, d is the donor-acceptor distance, θ is the donor-hydrogen-acceptor angle and ϕ is the hydrogen-acceptor-base atom angle, where the carbon is the base atom in this example.

3.1 and 3.2, respectively.

$$E_{HB} = V_0 \left\{ 5 \left(\frac{R_0}{R} \right)^{12} - 6 \left(\frac{R_0}{R} \right)^{10} \right\} F(\theta, \phi, \varphi) \tag{3.1}$$

with

V_0 = 8 kcal/mol	R_0 = 2.80 Å
sp^3 donor - sp^3 acceptor	$F = cos^2\theta e^{-(\pi-\theta)^6} cos^2(\phi - 109.5)$
sp^3 donor - sp^2 acceptor	$F = cos^2\theta e^{-(\pi-\theta)^6} cos^2\phi$
sp^2 donor - sp^3 acceptor	$F = cos^4\theta(e^{-2(\pi-\theta)^6})$
sp^2 donor - sp^2 acceptor	$F = cos^2\theta e^{-(\pi-\theta)^6} cos^2(\max[\phi, \varphi])$

$$E_{SB} = V_S \left\{ 5 \left(\frac{R_S}{R+x} \right)^{12} - 6 \left(\frac{R_S}{R+x} \right)^{10} \right\} \tag{3.2}$$

with

V_S = 10 kcal/mol, R_S = 3.2 Å, and x = 0.375 Å.

In each equation, R is the distance between the donor and acceptor atoms. The θ angle is the donor–hydrogen–acceptor angle, and ϕ is the hydrogen–acceptor-base atom angle, where the base atom is the atom bonded to the acceptor (e.g., carbonyl carbon for a carbonyl oxygen acceptor atom). The angle φ is an out-of-plane angle that arises when both the donor and acceptor have sp^2 hybridization. For the salt-bridge energy function, the values of V_S, R_S, and x were selected such that the computed energies matched those of experimental results on salt bridges (Xu et al., 1997). Because salt bridges are essentially a special case of hydrogen bonds in which the donor and acceptor are charged, for simplicity, hydrogen bonds and salt bridges will both be referred to as hydrogen bonds.

To determine a reasonable default energy cutoff for hydrogen bonds, the threshold that best conserves the hydrogen bonds within a family of protein structures was evaluated (Jacobs et al., 2001). Multiple structures within four different protein families were studied to find such a threshold. The PDB codes used for each family are as follows: trypsin (1tpo, 2ptn, 3ptn), trypsin inhibitor (4pti, 5pti, 6pti, 9pti), adenylate kinase (1zin, 1zio, 1zip), and HIV protease (1dif, 1hhp, 1htg). Figure 3.5 shows the hydrogen-bond energy distribution for one of these families, namely the three HIV protease structures. A large spike appears in the distribution between -0.1 and 0.0 kcal/mol. This spike is largely due to the fact that quite generous definitions of hydrogen bonds are allowed initially (donor–hydrogen–acceptor angle, $\theta \geq 90°$ and donor–acceptor distance, $d \leq 3.6$ Å, as shown in Figure 3.4). The inset of Figure 3.5 expands the region near 0.0 kcal/mol, demonstrating how a large number of very weak hydrogen bonds, often with θ angles near 90°, can be removed by setting $E_{cut} \leq -0.1$ kcal/mol. Thus, the generous hydrogen bond distance and angle screening criteria can be effectively filtered by setting E_{cut}. When these geometric criteria and an energy threshold of -0.1 kcal/mol are applied to analyze the hydrogen bonds and salt bridges in five neutron diffraction structures, a Gaussian distribution is observed for the number of hydrogen bonds as a function of donor–acceptor distance, with virtually all hydrogen bonds and salt bridges having distances between 2.6 and 3.6 Å. The distribution in donor–hydrogen–acceptor angles is bimodal, with a strong, Gaussian peak between 130 and 180° and a weaker peak between 90 and 130°. An energy cutoff of -0.1 kcal/mol is used in all subsequent FIRST analyses.

3.3.4 Identifying and Modeling Hydrophobic Interactions

The hydrophobic effect observed in protein folding describes the tendency for nonpolar residues to bury themselves within the interior of the protein structure. This process frees many solvent DOF, which would necessarily form hydrogen bonded ice-like structure around an exposed hydrophobic group in an attempt to compensate for the loss of entropy by increasing the enthalpy. Buried within the protein, the hydrophobic groups interact weakly in what can be appropriately described as a slippery or "greasy" manner. It has been shown that these hydrophobic interactions contribute significantly to protein stability and are generally believed to be critical in driving the protein folding process (Dill, 1990).

As with covalent bonds and hydrogen bonds, hydrophobic interactions must be modeled as a connection between two atoms due to the graph-theory nature of the FIRST program. Hydrophobic interactions are identified as contacts between pairs of carbon atoms or between carbon and sulfur atoms. Van der Waals radii of 1.7Å and 1.8Å were assigned to carbon and sulfur atoms, respectively (Bondi, 1964). A pair of carbon and/or sulfur atoms were determined to be in hydrophobic contact if the distance between their atom centers was $\leq r_a + r_b + R$, where r_a is the van der Waals radii of atom a, and r_b is the van der Waals radii of atom b (Figure 3.6A). R was set to 0.25Å as this value was empirically determined

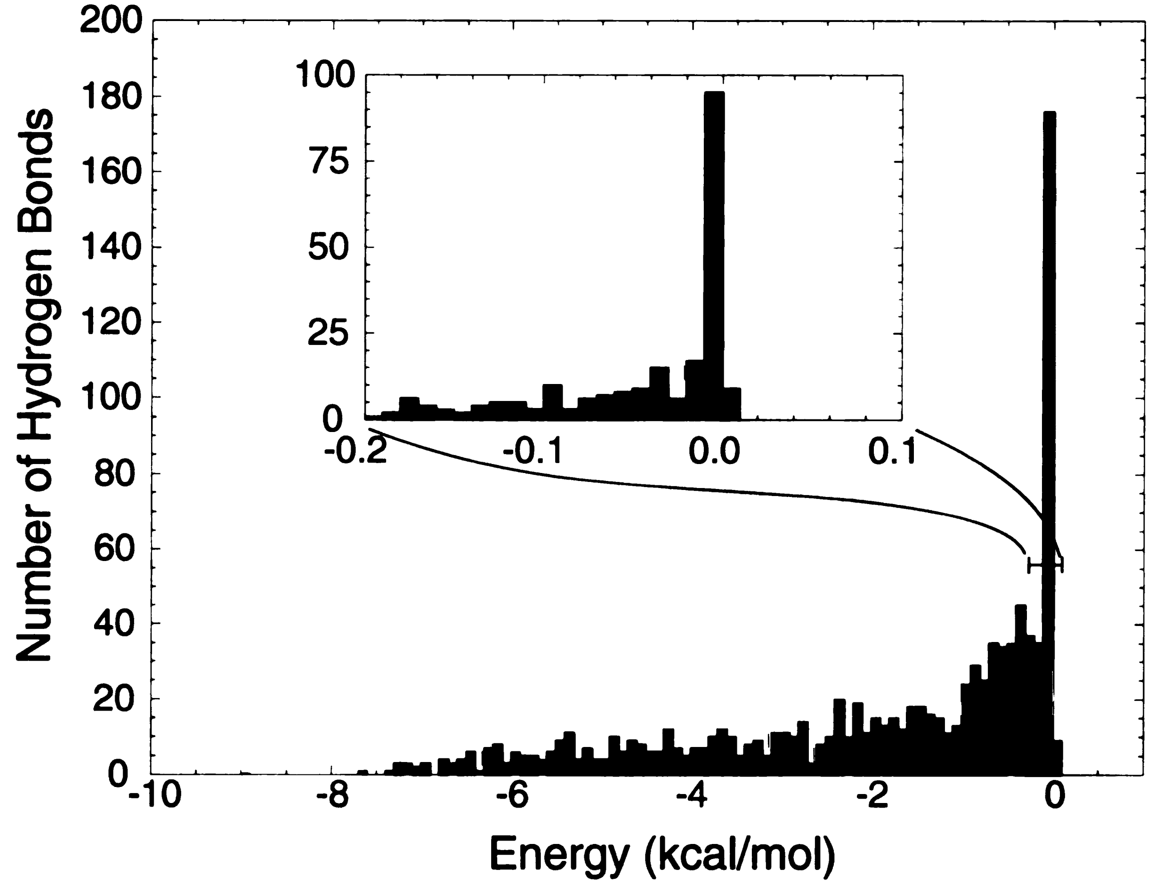

Figure 3.5: Histogram of hydrogen bond energies from three structures of HIV protease. Hydrogen atom positions in each of the three structures (PDB codes: 1dif, 1hhp, 1htg) were computed using the program WhatIf. The inset expands the low-energy region between −0.2 and 0 kcal/mol. An energy cutoff of −0.1 kcal/mol is used to eliminate the large number of very weak hydrogen bonds in the spike near 0 kcal/mol.

to yield the best result when predicting protein folding cores in a test set of ten proteins (described in Chapter 4) when sampling over many values of R.

The net effect of hydrophobic interactions on the flexibility of a protein structure is to restrict motion. That is, they impose distance constraints between hydrophobic groups and therefore remove DOF from the system. However, due to the nonspecific nature of hydrophobic interactions, they will have a less constraining effect on protein motion than hydrogen bonds. Therefore, hydrophobic interactions are modeled such that they introduce less constraints on a protein than hydrogen bonds. This is accomplished by connecting a pair of hydrophobic atoms via a series of three *pseudoatoms*, as shown in Figure 3.6. The sole purpose of the pseudoatoms is to attenuate the number of DOF consumed by a hydrophobic tether. For example, if we were to simply connect two hydrophobic atoms, the single bond would generate one central-force constraint (the actual bond) and four bond-bending constraints (due to four new bond angles). The net effect would be to remove 5 DOF from the system, a result similar to how covalent bonds are modeled. By introducing three pseudoatoms in between the hydrophobic atoms, we first add 9 DOF to the system (each pseudoatom adds 3 DOF). The intervening bonds generate 4 central-force constraints and 7 bond-bending constraints, for a net loss of 2 DOF (9 (DOF) - 11 (constraints) = -2) for each hydrophobic tether introduced into the protein. By comparison, each hydrogen bond removes 3 DOF from the system, and therefore, hydrophobic tethers are less constraining than hydrogen bonds.

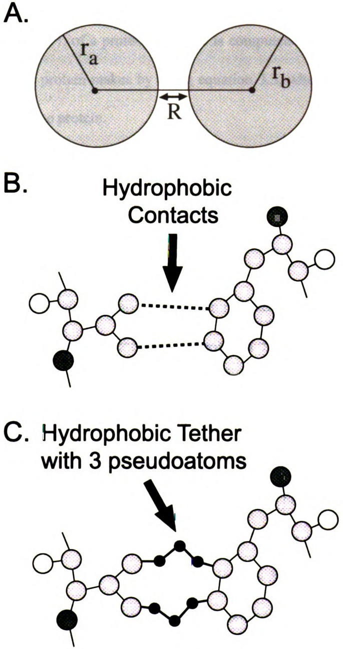

Figure 3.6: Identifying and modeling a hydrophobic tether distance constraint. A hydrophobic interaction is identified between a pair of carbon and/or sulfur atoms if $r_a + r_b + R \leq 0.25$Å, where r_a is the van der Waals radii of atom a and r_b is the van der Waals radii of atom b. R was empirically defined to be 0.25Å. Van der Waals radii of 1.7Å and 1.8Å were assigned to carbon and sulfur atoms, respectively. Hydrophobic tethers are modeled using three pseudoatoms, which results in a loss of 2 DOF per hydrophobic tether.

3.3.5 Computing the Mean Coordination of a Protein Structure

The mean coordination, $\langle r \rangle$, of a protein structure is computed as the average number of bonds each atom in the protein makes by using equation 3.3, where n_r is the number of r-coordinated atoms in the protein.

$$\langle r \rangle = \frac{\sum\limits_{r=2}^{r_{max}} r n_r}{\sum\limits_{r=2}^{r_{max}} n_r} \tag{3.3}$$

The mean coordination gives a partial description of the protein bond network, and is strongly dependent on how many bonds are present in a protein at any given time. For overconstrained systems in which bonds are being diluted, the mean coordination can be used to describe the state of the system when the rigid → flexible transition occurs. Below, a method for simulating the thermal denaturation of proteins is presented in which hydrogen bonds are repeatedly removed from the protein structure, beginning with an overconstrained native state through to a flexible denatured state. The mean coordination is shown to be a useful number with which to compare the rigid → flexible transition in different proteins that occurs during the simulated thermal denaturation. Additional detail can be found in the supplementary material of Rader et. al., 2001.

3.3.6 Computing the Fraction of Floppy Modes

A key quantity computed by FIRST when analyzing the flexibility in a protein structure, or any 3D bond-bending network, is the number of floppy modes, F, also known as the number of independent bond-rotational DOF. This number can be used to compute the fraction of floppy modes, f, by using equation 3.4, where the term in the denominator, $3N$, represents the total number of DOF in the protein (N is the number of atoms in the protein).

$$f = \frac{F}{3N} \tag{3.4}$$

The fraction of floppy modes will necessarily increase as bonds are removed from the bond-bending network representation of a protein or a glass. An example of f plotted versus $\langle r \rangle$ for random dilution of a glass network is shown in Figure 3.7A. As bonds are randomly removed from the glass network the rigid $\rightarrow$ flexible phase transition occurs when the slope in the line changes sign. This point can be identified as the inflection point in a first derivative plot of f' vs. $\langle r \rangle$, and as a peak in the second derivative plot, f'' vs. $\langle r \rangle$. As in glass networks, the rigid $\rightarrow$ flexible phase transition observed during simulated protein unfolding (described in the next section) can be tracked using f vs. $\langle r \rangle$.

3.3.7 Simulating Denaturation

As a protein is gradually thermally denatured, the covalent bonds remain intact, whereas hydrogen bonds will begin to break. The flexibility in the protein will increase as the number of hydrogen bonds in the protein decreases. Our hypothesis is that information about the protein unfolding/folding pathway is encoded in the network of hydrogen bonds present in the native state of a protein. This hypothesis was tested by removing hydrogen bonds from a protein structure to simulate thermal denaturation, then using FIRST to observe

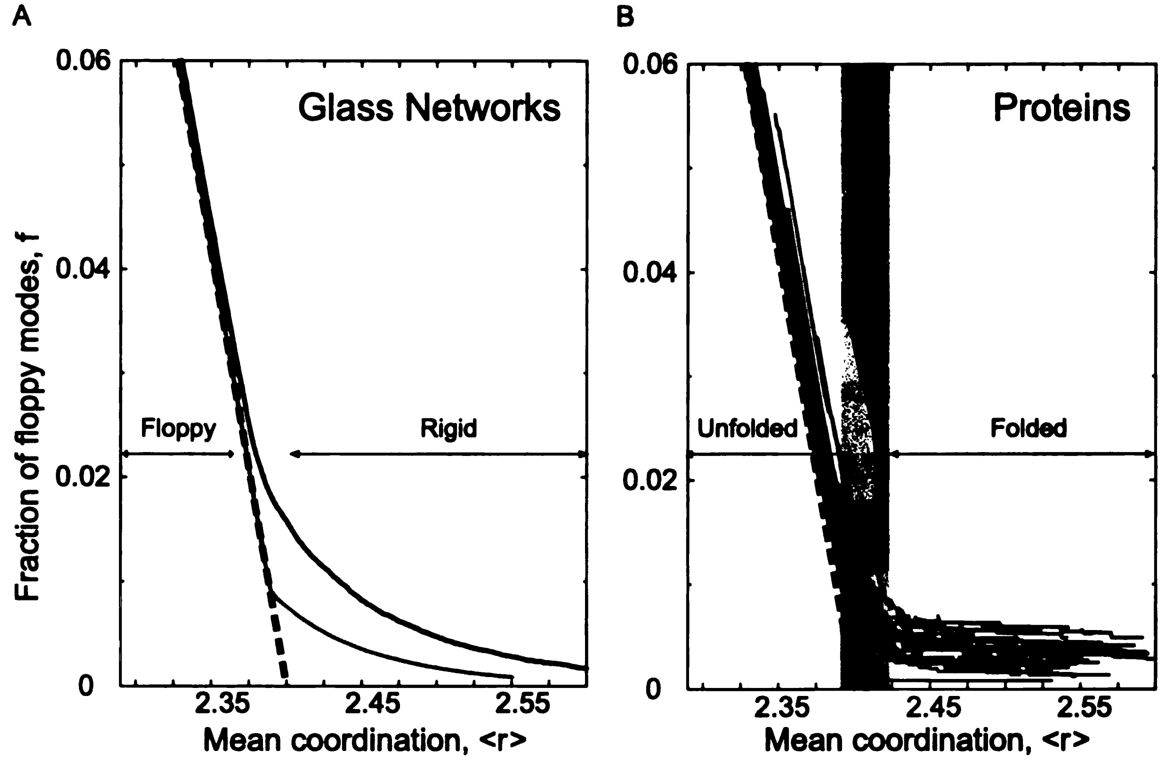

Figure 3.7: The fraction of floppy modes, $f = F/3N$, as a function of the mean coordination in two glass models and a set of 26 proteins. The mean-field Maxwell approximation to computing the number of floppy modes is shown as a black dashed line in each panel. A. The results for random bond dilution of a glass network. The purple line shows results in which the rigid → flexible transition is second-order. The orange line represents a first-order transition that arises in glass networks that lack small rings. B. Results for a representative set of 26 structurally and functionally diverse proteins. The blue lines are monomers; red lines, dimers; green lines, tetramers. The gray shaded region indicates the range in which protein folding/unfolding occurs.

where the resultant change in flexibility occurs. The results will depend upon the order in which hydrogen bonds are removed. Because hydrophobic interactions actually become somewhat stronger with moderate temperature increases (Tanford, 1980), these interactions are maintained throughout the simulation.

