ORGANIZING FORCES AND CONFORMATIONAL ACCESSIBILITY IN THE UNFOLDED STATE OF PROTEINS

by

Nicholas C. Fitzkee

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

October 2005

Abstract

For over fifty years, the unfolded state of proteins had been thought to be featureless and random. Experiments by Tanford and Flory confirmed that unfolded proteins possessed the same dimensions as those predicted of a random flight chain in good solvent. In the late eighties and early nineties, however, researchers began to notice structural trends in unfolded proteins. Some experiments showed that the unfolded state was very similar to the native state, while others indicated a conformational preference for the polyproline II helix in unfolded proteins. As a result, a paradox developed. How can unfolded proteins be both random and nonrandom at the same time?

Current experiments and most theoretical simulations cannot characterize the unfolded state in high detail, so we have used the simplified hard sphere model of Richards to address this question. By modeling proteins as hard spheres, we can not only determine what interactions are important in the unfolded state of proteins, but we can address the paradox directly by investigating whether nonrandom behavior is in conflict with random coil statistics.

Our simulations identify hundreds of disfavored conformations in short peptides, each of which proves that unfolded proteins are not at all random. Some interactions are important for the folded state of proteins as well. For example, we find that an α-helix cannot be followed directly by a β-strand because of steric considerations. The interactions outlined here limit the conformational possibilities of an unfolded protein far beyond what would be expected for a random coil. For a 100-residue protein, we find that approximately 9 orders of magnitude of conformational freedom are lost because of local chain organization alone. Furthermore, we show that the existence of this organization is compatible with random coil statistics.

Although our simulations cannot settle the controversy surrounding the unfolded state, we can conclude that new methods of characterizing the unfolded state are needed. Since unfolded proteins are not random coils, the methods developed for describing random coils cannot adequately describe the complexities of this diverse structural ensemble.

Thesis Advisor:	George D. Rose
Second Reader:	Douglas E. Barrick
Thesis Committee:	Gregory D. Bowman
	Richard A. Cone
	Karen G. Fleming
	Betrand García-Moreno E.
	Neville R. Kallenbach (New York University)
	Eaton E. Lattman
	Sarah A. Woodson

To my mother:

Thank you for never letting me give up.

ACKNOWLEDGMENTS

Many kind people have helped me in the creation of this project, both personally and professionally. While I will not be able to thank everyone here, I would like to name a few specific friends and colleagues who have been most important in influencing my thoughts during the past several years.

Of course, none of this would have happened without my thesis advisor George Rose. I thank him not only for his guidance and instruction, but also for his patience. His insightful criticism and suggestions on how to improve my work have impacted me far more than any document could possibly reveal. He encouraged me to see beyond the scientific details and simply marvel at the wonder of it all, and many times he lifted my head out of the code and helped me see the significance of a scientific result. Most importantly, he taught me the beauty of simplicity and elegance in science, and I will never forget that science is a discipline best practiced by trying to serve and not by trying to impress.

I also thank the Jenkins faculty for their guidance and constructive criticism during my thesis reviews. This project would have been much less thorough without their help, and I appreciate their additional commitment to serve on my thesis committee along with Neville Kallenbach. They have been unwavering in their encouragement, and I am thankful for the high standards they have set.

The members of the Rose lab have been very helpful as well. Although he left for India shortly after I joined the lab, I thank Raj Srinivasan for his commitment to both excellence in programming and science. I am thankful for his example and have tried, often unsuccessfully, to model my own code and thinking after his. I am also thankful to

Pat Fleming for sharing his ideas with me, particularly the model for hard sphere solvation. I thank Timo Street for his keen eye and ingenuity when contemplating control experiments. I also thank Nick Panasik, Haipeng Gong, and Teresa Przytycka for their helpful comments and for making the Rose Lab a wonderful environment for scientific investigation.

I am grateful for all the friends I have in the biophysics department, and in particular I thank Brian Cannon for his unceasing encouragement. I aspire to be as good a friend to others as he has been to me.

I thank Ranice Crosby, Jerry Levin, Lisa Jia, and Ken Rutledge for their administrative support in the biophysics department. They are first-rate in their commitment to graduate students, and I have benefited from their close attention to details.

I am particularly thankful for the members of *Adoremus* and my church's care group. I thank them for the perspective and support they provided while I was occupied with my research. I also thank them for their prayers and their example of humility. During a time when it is tempting to think otherwise, they have taught me that the measure of a man is not whether he has a Ph.D.

I also thank my family and friends for their love and support. In particular, I thank my wife Jen Fitzkee for her unending patience, her kind words, and her wise counsel as I have worked through graduate school.

Finally, and most importantly, I thank Jesus Christ, who sustains all things by His powerful word. He has opened my eyes to salvation, and more than any other He has led me to where I am now. May this dissertation be an exaltation of praise for Him.

TABLE OF CONTENTS

ABBREVIATIONS

C	carbonyl carbon
C_α/CA	α-carbon
C_β/CB	β-carbon
CD	circular dichroism
CPK	molecular modeling scheme developed by Robert Corey, Linus Pauling, and Walter Koltun which represents atoms as spheres
D	denatured state
IPH	Flory's isolated pair hypothesis
ϕ	backbone phi torsion angle (defined by C-N-C_α-C)
Φ	kinetic phi-value
GmHCl	guanidinium hydrochloride
N	native state
NMR	nuclear magnetic resonance
NOE	nuclear Overhauser effect
O	carbonyl oxygen
PCL	protein coil library
PDB	protein data bank
P_{II}	polyproline II helix
RDCs	residual dipolar couplings
R, *r*	correlation coefficient
R_G	root-mean-squared radius of gyration
RMSD	root-mean-square difference

SANS	small angle neutron scattering
SAXS	small angle X-ray scattering
τ	backbone tau scalar angle (defined by N-C_α-C)
U	unfolded state
ω	backbone omega torsion angle (defined by C_α-C-N-C_α), normally planar
ψ	backbone psi torsion angle (defined by N-C_α-C-N)

LIST OF TABLES

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

CHAPTER 1

Introduction

Protein folding is a field rife with intensely held opinions, vastly differing theories, and many unanswered questions. It is also a field with much beauty and elegance, both in the molecular structures that it studies as well as in the theories and ideas that have withstood the tests of time and scrutiny. But use of the term "protein folding" implicitly dictates that, in addition to a folded form of proteins, there must also exist an unfolded form. Both the folded and unfolded states of proteins have been the subject of intense study for over fifty years, and yet the nature of the unfolded state remains mysterious. While to date the folded structures of nearly 32,000 proteins have been determined, there is no database of unfolded protein structures, nor can there be. Instead, the size of the unfolded ensemble requires us to form models for the unfolded state and carefully interpret the experimental data in light of these models.

As early as the 1930's, the unfolded state of proteins drew interest as the disordered counterpart to their regularly structured, biologically active form (Wu 1931; Mirsky and Pauling 1936). Because the unfolded state is difficult to observe experimentally under biological conditions, changes in temperature, osmolyte concentration, and pH have been used to unfold—or denature—proteins. This *denatured* state is assumed to be thermodynamically equivalent to the biologically relevant unfolded state (Pace and Shaw 2000), and here we will use the terms interchangeably. Studying the unfolded protein yields insight on aspects of protein folding, including how proteins fold thermodynamically and kinetically, as well as what forces are important in protein folding. When these topics are understood, it then is possible to design models for the unfolded state as well as folding itself. Studies of the unfolded state also shed light on the set of proteins that are normally unfolded in the cell (Dunker et al. 2001), methods for transporting and breaking down proteins within the cell (Matouschek 2003), and interactions between folded and unfolded proteins (Gunasekaran et al. 2004).

1.1 Protein Folding and Thermodynamics

Two-State Folders

The simplest conceivable thermodynamic approach to protein folding is the two-state unfolding reaction, given by

$$N \leftrightarrow D \qquad K = \frac{[D]}{[N]} \tag{1.1}$$

where *[N]* and *[D]* represent the native and denaturant concentrations, respectively, and *K* is the equilibrium constant. In this case, the Gibbs free energy of unfolding for the reaction is given by the equation:

$$\Delta\overline{G} = \Delta\overline{G}^0 + RT\ln K \tag{1.2}$$

Here, $\Delta\overline{G}$ is the molar Gibbs free energy and $\Delta\overline{G}^0$ is the standard-state molar Gibbs free energy. At equilibrium, the free energy is zero and the equation can be rearranged to yield the familiar form for determining the standard-state free energy:

$$\Delta\overline{G}^0 = -RT\ln K \tag{1.3}$$

This approach to protein folding was largely advocated by Christian Anfinsen, who won a Nobel Prize for his efforts in determining the thermodynamic reversibility of ribonuclease folding (Anfinsen 1973). His research proved that the folded structure is encoded entirely by the amino acid sequence of the protein and that the folded conformation lies at the minimum of Gibbs free energy under folding conditions. Since then, many proteins have been found to fold in a reversible, two-state manner.

As Tanford pointed out (Tanford 1968), two-state folding can be identified by any of three methods. The first and simplest method of identifying two-statedness is the concidence of unfolding curves when observed via differing techniques. For example, if the normalized circular dichroism (CD) unfolding curve with respect to urea superimposes on the normalized fluorescence unfolding curve with respect to urea, strong evidence exists for a two-state reaction. This is because CD monitors helix formation in the protein (a global property) while fluorescence monitors the environment of a fluorophore (a local property). If both methods report an identical unfolding transition, it is highly likely, though not always guaranteed, that unfolding is occurring in a highly cooperative, concerted fashion. Experimental error can be significant for unfolding curves, and greater confidence can be obtained as more techniques are used to observe unfolding. On the other hand, as pointed out by Dill and Shortle (Dill and Shortle 1991), non-coincidence does not necessarily violate the two-state folding if the unfolded state's properties change with added denaturant.

The second method used to identify two-state folding is agreement between calorimetric and van't Hoff enthalpies (Tanford 1968). The calorimetric enthalpy $\Delta\overline{H}^0_{cal}$ can be estimated using differential scanning calorimetry (Privalov 1979). This enthalpy is model-independent and applies to the entire unfolding reaction regardless of whether it is two-state. The van't Hoff enthalpy $\Delta\overline{H}^0_{vH}$ is determined by fitting data to a two-state model. This is accomplished by a simple derivation from thermodynamic relations.

Since the Gibbs free energy relates to the enthalpy $\Delta\overline{H}^0$ and entropy $\Delta\overline{S}^0$, one can use the following formula as a starting point, given that *T* is absolute temperature:

$$\Delta\overline{G}^0 = \Delta\overline{H}^0 - T\Delta\overline{S}^0 \tag{1.4}$$

By combining equation (1.4) with the two-state equation (1.3), one determines the following relationship:

$$R\ln K = -\Delta\overline{H}^0\left(\frac{1}{T}\right) + \Delta\overline{S}^0 \tag{1.5}$$

This plot, called a van't Hoff plot, can be used to determine the two-state enthalpy near the denaturant transition (Becktel and Schellman 1987). If *R* ln *K* is plotted versus T^{1}, the van't Hoff enthalpy is the negative slope of the line. By calculating the ratio of $\Delta\overline{H}^0_{vH} / \Delta\overline{H}^0_{cal}$ one can identify whether the thermodynamics are behaving as the two state model would predict. If the ratio is near unity, the transition is two-state, whereas a ratio less than one indicates the existence of populated intermediates in the reaction.

A third method of determining the presence of stable intermediates in the reaction involves the use of kinetics (Tanford 1968). The equilibrium constant *K* in a two-state reaction has the following form:

$$K = \frac{k_u}{k_f} \tag{1.6}$$

where k_u is the unfolding rate and k_f is the rate of folding. A comparison of the equilibrium constant with rhe ratio of these two experimentally determined kinetic parameters can also indicate that the unfolding has no populated intermediates: if the reaction is two-state, the ratio of folding rates should be equivalent to the apparent equilibrium constant.