During thermal denaturation, the hydrogen bonds are expected to break in an energy-dependent manner. This process is simulated by using the following procedure. Initially, the flexibility of the native protein structure is analyzed with all its covalent and noncovalent interactions included (hydrogen bonds and hydrophobic interactions). The weakest hydrogen bond in the structure is then broken by removing any distance constraints created by that bond. The effect of removing this bond is then observed by applying FIRST to identify the flexible regions in the protein. We continue this process of breaking the weakest hydrogen bond remaining in the structure and updating the identification of flexible regions until all the hydrogen bonds have removed.

3.3.8 Visualizing Results: The 3D Rigid Cluster Decomposition

The results of FIRST indicate for each bond in the protein structure whether it is flexible (free to rotate) or rigid (not rotatable). Groups of atoms coupled to each other via rigid bonds form a rigid cluster. One or more independent rigid clusters with intervening flexible regions may be found in a protein structure. The distribution of rigid clusters and flexible bonds identified by FIRST is called a *rigid cluster decomposition* (RCD) and can be viewed graphically by color-mapping the results onto the 3D structure of the protein. Figure 3.8A displays the results for CI2 when the 18 weakest hydrogen bonds have been diluted from the structure. Flexible bonds are shown as thin black tubes, while rigid clusters are depicted by thick, colored tubes, with each independent rigid cluster distinguished by a different color. Hydrogen bonds and hydrophobic interactions are shown as dark gray lines. It is generally easier to interpret the results by removing the side chains from the graphical depiction of the results. The results shown at the top of Figure 3.8B are identical to those in Figure 3.8A, except that the side chains have not been displayed. It is much easier to identify common secondary structural elements, such as the α-helix (colored blue), a β-strand (colored red), a β-turn (colored orange) and a loop region (colored yellow), when viewing only the main chain bonds.

3.3.9 Visualizing Results: The 1D Rigid Cluster Decomposition

The hydrogen bond dilution method to simulate denaturation produces a RCD each time a hydrogen bond is removed from a protein structure. Interpreting the 3D results requires flipping through a large number protein structures, and keeping track of where flexibility occurs in the structure as a function of the hydrogen-bond dilution. To overcome this visualization problem, we employ the reduced 1-dimensional (1D) representation of the data depicted graphically in Figure 3.8B. In the 1D representation, the only results shown are for the backbone N–C_α and C_α–C bonds. As in the 3D figures, each backbone bond is represented as a thin black line if it is flexible (rotatable), or as a colored block if it is rigid.

The 1D mapping of the flexibility data is a convenient means of reducing the amount of information generated in a hydrogen bond dilution experiment to a tractable level. A complete denaturation simulation can now be viewed as a series of horizontal lines, ordered

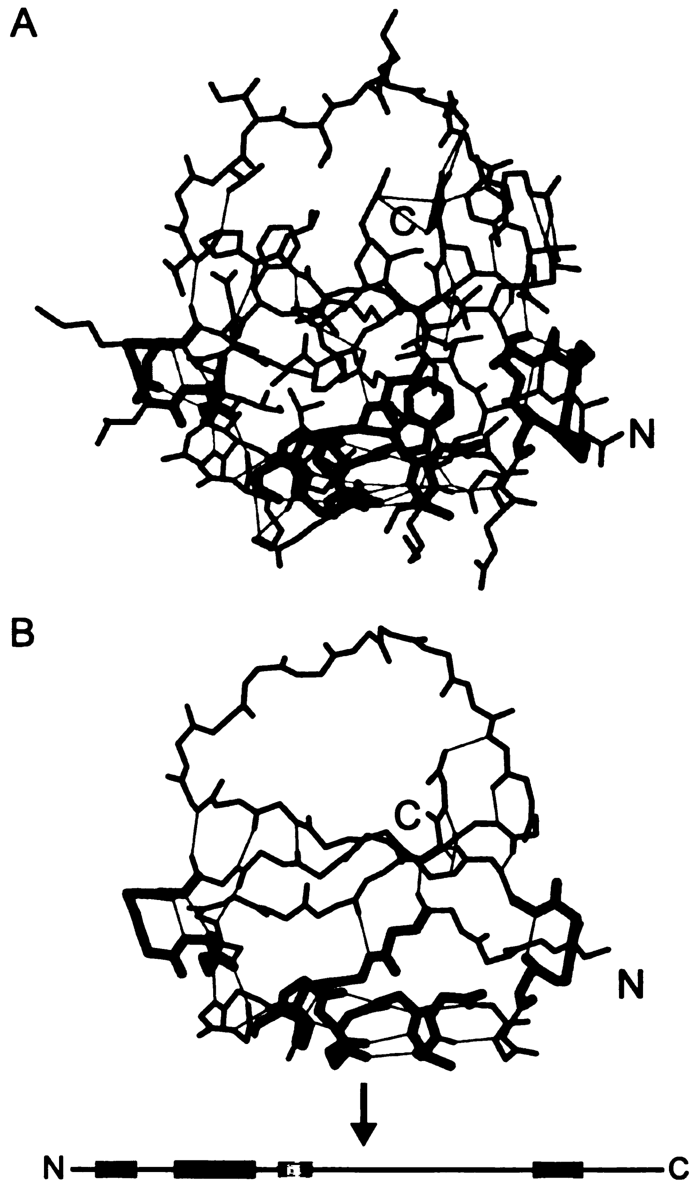

Figure 3.8: Rigid cluster decomposition results for CI2 when 67% of the weakest hydrogen bonds have been removed. A. This panel shows all non-hydrogen atoms present in the structure (excluding water molecules). There are four independent rigid clusters, as computed by FIRST. The rigid clusters are depicted by thick colored tubes (blue, red, orange and yellow, from largest to smallest). Each thin black tube represents a rotatable or flexible bond. The thin, dark gray lines show the location of hydrogen bonds and hydrophobic tethers. B. The same results of FIRST analysis for CI2, showing only the main-chain atoms. Because the main-chain for a protein monomer is an unbranched linear polymer, the flexibility results for the main chain can be mapped onto a 1D line. From the N-terminus to the C-terminus, each backbone bond is represented as a thin black line if it is flexible or a thick colored block if it is rigid. Independently rigid clusters are assigned different colors.

Figure 3.9: Results of the complete hydrogen bond dilution for c-SRC SH3 domain. A. The top line in this figure shows the results for the native state of the SH3 domain. There are 44 hydrogen bonds present. Each successive line shows the 1D rigid cluster decomposition as hydrogen bonds are removed from the structure. The lines shaded gray indicate results with identical 1D RCDs. These lines can be identical because the 1D rigid cluster decomposition only shows changes in the flexibility of the backbone bonds of a protein. B. By removing the redundant lines from panel A, we are left with results that show only when a change in the flexibility of the main chain occurred.

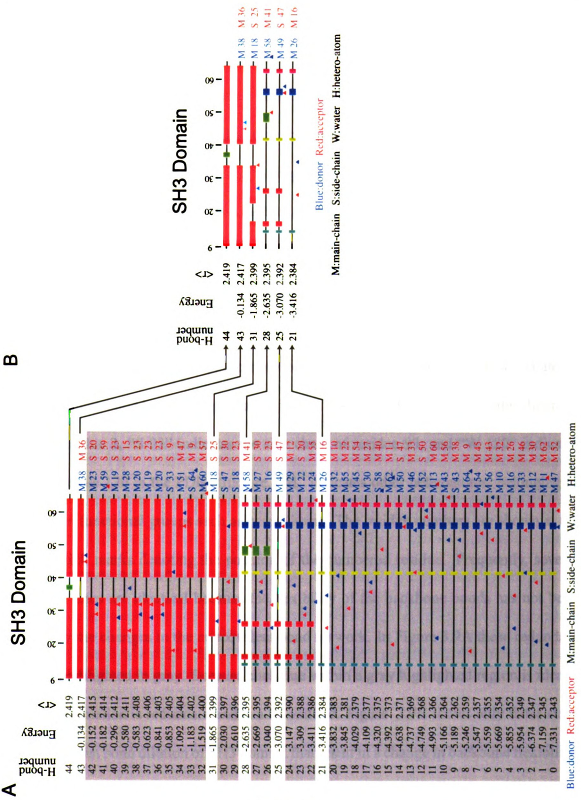
A
B
SH3 Domain
H-bond number
Energy
<r>
9
20
30
40
50
60
44 2.419
43 -0.134 2.417 M 38 M 36
31 -1.865 2.399 M 18 S 25
28 -2.635 2.395 M 58 M 41
25 -3.070 2.392 M 49 S 47
21 -3.416 2.384 M 26 M 16
Blue:donor Red:acceptor M:main-chain S:side-chain W:water H:hetero-atom

Figure 3.9

from a native state (at the top) to a denatured state (at the bottom). Each line shows the regions of structural stability and flexibility for the backbone atoms after a specific hydrogen bond has been removed during the denaturation process. Figure 3.9A provides an example of a complete thermal denaturation simulation for the SH3 domain of human tyrosine kinase c-SRC (PDB code: 1fmk). The three columns on the left-hand side of Figure 3.9A describe: 1) the number of remaining hydrogen bonds in the protein at each step; 2) the energy of the just-broken bond (in kcal/mol), according to the modified Mayo potential (Dahiyat et al., 1997); and 3) the mean coordination, $\langle r \rangle$, of the atoms in the network at that step. Regular secondary structure content is shown at the top, as determined by DSSP (Kabsch and Sander, 1983). The right-hand columns, together with the solid triangles beneath each line, show the residue locations of the donor (blue) and acceptor (red) atoms of the hydrogen bond broken to generate this step. For example, "M 2" indicates the main chain of residue 2, "S 93" indicates the side chain of residue 93, and "W 120" indicates water molecule 120 in the PDB structure. "H" indicates other heteroatoms, belonging to non-protein functional groups such as bound heme. The residue numbers are shown at the top, with tick marks denoting the position of the numbered residue. Frequently, several successive lines are identical because the flexibility of the backbone bonds has not been affected by the changes in the noncovalent bond network. In Figure 3.9A these redundant lines are highlighted in gray. Because each line within a gray highlighted region is identical, the information is redundant and can be omitted. Figure 3.9B shows only those steps in the hydrogen bond dilution of SH3 domain that result in a change in backbone flexibility. Images in this thesis are presented in color.

3.4 Results

3.4.1 Native State Flexibility Analysis: Open and Closed Structures of HIV Protease

Given a protein's native-state structure, all of the covalent bonds, hydrophobic tethers, hydrogen bonds and salt bridges are used to define the distance constraint network for the protein. Given these constraints, FIRST identifies all the rigid and flexible regions within a protein, and these results have been shown to correlate well with experimental measures of flexibility for a range of proteins (Jacobs et al., 1999, 2001; Thorpe et al., 2000, 2001).

The FIRST results for HIV protease, in both unbound (Figure 3.10A) and inhibitor bound forms (Figure 3.10B), have been compared with experimental measures of protein flexibility. The major peaks in main-chain thermal mobility (B-value), measured crystallographically, correlate directly with the α, β, and γ flexible regions predicted by FIRST (Figure 3.10) (Jacobs et al., 2001). It should be noted that for proteins with mobile domains or other moving rigid bodies, such as α-helices, the crystallographic mobility and FIRST results will not necessarily compare well with B-values. Crystallographically, they appear as mobile regions, whereas in FIRST they appear as rigid regions flanked by flexible loops, which allow the rigid-body motion.

HIV protease has also been crystallized with various inhibitors bound, resulting in a closed conformation with the flaps lowered. The main-chain dihedral angle changes (similar to the analysis of Korn and Rose (1994) observed for crystal structures of the open (PDB

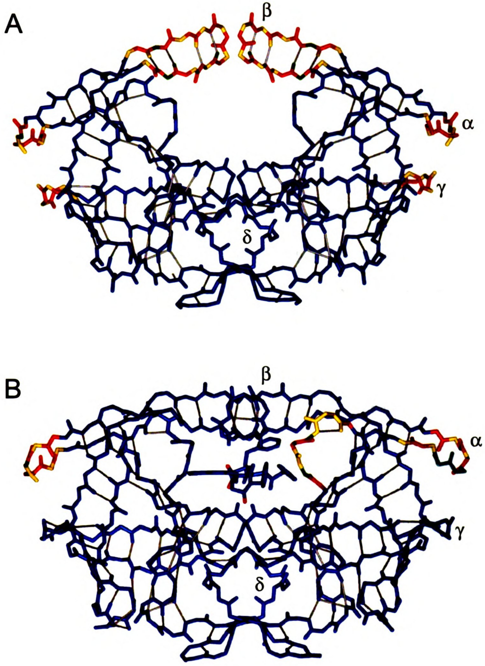

Figure 3.10: Rigid cluster decomposition for HIV protease. A. 3D RCD of HIV protease in an unbound, open conformation (PDB code: 1hhp) The "flaps" at the top of the structure are determined to be flexible in the open conformation (indicated by the red and yellow bonds), providing ligand access to the active site. B. 3D RCD of HIV protease in a ligand-bound, closed conformation (PDB code: 1htg). Upon ligand binding, the flaps become part of the large rigid cluster, colored blue.

code: 1hhp) and closed (PDB code: 1htp) have been computed. The FIRST-predicted flexible regions directly correspond with the regions of greatest dihedral angle change (Jacobs et al., 2001). In the three flexible regions (α, β, and γ), the flexibility is associated with a flip in at least one dihedral angle (defined as a change of more than 60 degrees) within a rigid β-turn in the center of each flexible region. The results are consistent with the motion observed by interpolation between different HIV protease crystal structures (Gerstein and Krebs, 1998) and an earlier dihedral analysis for a different pair of HIV protease structures (Korn and Rose, 1994) indicating that large dihedral angle changes at residues 40, 50, and 51 in the α and β regions result in a large, concerted movement of the flaps. Flexibility of the γ region has not been emphasized in other studies of HIV protease; however, it is known that drug-resistant mutants of the protease include two residues that pack against the γ region, 63 and 71, with residue 63 proposed to induce a conformational perturbation (Chen et al., 1995; Patrick et al., 1995). Thus, conformational coupling between the γ region and the flaps, through the γ–α loop interactions, may explain why mutations in the γ region, which are distal from the active site, cause resistance to drug binding.

Ligand binding restricts the motion of the flaps through new hydrogen bonds linking the two flaps to each other and to the ligand. Some of these hydrogen bonds between the flaps and ligand are mediated by a conserved water molecule found in retroviral but not mammalian homologs of HIV protease (Wlodawer and Erickson, 1993), providing a useful basis for designing more HIV-specific drugs. To compare the influence of ligands on HIV protease flexibility, there were a number of ligand-bound structures of good stereochemistry from which to choose. For brevity, only the results from PDB entry 1htg are shown, with inhibitor GR137615 bound to represent the closed form of HIV protease. (Two other ligand-bound structures, 1hiv and 1dif, have also been analyzed by FIRST, and the ligands' influence on protein flexibility was found to be substantially similar.) Unlike the open form, the closed structures were resolved crystallographically as dimers, and thus independent structural information is available for the two subunits of the dimer. This means it is possible to assess the influence of different side-chain conformations in the two halves (due to thermal fluctuations and environmental differences) in terms of their effects on the hydrogen-bonding network and flexibility. The left and right sides of HIV protease in Figure 3.10B indicate that the only substantial difference in their flexibility is caused by the asymmetry of the bound ligand.

Comparison of the ligand-bound structure with the open HIV protease also demonstrates how a ligand can rigidify part of the protein through new hydrogen bonds even though the ligand itself is not rigid, while making other parts of the protein more flexible. Particularly note the dimer interface, where inter-subunit rotation occurs upon ligand binding, breaking some of the interfacial stabilizing hydrogen bonds, and the loop to the right of the binding cavity, shown as a flexible region of the main-chain ribbon in Figure 3.10B. This loop flexibility is not reflected in the other HIV protease subunit, due to ligand asymmetry. Flexibility of the dimer interface in a ligand-bound structure is also a prominent feature found by NMR (Ishima et al., 1999) and MD analyses (Scott and Schiffer, 2000); MD also predicts flap flexibility in the ligand-free conformation.

Native-state flexibility analysis results for dihydrofolate reductase and adenylate kinase have also been performed. The FIRST results for these proteins have been shown to be consistent with experimentally observed conformational flexibility in the native states of these two proteins (Jacobs et al., 2001).