While many proteins have been behave in the manner two-state folders, it is clear that many states must actually exist in the transition. The protein must populate certain intermediate structures between the native and denatured states since no covalent bonds are broken during folding. Experimental evidence for two-state folding simply indicates that such intermediate states are not highly populated. Stated another way, two-state folding reactions are highly cooperative but not perfectly cooperative.

With the two-state nature of folding established, it becomes relevant to consider the thermodynamic stability of folding. There are several experimental techniques used to determine the stability of protein structures. Two are of particular importance: thermal denaturation and chemical denaturation with urea or guanidinium salts. Because proteins are highly stable under native conditions (with free energies of unfolding of 5-10 kcal/mol), both methods must determine protein stability under highly denaturing conditions and then extrapolate back to native conditions of temperature of denaturant concentration. The fact that both methods yield similar extrapolated free energies lends support to the thermodynamic equivalence of both thermally and chemically denatured states (Pace et al. 1998).

Thermal unfolding utilizes the observation that the change in heat capacity at constant pressure, $\Delta\overline{C}_p$, is mostly constant, large, and positive over the range of observed unfolding transitions (Privalov and Khechinashvili 1974). The large and positive value of $\Delta\overline{C}_p$ means that the denatured state can absorb more energy than the native state before an increase in temperature. The constancy of $\Delta\overline{C}_p$ allows integration of thermodynamic relations to arrive at the Gibbs-Helmholtz equation (Becktel and Schellman 1987):

$$\Delta\overline{G} = \Delta\overline{H}^0\left(1-\frac{T}{T_m}\right) + \Delta\overline{C}_p\left[T - T_m - T\ln\left(\frac{T}{T_m}\right)\right] \tag{1.7}$$

Here, T_m is the midpoint temperature of the unfolding transition. In thermodynamic unfolding experiments, the above equation is fit to the data and extrapolated to determine the unfolding free energy at room temperature. While $\Delta\overline{C}_p$ can be fit from the unfolding data, a better approach is to determine it separately, either by using calorimetric methods or by using a technique outlined by Privalov where a plot of $\Delta\overline{H}$ vs. T yields $\Delta\overline{C}_p$ as the slope (Privalov 1979).

Denaturant unfolding must also address the problem that high concentrations of denaturants are needed to shift equilibrium to the unfolded state. Although Tanford originally proposed the denaturant binding model to this end (Tanford 1970), his method involves assumptions and estimations of binding constants that complicate the extrapolation process. A more contemporary approach makes use of the observation that the free energy varies linearly with denaturant concentration (Greene and Pace 1974; Pace and Shaw 2000):

$$\Delta\overline{G}^0 = \Delta\overline{G}^0(\mathrm{H_2O}) - m[\mathrm{denaturant}] \tag{1.8}$$

The free energy of unfolding in the limit of zero denaturant concentration, $\Delta\overline{G}^0(\mathrm{H_2O})$, is determined by a fitting the free energies of unfolding in various concentrations of denaturant. Also determined in this method is the *m*-value, a measure both of the denaturant strength as well as a reflection of the amount of surface area exposed upon unfolding (Greene and Pace 1974). Of the two most common denaturants, guanidinium hydrochloride (GmHCl) is a stronger denaturant than urea, although urea is preferred because it is uncharged and produces more consistent results than GmHCl (Pace and

Shaw 2000). When extrapolated to infinite dilution, however, both denaturants typically yield equivalent values of the free energy of unfolding within error (Greene and Pace 1974; Ahmad and Bigelow 1982; Santoro and Bolen 1988), although there are exceptions (Ropson et al. 1990). Here again is evidence that the denatured states produced by both urea and GmHCl are thermodynamically equivalent and that the extrapolated value of $\Delta\overline{G}^0(\mathrm{H_2O})$ is not denaturant-dependent.

In 1970 Tanford introduced a transfer model for predicting the unfolding free energies of proteins (Tanford 1970). Using a thermodynamic cycle, he reasoned that it should be possible to predict the differences in folding free energies in two different solvents provided the transfer free energies of the individual residues were known. This line of reasoning led to the following equation:

$$\Delta\overline{G}^0 = \Delta\overline{G}^0(H_2O) + \sum_{\text{amino acids}} \Delta\alpha_i n_i \delta g_i \tag{1.9}$$

In this equation, $\Delta\alpha_i$ is the fraction of a particular amino acid *i* that is exposed upon denaturation, n_i is the number of residues of type *i*, and δg_i is the transfer free energy of the amino acid *i* from water to denaturant. If all the $\Delta\alpha_i's$ are similar, it is possible to factor that term out of the summation, and then a direct correspondence between equations (1.8) and (1.9) are observed. This is why the *m*-value is said to measure the solvent exposure upon denaturation (Pace and Shaw 2000). Indeed, the *m*-value has been shown to correlate with accessible surface area (Myers et al. 1995). Although a high degree of uncertainty is associated with the calculation of transfer free energies and the application of the transfer model above, estimations of $\Delta\alpha$ can be made to determine how much of the protein is exposed upon denaturation. An average value for $\Delta\alpha$ is 0.39

(Pace and Shaw 2000), which is less than average lower-bound estimates of 0.45 made by Creamer and Rose (Creamer et al. 1995; 1997). This is one indication that significant structure may exist in the denatured state. Another indication are m-values which nearly double upon mutation of staphylococcal nuclease. Such a change in m-value can be accounted for when the mutant alters residual structure in the denatured state (Dill and Shortle 1991).

Recently, Auton and Bolen have revisited the calculation of transfer free energies and dramatically reduced the experimental uncertainty associated with the calculation of transfer free energies (Auton and Bolen 2004). To do this, they incorporated three improvements in their technique: First, they used model compounds with minimal end effects, such as cyclic glycylglycine. Second, they accounting for activity coefficients. Finally, they developed correction factors for various concentrations units as motivated by Tanford (Tanford 1970). As a result, they have been able to show a highly linear correlation between peptide length and transfer free energy in solutions of various osmolytes (including urea). Such a correlation confirms the validity of equation (1.9). With these new corrected data in hand, it is hoped that investigation will continue into values for $\Delta\alpha$, as improved values will clarify the extent of collapsed structure in the denatured state.

Non Two-State Folders

While the majority of small, single domain proteins investigated to date fold via a two-state mechanism, two-statedness is not a universal rule. Proteins have been identified with three or even more states. For example, Barrick and Baldwin characterized the molten globule intermediate of apomyoglobin in a three state transition (Barrick and Baldwin 1993), and Riddiford observed four states when denaturing paramyosin with GuHCl (Riddiford 1966). The thermodynamic relations for these higher-state folders are increasingly more complex than for two-state folders. Similarly, expressions for continuous downhill folding can be derived where the transition from folded to unfolded is smooth and barrierless (Muñoz and Sanchez-Ruiz 2004). This type of folding has been observed in the small helical protein BBL (Garcia-Mira et al. 2002), and has been confirmed under close scrutiny (Ferguson et al. 2004; Naganathan et al. 2005).

It is clear that the two state model does not describe all proteins under *in vitro* conditions, nor does it likely describe all proteins *in vivo*. However, the two state model provides a useful framework for addressing the protein folding question, as it is the simplest way to calculate thermodynamic variables. Additionally, regardless of the number of states, all protein folding reactions must have an unfolded state that is of biochemical interest. Anything learned from the denatured state in a two-state reaction will likely be relevant for non two-state reactions as well. Therefore, the focus of this work will be primarily on two-state folders.

Statistical Mechanics of Folding

The elegance and simplicity of classical statistical thermodynamics has also been used to address the protein folding problem. The simplest approach models the statistical transition between helix and coil. These models were initially developed in the late 50's and early 60's by Schellman (Schellman 1955), Zimm and Bragg (Zimm and Bragg

1959), and Lifson and Roig (Lifson and Roig 1961). Here, we will use the formalism of Poland and Scheraga (Poland and Scheraga 1970). The approach is to model peptide chains as strings of characters, where the alphabet is either *c*, representing the unfolded or "coil" conformation, or *h*, representing the α-helical conformation. Thus, the strings *hhcch*, *ccccc*, and *hcchc* are all valid conformations for a five-residue protein in the helix-coil model. One important consideration is that all combinations of *h* and *c* are valid: only the hydrogen bonds formed by the helix determines which conformations will dominate. Thus, individual conformations are assumed to obey of the Flory isolated pair hypothesis (IPH), which states that conformational preferences of any given residue are largely independent of its local neighbors (Flory 1969).

Two factors play a role in the formation of a hydrogen bond in a helix. First, the six dihedral angles between the carbonyl oxygen and amino hydrogen must align themselves in a position to form the bond. This conformational rearrangement comes at a substantial entropic cost, but once it has been made, the hydrogen bond can form, possibly giving an enthalpic benefit to helix formation. For subsequent adjacent hydrogen bonds, the entropic cost is much less, as only two dihedral angles must be fixed. In terms of the helix-coil model, *s* represents the microscopic equilibrium constant between a residue in helix and a residue in coil when a hydrogen bond can be made:

$$c \leftrightarrow h \qquad s = \frac{[h]}{[c]} \tag{1.10}$$

Furthermore, σ quantifies the initial penalty for helix formation, taking into account that for the first two residues in a helix no hydrogen bond is made. Both *s* and σ are taken relative to the statistical weight for coil, which is set to a value of unity. Thus, the relative statistical weight for a string of *h* and *c* is given by the product of the contributions of *s*, *σ*, and 1. For example, the string *hhhccccchhhh*, has a statistical weight of:

$$(\sigma s)(s)(s)(1)(1)(1)(1)(1)(\sigma s)(s)(s)(s) = \sigma^2 s^7$$

To calculate statistics on the number of helical segments or the fraction of helices, one must normalize terms such as the one above by summation across all possible conformations of helix and coil. This sum defines the partition function *Z* for proteins with *N* residues:

$$Z(N) = 1 + \sum_{N_{hc}=1}^{N/2} \sum_{N_h=N_{hc}}^{N-N_{hc}} \Omega(N_{hc}, N_h, N)\sigma^{N_{hc}} s^{N_h} \qquad (1.11)$$

In this equation, the number of helices is N_{hc}, the number of helical residues is N_h, and Ω is the number of ways to arrange N_h helical residues and N-N_h coil residues if there are N_{hc} helices. Evaluation of this partition function can be complex, but it is tractable for short peptides, and a matrix formalism has been developed for longer peptides (Zimm and Bragg 1959).

Given the model parameters described above, it is possible to determine both the average fraction of the protein that is helical (*θ*) as well as the average number of helical segments by taking derivatives of the partition function. In this case,

$$\theta = \frac{1}{N}\frac{\partial \ln Z}{\partial \ln s} \qquad (1.12)$$

$$\langle N_{hc} \rangle = \frac{\partial \ln Z}{\partial \ln \sigma} \qquad (1.13)$$

These parameters are observable through experiment and simulation and can be used to calculate values for *s* and *σ* by nonlinear least squares fitting. Theta, in particular, can be tracked by observing the CD signal at 222 nm (Richardson and Makhatadze 2004).

When this is done, good agreement is generally observed between the fitted curves and the experimental data (Zimm and Bragg 1959), with values of *s* and *σ* ranging from 0.19 to 1.35 and 0.1 x 10^{-4} to 100 x 10^{-4} at 20^{o}, respectively, for the 20 naturally occurring amino acids (Wojcik et al. 1990).