3.4.2 The Folding Transition State

The results of simulating denaturation can be tracked quantitatively along the unfolding pathway in terms of the change in number of fractional floppy modes, f (bond-rotational DOF) as the mean coordination decreases. A plot of f as a function of $\langle r \rangle$ for 26 structurally diverse proteins (listed in Table 3.1) and for two limiting models of network glasses are shown in Figure 3.7. The overall similarity in the flexibility transition behavior of f for the diverse proteins and glasses is striking.

To examine these results in more detail, in particular the phase transition region shown in gray in Figure 3.7, A. J. Rader, a graduate student of Dr. M. F. Thorpe in the Department of Physics and Astronomy at Michigan State University, has obtained the first and second derivatives of f versus $\langle r \rangle$ (Figures 3.11 and 3.12, respectively). The derivatives were calculated numerically by fitting a cubic equation over an interval corresponding to $\Delta\langle r \rangle$ = 0.75, which contained typically from 90 to 2,000 data points. In Figure 3.11, we see the sharp rise of the first derivative through the transition region, again marked in gray. One of the glass models (orange line) shows a first-order transition as indicated by the discontinuity at $\langle r \rangle = 2.389$. The insert in Figure 3.11 is adapted from several folding experiments (Creighton, 1993), showing that as the temperature increases, the fraction of folded protein decreases. The fraction of floppy modes plays the role of a free energy as the transition is traversed (Duxbury et al., 1999), and as such the second derivative couples

Table 3.1: Set of 26 structurally diverse protein analyzed using FIRST. The PDB code, protein name, and CATH (Orengo et al., 1997) structural class are listed in the first three columns. N_{res} is the number of residues in the protein; N_{H_2O} is the number of buried water molecules in the protein. $\langle r \rangle_T$ is the mean coordination of the protein in the transition state of the protein, identified as the inflection point in the plot of f' vs $\langle r \rangle$. $\langle r \rangle_{FC}$ is the mean coordination of the protein when the folding core has been identified (described in Chapter 4).

Code	Protein Name	Class	N_{res}	N_{H_2O}	$\langle r \rangle_T$	$\langle r \rangle_{FC}$
	monomers					
1a2p	barnase	$\alpha\beta$	108	5	2.41	2.39
1a3k	galectin	β	137	5	2.40	–
1a6m	myoglobin	α	151	7	2.40	2.37
1ake	adenylate kinase	$\alpha\beta$	214	14	2.40	–
1bpi	bovine pancreatic trypsin inhibitor	*few*	58	4	2.39	2.38
1bu4	ribonuclease T1	$\alpha\beta$	104	0	2.40	2.39
1hml	α-Lactalbumin	α	123	4	2.40	2.38
1hrc	cytochrome c	α	105	4	2.42	2.38
1nkr	killer cell inhibitor receptor	β	201	5	2.39	–
1ruv	ribonuclease A	$\alpha\beta$	124	3	2.41	2.40
1rx1	DHFR	$\alpha\beta$	159	0	2.41	–
1ten	tenascin	β	90	0	2.40	–
1ubi	ubiquitin	$\alpha\beta$	76	1	2.39	2.40
2chf	CheY	$\alpha\beta$	128	7	2.39	–
2ci2	chymotrypsin inhibitor	$\alpha\beta$	83	0	2.40	2.41
2liv	LIV-binding protein	$\alpha\beta$	344	7	2.40	–
3lzm	T4 lysozyme	α	164	7	2.41	2.38
4i1b	interleukin 1-β	β	153	9	2.40	2.39
	dimers					
1bif	PFKinase/FBPase	$\alpha\beta$	864	242	2.40	–
1cku	electron transfer protein	*few*	170	4	2.40	–
1hhp	HIV-1 protease	β	198	0	2.39	–
1vls	aspartate receptor	β	292	32	2.39	–
	tetramers					
1ice	interleukin 1-β converting enzyme	$\alpha\beta$	514	19	2.41	–
1ids	Fe-SOD	$\alpha\beta$	792	43	2.40	–
1szj	GAPDH	$\alpha\beta$	1332	105	2.40	–
2cts	citrate synthase	α	874	60	2.40	–

to the fluctuations and reaches a maximum at the transition point as shown in Figure 3.12.

The second derivative, shown in Figure 3.12, is noisier, due to the numerical differentiations, but nevertheless shows similar behavior for the 26 proteins, with the peak that defines the transition state occurring at $\langle r \rangle = 2.405 \pm 0.015$. There is no obvious pattern in size, architecture, oligomeric state, or ligand content for the few proteins with irregular curves. Cytochrome *c* (PDB code: 1hrc) is the one protein with a bimodal curve that decreases near the transition region, and this behavior occurs both when the heme group is included or excluded from the calculation. Proteins with somewhat broad peaks and a shoulder at lower $\langle r \rangle$ values are α-lactalbumin (PDB code: 1hml), barnase (PDB code: 1a2p), and glyceraldehyde-3-phosphate dehydrogenase (PDB code: 1szj). The behavior of all proteins becomes predictably noisier at low mean coordination values, as more and more hydrogen bonds are removed from the native structure. The insert in Figure 3.12 compares these results with the specific heat curve for a typical protein (Privalov, 1996; Angell, 1999). The shape of the second derivative in Figure 3.12 is suggestive of a relationship with the specific heat, as sketched in the insert. The two quantities are similar in that both are related to fluctuations, with specific heat reflecting fluctuations in the energy, and f'' representing fluctuations in conformational flexibility. It is unclear whether the width of the measured specific heat, as typically measured experimentally, is associated with a single protein, or whether it is broadened due to monitoring an ensemble of unfolding proteins. The specific heat of a single protein as it unfolds thus could be considerably narrower than the measured specific heat, which will be known once experiments can be done on single proteins.

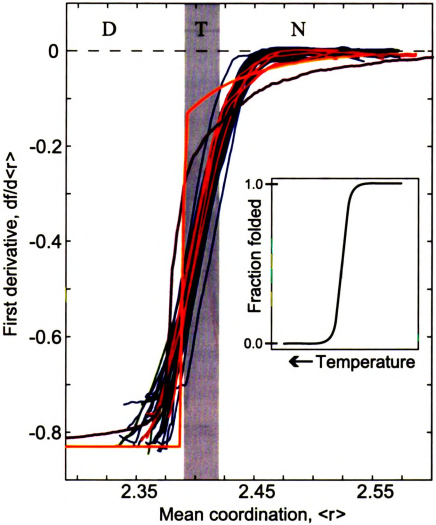

Figure 3.11: Change in the fraction of floppy modes, f', as a function of mean coordination, $\langle r \rangle$, for the set of 26 proteins listed in Table 3.1. The gray shaded region shows the location where the folding transition takes place. The curves for two kinds of glass networks from Figure 3.7 (thick orange and purple lines) are shown superimposed on the protein curves. The notation at the top indicates the **D**enatured state, **T**ransition state, and the **N**ative states of the proteins. For comparison with results for a typical thermal denaturation experiment, the inset sketches the decrease in fraction of folded protein as temperature increases (adapted from Figure 7.11 in (Creighton, 1993))

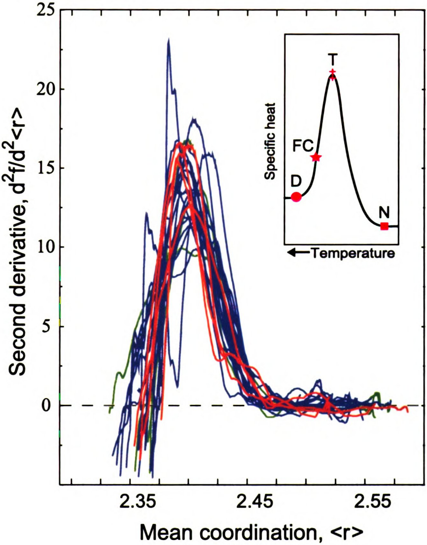

Figure 3.12: The second derivative of the fraction of floppy modes, f'', as a function of mean coordination, $\langle r \rangle$, for the set of 26 proteins listed in Table 3.1. The inset shows a sketch of the specific heat as a function of temperature for a protein, with the location of the **D**enatured state, **F**olding core, **T**ransition state, and the **N**ative state of the protein indicated. The x–axis of the inset has the temperature increasing to the left.

3.5 Conclusions

In this chapter, a novel distance constraint approach for characterizing the intrinsic flexibility of a protein structure has been presented. Hydrophobic interactions and the strong bond forces, covalent bonds, salt bridges and hydrogen bonds, impose constraints on the allowed motion in a protein structure. FIRST uses these constraints to decompose a structure into rigid clusters, consisting of nonrotatable dihedral angles, and flexible regions. There are several advantages of FIRST relative to previous methods for analyzing protein flexibility. FIRST calculations can be done virtually in real time (a few seconds of CPU time) once the network of distance constraints has been defined. Analysis of a native-state protein structure indicates regions likely to undergo conformational change as part of the protein's function. For a given set of distance constraints, the rigid regions and the flexible joints between them are determined exactly. The ability to very quickly determine coupled motions among the dihedral angles of a flexible region gives FIRST an advantage over other methods. Collective motions, in which changing one flexible dihedral angle will influence the other flexible dihedral angles within the region, are identified within the protein. Analysis of the relative flexibility within HIV protease (presented in this chapter), dihydrofolate reductase, and adenylate kinase (Jacobs et al., 2001), even when performed on a single structure, captures much of the functionally important conformational flexibility observed experimentally between different ligand-bound states.

In addition to native-state flexibility analysis, a simple model of protein unfolding by thermal denaturation was presented. In this model, it is assumed that the rigid clusters defined by FIRST represent regions of the protein that are folded. Because flexible regions contain rotatable bonds they are able to sample conformational space, and therefore represent unfolded regions of the protein. Thermal denaturation was simulated by breaking hydrogen bonds in order of their energy, weakest first. The resulting protein folding transition can be viewed as a flexible to rigid phase transition, similar to that observed for network glasses. The mean coordination, $\langle r \rangle$, of atoms in the protein, including non-covalent interactions, can be regarded as the reaction coordinate controlling protein folding, and provides a unifying treatment of the many dynamic and structural processes involved. Proteins are self-organized networks, due to the special nature of the cross-linking of the polypeptide chain via hydrophobic contacts and hydrogen bonds. This transition is shared among diverse proteins ranging from all-α to all-β folds, and from monomers to tetramers, and occurs once the protein denatures to a mean coordination of $\langle r \rangle \cong 2.41$, which is very similar to the value found in network glasses (Thorpe et al., 1999).

Chapter 4

Identifying Protein Folding Cores from the Evolution of Flexible Regions During Unfolding

Research presented in this chapter is being published as the following reference:

B. M. Hespenheide, A. J. Rader, M. F. Thorpe, and L. A. Kuhn. Identifying protein folding cores from the evolution of flexible regions during unfolding. *J. Mol. Graph. Model.*, In press.

4.1 Abstract

The unfolding of a protein can be described as a transition from a predominantly rigid, folded structure to a denatured state, or an ensemble of denatured states. During unfolding, the hydrogen bonds and salt bridges break, destabilizing the secondary and tertiary structure. Previous work (described in Chapter 3) shows that the network of covalent bonds,

salt bridges, hydrogen bonds, and hydrophobic interactions, forms constraints that define which regions of the native protein are flexible or rigid (structurally stable). Here, thermal denaturation of protein structures is simulated by diluting the network of salt bridges and hydrogen bonds, breaking them one by one, from weakest to strongest. The structurally stable and flexible regions are identified at each step, providing information about the evolution of flexible regions during denaturation. This approach is used to test the hypothesis that the folding core is the region of strongest tertiary interactions, and greatest structural stability. For ten diverse proteins, the folding core is identified as the region formed by two or more regular secondary structures that is most stable against thermal denaturation. For the ten proteins with different architectures the predicted folding cores from this flexibility/stability analysis are in good agreement with those identified by native-state hydrogen-deuterium exchange experiments.

4.2 Introduction

Understanding protein folding pathways has been the subject of many recent theoretical and experimental studies (Onuchic et al., 1997; Gruebele, 1999; Shea and Brooks III, 2001; Mirny and Shakhnovich, 2001; Jackson, 1998; Englander, 2000; Eaton et al., 2000; Vendruscolo et al., 2001). These studies often focus on processes that occur early in folding, and models such as nucleation-condensation (Fersht et al., 1992; Clarke and Itzhaki, 1998; Fersht, 2000) and diffusion-collision (Karplus and Weaver, 1994) have been used to describe the initial step(s). Whether folding is initiated by nucleation of tertiary interactions or diffusion-controlled coalescence of already folded secondary structures is being debated, and a single model may or may not hold for all proteins. However, a unifying theme is that the initial steps in the folding process involve the interaction of non-local regions in the protein sequence forming a substructure that is substantially preserved in the fully folded protein. Several experimental techniques have been designed to identify early folding substructures (Galzitskaya and Finkelstein, 1999; Torshin and Harrison, 2001; Hilser et al., 1998). These techniques are unique in that the analysis is performed solely on the native-state conformation, instead of following the folding reaction from a denatured state to the native state. The advantage of using the native state is that this conformation is largely ordered, whereas the denatured state is typically an ensemble of dissimilar, unfolded conformations.

An experimental technique that gives detailed structural information about unfolding is hydrogen-deuterium exchange NMR (H-D exchange). Under native conditions, rotation about main-chain Φ/Ψ dihedral angles leads to fluctuations in which a protein can locally explore conformational space. H-D exchange occurs when the amide and carbonyl groups involved in a hydrogen bond separate enough for deuterated water to intervene, allowing the shared proton to be replaced by a deuteron (Englander et al., 1997). Because deuterium does not produce a signal in proton NMR experiments, it is possible to identify which amides undergo hydrogen exchange by comparing the NMR spectra before and after the exchange. By allowing the experiment to run for different time steps, individual exchange rate constants can be assigned to each of the main-chain amide protons identified in the NMR spectra. Woodward has proposed that amide protons that exchange only after long periods of exposure to deuterated water define the slow-exchange core of a protein

(Woodward, 1993). Li and Woodward compiled the results from a number of studies on native-state H-D exchange for different proteins, tabulating the residues forming the slow-exchange core in each protein (Li and Woodward, 1999). They have proposed that the secondary structures to which these residues belong define the folding core for the protein. Additionally, they have shown for barnase and chymotrypsin inhibitor 2 (CI2) that the folding core identified by H-D exchange consists of residues with high Φ-values (Oliveberg and Fersht, 1996), indicating that slow-exchange core residues contribute to the stabilization of the folding transition state.

For H-D exchange to occur in main-chain amides involved in hydrogen bonds, flexibility in the protein structure is required to allow access to deuterated water. Given that residues in the folding core have small exchange rates, it is reasonable to assume that the folding core protons either are not accessible to solvent or are in regions that are sufficiently rigid that the hydrogen bond donor and acceptor cannot move apart enough to allow H-D exchange. This could be probed by observing how the flexibility of a protein structure changes as it is gradually denatured.

The hypothesis is that the folding core is stabilized by a network of particularly dense and/or strong noncovalent interactions, which tend to resist unfolding or denaturation. Following this hypothesis, a novel computational method for predicting the folding core of a protein is presented. This approach employs the FIRST software, which accurately predicts flexible regions in proteins (Jacobs et al., 1999; Thorpe et al., 2000; Jacobs et al., 2001) by analyzing the constraints on flexibility formed by the covalent and noncovalent bond network. Covalent bonds, salt bridges, hydrogen bonds, and hydrophobic interactions are included in the protein model. Because thermal denaturation or unfolding involves the breaking of hydrogen bonds and salt bridges, we compare several methods for simulating thermal denaturation, and observe how the removal of these bonds affects the stability and flexibility of the protein. As hydrogen bonds are removed, the protein structure becomes more and more flexible as the stable regions decrease in size. The folding core can then be predicted as the most stable region involving at least two secondary structures. The thermal denaturation model in which hydrogen bonds and salt bridges are removed from weakest to strongest predicts folding cores that correlate best with the experimentally observed folding cores. The ability to predict an early state in folding indicates that information about the folding pathway is encoded in the structure of the native state.

4.3 Methods

4.3.1 Selection of Proteins for Analysis

Crystallographic structures for ten monomeric proteins (Table 4.1) were selected from the PDB (Berman et al., 2000) for analysis. These proteins were chosen based on their diversity of structure and the availability of native state H-D exchange data for comparison (Li and Woodward, 1999). A 3D structure was not available for apo-myoglobin (which lacks heme), though qualitative data show its fold is very similar to that of holo-myoglobin (with heme), except for dynamic fluctuations of the F helix (Fontana et al., 1997). As an approximation to the apo-myoglobin structure, we analyzed the holo structure upon removal of its heme group. For this structure, FIRST analysis also found the F helix to be one of

Table 4.1: Dataset of 10 proteins used to identify folding cores. The PDB code and number of residues are listed for each protein. The fourth column gives the CATH (Orengo et al., 1997) structure classification for each protein. The mean coordination of each protein at that point in the hydrogen bond dilution when the folding core is found, $<r>_{core}$ is listed in column 5. Number of H_2O lists the number of buried water molecules identified by PRO_ACT (Williams et al., 1994)

Protein Name	PDB Code	Size (Res.)	Stuct. Class	$<r>_{core}$	Number of H_2O	Number of S-S Bonds
BPTI	1bpi	58	few	2.38	4	3
Ubiquitin	1ubi	76	α-β	2.40	1	0
CI2	2ci2	83	α-β	2.41	0	0
Ribonuclease T1	1bu4	104	α-β	2.39	0	2
Cytochrome *c*	1hrc	104	α	2.39	4	0
Barnase	1a2p	110	α-β	2.39	5	0
α-Lactalbumin	1hml	123	α	2.38	4	0
Apo-myoglobin	1a6m	151	α	2.37	11	0
Interleukin-1β	1i1b	153	β	2.39	9	0
T4 Lysozyme	3lzm	164	α	2.38	7	0

the two most flexible helices in the protein (data not shown). The experimental results of H-D exchange used for comparison in this study are for apo-myoglobin. The proteins were preprocessed as described in Chapter 3 under Methods: Preprocessing Protein Structures for Analysis.