Far from the simplified approach taken by the helix-coil model, the denatured state is likely to be highly complex in its structure and form, despite the good experimental agreement for short peptides. The helix-coil model does, however, highlight a critical point in understanding the denatured state: if a statistical mechanical model of protein folding is desired, one requires intimate knowledge of the denatured state. Since the two-state equilibrium constant *K* can be expressed as a ratio of the unfolded and folded partition functions (Hill 1960), before one can predict the folding equilibrium one must have enough detailed knowledge about the denatured state to construct its partition function. Simply put, in order to understand the folding transition one needs to understand both sides of the equilibrium equation. This idea was stated concisely by Becktel (Becktel and Schellman 1987):

> There is a temptation, especially with proteins of known crystal structure, to relate changes in stability exclusively to features of the native structure of the molecule. This mode of thought must be avoided because it is likely that a large component of the free energy of stabilization as defined above stems from the increased solvation of the unfolded chain relative to the folded one.

Because the unfolded state can contribute significantly to the thermodynamic equilibrium between folded and unfolded, it warrants at least as much scientific investigation as the folded state.

1.2 Protein Folding Kinetics

Whereas thermodynamics describes the equilibrium balance between the folded and unfolded state, kinetics characterizes the rate of transition between the two states (Tanford 1968; Nölting 1999). Kinetics is highly pertinent to the study of unfolded proteins: the folding rate should reflect at a basic level the complexity of the folding process. Larger rates imply a more difficult search to find the native state.

Basics of Folding Kinetics

The fundamental equation for the two-state kinetic transition from folded to unfolded is much the same as equation (1.1). Here, we represent the folding rate as k_f and the unfolding rate as k_u. When denaturant is rapidly added to a system of native protein and the concentration of native protein is measured (typically indirectly through an optical probe), the resulting decay is exponential and has the following form:

$$[N] = \frac{[N]_0}{k_u + k_f}\left\{k_f + k_u e^{-(k_u+k_f)t}\right\} \tag{1.14}$$

Here, as was also the case with thermodynamic parameters, k_f and k_u are concentration dependent and must be extrapolated back to zero concentration of denaturant using simple linear extrapolation (Jackson and Fersht 1991). Values for k_f and k_u in infinite dilution are then determined by examining the characteristic "chevron plot" of $\ln(k_f + k_u)$ versus denaturant concentration and fitting to the appropriate two state equation (Matthews 1987; Jackson and Fersht 1991).

The observed folding rates for small proteins are generally very fast, and even some very large proteins fold quickly. The 62-residue IgG binding domain of protein L folds at a rate of 61 s^{-1} at pH 7.0 (Scalley et al. 1997). A tryptophan-containing mutant of Ubiquitin, which contains 76 residues, folds at a rate of 1.53 x 10^3 s^{-1} at pH 5.0

(Khorasanizadeh et al. 1993). Even the 151-residue CheW protein folds at the surprisingly fast rate of 1.70 x 10^3 s^{-1} at pH 7.0 (Maxwell et al. 2005). The recently established kinetic dataset of 30 proteins under standard condition reveals that, for many proteins, folding is a process that takes very little time. Proteins can easily navigate between the unfolded and folded states.

Relating Kinetics to Protein Conformations

In 1998, Plaxco and coworkers identified a significant relationship between the folding rate and the structure of a protein (Plaxco et al. 1998). They defined a numerical construct called *relative contact order* as:

$$RCO = \frac{1}{LN}\sum_{N}\Delta S(i,j) \qquad (1.15)$$

In this equation, L is the number of residues and N is the total number of residue-residue contacts. For a contact involving residues i and j, $\Delta S(i,j)$ is the number of residues between i and j. Simply stated, for any two residues that contact each other in the protein, the relative contact order measures the average number of residues that separate the two, divided by the total number of residues in the protein. This is a measure of fold complexity, since a larger separation means that the protein must create more non-local contacts. Interestingly, Plaxco *et. al.* found that proteins with high contact order had slow folding rates, with a correlation coefficient of 0.81 (Plaxco et al. 1998). Never before had such a dramatic relationship between protein structure and kinetics been illustrated.

In the years since, folding rates have been shown to correlate with other structural properties. Naganathan and Muñoz have shown that folding rates scale linearly with the square root of protein size (Naganathan and Munoz 2005). Folding rates also scale with the secondary structure composition of proteins, both determined by simulation (Gong et al. 2003) and predicted by neural networks (Ivankov and Finkelstein 2004). These experiments show that, although the kinetics are related to the structure of the final folded protein, they may be related the structure of the unfolded protein as well. If, as some have suggested, the unfolded state has significant native secondary structure content, it would make sense for folding rates to correlate with the number of helices, strands, turns and loops (Gong et al. 2003). Studies that do not rely on contact order also resolve a paradox, since the final folded contact order in the unfolded state is likely undetermined.

Kinetics and the Levinthal Paradox

Because protein kinetics are intimately related to the structure of both the native and denatured states, a question arises: if the unfolded state is a highly random ensemble of many featureless conformations, how can the protein fold so quickly? To put this more concretely, consider a simple 100-residue protein where each residue is either folded or unfolded. If the unfolded state is a random collection of these hypothetical conformations, there are $2^{100} \approx 1 \times 10^{30}$ possible conformations in the unfolded state. Clearly, this is a simple model, and there are likely many more conformations since real proteins have more than two states per residue. If, however, we continue with this simple model and assume that a protein can sample one conformation every 10^{-13} seconds (Cohen and Sternberg 1980), the protein would take approximately 10^{17} seconds to fold, or more than three billion years. Since this is an underestimate, how is it that the protein folds so quickly—1×10^{3} s^{-1} versus 1×10^{-17} s^{-1}?

This line of reasoning has come to be known as the Levinthal paradox, after Cyrus Levinthal (Levinthal 1969). Levinthal reasoned that proteins fold not by a random search through a vast number of featureless conformations, but rather by a directed search along *pathways* of folding. The idea of a directed search simplifies the folding process enormously, and many of the folding theories that exist today attempt to identify the forces that direct proteins along their folding pathway (see below).

1.3 Interactions in Protein Folding

Interpretation of the thermodynamic and kinetic data of protein folding is difficult without a thorough understanding of the forces involved in the reaction. In addition, modeling proteins in both the folded and unfolded states requires a functional form—often simplified—for the dominant forces. The authoritative article on the forces involved in protein folding was published by Kauzmann in 1959. Armed with only the most basic experimental data on the structure and properties of proteins, and having only the myoglobin crystal structure from which to draw conclusions (Kendrew et al. 1958), Kauzmann was able to catalog the forces in protein folding with such accuracy that his review remains relevant over 40 years later (Kauzmann 1959). Here, we will follow Kauzmann's approach and briefly catalog the forces involved in the denatured state.

Covalent Forces

The most important forces in protein folding are paradoxically also the least interesting. Clearly, without the covalent bonding of atoms within the protein itself giving rise to the proteins' primary structure, there would be no protein folding problem.

Yet these forces are common to all polymer chains and are not very different in the protein molecule. Although covalent bonds are not broken during the transition from folded to unfolded in non-disfulfide containing proteins, bond character in proteins is nonetheless important both from an experimental and theoretical perspective.

Since covalent bonding is an electronic phenomenon, optical methods can be used in some cases to determine the orientation and behavior of covalent bonds. Infrared spectroscopy measures the frequencies of bond stretching and bending and shows that the bonds in proteins are slightly flexible (Schellman and Schellman 1964). With the exception of the torsion angle of the peptide group itself (Pauling et al. 1951), most bonds are free to rotate and can be slightly distorted from their ideal bond geometries. In the unfolded state, where structures are thought to fluctuate and forces are stochastically directed, distortions in geometries should be rare events due to the lack of compensating forces.

The infrared studies described above allow theorists to determine the energetics of bond stretching, bending, and so forth. These values can be used to parameterize computer simulations. For example, the functional form of the bond-stretch energy used in the CHARMM simulation package is:

$$E = k(b - b_0)^2 \tag{1.16}$$

Here, k is the spring constant for the stretching interaction, and b_0 is the equilibrium bond length. For C_α-C_β bond stretch in alanine, these parameters are 222.5 cal/mol/Å^2 and 1.54 Å, respectively (MacKerell et al. 1998). Calculations of this nature are time-consuming, and it is often far simpler to assume that bond lengths and angles are rigid. This approach has been used with success in simulations (Srinivasan et al. 2004), but it is often difficult to rebuild native PDB structures with idealized bond lengths and angles (Holmes and Tsai 2004).

Atomic Overlap: Sterics

The Pauli exclusion principle establishes that no two electrons can occupy the same orbital with the same spin. Accordingly, non-covalently bonded atoms are resistive to atomic overlap. At the same time, the nature of the electron cloud allows induced dipoles to form and create an attractive force. Because the quantum mechanics of this behavior are difficult to quantify, particularly in large simulations of protein molecules, simpler forms have been developed to determine the energy of interaction between two closely spaced, nonbonded atoms. Of these, the most well-known is the Lennard-Jones equation, which counters the attractive nature of an atomic induced-dipole with a repulsive energy:

$$E_{i,j}(r) = \varepsilon_{i,j}\left[\left(\frac{R_{i,j}}{r}\right)^{12} - 2\left(\frac{R_{i,j}}{r}\right)^{6}\right] \tag{1.17}$$

In this equation, the energy of interaction at a distance r between two nonbonded atoms i and j is calculated from a energetic parameter ε, and the contact distance R. This "soft-sphere" potential has been very successful in modeling nonbonded interactions theoretically (MacKerell et al. 1998; Pappu and Rose 2002).

An even simpler approach is to model atoms as hard spheres, ignoring the induced dipole forces altogether. This method has the advantage of identifying exactly which atoms are involved in an unfavorable steric clash. Figure 1.1 depicts the hard sphere collisions that occur in the alanine Ramachandran plot. Because of the simplifying nature of the hard sphere model, one can immediately tell which atomic collisions are responsible for limitations in the backbone dihedral angles φ and ψ. There is some question, however, about the validity of the hard sphere model. Certainly it was appropriate when Ramachandran, Ramakrishnan, and Sasisekharan determined the allowed conformations of an alanine dipeptide over four decades ago (Ramachandran et al. 1963; Ramachandran and Sasisekharan 1968), but is it appropriate today? As championed by Richards (Lee and Richards 1971; Richards 1977), the hard sphere has been tremendously successful in identifying critical properties of folded proteins, such as packing densities. Other uses for the hard sphere model include calculating surface areas (Lee and Richards 1971), locating cavities in proteins (Eriksson et al. 1992), fitting side chain conformations (Bower et al. 1997), and identifying irregular protein structures (Laskowski et al. 1993b). The verdict seems to be that, while the hard sphere model may be an oversimplification, it works quite well for calculating many properties of the protein chain.

Because the hard sphere model simplifies the complexities of atomic shape into one parameter (a radius), there is disagreement about the best set of radii to use. While densities and small molecule data may yield one answer (Bondi 1964), contact distances from actual protein crystal structures may give another (Li and Nussinov 1998). It is often the case that the right set of radii will be different depending on the property being examined. Fortunately, many properties are tolerant to small changes in van der Walls radii (Shrake and Rupley 1973), and several general-purpose sets of radii and contact distances exist (Hopfinger 1973).

The hard sphere model is well suited for describing the unfolded state of proteins, not only because of the statistical nature of the unfolded state, but also because of the lack of perturbing interactions which may induce the protein to violate hard sphere interactions. Accordingly, many theoretical models of the denatured state have modeled atoms as hard spheres. Some of these experiments will be described below.

Water and the Hydrophobic Effect

Simply stated, the hydrophobic effect is the fact that nonpolar solutes are less soluble in water than in nonpolar solvents. This simple fact, however, has profound consequences on both native and denatured protein structure, and the hydrophobic effect is thought to be the dominant force involved in chain collapse (Kauzmann 1959; Dill 1990). It is well known that the nature of the hydrophobic effect differs between low temperature and high temperature. At room temperature, the effect is entropically driven, whereas at higher temperatures (~100 ^{o}C), the effect is driven by loss of enthalpy (Privalov and Gill 1988). This is thought to reflect a structuring of water near a nonpolar solute: at lower temperatures, the water is conformationally restricted and hence entropically unfavorable, while at higher temperatures, the water remains fluid but loses the enthalpic benefit of hydrogen bonds near the nonpolar solute (Dill 1990). While this is a satisfying description, it may not be complete, as Lee notes that the enthalpy of solvent reorganization during transfer is unfavorable, whereas structured water cages should exhibit a favorable enthalpy change (Lee 1991). Lee suggests that at lower temperatures the entropic contribution may simply result in the conformational limitations that arise in forming a cavity in water (Lee 1991).

While the hydrophobic effect is a dominant force for protein folding, it is equally important in the denatured state alone. For thermally or acid denatured proteins, the denatured state must overcome the unfavorable transfer of apolar side chains from the hydrophobic core to solvent. In denaturant-induced unfolding, the nature of the hydrophobic effect will determine the residual structure, if any, in the denatured state (Tanford 1968). Because the hydrophobic transfer free energies have been shown to correlate with side chain accessible surface area (Chothia 1974), understanding the nature of the denatured state is a prerequisite to understanding the change in accessible surface area. Creamer and Rose have shown that assumptions about the denatured state can have a significant impact on determining the accessible surface area changes upon folding, and thus the ability to predict free transfer free energies are only as good as our models for the denatured state of proteins (Creamer et al. 1995; 1997).

Because of the strength of the hydrophobic effect and its importance in the denatured state, there is much speculation as to whether collapsed structure persists in the denatured state of proteins. Shortle *et. al.* addressed this question in staphylococcal nuclease by removing the large hydrophobic amino acids through mutation (Shortle et al. 1990). Using the linear extrapolation method, they were able to examine the changes in *m*-value with each mutation. As discussed above, the *m*-value is a measure of solvent exposure upon denaturation. When several mutations were made, a dramatic increase in *m*-values were observed, corresponding to an expansion of the denatured state (Shortle et al. 1990). Such an expansion suggests that the wild-type denatured state retains a significant amount of compactness, if not native-like structure. Thus, removing the large hydrophobic groups was speculated to destabilize the native state by lowering the stability of the denatured state.

From a theoretical perspective, it is highly desirable to have a model of water which accurately accounts for the solvability of peptides and proteins. Unfortunately, no such model exists, and the models that do exist often disagree on the fundamental properties of water itself (Guillot 2002). The two most popular models for protein simulations are SPC and TIP3P, both of which model water as a planar molecule with three partial charges (Berendsen et al. 1981; Jorgensen et al. 1983). A comparison of simulations on the alanine dipeptide using both of these water models shows that, although the trends in solvation are qualitatively consistent, the quantitative values for energies, average ϕ and ψ torsions, and other peptide properties are not (Anderson and Hermans 1988; Tobias and Brooks 1992; Hu et al. 2003). Explicit models of water that can accurately account for all of the experimental measurements on protein solvation are difficult to parameterize, and to date no satisfying model has been developed.

Because explicit models of water are at present imperfect and computationally expensive, simplifying models of solvation have been developed to account for the hydrophobic effect. These models generally take advantage of the relation between hydrophobicity and accessible surface area. The method of Honig, for example, uses accessible surface areas to determine the contributions of nonpolar groups while modeling water as a constant dielectric for the contributions of polar and charged groups (Sitkoff et al. 1994). The transfer free energies calculated by this method, when compared to experimental data, have a correlation coefficient of 1.00. Another simplifying model for water was developed by Fleming *et. al*. (Fleming et al. 2005). This model assumes that, upon solvation of the peptide backbone, certain sites will be preferentially solvated, namely, the carbonyl oxygen and amino nitrogen. Because of the preferential solvation, hydrophobic accessible surface area is not uniform across the surface of the molecule. In this model, the calculation of accessible surface area must take into account the sites that will already be solvated. This method, termed conditional hydrophobic accessible surface area, or CHASA, has been parameterized to agree with much more sophisticated simulations of backbone solvation (Mezei et al. 2004), and can accurately predict conformational propensities in a database of protein structures (Fleming et al. 2005). The success of simple hydration models in quantifying hydrophobicity makes them an attractive alternative to the all-atom models of water. Given the inherent complexity in the unfolded protein chain, simplified water models gain an additional degree of attractiveness.

Hydrogen Bonding

The first suggestion that hydrogen bonding may be favorable in folded proteins came from Linus Pauling's proposals for the α-helix and β-strand (Pauling and Corey 1951; Pauling et al. 1951). Originally, it was thought that hydrogen bonding strongly drove protein folding, but Kauzmann's review suggested that the hydrophobic force, and not hydrogen bonding, was responsible for driving folding (Kauzmann 1959). Kauzmann's reasoning followed from dimerization experiments on urea performed by John Schellman which showed that the free energy of hydrogen bond formation is 1.9 kcal/mol and the enthalpy of hydrogen bond formation is –2.1 kcal/mol (Schellman

1955). Such a small favorable enthalpy, Kauzmann believed, could not drive the protein folding reaction, given the other forces involved.

Subsequent experiments added confusion to the issue, as the energetic stability of peptide hydrogen bonds could not be shown to be favorable or unfavorable with respect to water hydrogen bonds. Shortly after Kauzmann's review, Klotz and Franzen published results from N-methylacetamide in water which showed that the enthalpy of peptide hydrogen bond formation was near zero (Klotz and Franzen 1962). Similarly, Honig used quantum mechanics calculation to determine that the free energy of hydrogen bond formation in the interior of a protein was unfavorable by 2.5 kcal/mol (Ben-Tal et al. 1997). On the other hand, equally compelling experimental evidence suggests that peptide hydrogen bonds are favorable. Hydrogen bonds abound in the interior of proteins (Stickle et al. 1992), and several studies on the helix-coil transition have indicated that hydrogen bonding in proteins is favorable (Scholtz et al. 1991; Richardson et al. 2005). To date, no satisfying reconciliation has been made between the experiments that favor a peptide hydrogen bond and those that disfavor it.

Fortunately, for the unfolded state, the situation is much simpler. As pointed out by Fleming and Rose, the important question is not whether peptide-peptide hydrogen bonds are more favorable than peptide-water hydrogen bonds, but rather whether peptide-peptide hydrogen bonds are more favorable than no hydrogen bonds at all (Fleming and Rose 2005). In response to this question, the data consistently show that a non-hydrogen bonded donor or acceptor is highly unfavorable by 6 kcal/mol or more (Ben-Tal et al. 1997). It follows that, in the unfolded state, all hydrogen bonds will be satisfied, either by an intra-peptide hydrogen bond or a hydrogen bond with solvent water. This idea has strong experimental support (Stickle et al. 1992; McDonald and Thornton 1994; Fleming and Rose 2005), and it suggests a means by which the unfolded state may be organized: conformations not only must adhere to steric constraints, but they are required to exhibit proper hydrogen bonding as well.

Electrostatic Interactions

It is clear that electrostatics plays an important role in the stability of proteins, both on a local (Kauzmann 1959) and global (Ripoll et al. 2005) scale. This is because of the pH titration behavior of most proteins: the native state is generally disfavored at both extremes of pH. What is less clear is the significance of electrostatics in the denatured state. The high dielectric constant of water, $\varepsilon = 78.4$ (Fernandez et al. 1995), will effectively mask all electrostatic interactions in a randomly structured denatured state because bulk solvent will cover large portions of the peptide chain. On the other hand, if the denatured state contains residual, compact structure, electrostatic interactions may be significant.

Evidence for the second view was given recently by Whitten and García-Moreno (Whitten and Garcia-Moreno 2000). They measured the pH dependence of unfolding of staphylococcal nuclease using two methods: first, they used chemical denaturants and temperature to denature the protein and then extrapolated to native conditions to obtain stability as a function of different pH environments. Second, they obtained the pH dependence of stability potentiometrically. At pH 7.0, there was an almost 4 kcal/mol difference between the two stabilities. The authors interpreted this to mean that the $\mathrm{pK_a}$'s of several groups were depressed in the unfolded state compared to their model-compound values. Because pK_a's are sensitive to local environment, it was proposed that the denatured state retained a significant degree of compact structure (Whitten and Garcia-Moreno 2000). As a result, the authors concluded that electrostatic interactions play a significant role in stabilizing both the native and denatured states.

As more and more structures become available in the protein data bank (PDB) (Berman et al. 2000), researchers have sought to correlate protein energetics with conformational distributions contained therein. If, for example, electrostatics can explain the distribution of a particular set of residues in the PDB, the electrostatics should be a dominant force for that set of structures. One such study of this type was done by Avbelj and Baldwin, and it utilized the coil library of structures—a subset of non-helix, non-strand fragments of the PDB (Avbelj and Baldwin 2003). The coil library has been shown to possess similar backbone ϕ, ψ's as unfolded proteins (Serrano 1995; Swindells et al. 1995). Accordingly, Avbelj and Baldwin examined whether the electrostatic dipole moment of the peptide bond could explain the distributions of ϕ for each residue. Using a simple torsional energy with an electrostatic component, they were able to reproduce ϕ distributions better than other models for the denatured state which did not include an explicit electrostatic component (Avbelj and Baldwin 2003). From this, it can be concluded that the electrostatic contribution of the peptide dipole is important in determining the conformation of the backbone in the denatured state. Similar research by Ho *et. al.* showed that one cannot reconstruct the distribution of ϕ, ψ's without including an electrostatic energy (Ho et al. 2003). Thus, while it has been thought that electrostatics plays a minimal role in the denatured state, recent experimental and theoretical evidence suggests otherwise: clearly the contribution is worth further investigation.

Summary of Interactions

All of the interactions listed above are undoubtedly important in determining the conformations of the denatured state. Is it possible, however, to order the forces by rank of importance? Many scientific minds have attempted to address this question (Kauzmann 1959; Tanford 1970; Dill 1990), with sometimes contradictory conclusions. It is difficult to deny the importance steric exclusion, however. When one considers the size of protein conformational space, the possible conformations that are eliminated as a consequence of steric overlap is truly mind-boggling. Some calculations estimate that the fraction of conformational space eliminated upon chain collapse by sterics alone is 10^{-44} for a 100-residue protein (Dill 1985). While other forces will surely influence the size of conformational space further, it is doubtful that they will be more significant than the simple fact that two atoms cannot occupy the same space at the same time.

1.4 Models for the Unfolded State

The influence of models on our understanding of the unfolded state cannot be understated. An accurate understanding of the forces in the denatured state is useless without a conceptual framework for how those forces shape the denatured ensemble. Today, three theories about the conformational properties of the denatured state dominate the field. Unfolded proteins have been modeled as random coils, native-like chains, and fluctuating segments of polyproline II helix. While these are not mutually exclusive models, in the next three sections we will address each model separately, discussing the experiments and theory that have led to the development of each model.

1.5 The Random Coil Model

Theoretical Overview

The random coil model is the oldest and well-established model for the denatured state. Developed primarily by Flory in the 1950's and 1960's (Flory 1953; 1969), this model is also the most theoretically robust of models for the denatured state today. This is primarily because of the statistical nature of a random coil: by modeling unfolded proteins as stochastic chains, it is possible to extract a concise mathematical formalism for chain properties, whereas models with nonrandom behavior are more difficult to describe mathematically. Because of its statistical tractability and its simplicity in interpreting experimental results, the random coil model has withstood the test of time, and it will likely exist as a model for denatured proteins for some time to come.