4.3.2 FIRST Flexibility Analysis

The structural flexibility of a protein structure is a property that depends upon how the motion of each atom is restricted by bond forces. In the absence of noncovalent forces, the single covalent bonds in a protein could rotate about any dihedral angle that did not result in steric overlap. The protein would be free to adopt a large number of conformations with comparable energies. Thus, it is the *non*covalent forces that largely define the secondary, tertiary, and quaternary structure observed in proteins. The noncovalent interactions, such as hydrogen bonds and hydrophobic interactions, impose constraints on bond rotation that can be observed by identifying the stable and flexible regions in a protein structure. The software FIRST (Floppy Inclusions and Rigid Substructure Topography) is used to represent the covalent and noncovalent constraints present in a protein and to compute the resulting flexibility of the main chain and side chains (Thorpe et al., 2000; Jacobs et al., 2001). Because it is the macroscopically significant flexibility that I am interested in, rather than the high-frequency fluctuations associated with thermal motion, bond lengths and angles are assigned their equilibrium values as observed in the crystal structure. These fixed bonds lengths and angles give rise to distance constraints between pairs of atoms in the protein, either explicitly from chemical bonds or implicitly from other local bond lengths and angles. For example, each of the covalent bonds between adjacent N, C_α, and C atoms in the backbone has a constant bond length and forms a constant bond angle (Figure 4.1).

This fixes the distance, shown as a dashed gray line in Figure 4.1, between the second nearest neighbor N and C atoms. All such fixed bond angles can be represented by the associated distance constraints. In this manner, all the distance constraints that arise due to covalent bonds and angles are identified, and constraints for nonrotatable peptide and other double or partial double bonds, as well as those arising from salt bridges, hydrogen bonds, and hydrophobic interactions are added, as described above (detailed in (Rader et al., 2001)). FIRST uses 3D constraint counting (Jacobs et al., 2001) on this network of distance constraints to identify the flexible and rigid (structurally stable) regions within a protein. The results of FIRST native-state flexibility analysis have been shown to com-

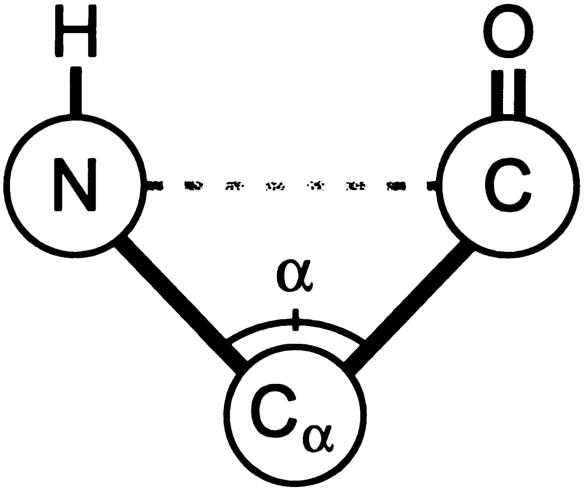

Figure 4.1: Example of bond-length and bond-angle distance constraints for the main-chain atoms of an amino acid. The positions of the N, C_α, C atoms are crystallographically defined, and the sp^3 hydridization of the C_α atom defines the bond angle α. Because the angle α is constant, the distance between the N and C atoms, shown as dashed, gray line, is also constant. The thick black lines between the N–C_α and C_α–C atoms represent bond-stretching distance constraints that arise from the backbond covalent bonds.

pare well with experimental definitions of flexible regions in a series of proteins including lysine-arginine-ornithine binding protein (Jacobs et al., 1999), cytochrome *c* (Thorpe et al., 2000), HIV protease, adenylate kinase, and dihydrofolate reductase (Jacobs et al., 2001).

4.3.3 Simulating Denaturation

As a protein is gradually thermally denatured, the covalent bonds remain intact, whereas hydrogen bonds begin to break. The flexibility in the protein will increase as the number of hydrogen bonds in the protein decreases. Our hypothesis is that the folding core is the region that will remain structurally stable the longest under denaturing conditions. This hypothesis was tested by removing hydrogen bonds from a protein structure to simulate thermal denaturation, then using FIRST to observe where the resultant flexibility occurs. The results will depend upon the order in which hydrogen bonds are removed. Because hydrophobic interactions actually become somewhat stronger with moderate temperature increases (Tanford, 1980), these interactions are maintained throughout the simulation. Three methods for diluting the hydrogen bond network of a protein are presented, each designed to test the importance of the strength and/or density of the hydrogen bonds when selecting which bond to remove next.

1. Thermal Denaturation. As the temperature of a protein is gradually increased, the hydrogen bonds are expected to break in an energy-dependent manner. This process is simulated by using the following procedure. Initially, the flexibility of the native protein structure is analyzed with all its covalent and noncovalent interactions included (hydrogen bonds and hydrophobic interactions). The weakest hydrogen bond in the structure is then

broken by removing any distance constraints imposed by that bond. The effect of removing this bond is then observed by applying FIRST to identify the flexible regions in the protein. This process of breaking the weakest hydrogen bond remaining in the structure and updating the identification of flexible regions is continued until all the hydrogen bonds have been removed.

2. Random Removal of Noncovalent Bonds Over a Small Energy Window. The thermal denaturation method described in (1) removes hydrogen bonds strictly in order of energy. To introduce some noise into the simulation, the next hydrogen bond to be removed is randomly selected from the 10 weakest bonds remaining in the protein. This modification was designed to reflect the stochastic nature of thermal denaturation and to test the effect of inaccuracies in the hydrogen–bond energy function. The results of this simulation should also indicate that the small fluctuations expected to occur during thermal denaturation do not significantly affect the flexibility or folding core predictions.

3. Completely Random Removal of Noncovalent Bonds. To check whether the relative energies of hydrogen bonds, and not just their density in the structure, are indeed important in thermal denaturation, completely random dilutions of the hydrogen bonds in a protein, without respect to their energies, have been performed. In this case, the next hydrogen bond to be removed from the protein is selected randomly from all remaining hydrogen bonds.

4.3.4 Identifying the Folding Core

The native-state flexibility of a protein structure is computed using FIRST with all noncovalent interactions present. Generally, in the native state, most of the residues belonging to an α-helix or β-strand are rigid, and the secondary structures are mutually rigid. As the hydrogen bonds are removed from the protein, parts of the secondary structures may become flexible, such as the ends of a helix or strand. Also, the secondary structures tend to become independently rigid at intermediate steps in denaturation, due to loss of tertiary hydrogen bonds.

The protein folding core is defined in this study as the set of secondary structures that remain mutually rigid the longest in the simulated denaturation. The secondary structures for the native states of each of the ten proteins were identified by using the program DSSP (Kabsch and Sander, 1983) and tracked during the unfolding simulation. Not all residues in the secondary structure are required to be rigid when identifying the folding core. An α-helix is considered to be rigid if at least 5 consecutive residues, corresponding to one complete turn of an α-helix, belong to the rigid cluster. If a helix is defined by DSSP to contain fewer than 5 residues, as can occur with 3_{10} helices, all its residues must be mutually rigid to be considered a rigid secondary structure. The β-strands are required to have at least 3 consecutive residues rigid to be considered as part of the folding core. This criterion of three consecutive rigid residues allows for at least 2 hydrogen bonds to an adjacent strand. If a strand is defined by DSSP as consisting of less fewer than 3 residues, the entire strand is required to be rigid to be counted as part of the folding core.

4.4 Results

4.4.1 Thermal Denaturation

For cytochrome *c*, the native state is composed of a single, structurally stable region represented by the top line in Figure 4.2, and the 3D structure shown at the right. When hydrogen bonds 114 through 65 (the weakest 50) were removed, the large rigid cluster (colored red) significantly decreased in size (at the fifth line in panel A), resulting in new flexibility in those residues between the N- and C-terminal helices. These helices formed the only significantly rigid region in the protein. The folding core was predicted as the last point in the denaturation when at least two secondary structures formed a single rigid region. This point in cytochrome *c* occurred in the fifth-to-last line, where the N- and C-terminal helices are mutually rigid. On the next line, no single rigid cluster contained more than one secondary structure. The predicted folding core is shown structurally at bottom right, and summarized in a 1D representation just below the denaturation results, along with the folding core determined by H-D exchange (Li and Woodward, 1999; Jeng et al., 1990), shown in orange. The predicted and observed folding cores correspond well, both indicating that the N- and C-terminal helices together form a stable folding core.

Detailed unfolding pathway and folding core predictions upon thermal denaturation are shown for barnase in Figure 4.3. There was a significant change in the flexibility of the protein observed after 34 hydrogen bonds had been removed (fourth line from the top), in this case resulting in several small rigid regions that could move independently of one another (as indicated by their different colors in the plot), and one large rigid region (shown

Figure 4.2: Results of simulated thermal denaturation for cytochrome *c*. This figure shows how the structure fragments into smaller rigid regions, with intervening flexible bonds, as the hydrogen bond network denatures with increasing temperature. α-helices within the native structure are indicated as red zigzags at the top. Shown at right is the 3D RCD representation of the largest rigid cluster (colored red) in the protein for the native state (top), and intermediate, partially unfolded state (middle) and the folding core (bottom), defined here as the last point in denaturation at which the largest rigid region consists of more than one secondary structure. The summary of the folding core prediction, shown at the bottom, indicates that there is close correspondence between the prediction of the folding core as the most stable supersecondary region and the folding core as defined by protection from H-D exchange (Li and Woodward, 1999)

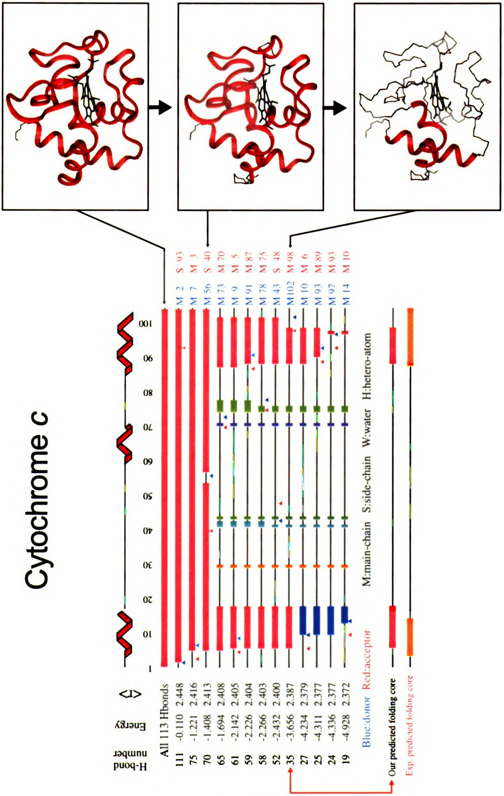
Cytochrome c
H-bond number
Energy
<I>
1
10
20
30
40
50
60
70
80
90
100
All 113 Hbonds
111 -0.110 2.448 M 2 S 93
75 -1.221 2.416 M 7 M 3
70 -1.408 2.413 M 56 S 40
65 -1.694 2.408 M 73 M 70
61 -2.142 2.405 M 9 M 5
59 -2.226 2.404 M 91 M 87
58 -2.266 2.403 M 78 M 75
52 -2.432 2.400 M 43 S 48
35 -3.656 2.387 M102 M 98
27 -4.234 2.379 M 10 M 6
25 -4.311 2.377 M 93 M 89
24 -4.336 2.377 M 97 M 93
19 -4.928 2.372 M 14 M 10
Blue:donor Red:acceptor
M:main-chain S:side-chain W:water H:hetero-atom
Our predicted folding core
Exp. predicted folding core

Figure 4.2

in red). Our study of folding transition states has shown that the rigid core of proteins disintegrates into several independent rigid regions when the mean coordination decreases below ~2.415. This is seen for both barnase and cytochrome *c* in this figure, yielding a transition between rigid and flexible states that is also found for network glasses near this same mean coordination (Rader et al., 2001). An intermediate structural state in barnase is formed by the packing of an α-helix against the β-sheet (second structural panel at right in Figure 4.3). This super-secondary structure recedes to form the folding core itself, consisting of the α-helix packed against part of the β-sheet (fourth line from bottom in Figure 4.3, with structure shown in last panel at right). The H-D exchange folding core, shown at bottom (orange), matches the predicted folding core (red) well, with the exception of the short, C-terminal β-strand. Figure 4.4 shows the unfolding pathway for interleukin-1β, a protein whose secondary structure content consists entirely of β-strands. The structure shows little breakup during the initial steps of the unfolding simulation. A significant event occurred when hydrogen bond 106 was broken, resulting in flexibility for a large portion of the structure. The β-strands formed by residues between 50 and 130 remain rigid, and eventually are identified as the folding core on the fourth line from the bottom. A comparison to the experimental folding core, shown at the bottom in orange, shows good overlap.

For completeness, the hydrogen bond dilution results for bovine pancreatic trypsin inhibitor (BPTI) are shown in Figure 4.5. BPTI is a member of the DSSP class "few" due to its small size and few secondary structures. The unfolding path represented in Figure 4.5 shows a gradual breakup of the structure into small flexible regions. The N-terminal helix becomes flexible when hydrogen bond 29 is broken, followed by the C-terminal helix

Figure 4.3: Results of simulated thermal denaturation for barnase. The secondary structure content of barnase is depicted at the top of the figure; α-helices, red zigzags; β-strands, yellow arrows. The 3D RCD of the largest rigid cluster (colored red) is shown on the right side for barnase in the native state (top), a transition state (middle) and the folding core (bottom). The predicted folding core, identified on the fourth line from the bottom, is compared to the experimentally defined folding core (colored orange) at the bottom. There is good overlap between the predicted and experimental results.

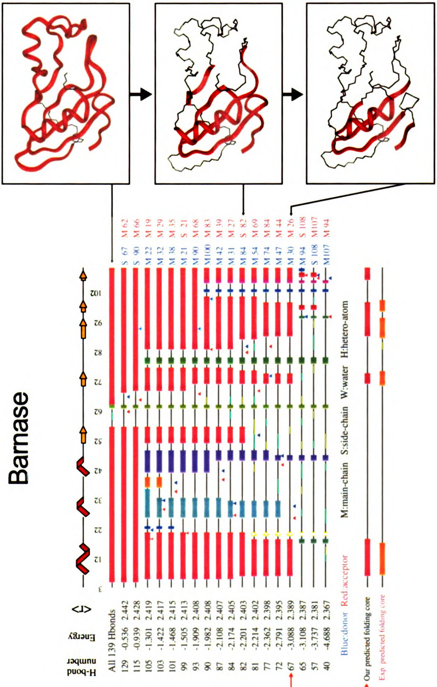
Barnase
H-bond number
Energy
<r>
3
12
22
32
42
52
62
72
82
92
102
All 139 Hbonds
129 -0.536 2.442 S 67 M 62
115 -0.939 2.428 S 90 M 66
105 -1.301 2.419 M 22 M 19
103 -1.422 2.417 M 32 M 29
101 -1.468 2.415 M 38 M 35
99 -1.505 2.413 M 21 S 21
93 -1.909 2.408 M 90 M 68
90 -1.982 2.408 M100 M 83
87 -2.108 2.407 M 42 M 39
84 -2.174 2.405 M 31 M 27
82 -2.201 2.403 M 84 S 82
81 -2.214 2.402 M 54 M 69
77 -2.362 2.398 M 74 M 84
72 -2.791 2.395 M 47 M 44
67 -3.088 2.389 M 30 M 26
65 -3.108 2.387 M 94 S 108
57 -3.737 2.381 S 108 M107
40 -4.688 2.367 M107 M 94
Blue:donor Red:acceptor
M:main-chain S:side-chain W:water H:hetero-atom
Our predicted folding core
Exp. predicted folding core

Figure 4.3

Figure 4.4: Results of simulated thermal denaturation for interleukin-1β. The secondary structure content of this protein is entirely β-strands. Their location is indicated by the yellow arrows at the top of the figure. The folding core for interleukin-1β, identified on the fourth line from the bottom, is compared to the experimentally determined folding core (depicted in orange) at the bottom of the figure. There is good overlap between the two.