The simplest (and most unrealistic) class of random coil model is that of the freely jointed chain. The freely jointed chain represents the protein as a chain of identical residues with no excluded volume constraints. There are no restrictions between the orientation of residues—only that the distance between residues corresponds to one bond length, typically designated *l* (in its vector form $\vec{l}$). If there are *n* bonds in the chain, then the equation for the end-to-end vector $\vec{r}$ of the chain is simply:

$$\vec{r} = \sum_{i=1}^{n} \vec{l}_i \tag{1.18}$$

However, this value is not very useful: a truly random chain is simply a random walk through space, and thus the ensemble-averaged chain displacement is zero. A measure that is useful, both from a theoretical perspective and in the fact that it can be observed experimentally, is the mean-squared end-to-end distance, $\left\langle r^2 \right\rangle$:

$$\left\langle r^2 \right\rangle = \left(\sum_{i=1}^{n} \vec{l}_i \right) \cdot \left(\sum_{i=1}^{n} \vec{l}_i \right) = nl^2 + 2\sum_{i=1}^{n-1} \sum_{j=i}^{n} \left\langle \vec{l}_i \cdot \vec{l}_j \right\rangle = nl^2 \tag{1.19}$$

In the third expression, the double-sum term is zero because of the random orientation of bond vectors. Equation (1.19) shows that the mean square end-to-end distance is proportional to the number of bonds in the chain, or that the root mean square end-to-end distance is proportional to $n^{0.5}$. Another observable property of unfolded chains is the radius of gyration. The radius of gyration is akin to the statistical standard deviation on the geometric center of a protein (it can also be weighted by mass or scattering factors). For a protein with *m* atoms with positions $\vec{x}$ and a geometric center at position $\bar{\vec{x}}$ the square radius of gyration, R_G^2, is defined as:

$$R_G^2 = \frac{1}{m} \sum_{i=1}^{m} \left(\vec{x}_i - \bar{\vec{x}} \right)^2 \tag{1.20}$$

For a freely jointed chain of infinite length, it can be shown that (Flory 1969):

$$R_G^2 = \left\langle r^2 \right\rangle / 6 \tag{1.21}$$

Real proteins, of course, have excluded volume and restrictions on their bond orientations. These considerations can be approximated in the random coil model: the chain can be given a realistic geometry and local energy functions can approximate the contribution of electrostatics and van der Walls contacts. This model, called the rotational isomeric state model (Flory 1969), works very well for short peptides, but because of the difficulties involved in calculating accurate energy functions, it does not accurately estimate chain dimensions for proteins. Flory devised a simple means to estimate the scaling properties for long polymer chains: by assuming that the forces of excluded volume and entropic disorder are at odds with one another in the chain, he determined that the radius of gyration of minimum energy is (Flory 1953):

$$R_G = R_0 N^{\nu} \tag{1.22}$$

Here, R_0 is a constant that depends on the chain geometry and solvent, and N is the number of residues in the protein chain. The exponent, v, is often used as a measure of solvent quality: larger values of v indicate that chain-solvent interactions are more favorable than chain-chain interactions, whereas smaller values of v indicate a preference for chain-chain interactions. Flory predicted that at $v = 0.6$ the two forces would exactly match each other, and he termed solvents that exhibit this behavior θ-solvents. More contemporary calculations have estimated the value of the θ-solvent exponent to be 0.588 (Le Guillou and Zinn-Justin 1977), and a recent survey of proteins under strong denaturing conditions has corroborated this value (Kohn et al. 2004). Indeed, the observation that denatured proteins exhibit random coil behavior for v is one of the strongest arguments in favor of the random coil model, although it should be emphasized that random coil behavior will be displayed for any chain if the length scales are long enough (Tanford 1968).

Experimental Studies

The early experimental studies on the denatured state were heavily influenced by the theoretical work of Flory. Many of these studies were done in Tanford's lab. Using intrinsic viscosity, Tanford was able to determine the scaling properties of unfolded proteins experimentally. The fact that v for unfolded proteins was found to be approximately 0.6 was strong experimental evidence in favor of the random coil model (Tanford 1968), and Tanford concluded that unfolded proteins were indeed random coils based on this evidence. Similar work by Brant and Flory also found that unfolded protein dimensions scaled as random coils. Furthermore, they determined that, for peptides with non-proline and non-glycine residues, the side chain composition of the chain only marginally affects scaling properties (Brant and Flory 1965a). These experiments and those like them helped to establish the random coil model as the dominant model for denatured proteins, both then and now.

In addition to investigating scaling properties of denatured proteins, Tanford also studied whether the unfolded state differs under different conditions. It was found that a thermally denatured protein could undergo a further optical transition when treated with GmHCl (Aune et al. 1967). This was taken to be evidence that thermally denatured proteins are not as unfolded as those denatured with GmHCl, and Tanford advised that all unfolding experiments should be done with a strong denaturant like GmHCl rather than by temperature or pH titration (Tanford 1968). Subsequent research, however, found this conclusion to be inaccurate. Privalov points out that the interpretation of the optical data was flawed, and notes that intrinsic viscosity cannot be used for high temperatures without the appropriate correction factors (Privalov 1979). When these factors are accounted for, thermally denatured proteins appear to behave identically to chemically denatured proteins.

Contemporary experimental work continues to rely heavily on the Flory scaling factor *v*, but the experimental method of choice in now small angle X-ray scattering (SAXS) (Doniach 2001). SAXS can provide two useful criteria for determining the dimensions of a denatured protein. First, using the Guinier equation, it is possible to extract a model-free radius of gyration from the SAXS profile. Radius of gyration data can be amassed from many different experimental studies to examine Flory's scaling law over a large range of protein sizes. Recently, this has been done (Millett et al. 2002; Kohn et al. 2004), and it is found that, even up to 549 residues, unfolded proteins exhibit random coil scaling, with $v = 0.589 \pm 0.030$. The other useful method that SAXS provides for measuring chain compactness is the Kratky plot. In this plot, the scattering profile is rescaled so that the intensity I is multiplied by the scattering factor, s^2. For a random chain, a plot of $s^2I(s)$ versus s should be monotonically increasing, whereas a compact chain will exhibit a maximum in this plot (Doniach 2001). Random coil Kratky plots are observed for a wide array of unfolded proteins, providing additional evidence that denatured proteins are random coils (Semisotnov et al. 1996).

Experimental scaling evidence for the random coil model is convincing, but it may be unreasonable to expect that one number, the Flory exponent, will account for all of the complexities of denatured proteins. Although the random coil model remains to date the most popular and well-characterized model for the denatured state, other experiments as well as simulations have begun to shed doubt on whether unfolded proteins are really random coils.

Random Coils in Simulation

Because of the thoroughness of Flory's original theoretical development, it is not surprising that many computational simulations confirm his results. A recent example is work by Goldenberg, who performed simulations on four proteins ranging in size from 26 to 268 residues (Goldenberg 2003). He finds that these proteins in simulation exhibit an end-to-end distance distribution that is expected for random chains with excluded volume. In addition, the scaling law he derives agrees closely with the predictions by Flory and findings of experiment, with $v = 0.58 \pm 0.02$. Goldenberg's simulations, however, are limited by his handling of excluded volume. Because of the difficulty in simulating long random chains with independent conformations, each trial in his simulation is generated without consideration of excluded volume and then minimized to remove hard sphere bumps. A close examination of the distribution of Ramachandran angles reveals that these simulations fail to capture the observed conformations of real proteins, folded or unfolded (Hovmöller et al. 2002; Hu et al. 2003). Thus, these simulations are not as compelling as originally hoped, and the simulation of random coils remains quite controversial (Dinner and Karplus 2001; van Gunsteren et al. 2001a; b).

Other simulations have found direct violations to Flory's original theory of random coils. One of the assumptions of the random coil model is that each residue's conformational distribution assorts independently, i.e. the conformation of a given residue is not affected by the conformation of an adjacent residue except for those restrictions determined in the Ramachandran plot (Flory 1969). This assumption—the isolated pair hypothesis—was tested rigorously by Pappu *et. al*. in a simple hard-sphere simulation of short peptides (Pappu et al. 2000). In this work, the authors tiled the Ramachandran plot into box-shaped bins called *mesostates*, and examined in detail the fraction of allowed conformations for a dipeptide in each mesostate. If each residue were independent, then the fraction of allowed conformations for a longer peptide would equal the product of the fractions from its component mesostates. For example, if the allowable fraction for the helix mesostate O is 0.6, then the fraction for three O residues should be $0.6^3 = 0.22$ if each residue assorts independently. Instead, it was found that this property was not satisfied, and the allowable fraction was often much less than the fraction predicted by the isolated pair hypothesis, particularly for conformations that intermixed helix and strand conformations.

Subsequent simulations have confirmed the original findings of Pappu *et. al*. Langevin dynamics simulations by Zaman *et. al.* showed that the conformational transitions between regions in the Ramachandran plot are not symmetric as they should be if conformations assorted independently (Zaman et al. 2003). Other work by Brooks' group has used simulation to show that the helix-coil parameters s and σ are size dependent for short peptides, indicating that conformational independence is not a valid assumption for peptides shorter than 6 residues (Ohkubo and Brooks 2003). Furthermore, a recent survey of protein conformational space for tripeptides, tetrapeptides, and pentapeptides has shown that, rather than being conformationally independent, the actual conformational space is quite constrained and can even be mapped sensibly in 3 dimensions rather than $2^5 = 32$ (Sims et al. 2005). While on a large scale unfolded proteins may behave as random coils, all of this evidence suggests that something more complex may be occurring on the per-residue level. Exactly what is still uncertain.

1.6 The Residual (Native-Like) Structure Model

The next model to develop in studying the denatured state posits that unfolded proteins contain a certain degree of native-like structure or topology. The degree to which the unfolded chain has native structure is undetermined, but it is clear that inasmuch as unfolded proteins resemble their folded counterparts it will be easier to fold. This idea of residual structure grew in popularity during the late 1980's as researchers observed aberrant *m*-values for protein denaturation (Dill and Shortle 1991). Because of the uncertainty in the native-like bias, this model does not have the same highly developed theoretical framework that the random coil model has. As a result, this model remains highly controversial, particularly because it is difficult to conceptualize how an unfolded chain could possess native-like topology and yet exhibit random-coil statistics.

Experimental Evidence for Residual Structure

Although the initial support for collapsed denatured states came from protein *m*-values, the primary technique used to measure native-like structure has been nuclear magnetic resonance spectroscopy (NMR). NMR is uniquely suited for observation of the denatured state: as a spectroscopic technique, it can observe the entire denatured ensemble, but it does this in a unique way. NMR is sensitive to the magnetic environment of atomic nuclei, and unlike other spectroscopies individual atoms can be identified in a straightforward way (Levitt 2001). Several labs have used NMR to identify residual, native-like structure in the denatured state.

Since atomic nuclei exchange energy through quantum mechanical coupling, a natural application of NMR to unfolded proteins is to measure distances between atoms or residues. This approach has been applied to several protein systems, most notably by the Shortle group. Gillespie and Shortle performed this type of experiment on Δ131Δ, a fragment of staphylococcal nuclease with residues removed from each of the N- and C-termini. Previously, it had been shown that Δ131Δ was a good model system for unfolded staphylococcal nuclease (Alexandrescu et al. 1994). In this study, Gillespie and Shortle introduced 14 spin labels individually through cysteine mutagenesis (Gillespie and Shortle 1997a). The spin labels allowed them to obtain almost 700 distances between the labels themselves and the coupled nitrogen atoms in the protein. Using these distances, they were able to reconstruct a model for denatured staphylococcal nuclease, and this model appeared quite similar to the native nuclease: α-helices remained approximately cylindrical, hydrophobic regions retained their hydrophobic cores, and β-strands continued to be extended (Gillespie and Shortle 1997b). The unfolded ensemble appeared similar to folded nuclease, except that the unfolded structures were expanded and less rigid.