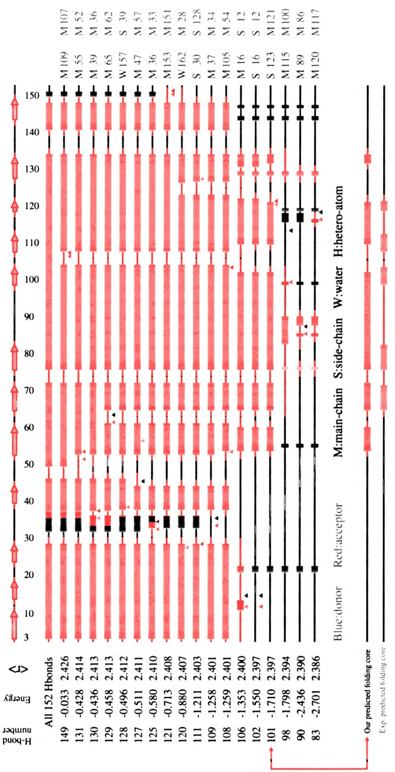
H-bond number
Energy
<J>
All 152 Hbonds
149 -0.033 2.426
131 -0.428 2.414
130 -0.436 2.413
129 -0.458 2.413
128 -0.496 2.412
127 -0.511 2.411
125 -0.580 2.410
121 -0.713 2.408
120 -0.880 2.407
111 -1.211 2.403
109 -1.258 2.401
108 -1.259 2.401
106 -1.353 2.400
102 -1.550 2.397
101 -1.710 2.397
98 -1.798 2.394
90 -2.436 2.390
83 -2.701 2.386
3
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
M 109 M 107
M 55 M 52
M 39 M 36
M 65 M 62
W 157 S 39
M 47 M 57
M 36 M 33
M 153 M 151
W 162 M 28
S 30 S 128
M 37 M 34
M 105 M 54
M 16 S 12
S 16 S 12
S 123 M 121
M 115 M 100
M 89 M 86
M 120 M 117
Blue:donor Red:acceptor
M:main-chain S:side-chain W:water H:hetero-atom
Our predicted folding core
Exp. predicted folding core

Figure 4.4

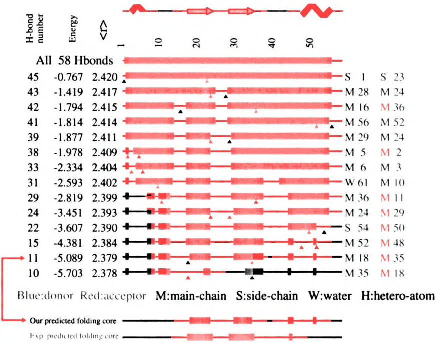

Figure 4.5: Results of simulated thermal denaturation for bovine pancreatic trypsin inhibitor. This small protein with few secondary structures shows a gradual rigid → flexible transition as hydrogen bonds are diluted from the structure. The position of the secondary structures is indicated at the top of the figure; α-helices, red zigzags; β-strands, yellow arrows. The predicted folding core is identified on the second line from the bottom, and is compared to the experimental folding core (in orange) at the bottom of the figure. There is very good agreement between the two.

when hydrogen bond 15 is broken. The remaining two secondary structures remain mutually rigid, along with residues 45 and 51, to form the predicted folding core of BPTI. The overlap between the predicted and the experimental folding cores, shown at the bottom, is good.

Thermal denaturation simulations were performed to predict the folding core for each protein in our dataset. Figure 4.6 summarizes the results from these simulations, comparing the predicted folding core to the observed folding core. For a majority of the proteins (8 out of 10), the folding core predictions agree well with folding cores predicted by regions of slow H-D exchange, and often involve tertiary interactions between sequence-distant secondary structures. For α-lactalbumin, half of the folding core region is in agreement, and for T4 lysozyme, the folding core identified by experiment is much larger than that identified by flexibility analysis. Given that different experimental conditions can also produce different results, it is planned to consult a broader range of experimental probes of T4 lysozyme folding, as well as doing further structural analysis. However, given the diverse structures and folding mechanisms for these ten proteins, the overall good agreement between theory and experiment suggests that flexibility analysis is a useful tool for probing the stability of substructures, in particular the folding core, along the folding/unfolding pathway. This approach provides explicit 3D structural maps of the stable regions predicted in the protein at each step during denaturation, as well as providing a model for the interactions important in stabilizing folding cores: a dense network of hydrogen–bond interactions that augment the ubiquitous, but less specific, hydrophobic interactions.

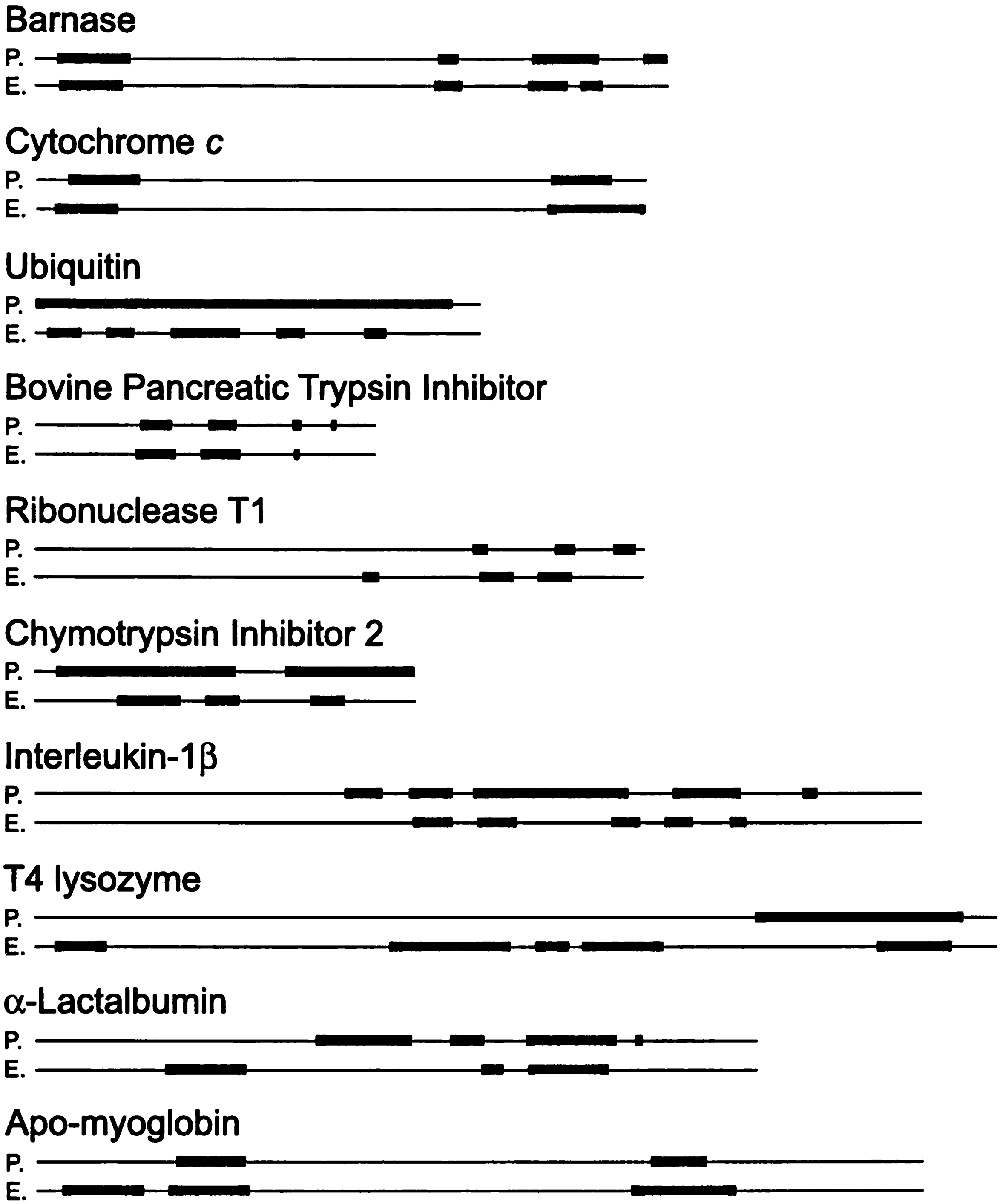

Figure 4.6: Comparison of the folding core predicted by FIRST flexibility analysis (P) to the observed folding core of H-D exchange experiments (E) for barnase (Perrett et al., 1995), cytochrome *c* (Jeng et al., 1990), ubiquitin (Pan and Briggs, 1992), BPTI (Woodward and Hilton, 1980), ribonuclease T1 (Mullins et al., 1997), CI2 (Neira et al., 1997), interleukin-1β (Driscoll et al., 1990), T4 lysozyme (Anderson et al., 1993), α-lactalbumin (Schulman et al., 1995) and apo-myoglobin (Hughson et al., 1990)

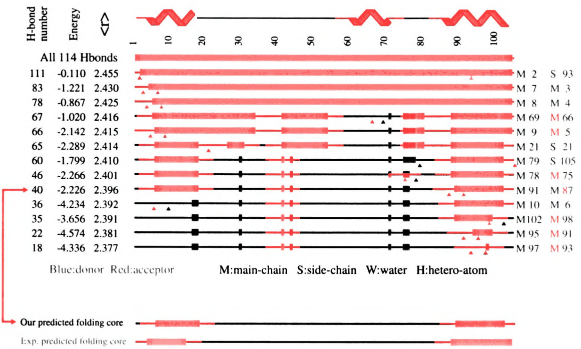

Figure 4.7: Results of random hydrogen bond dilution over a window of 10 hydrogen bonds for cytochrome *c*. Denaturation is simulated by removing hydrogen bonds as in the thermal denaturation method, however, instead of always removing the weakest hydrogen bond next, a hydrogen bond is randomly selected from the 10 weakest hydrogen bonds in the protein. Beneath the figure the predicted folding core (red) is compared to the observed folding core (orange). The similarity in folding core predictions between this result and that of thermal denaturation simulation (Figure 4.2) indicate that the results of simulated thermal denaturation are robust.

4.4.2 Evaluating Other Models of Denaturation

Figure 4.7 shows the result of simulating cytochrome *c* denaturation by removing a hydrogen bond randomly from the ten lowest-energy bonds in the protein at each step. It can be seen in the second column on the left that the energies of the bonds being removed are generally becoming more negative (stronger), however they are not removed strictly from weakest to strongest energy as in the thermal denaturation (Figure 4.2). This approach tests the robustness of the thermal denaturation scheme to thermal fluctuations or some inaccuracy in the calculation of hydrogen–bond energies. Comparing Figure 4.2 to Figure 4.7 shows that introducing some randomness into the thermal denaturation has little effect on accurate prediction of the folding core for cytochrome c, and mainly predicts a more rigid unfolding intermediate state between -1.4 and -2.2 kcal/mol. Twenty separate runs were performed with different random selection of the hydrogen bonds removed, and all runs predicted the same folding core (data not shown).

As an extreme example of a random dilution, we simulated denaturation in which the hydrogen bond energies were not taken into account. Each hydrogen bond was weighted equally, and the next bond to be removed was chosen randomly from all hydrogen bonds remaining in the protein. If the folding core of a protein could be identified solely by having the highest density of covalent bonds, hydrogen bonds and hydrophobic interactions, regardless of their energy, the results for this approach would be accurate. Four separate, random denaturation simulations for cytochrome *c* are shown in Figure 4.8. Below each panel, a comparison between the folding core predicted from this simulation and the experimentally observed folding core is shown. Panel C in Figure 4.8 shows that a completely random simulation can produce a correct folding core prediction and have similar intermediate features to thermal denaturation according to hydrogen- bond energy (compare with Figure 4.2). However, the other panels in Figure 4.8 indicate that a random hydrogen bond removal scheme most commonly mispredicts the folding core. These results show that the energy of hydrogen bonds is a significant factor in simulating the denaturation and unfolding of proteins, as validated by folding core prediction.

Figure 4.8: Four completely random dilutions of the hydrogen bonds in cytochrome *c*. Each panel represents a single unfolding simulation in which the hydrogen bonds were removed in random order. The secondary structures are shown at the top of each panel (the red zigzags represent α-helices). The predicted folding core from each panel is compared to the observed folding core (in orange) at the bottom of each panel. The panel at the lower left shows that an accurate folding core prediction can by chance be obtained from a completely random hydrogen bond removal scheme. However, the results in the other three panels are in poor agreement with the observed folding core. These data indicate that density of hydrogen bonds alone is not the sole determinant when forming a folding core.

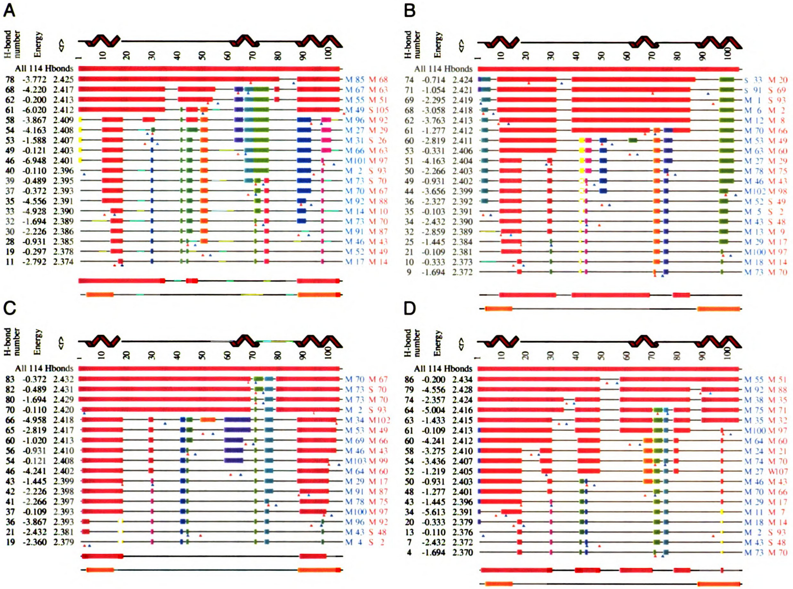
A
H-bond number
Energy
<r>
1
10
20
30
40
50
60
70
80
90
100
All 114 Hbonds
78 -3.772 2.425 M 85 M 68
68 -4.220 2.417 M 67 M 63
62 -0.200 2.413 M 55 M 51
61 -6.020 2.412 M 49 S 105
58 -3.867 2.409 M 96 M 92
54 -4.163 2.408 M 27 M 29
53 -1.588 2.407 M 31 S 26
49 -0.121 2.403 M 66 M 63
46 -6.948 2.401 M101 M 97
40 -0.110 2.396 M 2 S 93
39 -0.489 2.395 M 73 S 70
37 -0.372 2.393 M 70 M 67
35 -4.556 2.391 M 92 M 88
33 -4.928 2.390 M 14 M 10
32 -1.694 2.389 M 73 M 70
30 -2.226 2.386 M 91 M 87
28 -0.931 2.385 M 46 M 43
19 -0.297 2.378 M 52 M 49
11 -2.792 2.374 M 17 M 14
B
H-bond number
Energy
<r>
1
10
20
30
40
50
60
70
80
90
100
All 114 Hbonds
74 -0.714 2.424 S 33 M 20
71 -1.054 2.421 S 91 S 69
69 -2.295 2.419 M 1 S 93
68 -3.058 2.418 M 6 M 2
62 -3.763 2.413 M 12 M 8
61 -1.277 2.412 M 70 M 66
60 -2.819 2.411 M 53 M 49
53 -0.331 2.406 M 63 M 60
51 -4.163 2.404 M 27 M 29
50 -2.266 2.403 M 78 M 75
49 -0.931 2.402 M 46 M 43
44 -3.656 2.399 M102 M 98
36 -2.327 2.392 M 52 S 49
35 -0.103 2.391 M 5 S 2
34 -2.432 2.390 M 43 S 48
32 -2.859 2.389 M 13 M 9
25 -1.445 2.384 M 29 M 17
21 -0.109 2.381 M100 M 97
10 -0.333 2.373 M 18 M 14
9 -1.694 2.372 M 73 M 70
C
H-bond number
Energy
<r>
1
10
20
30
40
50
60
70
80
90
100
All 114 Hbonds
83 -0.372 2.432 M 70 M 67
82 -0.489 2.431 M 73 S 70
80 -1.694 2.429 M 73 M 70
70 -0.110 2.420 M 2 S 93
66 -4.958 2.418 M 34 M102
65 -2.819 2.417 M 53 M 49
60 -1.020 2.413 M 69 M 66
56 -0.931 2.410 M 46 M 43
54 -0.121 2.408 M103 M 99
46 -4.241 2.402 M 64 M 60
43 -1.445 2.399 M 29 M 17
42 -2.226 2.398 M 91 M 87
41 -2.266 2.397 M 78 M 75
37 -0.109 2.393 M100 M 97
36 -3.867 2.393 M 96 M 92
21 -2.432 2.381 M 43 S 48
19 -2.360 2.379 M 4 S 2
D
H-bond number
Energy
<r>
1
10
20
30
40
50
60
70
80
90
100
All 114 Hbonds
86 -0.200 2.434 M 55 M 51
79 -4.556 2.428 M 92 M 88
74 -2.357 2.424 M 38 M 35
64 -5.004 2.416 M 75 M 71
63 -1.433 2.415 M 35 M 32
61 -0.109 2.413 M100 M 97
60 -4.241 2.412 M 64 M 60
58 -3.275 2.410 M 24 M 21
54 -3.436 2.407 M 74 M 70
52 -1.219 2.405 M 27 W107
50 -0.931 2.403 M 46 M 43
48 -1.277 2.401 M 70 M 66
43 -1.445 2.396 M 29 M 17
34 -5.613 2.391 M 11 M 7
20 -0.333 2.379 M 18 M 14
13 -0.110 2.376 M 2 S 93
7 -2.432 2.372 M 43 S 48
4 -1.694 2.370 M 73 M 70

Figure 4.8

4.5 Conclusions

Several theoretical techniques have been developed to probe protein folding pathways through an analysis of the native state. Galzitskaya and Finkelstein (1999) have developed a technique to computationally analyze the energetics of all possible substructures in the native-state conformation and define a subset of these structures as the transition state ensemble. Computed Φ-values, which measure the similarity between transition-state structure and native-state structure for a given residue, from their ensemble show good correlation to experimentally determined values. Hilser et al. (1998) partition the protein into blocks along the sequence, then generate alternative partitions by shifting these blocks. The blocks are then kept folded or unfolded in all possible combinations to generate an ensemble of states. Folding cooperativity between one residue and all other residues in the protein is assessed by performing an energy-perturbing mutation of the residue, in all its occurrences within folded states, and observing the effects on all other residues. An alternative approach is that of Tsai et al. (1997,2000) in which the native state structure is also partitioned, first into domains (visually), then into potential hydrophobic folding units based upon a scoring function measuring compactness, degree of isolation, and hydrophobicity. A combinatorial approach is then used to reassemble possible folded states from these folding units. Similarly, Wallqvist et al. (1997) partition the structure by using a sequence mask, and assess pair wise and higher-order interactions in a unified-atom representation of the protein by using a knowledge-based folding potential. Essentially, all these approaches exhaustively partition the structure into substructures, and use a potential or scoring function to assess the interactions between substructures as potential intermediate states in folding.