The work of Gillespie and Shortle has been reproduced several times with similar results. Yi *et. al*. subsequently performed a labeling experiment with protein L and found a similar result: the couplings observed by paramagnetic labeling in the GmHCl-denatured state were roughly consistent with residual native-like structure (Yi et al. 2000). Another experiment by Lindorff-Larsen *et. al*. addressed the possibility of bias in the solution of structures using NMR distance constraints (Lindorff-Larsen et al. 2004). They developed a more efficient sampling method to produce structures consistent with the constraints in a spin labeled sample of bovine acyl-coenzyme A. Although their ensemble of structures displays much less native-like character than the ensemble of

Gillespie and Shortle, they nevertheless observe a nonrandom distribution of couplings consistent with some degree of local native structure.

NMR residual dipolar couplings (RDCs) have also provided evidence for residual structure in the denatured state. These RDCs normally average to zero in an isotopic solution of proteins, but they can be observed if the proteins are aligned in the magnetic field, using gels, bicelles, or phage particles (Prestegard et al. 2000). Here again, Shortle's group was influential in developing techniques for using RDCs to study denatured proteins. Using Δ131Δ, Shortle and Ackerman showed that RDCs in 8M urea correlate well with RDCs in water (Shortle and Ackerman 2001). Since RDCs measure the shape and distance properties of the molecule, this was interpreted to mean that, even in high concentrations of denaturant, the native-like structure of Δ131Δ persisted. Further investigation by Shortle and Ackerman found that this persistence of structure was robust, both to mutation of the protein (Ackerman and Shortle 2002b) and to the type of alignment media used (Ackerman and Shortle 2002a). As other labs investigated RDCs in the denatured state, two observations were made: First, the proteins studied so far exhibit nonuniform RDCs in the denatured state, a property that might not be expected of an isotopically fluctuating random coil (Mohana-Borges et al. 2004; Ohnishi et al. 2004). Second, with the exception of eglin C (Ohnishi et al. 2004), it is generally not the case that native RDCs correlate directly to denatured RDCs. If it is generally true the denatured proteins retain native-like topology, the second observation may be explained by the fact that RDCs depend not only on topology but also alignment and overall shape.

Other experimental data can also be interpreted within the residual structure model. First, it is clear that fluorescence energy transfer studies can also be used to measure intraresidue distances in the denatured state. Although it is much harder to obtain a large set of distances in theses studies, the data indicates a similar heterogeneity to what is observed in NMR experiments (Pletneva et al. 2005). The residual structure model has also been used to interpret experimental data on the unfolded state. One example of this was done by Calmettes *et. al.* (Calmettes et al. 1993). Using molecular simulations, they sought to reproduce the small angle neutron scattering (SANS) profile of denatured phosphoglycerate kinase. Their results showed that, while only random distributions could reproduce the observed SANS profiles, the smallest independent segment of structure could be almost 17Å in diameter. In other words, large rigid globular proteins could not reproduce the profile, but a chain of smaller "spheres" of native structure could model the distribution quite well. By modeling phosphoglycerate kinase as a chain of 17Å non-overlapping spheres, they were able to fit segments of native structure in to the spheres and develop structural models for the denatured state. The resulting structures were random on the global level but native-like locally, indicating that, at least for one example, denatured states could be modeled as segments of locally native structure.

Residual Structure in Simulations

Several researchers have addressed the issue of structural biases in the denatured state using simulations. Even for short peptides, such biases may provide physical clues to the determinants of residual structure in the native state, as the side chains of neighboring residues influence one another to form a native bias in the unfolded state. Such biases have been observed experimentally for a tripeptide (Eker et al. 2004), but have remained difficult to reproduce with simulations. One attempt to explain nearest-neighbor biases was recently performed by Avbelj and Baldwin (Avbelj and Baldwin 2004). Using calculations of electrostatic solvation free energy, they can successfully predict conformational trends for a residue given its nearest neighbors, supporting the idea that solvation is a primary factor in determining conformational bias in the denatured state.

Larger simulations have also shown a bias for residual structure in the native state. Although simulations of this type are exceedingly difficult to perform at present, simulations of small proteins using large scale distributed computing approaches can provide one means of simulating the denatured state. When such an approach is used, it is found that the average conformation of unfolded proteins is very similar to the native structure as measured by a contact distance matrix (Zagrovic et al. 2002). While it has been proven that averaging contact distance matrices may produce a misleading similarity between native and unfolded structures, Zagrovic and Pande have demonstrated the robustness of this *mean-structure* hypothesis and are convinced that the similarity represents a legitimate relationship between the native and denatured states (Zagrovic and Pande 2004). If valid, these results further corroborate the residual structure observed in the denatured state by experimental methods. Additionally, these simulations give one example of how chains with native-like structure can appear to be random coils: the individual members of a protein ensemble appear quite random, but the conformational bias results in a native-like structure that would be observed by NMR or other distance-sensitive techniques (Zagrovic and Pande 2003).

Another simulation that has found evidence for residual structure in the denatured state has been performed by Wong *et. al.* on the protein barnase, a small ribonuclease with 110 residues. *In silico* thermal denaturation of barnase yields an ensemble of structures with persistent native contacts and a dynamic native-like topology (Wong et al. 2000). The persistent contacts are observed to be hydrophobic in nature, and helices fluctuate between helical and non-helical forms. A comparison with the NMR distance constraints for unfolded barnase reveals good overall agreement about which regions are partially ordered, but it is cautioned that the short timescale of the molecular dynamics simulation may not have sampled all of the available conformations adequately. It is proposed by the authors that the role of residual structure in the unfolded state is to serve as folding initiation sites for the folding transition, thus speeding up the kinetic search for the native state. If indeed residual native-like structure in the unfolded state exists, this is a highly plausible explanation for its utility.

Objections to the Residual Structure Model

The suggestion that the denatured state retains a native-like bias stands in stark contrast to the random coil model, which states that no such bias should exist. It is not surprising therefore that both theoretical and experimental work have both questioned the validity of this model. Much of this work has sought to identify possible artifacts in the residual dipolar coupling data. If it can be shown that random or nearly random chains produce the same residual dipolar couplings as native-like unfolded chains, a primary argument in favor of the residual structure model would be eliminated.

Theoretical work has in fact shown that RDCs can be expected from random coil chains. Because the chains are aligned in a stretched gel or other alignment media, there will be a certain organization to a random chain based on the simple fact that it cannot penetrate the surrounding barriers. The initial calculations using this idea showed that random coil RDCs should be nonzero and uniform (Louhivuori et al. 2003). Later, the model was revised to explain the non-uniform nature of RDCs from real proteins, but this work could not rule out structure in the denatured state (Louhivuori et al. 2004). A related project has been more successful in reproducing the observed RDCs from native proteins. Jha *et. al*. have been able to back calculate the RDCs for ubiquitin, eglin C, and Δ131Δ by constructing random-flight chains based on conformational preferences stored in the coil library (Jha et al. 2005). When nearest neighbor biases are included, they observe a correlation of R=0.70 between the observed and predicted RDCs in apomyoglobin, but the correlation deteriorates to R=0.42 if no biases are included. Such a result may indicate the existence of very weak native-like bias in the denatured state, but it is doubtful that a bias this weak will significantly affect the folding transition.

A recent experiment also sheds doubt on the existence of native-like topology in the denatured state. Alexandrescu's group has examined the structure in the native state by comparing the residual dipolar couplings of native staphylococcal nuclease and a fragment thereof which is missing 47 C-terminal residues (Sallum et al. 2005). No correlation is observed between the fragment and wild-type nuclease, but a strong correlation is revealed when wild-type nuclease is denatured. Because the fragment and wild-type protein are structurally different under native conditions, it is reasoned that the denatured RDCs should reflect this difference if the residual structure model holds. Since the RDCs are nearly identical, it is argued that something else must be happening in the denatured state. They propose that structural fragmentation is the cause of correlation: on a global scale, the protein lacks native like topology, but locally it retains some native-like structure.

It is unclear at this point whether the residual structure model will hold up to further scrutiny. Spin labeling experiments have shown that some degree of native structure exists within the denatured state, but the work described above indicates that only a small native bias may be sufficient for explaining the residual dipolar coupling experiments. Regardless of the recent scrutiny, however, the residual structure model remains to be a dominant model for denatured proteins, largely because of its simplifying nature in describing how proteins fold.

1.7 The Polyproline II Helix Model

The final model for denatured proteins has its origins in an observation made in 1968 by Tiffany and Krimm (Tiffany and Krimm 1968a; b). They measured CD on short chains of polyproline and polyglutamic acid and noted that the spectra were similar. Since the conformation of polyproline is fixed, it was supposed that polyglutamic acid had a similar conformation. Indeed, the characteristic CD spectrum for unfolded proteins is identical to the spectrum observed for peptides of polyproline. Many binding targets are found to be in the polyproline II (P_{II}) helical conformation, and a large fraction of the coil library is found to be in P_{II} conformation (Stapley and Creamer 1999). These facts have led to the idea that the unfolded state is a fluctuating statistical ensemble of short fragments of polyproline II helix (P_{II}), a 3_1 helix with $\phi = -75^0$ and $\psi = 145^0$ (Creamer and Campbell 2002). Although presently it is not clear how P_{II} may influence the transition between the folded and unfolded states, having a more uniform starting point in the folding transition may ease the kinetic search problem.

Experimental Evidence for a P_{II} Denatured State

Many experiments have demonstrated a tendency for disordered proteins to adopt P_{II} conformations in addition to those described by Tiffany and Krimm above. Generally, these experiments examine the conformational propensities of short peptides. While these experiments have the advantage of being tractable, they have the disadvantage of neglecting longer-range interactions, such as hydrophobic collapse, that could perturb a true fragmented P_{II} ensemble. Such long range interactions may be a natural consequence of the fact that long P_{II} helices are highly unlikely in a denatured protein.

Woutersen and Hamm have developed a novel spectroscopic technique for measuring the backbone conformation of trialanine (Woutersen and Hamm 2000). After exiting the peptide amide I transition with a focused pulse of energy, they quickly (within two picoseconds) measure an absorption spectrum of the sample. Quantum mechanical coupling between adjacent peptide groups will result in a change in the observed spectrum. The spectrum can then be compared with theoretical calculations, and ϕ, ψ can be determined for the residue in question. This technique, called two dimensional pulse probe infrared spectroscopy, was applied to trialanine fragments (Woutersen and Hamm 2000), and it was observed that P_{II} was the preferred conformation, with $\phi \approx -80^0$ and

$\psi \approx 150^0$. Moreover, this conformation was argued to be the exclusive conformation of trialanine for two reasons: First, the complex relationship between the spectrum and the fitted parameters would have likely resulted in an unrealistic ϕ, ψ value if the conformation was an ensemble average. Second, the cross-peak anisotropy was observed to be near a theoretical maximum, and it was reasoned, that for this to occur contributions from other conformations would necessarily be small. As a result, these authors suggest that the denatured state of proteins has a high propensity for P_{II} conformation.

If P_{II} is the preferred conformation for a tripeptide, it is reasonable to expect that longer peptides should also exhibit a preference for this conformation. This has been investigated on a seven residue alanine fragment by Shi *et. al*. using NMR (Shi et al. 2002a; Shi et al. 2002b). NMR provides two useful techniques for identifying backbone conformations in proteins, and both were employed in this study to examine the conformation of polyalanine in water. First, using the Karplus relation it is possible to relate the $J_{HN\alpha}$ coupling to the backbone ϕ angle. Second, it is known that when a peptide forms an α-helix, nuclear Overhauser effect (NOE) couplings are observable between the methyl protons of one residue and the protons of the nearby residues in the helix. When these two methods were employed on the polyalanine fragment, it was found that ϕ was approximately -70^0 and that no NOE couplings were present (Shi et al. 2002a). Because of this, it was reasoned that the dominant peptide conformation had to be far from α-helix, in the P_{II} region of the Ramachandran plot. It was estimated that the conformational contributions of α-helix and β-strand were both less than 10%, although the β contribution increased at higher temperatures. Later work using the same methodology attempted to fit this $P_{II} \rightarrow \beta$ transition to helix-coil model, and it was found that the fluctuations around P_{II} were not cooperative, with a value of σ of about 1 (Chen et al. 2004). Here again, experimental results on short peptides favors the P_{II} conformation for the unfolded state.