FIRST flexibility analysis also has the goal of identifying structurally stable states along the unfolding/folding pathway, and does so by decoding the hierarchy of structurally stable motifs within the native state. The FIRST program treats a protein structure as a network of atoms and bonds, and the analysis decomposes the structure into rigid regions and flexible regions. I propose that rigid regions represent residues that are folded in native conformation, and likewise flexible regions represent unfolded residues. Given that the experimentally identified folding core represents a region of structure that resists unfolding, FIRST has been used to identify the region of structure that resists becoming flexible as we simulate unfolding. The good correlation between the predicted and experimental folding cores shown in Figure 4.6 supports the concept that the native state structure of a protein, specifically the distribution and strength of the noncovalent forces, does encode information about the folding pathway. Furthermore, because FIRST requires that bond forces be represented as distance constraints, or connections between atoms, the manner in which hydrogen bonds and hydrophobic interactions are modeled appears valid.

The power of FIRST flexibility analysis lies in its simplicity, computational speed (all steps in thermal denaturation of a large protein can be calculated in a minute on a personal computer), and explicit structural description of which regions of the protein are flexible or structurally stable at each step along the unfolding pathway. Using this approach, the phase transition from folded to unfolded can be tracked structurally as rigidity in the protein is lost (Rader et al., 2001), and the folding cores can be identified and are shown to be in good agreement with NMR experimental results.

Chapter 5

Summary and Perspectives

5.1 Secondary Structure Packing

5.1.1 Summary

In Chapter 2, an analysis of observed secondary structure packing geometries in a set of nonhomologous proteins is presented. One way in which this analysis is unique in that a novel coordinate transformation is used to measure the geometry. Alternative methods have relied on measuring helix–sheet interactions by approximating the β-sheet as a plane using three consecutive β-strands, where the strand interacting with the helix is in the middle. Because interstrand hydrogen bonding is not uniform, the effect of projecting three adjacent strands onto a common plane can lead to an overly simplistic representation of the sheet surface. Here, the frame of reference is based solely on the strand and the line of closest approach between the helix and the interacting strand. Also, the analysis presented here explicitly takes into account the N- to C-terminal direction of each secondary structure. By looking for correlations between packing geometry and the N- to C-terminal direction of the structures involved, the possible role of dipole interactions can be evaluated.

As a result of including the N- to C-terminal direction of secondary structures in our analysis, five possible strand orientations must be considered. Each orientation is defined by the direction of a given strand relative to its neighbors, or neighbor in the case of a strand at the end of a sheet. The observed helix-sheet interactions are subdivided into five categories depending on the orientation of the strand. It is shown that for strand orientations that arise due to antiparallel hydrogen bonding between neighboring strands, no favorable packing angle Ω is observed. For strands in which the neighbors are hydrogen bonding in parallel, a strong preference for the helix to align antiparallel to the strand is observed.

The interesting feature of the helix-sheet packing results is that they can be divided into two categories, those in which dipoles are present in the sheet, and those in which dipoles are absent. Furthermore, the class of interactions involving dipoles show a strong preference for packing angle, whereas the geometries observed for sheets with no dipole have helix-sheet packing angles that are nearly random. It is clear from the hydrogen bonding pattern of helices and parallel strands (Figure 2.5B) that a net dipole exists in these structures. One could therefore hypothesize that these structures would interact such as to optimize the dipole interactions. In fact this is what the results show (Figure 2.2).

5.1.2 Perspective

The observation that almost all of the helix-sheet interactions that display a packing angle preference also have favorable dipole interactions may provide insight into the folding mechanism for these proteins. Since a net dipole will only arise after a secondary structure has folded, it suggests that the helix and the sheet (at least that part of the sheet local to the interaction site) were structured before they interacted. This scenario is consistent with the diffusion–collision or hierarchical folding models discussed in Chapter 1. Likewise, for those helix-sheet interactions where a dipole is not present in the sheet, no *a priori* formation of the secondary structures is required. These types of secondary structure interactions would perhaps best fit a nucleation type folding mechanism in which a nucleation site may form near the helix-sheet interaction site. Due to the importance of tertiary interactions in a nucleation model, these nonsequence local interactions between side chains will have an important role in the overall topology of the folded structure. These tertiary interactions then stabilize local bond formation, such as helices and β-turns. In this manner, the direction in which helices and strands propagate depends on the initial topology of the nucleus. Due to the absence of a dipole in the sheet, there is no energetic gain for adopting a specific helix-sheet orientation that optimizes dipole-dipole interactions, and it is likely that the side chains of the respective secondary structures will play the dominant role in determining the packing geometry.

5.1.3 Future Directions

The interpretation of the helix-sheet packing data in the context of the folding mechanism discussed in the last paragraph gives rise to multiple experimental and computational hypotheses which could answer the question, "does helix-sheet packing preference lead to a preferred folding mechanism?" While the DC model would seem to best fit the case where favored helix-sheet geometries are observed, the computational methods used to study this mechanism have generally only been used for all helical proteins. This limitation is due to the difficulty in modeling the diffusion of β structures through solvent. The "anatomy trees" produced by the building block method of Nussinov and coworkers may provide better insight (Tsai et al., 2000). We might expect that the helix and the sheet could be identified as individual building blocks using their method, and that the helix and sheet would then interact on the way to the native state. However, as helices and sheets in nonfavored geometries may also be dissected into building blocks, and the experimental validation of this method is sparse, the correctness of this building block method remains to be seen.

Experimentally, the mechanism of helix-sheet interaction could be elucidated by a variety of methods. For example, H-D exchange coupled with mass spectrometry has recently been shown to be an alternative means to determine H-D exchange rates in proteins (Simmons and Konermann, 2002; Yang and Smith, 1997). Analysis of the exchange rates within a helix and sheet that are interacting may provide some information on how early these structures form during folding. These rates could be compared to the rate at which the helix-sheet interface is formed, as determined by the fluorescence quenching of an intrinsic or engineered tryptophan. If it appears that the tryptophan is buried before the secondary structures form, this would suggest a nucleation type model. However, if the rates suggest that the secondary structures form before the tryptophan is quenched, it would suggest a diffusion–collision type model. If dipoles truly play a significant role in the favorable packing geometries that have been observed, the experiment proposed above should indicate that the secondary structures in these proteins form before the helix-sheet interface. If a correlation between experimentally observed folding mechanism and preferred versus random packing geometry could be found it would provide further evidence that native-state structure encodes information about folding pathways.

Finally, a further analysis of the fine details of the helix-sheet packing interface is certainly warranted. Features such as the amino acid composition of the interface or the percentage of buried surface area could yield additional information with which to correlate packing geometries, and could certainly be beneficial to the protein engineering community in the design of novel proteins or for the advancement of *ab initio* protein folding algorithms. Interestingly, a bridge between the helix-sheet packing analysis of Chapter 2 and the protein flexibility studies in Chapter 3 and 4 could be imagined. If tertiary interactions are more important in helix-sheet interactions with nonspecific geometries, I would expect these interfaces to be more rigid compared to helix-sheet packings that are stabilized by dipole interactions. The details of this experiment would need to be worked out, but it seems viable due to the relative ease of using the FIRST program.

5.2 Protein Folding and Flexibility Analysis

5.2.1 Summary

Chapter 3 outlined a computational approach to study protein folding through analysis of native state structures. The method relies on accurately predicting the changes in flexibility that occur in a protein as it is thermally denatured. Protein flexibility is computed using the program FIRST, which takes as input the covalent bonds, salt bridges, hydrogen bonds and hydrophobic interactions present in a protein and computes for each bond whether it is free to rotate (flexible) or locked (rigid). Predictions of native-state flexibility using FIRST have been shown to correlate well with experimentally observed flexibility in several proteins. These favorable native state correlations prompted the following line of reasoning; if the unfolding of a protein is a transition from a mostly rigid native state to a mostly flexible denatured state, and FIRST can accurately measure flexibility, can FIRST be used to study protein unfolding?

A means of simulating protein unfolding was developed, based on the simple assumption that as a protein is gradually thermally denatured, the noncovalent bonds, specifically the hydrogen bonds and the salt bridges, are expected to break in an energy dependent manner. Ideally, the bonds would break in exact order of their energy. This assumption led to the method of hydrogen bond dilution, in which hydrogen bonds were removed from a protein in order of energy from weakest to strongest, while maintaining a static covalent bond structure. The hydrophobic interactions are also maintained since they are modeled to have a degree of flexibility that we would expect in both the native state and the denatured state. The breaking of a bond is accomplished in FIRST by removing any constraints imposed on the bond network by that bond. After each bond is removed, FIRST analysis is performed, and any changes in the main chain flexibility are noted. After all the hydrogen bonds present in a protein have been broken, a novel graphical representation of the main chain flexibility changes provides a clear view of the structurally stable (folded) and flexible (unfolded) regions of the protein.

Two key comparisons to experimental data helped prove the validity of the hydrogen bond dilution method as a means to study protein unfolding. The first observation, presented in Chapter 3, was that the second derivative of the fraction of floppy modes (independent bond-rotational DOF), f, as a function of mean coordination, $\langle r \rangle$, exhibited behavior much like the specific heat curves measured from calorimetry experiments. This correlation was not surprising, as the number of floppy modes in a bond network has been shown to exhibit free energy like properties, and the second derivative should therefore be a specific heat like quantity. Based on this association, plots of f'' vs. $\langle r \rangle$ could be used to identify the transition state of the unfolding simulation, and the mean coordination value at which the transition occurs. The overall shape of the second derivative plots was shown to be consistent in a set of 26 proteins that varied significantly in structure class, size, and oligomeric state (Rader et al., 2001) Based on these results, it was proposed that the mean coordination of a protein is a relevant structural order parameter for studying protein folding.

The second experiment that supported the use of FIRST flexibility analysis as a probe of protein unfolding was presented in Chapter 4. The hypothesis tested was, that under conditions of gradual thermal denaturation, the supersecondary structure that resists unfolding the longest (our defined folding core) represents an early forming substructure along the folding pathway. To validate our folding core predictions, I compared our results to those of H-D exchange experiments. Native-state H-D exchange can isolate the subset of residues in a protein that exchange very slowly. Because exchange requires some degree of protein unfolding to expose a main-chain amide to solvent, a structural interpretation of slow-exchanging residues is that they resist unfolding. Defining an experimentally observed folding core as the set of secondary structures to which slow-exchange residues belong allows a direct comparison between our predictions and experiment. The results presented in Chapter 4 indicate that there is a strong correlation between our folding core predictions and H-D exchange experimental observations of protein folding cores for 8 out of 10 proteins.

5.2.2 Perspectives

The folding of a protein is under both thermodynamic and kinetic control; proteins fold quickly to a low energy conformation. While the native state of a protein may not be the global free energy minimum, it certainly is one of the deepest, as indicated by the faithful refolding to the same native state observed in many proteins after denaturation. It has been the goal of most theoretical and phenomenological model of protein folding to explain the kinetics and thermodynamics of folding in light of experimental results. I have proposed that FIRST analysis and hydrogen bond dilution can provide information about protein folding, and the results can be interpreted in the context of previous work in the field.

The biggest questions that arise from the data are, "what does each step in a hydrogen bond dilution experiment represent" and "why do the results provide accurate folding core predictions for some proteins and not others?" The simplest interpretation of the data is that each step represents an ensemble of structures that would be observed at equilibrium along a single unfolding pathway, where rigid regions are folded and flexible regions can sample conformational space (resulting in the ensemble). The experimental analog would be a process in which the temperature of a single protein were raised in small increments and allowed to come to equilibrium for a long time period, after which time multiple probes of the protein structure would be measured. The temperature would be continually raised, and properties measured, until the protein was completely denatured. One caveat to this hypothetical experiment is that it would have to ensure that the protein follows only a single pathway. While single molecule studies are becoming more feasible (Carrion-Vazquez et al., 1999), detailed experimental data on the structure of a single protein during a folding reaction remains elusive.

The key points to the interpretation presented above are that the analysis is done at thermodynamic equilibrium, and that they represent a single folding pathway. At equilibrium, the free energy of the protein depends only the structure, which is associated with the fraction of floppy modes. This structural quantity is independent of the path taken to get to this structure, which in this case is the order in which hydrogen bonds are removed. It is therefore impossible to determine the rate constants corresponding to each step in the dilution. Therefore, hydrogen bond dilution results give no information on the kinetics of protein folding. The second point mentioned above is that each dilution result represents a single plausible unfolding pathway for the protein. An important result of the single pathway interpretation of our data is that the folding pathway will be the reverse of the proposed unfolding pathway, regardless of kinetics of the reaction.

Given the above discussion, the following statements best describe the results of hydrogen bond dilution: rigid regions represent stable folded structure, flexible regions represent unfolded structure, and each step in the hydrogen bond dilution depicts an equilibrium structure on a single possible unfolding/folding pathway. In light of these statements, it is possible to discuss the reasons for good and poor correlations to experiment for each protein, the deficiencies in the model and possible improvements, and future experiments.

The first factor to consider when exploring why the results correlate well for some proteins and not for others is our set of assumptions about the folding mechanism. As discussed in Chapter 1, stating that the native-state topology of a protein structure encodes information about folding requires that the bonds forming key structures along the pathway are maintained throughout the folding reaction. For example, imagine a scenario in which the folding nucleus is stabilized by a specific hydrogen bond, but this bond breaks sometime later along the pathway. In this case, not all of the bonds forming the folding nucleus are conserved in the native state. No method of native-state analysis will be able to identify this folding nucleus because the formation of bonds cannot be realized in a static structure. This deficiency in the model may explain the poor correlation between predicted and experimental folding cores for T4 lysozyme, which is believed to require a nonnative contact for stabilization of the TSE (Klein-Seetharaman et al., 2002). The formation of a nonnative bond along a folding pathway should not be confused with an off-pathway intermediate, which is a kinetically accessible, nonnative conformation. Off-pathway intermediates must unfold and "get back on the pathway" in order to fold to the native state. The experimental observation of off-pathway intermediates does not invalidate hydrogen bond dilution results. In fact, there is some controversy over whether these alternative pathways, or off-pathway intermediates, are artifacts due to chemical environment used to induce unfolding. Additionally, the role of disulfide bonds may not be accurately represented in our model as cystine bonds are maintained throughout the hydrogen bond dilution process.

In the set of results for 8 proteins, which excludes T4 lysozyme and α-lactalbumin, the correlation between predicted and experimental folding cores is quite good. This suggests that these results resemble an unfolding/folding pathway, or least the dominant features common to parallel pathways. From these data I conclude that the native-state topologies of some proteins do indeed encode information about the mechanism of folding, and this information is stored in the energy and density of the noncovalent bonds present in the native state.

5.2.3 Future Directions

The wealth of experimental data available, and the ease with which FIRST analysis can be run, will allow future studies on protein folding in two directions: 1, modifying specific aspects of the model in an attempt to better predict the experimental observations, and 2, additional comparison of our predicted unfolding/folding data to those of experiments. The following is a wish list of computational projects that have been discussed as follow-ups to this work, but have not yet been fully realized.

An evaluation of the enthalpy/entropy compensation effect in our model. The number of floppy modes can be interpreted to represent the internal free energy of a system based on the physical properties of floppy modes as a function of mean coordination. Certainly, the shape of the f'' vs $\langle r \rangle$ plots provide convincing qualitative support. However, it is not straightforward how the individual energy and entropy terms can be computed. Also, the changes in solvent degrees of freedom during folding are not taken into account.