At this point it becomes interesting to ask why the polyproline II helix is the favorable conformation for short stretches of polyalanine. The experimental results to date support the idea that solvation is an important interaction favoring this conformation. In an elegant experiment by Chellgren and Creamer, it was shown that the P_{II} conformation is more favorable in D_2O than in H_2O (Chellgren and Creamer 2004). Because D_2O has a higher tendency toward hydrogen bonding and therefore is more ordered than H_2O, it was suggested that P_{II} perturbs water less than other conformations, such as helix or strand, does. It also follows that P_{II} is less disruptive to water than a random coil conformation. Other experiments have shown P_{II} to be highly sensitive to solvent composition, supporting the idea that water molecules are important in the stability of the P_{II} helix (Liu et al. 2004).

Simulations of the Polyproline II Helix

Computational modeling is a useful tool for identifying the fundamental interactions that favor a conformation, and it has been applied extensively to the P_{II} helix in solution (Creamer and Campbell 2002). Of the simulations that have been done, the simplest calculations are also the most compelling. These have been performed by Pappu and Rose and use a purely repulsive soft-sphere potential (Pappu and Rose 2002). Monte Carlo simulations on short peptide chains using this simple potential indicate that the formation of P_{II} may be a consequence of sterics alone: the soft sphere potential minimizes protein packing density and maximizes the exposure of the backbone to the solvent. Maximization of solvent exposure also rationalizes the experiments described above, since a peptide in P_{II} conformation is maximized its potential interactions with water and thus will experience the full effect of changes in solvent composition. Further work by Pappu's lab has shown that a conformational preference for P_{II} is not inconsistent with random coil statistics (Tran et al. 2005). Although the repulsive potential favors an extended chain, conformational entropy prevents long segments of P_{II} from forming, and thus the chain yet retain random coil statistics.

Traditional methods of simulation have also found a favorable preference for P_{II} conformation. Two recent molecular dynamics studies have shown that short alanine peptides favor P_{II}. The first study, by Mu and Stock, examined trialanine and found that nearly 80% of the time the central residue was either in a P_{II} or β conformation (Mu and Stock 2002). A second study extended their findings by examining an eight residue polyalanine fragment. Ramakrishnan *et. al.* found once again that the simulated fragments occupied the P_{II} conformation approximately 70% of the time, although their P_{II} had a smaller ψ torsion of about 85^0 (Ramakrishnan et al. 2004). They observed that, in the remaining 30% of the time, the fragments sampled β-turns and short fragments of α-helices. Their results indicate the existence of other conformational preferences for longer peptides and proteins and may provide evidence that both P_{II} and native-like residual structure are present in unfolded proteins.

A comprehensive study of polyalanine in solution was performed by Kentsis *et. al.* using alanine peptides of length 7 and 14 (Kentsis et al. 2004). They used a sophisticated Monte Carlo simulation technique to capture the detailed effects of solvent-chain interactions, and to ensure robustness of their results two separate simulations were performed with different force fields. Both simulations identify a similar trend that the conformational preference of the protein backbone is the P_{II} helix. The maximum size of P_{II} helices is observed to be 5 residues, and the chain fluctuates readily between P_{II} and other conformations. The simulations further indicate solvent entropy as the dominant cause for P_{II} stabilization, and they suggest that the P_{II} conformation facilitates α-helix formation by reducing the entropic penalty of helix nucleation. A related study also finds solvent entropy to be important in the stabilization of P_{II}. Mezei *et. al*. used a similar simulation technique to measure the solvation free energy of rigid P_{II}, α-helix, and β-strand conformations in solution (Mezei et al. 2004). They find that the free energy of solvation for P_{II} is much more favorable than α-helix or parallel β-strand, with values of –4.7, –2.0, and –3.9 kcal/mol, respectively. Together, these simulations provide strong support for modeling the unfolded state as short segments of P_{II} helix.

Summary of Unfolded State Models

The random coil, residual structure, and P_{II} helix model all have support from experiment and theory, but at this time it is difficult to synthesize all of the data in to one coherent model for the unfolded state. Not all of the experiments are contradictory, but it is clear that a “reconciliation problem” exists, since a random coil is not random if it contains native-like structure or polyproline II helix (Millett et al. 2002). The forces, as modeled by experiment, are also puzzling in that some simulations favor the residual structure model whereas other simulations favor the P_{II} model. As discussed below, each of these considerations for the unfolded state has a significant impact on our understanding of protein folding, and much of the debate in the protein folding field stems from a poor understanding of the unfolded state.

1.8 Models for the Folding Transition

Several models exist for the protein folding transition, and all must start with assumptions about the denatured state. As stated earlier, one of the major tasks of any model is to propose a mechanism by which the protein resolves the Levinthal paradox (Levinthal 1969). It seems reasonable that not all proteins will fold with identical mechanisms (Fersht 2000), and it may be that the competing models have more similarities than differences (Gianni et al. 2003). Here, we will briefly describe several of the major models for protein folding, focusing on how the assumptions about the unfolded state in each model determine how folding occurs.

The Diffusion-Collision Model and Hierarchic Folding

The diffusion-collision model is one of the earliest models for protein folding, proposed initially by Karplus and Weaver in 1976 (Karplus and Weaver 1976; Karplus and Weaver 1979). The model assumes an unfolded state of microdomains, short stretches of residues that flicker in and out of their native structure. During folding, microdomains diffuse freely, similar to tethered spheres, and as native-like contacts are made, they combine to form the final native structure. A simple-minded estimate of the time required for two separate microdomains to merge can be modeled using the following equation (Karplus and Weaver 1976):

$$\tau \approx \frac{1}{\beta}\left(\frac{l\Delta V}{DA}\right) \tag{1.23}$$

In this equation, l characterizes the length of the tether between the two domains, ΔV is the volume in which the microdomains can diffuse, D is the diffusion coefficient, and A is the area of the target microdomain. The constant β reflects that not all microdomains will be in the folded conformation at all times, and is thus related to an equilibrium constant.

The diffusion-collision model, when properly parameterized, is able to predict the kinetic rates of some proteins well (Karplus and Weaver 1994; Burton et al. 1998). Additionally, solvent viscosity has been shown to impact the folding rates for a few proteins (Karplus and Weaver 1994). The model is compatible with some of the ideas of the residual structure model of the denatured state: the presence of semi-native microdomains in the unfolded ensemble would explain the existence of native-like structure, although it is difficult to imagine that long-range native topology would exist in this model.

An extension of the diffusion-collision model is the hierarchic or framework model for protein folding (Baldwin and Rose 1999b; a). It is based on the observation that proteins are organized hierarchically: domains in a folded protein are organized in a hierarchy of locally interacting subdomains (Rose 1979). The hierarchic model extends the diffusion-collision model by specifying that the folded microdomains are the nascent secondary structures in the final folded chain. Furthermore, these secondary structures do not simply diffuse freely throughout a tethered volume; rather, they interact locally with other units of secondary structure, assembling hierarchically to form the final folded chain. The hierarchic model is supported by experimental evidence that, when helices are excised from native proteins, they fluctuate in and out of their native secondary structure (Baldwin and Rose 1999a). Simulations have also given evidence for this type of protein folding (Baldwin and Rose 1999b). From the perspective of the unfolded state, the hierarchic model also fits well with the residual structure model. If the secondary structures are fluctuating between native and P_{II} conformations, it may also expain why P_{II} helix is prevalent in unfolded proteins. This model would seem to be at odds, however, with the random coil model for the denatured state.

Recently, Englander's group has proposed another model for folding that is similar to both the diffusion-collision and hierarchic models. This model proposes that proteins fold as a stepwise process as *foldon units* form on top of one another (Maity et al. 2005). Hydrogen exchange experiments on cytochrome *c* indicate that there are five foldon units in this protein and suggest distinct order in their assembly: the N- and C-terminal helices form first in the pathway, followed by a coil region and the helix from residues 60-70, followed by several other foldon units. In this model of folding, the microdomains are the foldon units, and in cytochrome *c* the foldons roughly correspond to secondary structures as predicted by the hierarchic model. Unlike the standard diffusion-collision model, however, where several different combinations of microdomains or local secondary structure elements can combine to form the final structure, foldons are proposed to associate in a specific stepwise order. This model of folding for cytochrome *c* also differs from the hierarchic model in that it proposes that the N- and C-terminal ends, which are non-local in character, bind early in the folding pathway. As more proteins are examined in light of this new model, it will become more clear whether the association of the N- and C- terminal ends is a general feature of folding (Krishna and Englander 2005). If so, then native-like topology in the denatured state may be more compatible with this model than in other diffusion-collision models.

Heteropolymer Collapse

The heteropolymer collapse model for protein folding was developed primarily in response to the idea that hydrophobicity is the driving force for protein folding. In this model, a random coil denatured state collapses under folding conditions due to the hydrophobic driving force (Dill 1985). The collapsed chain then undergoes structural rearrangement and formation of secondary and tertiary contacts until the final native structure is determined. This structure resolves the Levinthal paradox through the use of excluded volume: as a chain of finite thickness is forced into a small volume, an exceedingly large amount of conformational space is eliminated because of steric considerations. Unfortunately, estimates have shown that even this large loss in conformational space may not be enough to overcome the Levinthal paradox (Karplus and Weaver 1994), and the absence of collapsed intermediates in many protein folding pathways seems to indicate that if a hydrophobic collapse does occur, it does so very early on in the folding process. Additionally, this model's approach the unfolded state is the classical random coil, and therefore it does not account for native structure or P_{II} helices in the denatured state unless those conformations are somehow incorporated in to the collapsed protein.

Funnels and Nucleation-Condensation

One of the more recent and popular models of protein folding is the nucleation-condensation model. This model has incorporated the idea that folding pathways may not be unique in nature: rather than comprising a set of stepwise events, protein folding is suggested to consist of multiple, parallel pathways—the folding funnel concept (Dill and Chan 1997). This idea complicates protein folding: rather than one clear transition between the unfolded and folded state, many pathways must be accounted for, and experimentally it may be difficult to characterize one "true" transition state during folding. Fortunately, simulations indicate that while many pathways contribute to folding, an average pathway is observed that can account for most of the transition (Lazaridis and Karplus 1997).

The distinguishing idea behind the nucleation-condensation model is that the folded structure forms around a few key nonlocal contacts in the native structure. Secondary and tertiary structure then form in a concerted manner around the nucleus of native contacts. This type of folding has been observed in simulations of chymotrypsin inhibitor 2 (CI2) and seems to be common based on kinetic Φ-value analysis (Daggett and Fersht 2003). In CI2, the nucleus appears to be a small portion of the helix and β-sheet that form the core of the protein. Once this is formed, the remaining structure condenses upon the loosely formed core. Although the unfolded state for the nucleation-model is generally assumed to be random in nature, CI2 simulations show fluctuating elements of secondary structure (Lazaridis and Karplus 1997), and a small native-like bias in the unfolded state is compatible with this model. Indeed, it has been suggested that hierarchic folding and nucleation-condensation differ only in the amount of native-like structure in the denatured state (Gianni et al. 2003).

Why should the energy landscape for folding be funnel shaped? This intriguing question was addressed in 1984 by Go (Go 1984). He proposed the consistency principle in protein folding: in a folded protein, all energy terms are at (or near) their optimal values. The consistency principle was based on the observation that structural deviations from equilibrium values in folded proteins are rare. From this idea, Go hypothesized that nonlocal interactions may be just as important in stabilizing native structure as local interactions. From a folding funnel perspective, the consistency principle dictates that, as local and non-local native contacts are formed, a protein's free energy will become more favorable, thus driving the unfolded chain down the funnel toward the folded state. Onuchic and Wolynes call this idea "the principle of minimal frustration:" interactions within the protein are not in conflict (Onuchic and Wolynes 2004). Proteins proceed down the folding funnel, generally hindered only by an entropic barrier when approximately 60% of the native structure is formed (Wolynes et al. 1995).