The basis for flexibility analysis and the concept of a floppy mode are rooted in the mathematics of rigidity theory. Advances in the understanding of the free energy of a protein and in particular energy/entropy compensation in terms of floppy modes and degrees of freedom may be elucidated, in a structural context, by these approaches. However, there are alternative means of computing the energy and entropy of a protein chain by using empirical free energy functions such as those used by Galzitskaya and Finkelstein (1999), discussed in Chapter 1. Their function was simplified in that the enthalpy depended only on the residues that were in native conformation, and their entropy depended solely on the size of disordered loops. Their function for computing conformational free energy could be applied to FIRST flexibility results by the direct analogy between rigid clusters and their native conformation, and flexible loops and their disordered loops. In this manner, a free energy like quantity could be computed for each step of the hydrogen bond dilution process and used to identify the transition state along the pathway in our results.

Alternative hydrogen bond dilution schemes. Three methods for removing hydrogen bonds from a protein structure were presented in Chapter 4: strictly according to energy, random selection within the ten weakest bonds, and completely random selection. Because hydrogen bonds were removed from weakest to strongest, the random scheme was an attempt to show that small changes to the order in which the bonds were broken did not affect the results. Unfortunately, for the two proteins in which the results correlate poorly with experiment, small changes in the order of hydrogen bond dilution still produce poor correlations.

A concern raised in the past is that the hydrogen-bond energies do not change over the course of a hydrogen bond dilution experiment, even though they would be expected to attenuate as the protein becomes more flexible. Furthermore, it is assumed that the probability of a bond breaking is directly proportional to its energy, whereas it will most likely depend to some extent on the stability of the protein in the vicinity of a given bond. For example, imagine a pair of β-strands with five hydrogen bonds between them, all with energies of -6.0 kcal/mol except for the middle hydrogen bond, which has an energy of -0.1 kcal/mol. In the current model, the middle hydrogen bond would be broken very early in the dilution process, imposing a degree of flexibility on the β-strands. However, this may be unrealistic as the strong hydrogen bonds can restrict the movement of the strands, disallowing the conformational flexibility required for the weak hydrogen bond to break. Based on this physical picture, the probability of the weak bond breaking would be apparently lower due to constraints imposed by the strong bonds in the local vicinity. A dilution scheme that addresses this problem would recompute the energy of each hydrogen bond every time a bond was removed, and this energy would depend to some extent on all the hydrogen bonds within a certain radius of the given bond. Alternative schemes such as this have been implemented, and analysis of the initial data looks promising.

Analysis of the TSE and comparison to Φ values. The transition state for protein folding/unfolding represented in the hydrogen bond dilution results can be identified as a peak in the second derivative of the fraction of floppy modes as a function of mean coordination. Because the results of simulated thermal denaturation represent a single pathway, this structure can be identified and used to determine which regions of the protein are folded and which are not. These data can be used to compute a predicted Φ-value for each residue, which can then be compared to Φ values determined experimentally from the mutagenesis experiments. Preliminary data on barnase and the p53 tetramerization domain have indicated a very good correlation between predicted and experimental Φ values; however, comparisons were less good for SRC SH3 domain or CI2. Even more than the folding core predictions, the predicted structure of the transition state will depend upon the order in which hydrogen bonds are broken during an unfolding simulation. Future work on predicting Φ values will be addressed concurrently with optimizing the new hydrogen bond dilution scheme.

In closing, it is clear that FIRST flexibility analysis provides a novel means for studying protein folding. It appears that native state protein structures do indeed encode information about the folding mechanism for many proteins, and that FIRST can decode much of this information. Analysis of folding with FIRST is still in its infancy, but the ease with which the program can be developed and the wealth of experimental data available for comparison ensure that future experiments are forthcoming, and that FIRST results will continue to contribute to our understanding of one of the most important phenomena in nature, the folding of proteins.

Appendix A

Summary of Publications Outside of the Scope of the Work Presented in this Dissertation

- Q. Yaun, J. J. Petska, B. M. Hespenheide, L. A. Kuhn, J. E. Linz and L. P. Hart. Identification of mimotope peptides which bind to the mycotoxin deoxynivalenol-specific monoclonal antibody. *Appl. Environ. Microbiol.* **65**:3279–86, 1999.

  Monoclonal antibody 6F5 (mAb 6F5), which recognizes the mycotoxin deoxynivalenol (DON) (vomitoxin), was used to select for peptides that mimic the mycotoxin by employing a library of filamentous phages that have random 7-mer peptides on their surfaces. Two phage clones selected from the random peptide phage-displayed library coded for the amino acid sequences SWGPFPF and SWGPLPF. These clones were designated DONPEP.2 and DONPEP.12, respectively. The results of a competitive enzyme-linked immunosorbent assay (ELISA) suggested that the two phage displayed peptides bound to mAb 6F5 specifically at the DON binding site. The amino acid sequence of DONPEP.2 plus a structurally flexible linker at the C terminus (SWGPFPFGGGSC) was synthesized and tested to determine its ability to bind to mAb 6F5. This synthetic peptide (designated peptide C430) and DON competed with each other for mAb 6F5 binding. When translationally fused with bacterial alkaline phosphatase, DONPEP.2 bound specifically to mAb 6F5, while the fusion protein retained alkaline phosphatase activity. The potential of using DONPEP.2 as an immunochemical reagent in a DON immunoassay was evaluated with a DON-spiked wheat extract. When peptide C430 was conjugated to bovine serum albumin, it elicited antibody specific to peptide C430 but not to DON in both mice and rabbits. In an in vitro translation system containing rabbit reticulocyte lysate, synthetic peptide C430 did not inhibit protein synthesis but did show antagonism toward DON-induced protein synthesis inhibition. These data suggest that the peptides selected in this study bind to mAb 6F5 and that peptide C430 binds to ribosomes at the same sites as DON

- B. Essigmann, B. M. Hespenheide, L. A. Kuhn and C. Benning. Prediction of the active-site structure and NAD$^+$ binding in SQD1, a protein essential for sulfolipid biosynthesis in *Arabidopsis*. *Arch. Biochem. Biophys.* **369**:30-41, 1999.

  Sulfolipids of photosynthetic bacteria and plants are characterized by their unique sulfoquinovose headgroup, a derivative of glucose in which the 6-hydroxyl group is replaced by a sulfonate group. These sulfolipids have been discussed as promising anti-tumor and anti-HIV therapeutics based on their inhibition of DNA polymerase and reverse transcriptase. To study sulfolipid biosynthesis, in particular the formation of UDP-sulfoquinovose, we have combined computational modeling with biochemical methods. A database search was performed employing the derived amino acid sequence from SQD1, a gene involved in sulfolipid biosynthesis of *Arabidopsis thaliana*. This sequence shows high similarity to other sulfolipid biosynthetic proteins of different organisms and also to sugar nucleotide modifying enzymes, including UDP-glucose epimerase and dTDP-glucose dehydratase. Additional biochemical data on the purified SQD1 protein suggest that it is involved in the formation of UDP-sulfoquinovose, the first step of sulfolipid biosynthesis. To understand which aspects of epimerase catalysis may be shared by SQD1, we built a three-dimensional model of SQD1 using the 1.8Å crystallographic structure of UDP-glucose 4-epimerase as a template. This model predicted an NAD(+) binding site, and the binding of NAD(+) was subsequently confirmed by enzymatic assay and mass spectrometry. The active-site interactions together with biochemical data provide the basis for proposing a reaction mechanism for UDP-sulfoquinovose formation

Bibliography

V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich. Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry*, 33:10026–10036, 1994.

J. Adams, S. Leestma, and L. Nyhoff. *C++: An introduction to computing*, chapter 1. Prentice-Hall, Inc., New Jersey, 1995.

A. Amadei, B. L. de Groot, M. A. Ceruso, M. Paci, A. Di Nola, and H. J. Berendsen. A kinetic model for the internal motions of proteins: diffusion between multiple harmonic wells. *Proteins:Struct. Func. Gen.*, 35:283–292, 1999.

A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen. Essential dynamics of proteins. *Proteins:Struct. Func. Gen.*, 17:412, 1993.

D. E. Anderson, J. Lu, L. McIntosh, and F. W. Dahlquist. *NMR ofproteins*, pages 258–304. CRC Press, 1993.

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

C. B. Anfinsen and E. Haber. Studies on the reduction and re-formation of protein disulfide bonds. *J. Biol. Chem.*, 236:1361–1363, 1961.

C. B. Anfinsen, R. R. Redfield, W. L. Choate, J. Page, and W. R. Carroll. Studies on the gross structure, cross-linkages, and terminal sequences in ribonuclease. *J. Biol. Chem.*, 207:201–210, 1954.

C. A. Angell. *Hydration Processes in Biology*, pages 127–139. IOS Press, Amsterdam, 1999.

Y. Bai, J. S. Milne, L. Mayne, and S. W. Englander. Primary structure effects on peptide group hydrogen exchange. *Proteins:Struct. Func. Gen.*, 17:75–86, 1993.

D. Baker. A surprising simplicity to protein folding. *Nature*, 405:39–42, 2000.

H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28: 235–242, 2000.

R. Bhaskaran and P. K. Ponnuswamy. Positional flexibilities of amino acid residues in globular proteins. *Int. J. Peptide Prot. Res.*, 32:241–255, 1988.

A. Bondi. Van der Waals volumes and radii. *J. Phys. Chem.*, 68:441–451, 1964.

J. U. Bowie. Helix packing angle preferences. *Nat Struct Biol*, 4(11):915–917, 1997.

B. Brooks and M. Karplus. Harmonic dynamics of proteins: Normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. Sci.*, 80:6571–6575, 1983.

B. R. Brooks, D. Janezic, and M. Karplus. *J. Comput. Chem.*, 16:1522–1542, 1995.

C. L. Brooks, M. Karplus, and B. M. Pettitt. *Proteins. A theoretical perspective of dynamics, structure, and thermodynamics*. Wiley, New York, 1988.

C. L. Brooks III, M. Gruebele, J. N. Onuchic, and P. G. Wolynes. Chemical physics of protein folding. *Proc. Natl. Acad. Sci.*, 95:11037–11038, 1998.

C. L. Brooks III, J. N. Onuchic, and D. J. Wales. Taking a walk on a landscape. *Science*, 293:612–613, 2001.

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

J. D. Bryngelson and P. G. Wolynes. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci.*, 84:7524–7528, 1987.

H. B. Bull and K. Breese. Surface tension of amino acid solutions: A hydrophobicity scale of the amino acid residues. *Arch. Biochem. Biophys.*, 161:665–670, 1974.

I. Bustos-Jaimes, A. Sosa-Peinado, E. Rudino-Pinera, E. Horjales, and M. L. Calcagno. On the role of the conformational flexibility of the active-site lid on the allosteric kinetics of glucosamine-6-phosphate deaminase. *J. Mol. Biol.*, 319:183–189, 2002.

M. Carrion-Vazquez, A. F. Oberhauser, S. B. Fowler, P. E. Marzalek, S. E. Broedel, J. Clarke, and J. M. Fernandez. Mechanical and chemical unfolding of a single protein: A comparison. *Proc. Natl. Acad. Sci.*, 96:3694–3699, 1999.

H. S. Chan and K. A. Dill. Protein folding in the landscape perspective: Chevron plots and non-Arrhenius kinetics. *Proteins:Struct. Func. Genet.*, 30:2–33, 1998.

Z. Chen, Y. Li, H. B. Schock, D. Hall, E. Chen, and L. C. Kuo. Three dimensional structure of a mutant HIV-1 protease displaying cross-resistance to all protease inhibitors in clinical trials. *J. Biol. Chem.*, 270:21433–21436, 1995.

C. Chothia. Coiling of beta-pleated sheets. *J. Mol. Biol.*, 163:107–117, 1983.

C. Chothia, M. Levitt, and D. Richardson. Helix to helix packing in proteins. *J. Mol. Biol.*, 145:215–250, 1981.

K.-C. Chou, G. Nemethy, S. Rumsey, R. W. Tuttle, and H. A. Sheraga. Interactions between an α-helix and a β-sheet. *J. Mol. Biol.*, 186:591–609, 1985.

J. A. Christopher, R. Swanson, and T. O. Baldwin. Algorithms for finding the axis of a helix: fast rotational and parametric least-squares method. *Comput. Chem.*, 20:339–345, 1996.

J. Clarke and L. S. Itzhaki. Hydrogen exchange and protein folding. *Curr. Opin. Struct. Biol.*, 8:112–118, 1998.

J. Clarke, L. S. Itzhaki, and A. R. Fersht. Hydrogen exchange at equilibrium: a short cut for analysing protein-folding pathways? *TIBS*, 22:284–287, 1997.

D. Cobessi, F. Tete-Favier, S. Marchal, G. Branlant, and A. Aubry. Structural and biochemical investigations of the catalytic mechanism of an NADP-dependent aldehyde dehydrogenase from streptococcus mutants. *J. Mol. Biol.*, 300:141–152, 2000.

F. E. Cohen, M. J. E. Sternberg, and W. R. Taylor. Analysis and prediction of the packing of α-helices against a β-sheet in the tertiary structure of globular proteins. *J. Mol. Biol.*, 156:821–862, 1982.

T. E. Creighton. *Proteins: Structures and Molecular Properties*, pages 287–291. W. H. Freedman, New York, 2nd edition, 1993.

V. Daggett, A. Li, L. S. Itzhaki, D. E. Otzen, and A. R. Fersht. Structure of the transition state for folding of a protein derived from experiment and simulation. *J. Mol. Biol.*, 257: 430–440, 1996.

B. I. Dahiyat, D. B. Gordon, and S. L. Mayo. Automated design of the surface positions of protein helices. *Prot. Sci.*, 6:1333–1337, 1997.

K. A. Dill. Dominant forces in protein folding. *Biochemistry*, 1990.

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

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

N. V. Dokholyan, L. Li, F. Ding, and E. I. Shakhnovich. Topological determinants of protein folding. *Proc. Natl. Acad. Sci.*, 99:8637–8641, 2002.

P. C. Driscoll, A. M. Wingfield, and G. M. Clore. Determination of the secondary structure and molecular topology of interleukin-1β by use of two- and three-dimensional heteronuclear 15N-1H NMR spectroscopy. *Biochemistry*, 29:4668–4682, 1990.

L. Duan, L. Wang, and P. A. Kollman. The early stage of folding of villin headpiece subdomain observed in a 200-nanosecond fully solvated molecular dynamics simulation. *Proc. Natl. Acad. Sci.*, 95:9897–9902, 1998.

P. M. Duxbury, D. J. Jacobs, and M. F. Thorpe. Floppy modes and the free energy: Rigidity and connectivity percolation on bethe lattices. *Phys. Rev. E*, 59(2):2084–2092, 1999.

W. A. Eaton, V. Muñoz, S. J. Hagen, G. S. Jas, L. J. Lapidus, E. R. Henry, and J. Hofrichter. Fast kinetics and mechanisms in protein folding. *Annu. Rev. Biophys. Biomol. Struct.*, 29:327–359, 2000.

S. W. Englander. Protein folding intermediates and pathways studied by hydrogen exchange. *Annu. Rev. Biophys. Biomol. Struct.*, 29:213–238, 2000.

S. W. Englander and L. Mayne. Protein folding studied using hydrogen-exchange labeling and two-dimensional NMR. *Annu. Rev. Biophys. Biomol. Struct.*, 21:243–265, 1992.

S. W. Englander, L. Mayne, Y. Bai, and T. R. Sosnick. Hydrogen exchange: The modern legacy of Linderstrøm-Lang. *Prot. Sci.*, 6:1101–1109, 1997.

D. M. Epstein, S. J. Benkovic, and P. E. Wright. Dynamics of the dihydrofolate reductase-folate complex: Catalytic sites and regions known to undergo conformational change exhibit diverse dynamical features. *Biochemistry*, 34:11037–11048, 1995.

A. Fadini and F.-M. Schnepel. *Vibrational Spectroscopy. Methods and Applications*. John Wiley and Sons, New York, 1989.

A. R. Fersht. Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.*, 7:3–7, 1997.

A. R. Fersht. Transition-state structure as a unifying basis in protein-folding mechanisms: Contact order, chain topology, stability, and the extended nucleus mechanism. *Proc. Natl. Acad. Sci.*, 97(4):1525–1529, 2000.

A. R. Fersht, A. Matouschek, and L. Serrano. The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.*, 224:771–782, 1992.

K. F. Fischer and S. Marqusee. A rapid test for identification of autonomous folding units in proteins. *J. Mol. Biol.*, 302:701–712, 2000.

P. J. Flory. *Statistical mechanics of chain molecules*. Wiley, New York, 1969.

A. Fontana, M. Zambonin, P. P. de Laureto, V. de Filippis, A. Clementi, and E. Scaramella. Probing the conformational state of apomyoglobin by limited proteolysis. *J. Mol. Biol.*, 266:223–230, 1997.

O. V. Galzitskaya and A. V. Finkelstein. A theoretical search for the folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl. Acad. Sci.*, 96:11299–11304, 1999.

B. Gavish. *The fluctuating enzyme*, pages 263–339. John Wiley and Sons, New York, 1986.

M. Gerstein and W. Krebs. A database of molecular motions. *Nucleic Acids Res.*, 26: 4280–4290, 1998.

N. Go, T. Noguti, and T. Nishikawa. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl. Acad. Sci.*, 80:3696–3700, 1983.

M. Gruebele. The fast protein folding problem. *Annu. Rev. Phys. Chem.*, 50:485–516, 1999.

Z. Guo and D. Thirumulai. The nucleation-collapse mechanism in protein folding: evidence for the non-uniqueness of the folding nucleus. *Folding and Design*, 2:377–391, 1997.

M. R. Hicks, J. Walshaw, and D. N. Woolfson. Investigating the tolerance of coiled-coil peptides to nonheptad sequence inserts. *J. Struct. Biol.*, 137:73–81, 2002.

V. J. Hilser, D. Dowdy, T. G. Oas, and E. Freire. The structural distibution of cooperative interactions in proteins: Analysis of the native state ensemble. *Proc. Natl. Acad. Sci.*, 95:9903–9908, 1998.

U. Hobohm, M. Scharf, and R. Schneider. Selection of representative protein data sets. *Prot. Sci.*, 1:409–417, 1993.

W. G. Hol, L. M. Halie, and C. Sander. Dipoles of the α-helix and β-sheet:their role in protein folding. *Nature*, 294:532–536, 1981.

B. Honig. Protein folding: From the Levinthal paradox to structure prediction. *J. Mol. Biol.*, 293(2):283–293, 1999.

R. Huber. Conformational flexibility in protein molecules. *Nature*, 280:538–539, 1979.

F. M. Hughson, P. E. Wright, and R. L. Baldwin. Structural characterization of a partially folded apomyoglobin intermediate. *Science*, 249:1544–1548, 1990.

R. Ishima, D. Freedber, Y.-X. Wang, J. Louis, and D. Torchia. Flap opening and dimer-interface flexibility in the free and inhibitor-bound HIV protease, and thise implications for function. *Struct. Fold. Des.*, 7:1047–1055, 1999.

L. S. Itzhaki, D. E. Otzen, and A. R. Fersht. The structure of the transition state for folding of chymotrypsin inhibitor 2 analyzed by protein engineering methods: Evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.*, 254:260–288, 1995.

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

D. Jacobs and M. F. Thorpe. Computer-implemented system for analyzing rigidity of substructures within a macromolecule. U.S. Patent number 1998:6,014,449. 1998.

D. J. Jacobs and B. Hendrickson. An algorithm for two-dimensional rigidity percolation: The pebble game. *J. Comp. Phys.*, 137:346, 1997.

D. J. Jacobs, L. A. Kuhn, and M. F. Thorpe. Flexible and rigid regions in proteins. In M. F. Thorpe and P. M. Duxbury, editors, *Rigidity theory and applications*, pages 357–384. Kluwer Academic/Plenum Press, 1999.

D. J. Jacobs, A. J. Rader, L. A. Kuhn, and M. F. Thorpe. Protein flexibility predictions using graph theory. *Proteins:Struct. Func. Gen.*, 44:150–165, 2001.

D. J. Jacobs and M. F. Thorpe. Generic rigidity percolation: The pebble game. *Phys. Rev. Letts.*, 75:4051, 1995.

J. Janin and C. Chothia. Packing of α-helices onto β-pleated sheets and the anatomy of α/β proteins. *J. Mol. Biol.*, 143:95–128, 1980.

M.-F. Jeng, S. W. Englander, G. A. Elöve, A. J. Wand, and H. Roder. Structural description of acid-denatured cytochrome *c* by hydrogen exchange. *Biochemistry*, 29(46):10433–10437, 1990.

W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.

M. Karplus and D. L. Weaver. Diffusion–collision model for protein folding. *Biopolymers*, 18:1421–1437, 1979.

M. Karplus and D. L. Weaver. Protein folding dynamics: The diffusion-collision model and experimental data. *Prot. Sci.*, 3:650–668, 1994.

S. L. Kazmirski, K.-B. Wong, S. M. V. Freund, Y.-J. Tan, A. R. Fersht, and V. Daggett. Protein folding from a highly disordered denatured state: The folding pathway of chymotrypsin inhibitor 2 at atomic resoluion. *Proc. Natl. Acad. Sci.*, 98(8):4349–4354, 2001.

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

A. Kippen, J. Sancho, and A. R. Fersht. Folding of barnase in parts. *Biochemistry*, 33: 3778–3786, 1994.

J. Klein-Seetharaman, M. Oikawa, S. B. Grimshaw, J. Wirmer, E. Duchardt, T. Ueda, T. Imoto, L. J. Smith, C. M. Dobson, and H. Schwalbe. Long-range interactions within a nonnative protein. *Science*, 295:1719–1722, 2002.

D. K. Klimov and D. Thirumalai. Stretching single-domain proteins: phase diagram and kinetics of force-induced unfolding. *Proc. Natl. Acad. Sci.*, 96:6166–6170, 1999.

A. P. Korn and D. R. Rose. Torsion angle differences as a means of pinpointing local polypeptide chain trajectory changes for identical proteins in different conformational states. *Prot. Eng.*, 7:961–967, 1994.

G. Laman. On graphs and rigidity of plane skeletal structures. *J. Eng. Math.*, 4:331–340, 1970.

J. Langrange. *Mečanique analytique*, 1788.

P. E. Leopold, M. Montal, and J. N. Onuchic. Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proc. Natl. Acad. Sci.*, 89:8721–8725, 1992.

C. Levinthal. Are there pathways for protein folding? *J. Chem. Phys.*, 65:44, 1968.

M. Levitt, M. Gerstein, E. Huang, S. Subbiah, and J. Tsai. Protein folding: The endgame. *Annu. Rev. Biochem.*, pages 549–579, 1997.

L. Li and E. I. Shakhnovich. Constructing, verifying, and dissecting the folding transition state of chymotrysin inhibitor 2 with all-atom simulations. *Proc. Natl. Acad. Sci.*, 98 (23):13014–13018, 2001.

R. Li and C. Woodward. The hydrogen exchange core and protein folding. *Prot. Sci.*, 8: 1571–1591, 1999.

K. Linderstrøm-Lang. Deuterium exchange and protein structure. In A. Neurberger, editor, *Symposium on protein structure*, London, 1958. Metheun.

M. Llinas and S. Marqusee. Subdomain interactions as a determinant in the folding and stability of t4 lysozyme. *Prot. Sci.*, 7:96–104, 1988.

J. Ma and M. Karplus. Ligand-induced conformational changes in ras p21: a normal mode and energy minimization analysis. *J. Mol. Biol.*, 274:114–131, 1997.

J. C. Maxwell. On the calculation of the equilibrium and stiffness of frames. *Philos. Mag.*, 27:294–299, 1864.

J. A. McCammon, S. H. Northrup, M. Karplus, and R. M. Levy. Helix-coil transitions in a simple polypeptide model. *Biopolymers*, 19:2033–2045, 1980.

I. K. McDonald and J. M. Thornton. Satisfying hydrogen bonding protein in proteins. *J. Mol. Biol.*, 238:777–793, 1994.

L. Mirny and E. Shakhnovich. Protein folding theory: From lattice to all-atom models. *Annu. Rev. Biophys. Biomol. Struct.*, 30:361–396, 2001.

R. S. Molday, S. W. Englander, and R. G. Kallen. Primary structure effects on peptide group hydrogen exchange. *Biochemistry*, 11:150–158, 1972.

L. S. Mullins, C. N. Pace, and F. M. Raushel. Conformational stability of ribonuclease T1 determined by hydrogen-deuterium exchange. *Prot. Sci.*, 6:1387–1395, 1997.

J. K. Myers and T. G. Oas. Mechanisms of fast protein folding. *Annu. Rev. Biochem.*, 71: 783–815, 2002.

J. L. Neira, L. S. Itzahki, D. E. Otzen, B. Davis, and A. R. Fersht. Hydrogen exchange in chymotrypsin inhibitor 2 probed by mutagenesis. *J. Mol. Biol.*, 270:99–110, 1997.

W. L. Nichols, G. D. Rose, L. F. T. Eyck, and B. H. Zimm. Rigid domains in proteins: An algorithmic approach to their identification. *Proteins:Struct. Func. Gen.*, 23:38–48, 1995.

B. Nölting, R. Golbik, J. L. Neira, A. S. S. G. Schreiber, and A. R. Fersht. The folding pathway fo a protein at high resolution from microseconds to seconds. *Proc. Natl. Acad. Sci.*, 94:826–830, 1997.

H. Nymeyer, A. E. García, and J. N. Onuchic. Folding funnels and frustration in off-lattice minimalist protein landscapes. *Proc. Natl. Acad. Sci.*, 95:5921–5928, 1998.

M. Oliveberg and A. R. Fersht. Thermodynamics of transient conformations in the folding pathway of barnase: Reorganization of the folding intermediate at low pH. *Biochemistry*, 35:2738–2749, 1996.

J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes. Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem.*, 48:545–600, 1997.

J. N. Onuchic, H. Nymeyer, A. E. García, J. Chahine, and N. D. Socci. *The energy landscape theory of protein folding: Insights into folding mechanisms and scenarios*, volume 53 of *Advances in Protein Chemistry*, chapter 3. Academic Press, San Diego, 2000.

C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. CATH-A hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, 1997.

S. B. Ozkan, I. Bahar, and K. A. Dill. Transition states and the meaning of Φ-values in protein folding kinetics. *Nat. Struct. Biol.*, 8(9):765–769, 2001.

Y. Pan and M. S. Briggs. Hydrogen exchange in native and alcohol forms of ubiquitin. *Biochemistry*, 31:11405–11412, 1992.

R. V. Pappu and D. L. Weaver. The early folding kinetics of apomyoglobin. *Prot. Sci.*, 7: 480–490, 1998.

A. Patrick, R. Rose, J. Greytok, C. Bechtold, M. Hermsmeier, P. Chen, J. Barrish, R. Zahler, P. Colonno, and P. Lin. Characterization of a human immunodeficiency virus type 1 variant with reduced sensitivity to an aminodiol protease inhibitor. *J. Virol.*, 69:2148–2152, 1995.

L. Pauling and R. B. Corey. The pleated sheet, a new layer configuration of polypeptide chains. *Proc. Natl. Acad. Sci*, 37:2451–2456, 1951.

Z. Peng and L. C. Yu. *Autonomous folding units*, volume 53 of *Advances in Protein Chemistry*, chapter 1. Academic Press, San Diego, 2000.

S. Perrett, J. Clarke, A. M. Hounslow, and A. R. Fersht. Relationship between equilibrium amide proton exchange behavior and the folding pathway of barnase. *Biochemistry*, 34: 9288–9298, 1995.

P. L. Privalov. Intermediate state in protein folding. *J. Mol. Biol.*, 258:707–725, 1996.

A. J. Rader, B. M. Hespenheide, L. A. Kuhn, and M. F. Thorpe. Protein unfolding: Rigidity lost. *Proc. Natl. Acad. Sci.*, 99:3540–3545, 2001.

R. Ragone, F. Facchiano, A. Facchiano, A. M. Facchiano, and G. Colonna. Flexibility plot of proteins. *Prot. Eng.*, 2(7):497–504, 1989.

D. Sabbert, S. Engelbrecht, and W. Junge. Functional and idling rotatory motion within F1-ATPase. *Proc. Natl. Acad. Sci.*, 94:4401–4405, 1997.

F. R. Salemme. Structural properties of protein beta-sheets. *Prog. Biophys. Mol. Biol.*, 42: 95–133, 1983.

B. A. Schulman, C. Redfield, Z. Peng, C. M. Dobson, and P. S. Kim. Different subdomains are most protected from hydrogen exchange in the molten globule and native state of human α-lactalbumin. *J. Mol. Biol.*, 253:651–657, 1995.

W. Scott and C. Schiffer. Curling of flap tips in HIV-1 protease as a mechanism for substrate entry and tolerance of drug resistance. *Struct. Fold. Des.*, 9:1259–1265, 2000.

E. I. Shakhnovich. Folding nucleus: Specific or multiple? ?Insights from lattice models and experiments. *Folding and Des.*, 3:R108–R111, 1998.

J.-E. Shea and C. L. Brooks III. From folding theories to folding proteins: A review and assessment of simulation studies of protein folding and unfolding. *Annu. Rev. Phys. Chem.*, 52:499–535, 2001.

D. A. Simmons and L. Konermann. Characterization of transient protein folding intermediates during myoglobin reconstitution by time-resolved electrospray mass spectrometry with on-line isotopic pulse labeling. *Biochemistry*, 41:1906–1914, 2002.

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

D. Stickle, L. Presta, K. Dill, and G. Rose. Hydrogen bonding in globular proteins. *J. Mol. Biol.*, 226:1143–1159, 1992.

C. Tanford. *The hydrophobic effect.* Wiley-Interscience, New York, second edition, 1980.

T. Tay and W. Whiteley. Recent progress in the generic rigidity of structures. *Struct. Topol.*, 9:31–38, 1984.

D. Thirumalai and D. K. Klimov. Fishing for folding nuclei in lattice models and proteins. *Fold. Des.*, 3:R112–R118, 1998.

A. Thomas, M. J. Field, and D. Perahia. Analysis of the low-frequency normal modes of the R state of aspartate transcarbamylase and a comparison with the T state modes. *J. Mol. Biol.*, 261:490–506, 1996.

M. F. Thorpe, B. M. Hespenheide, Y. Yang, and L. A. Kuhn. Flexibility and critical hydrogen bonds in cytochrome *c*. In R. B. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. E. Klein, editors, *Pacific Symposium on Biocomputing*, pages 191–202. World Scientific, New Jersey, 2000.

M. F. Thorpe, D. J. Jacobs, N. V. Chubynsky, and A. J. Rader. Generic rigidity of network glasses. In M. F. Thorpe and P. M. Duxbury, editors, *Rigidity theory and applications*, pages 239–278. Kluwer Academic/Plenum Press, 1999.

M. F. Thorpe, M. Lei, A. J. Rader, D. J. Jacobs, and L. A. Kuhn. Protein flexibility and dynamics using constraint theory. *J. Mol. Graph. Model.*, 19:60–69, 2001.

I. Y. Torshin and R. W. Harrison. Charge centers and formation of the protein folding core. *Proteins:Struct. Func. Gen.*, 43:353–364, 2001.

C. Tsai, J. V. Maizel Jr., and R. Nussinov. Anatomy of protein structures: Visualizing how a one-dimensional protein chain folds into a three-dimensional shape. *Proc. Natl. Acad. Sci.*, 97(22):12038–12043, 2000.

C. Tsai and R. Nussinov. Hydrophobic folding units derived from dissimilar monomer structures and their interactions. *Prot. Sci.*, 6:24–42, 1997.

C. Tsai, D. Xu, and R. Nussinov. Protein folding via binding and vice versa. *Fold. Des.*, 3: R71–R80, 1998.

R. M. Venable, B. R. Brooks, and F. W. Carson. Theoretical studies of relaxation of a monomeric subunit of HIV-1 protease in water using molecular dynamics. *Proteins:Struct. Func. Gen.*, 15(4):374–384, 1993.

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

M. Vihinen, E. Torkkila, and P. Riikonen. Accuracy of protein flexibility predictions. *Proteins*, 19:141–149, 1994.

R. L. von Montfort, T. Pijning, K. H. Kalk, J. Reizer, M. H. Saier Jr., M. M. Thunnissen, G. T. Robillard, and B. W. Dijkstra. The structure of an energy-coupling protein from bacteria, IIB cellobiose, reveals similarity to eukaryotic protein tyrosine phosphatases. *Structure*, 5:217–225, 1997.

G. Vriend. What if: A molecular modeling and drug design program. *J. Mol. Graph*, 8: 52–56, 1990.

A. Wallqvist, G. W. Smythers, and D. G. Covell. Identification of cooperative folding units in a set of native proteins. *Prot. Sci.*, 28(3):1627–1642, 1997.

D. Walther, F. Eisenhaber, and P. Argos. Principles of helix - helix packing in proteins: the helical lattice superposition model. *J. Mol. Biol.*, pages 536–553, 1996.

D. Walther, C. Springer, and F. E. Cohen. Helix-helix packing angle preferences for finite helix axes. *Proteins:Struct. Func. Gen.*, 33:457–459, 1998.

M. A. Williams, J. M. Goodfellow, and J. M. Thornton. Buried waters and internal cavities in monomeric proteins. *Prot. Sci.*, 3:1224–1235, 1994.

A. Wlodawer and J. Erickson. Structure-based inihibitors of HIV-1 protease. *Annu. Rev. Biochem.*, 62:543–585, 1993.

C. Woodward. Is the slow exchange core the protein folding core? *TIBS*, 18:359–360, 1993.

C. K. Woodward and B. D. Hilton. Hydrogen isotope exchange kinetics of single protons in bovine pancreatic trypsin inhibitor. *Biophys. J.*, 32:561–575, 1980.

D. Xu, C.-J. Tsai, and R. Nussinov. Hydrogen bonds and salt bridges across protein-protein interfaces. *Prot. Eng.*, 10(9):999–1012, 1997.

H. Yang and D. L. Smith. Kinetics of cytochrome *c* folding examined by hydrogen exchange and mass spectrometry. *Biochemistry*, 36:14992–14999, 1997.