Traditionally, Φ-value analysis has been used to show that native structure, if present, is minimal in the transition state (Daggett and Fersht 2003). This observation has led to the assumption that a nucleation-condensation model requires a random-coil denatured state. In fact, the nucleation-condensation model is rather insensitive to the details of unfolded proteins, mainly because the consistency principle supplies an organizational regime capable of overcoming the Levinthal paradox in both random and natively-biased denatured states. Some recent observations are of interest. First, it has been noted that a large experimental uncertainty is associated with experimentally determined Φ-values (Ingo Ruczinski, personal communication), and therefore transition states may be more structured than originally believed. Second, Frieden has observed that side chain stabilization forms very late in the folding process (Frieden 2003), so it seems unlikely that specific side chain interactions are participating in a nucleation event. Thus, the primary justification for a random denatured state in the nucleation-condensation model (Φ-values) is not as strong as once thought, and certain types of interactions (long range tertiary interactions), cannot drive the formation of a folding nucleus. A convenient explanation would be that nascent secondary structure elements form the nucleus, making this model more similar to the diffusion-collision model than distinct from it.

The Topomer Search Model

The final model we address is the topomer search model. This model was formulated by Plaxco's group based on the correlation of reduced contact order (equation 1.15) with folding rates in small proteins (Makarov and Plaxco 2003). In it, the unfolded state is largely random, and the rate-limiting step in folding is the formation of the appropriate native-like topology in space. A protein with a complex native topology as measured by reduced contact order will take proportionately more time to fold than a protein with a simple topology in the native state. The topomer search model does not exclude the existence of rapidly fluctuating units of native-like secondary structure, nor does it prohibit the formation of P_{II} in the unfolded stuate; however, in this model the existence of these conformations is largely irrelevant, because the rate-limiting step is the global formation of a native-like topomer, and the formation of this global topology is assumed to occur on much longer timescales than other chain fluctuations. The critical assumption is that the rate of folding is proportional to the probability that a random chain has a native-like topology, a decidedly polymer-theoretic approach that requires a random search through conformational space.

It was at first thought that the topomer search model resolved the Levinthal paradox because fast folding proteins have a high probability of forming native-like topomers in the denatured state. This idea was investigated by Wallin and Chan and found to be true: when clear definitions are given for native-like topology, faster folding proteins have a larger section of conformational space that is native-like than slower folding proteins (Wallin and Chan 2005). However, it was found that the Levinthal paradox still applies because the number of native-like conformations is still dwarfed by the size of the random search. Thus, the topomer search model cannot at present account for the Levinthal search problem.

Another problem faced by the topomer search model is the fact that other, simpler metrics than contact order have also been shown to correlate with folding rate. As discussed above, secondary structure content (Gong et al. 2003) and even protein size (Naganathan and Munoz 2005) can also predict folding rates. While the topomer search model may describe why contact order and folding rate are related, it does not explain why other metrics perform as well. Because of this, the topomer search model seems to be in transition: a new model for its unfolded state is needed that will account for the search problem while at the same time allowing for other factors in determining folding rates.

1.9 Overview of Thesis

Chapter 2. As mentioned above in section 1.5, the isolated pair hypothesis is a fundamental tenet of the random coil model for unfolded proteins. In this chapter, we build on the work of Pappu and Rose by examining in detail why the Flory IPH fails. For over 50 years, it has been assumed that no systematic steric clashes existed beyond the dipeptide. This chapter outlines an experiment using simple hard sphere simulations that identifies a fundamental interaction between α-helices and β-strands that exists in both the folded and unfolded states of proteins. In short, helices cannot be followed by strands without a turn or coil region between the two. We confirm the validity of the hard sphere model by showing that helices and strands are not juxtaposed in the PDB. Not only does this chapter firmly establish the hard sphere model as a valid means of investigating denatured proteins, it suggests that other disfavored conformations may exist that substantially reduce the size of conformational space in unfolded proteins.

Chapter 3. Another key behavior of random coils are radii of gyration that scale as equation (1.22). Although Tanford himself was aware that nonrandom conformations can exhibit random R_G's, in this chapter we address the sensitivity of random coil scaling directly. Starting with native protein structures, we construct an absurd model for the denatured state where only one in twelve residues are free to sample protein conformational space. The model should not scale as a random coil, because 92% of the protein structure is perfectly rigid. Nevertheless, a dataset of 33 proteins simulated in this way exhibit near-perfect random coil statistics. As in chapter 2, we identify another reason to doubt the assertion that unfolded proteins are random coils. Since the random coil model is not unique in its ability to predict the observed scaling of chain dimensions, the residual structure and P_{II} models cannot be discounted solely on the basis of scaling data.

Chapter 4. Here, we present an aside in preparation for an exhaustive enumeration of protein conformational space. The protein coil library (PCL), as discussed above, has proven to be a useful tool for modeling unfolded protein conformations. Unfortunately, no standard form of the coil library has been developed to date. In this chapter, we develop a standardized implementation of the coil library and make that implementation available on the World Wide Web. Turns are included in this library, but helices and strands are removed. The resulting database can be examined for conformational preferences, and the functionality is provided to search the database using the output from several different culling servers. This library will be used as a control for the simulations in chapter 5: disfavored conformations identified in simulation should not be present in the PCL.

Chapter 5. As a culmination to the previous chapters, the goal of chapter of five is to exhaustively identify all disfavored conformations in polyalanine peptides of length 1-6. Section 1.3 presented two potential organizing forces in protein folding: hard sphere sterics and solvation. Both of these forces are implemented here, and we examine the extent to which these forces reduce the size of conformational space. We identify many disfavored conformations and show that each occurs rarely, if at all, in the coil library. To supplement the information presented in chapter 1, all of the conformations identified in this chapter collected and placed onto a searchable web database. Because local chain interactions should diminish beyond six residues, this database represents a complete description of how sterics and solvation cause the IPH to fail. It is also a goal of this chapter to identify the significance of these local interactions. How much do local sterics and solvation reduce the size of conformational space compared to the random coil model? To answer this question, we develop a statistical model to estimate the significance of local interactions in reducing the size of conformational space and find that approximately nine orders of magnitude are eliminated for a 100-residue unfolded chain. Finally, we address the significance of the disfavored conformations: it is clear that these conformations are in opposition to a random coil model for unfolded proteins, but the significance in the other two models is less clear. We examine all the conformations together and rationalize how they may be significant in both the residual structure and P_{II} models for the denatured state.

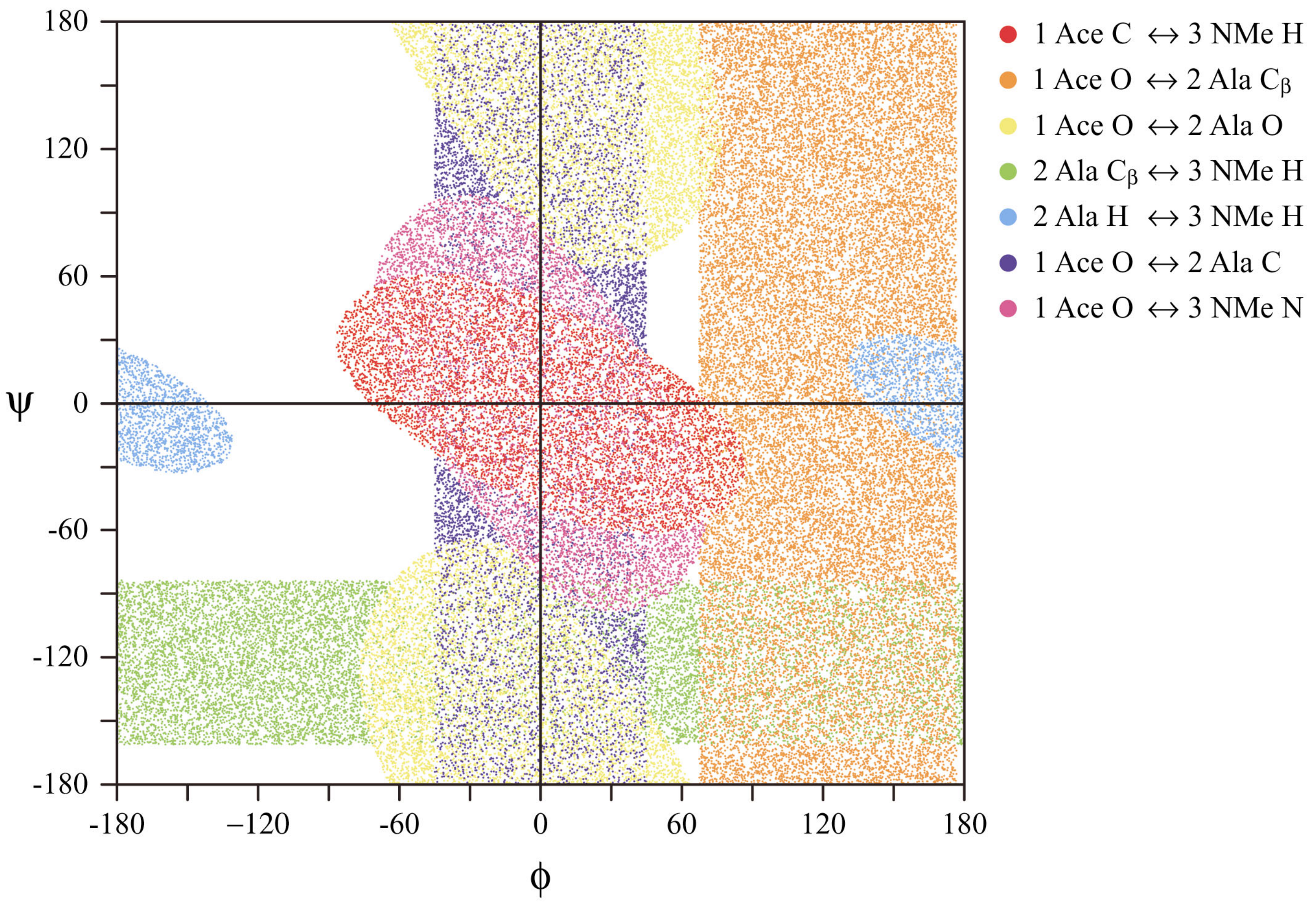


Figure 1.1. Hard sphere collisions involved in the Ramachandran plot. Each colored region represents a different atomic collision in N-actetyl-Ala-N-methylamide, as indicated in the legend. In this structure, N-acetyl (Ace) is residue one, alanine (Ala) is residue two, and N-methylamide (NMe) is residue three.

CHAPTER 2

Steric Restrictions in Protein Folding: An α-helix Cannot be Followed by a Contiguous β-strand[*]

2.1 Abstract

Using only hard-sphere repulsion, we investigated short polyalanyl chains for the presence of sterically-imposed conformational constraints beyond the dipeptide level. We found that a central residue in a helical peptide cannot adopt dihedral angles from strand regions without encountering a steric collision. Consequently, an α-helical segment followed by a β-strand segment must be connected by an intervening linker. This restriction was validated both by simulations and by seeking violations within proteins of known structure. In fact, no violations were found within an extensive database of high-resolution X-ray structures. Nature's exclusion of α-β hybrid segments, fashioned from an α-helix adjoined to a β-strand, is built into proteins at the covalent level. This straightforward conformational constraint has far-reaching consequences in organizing unfolded proteins and limiting the number of possible protein domains.

[*] This chapter is reprinted with permission from the authors as allowed by the copyright agreement. Fitzkee, N. C. and Rose, G. D. (2004) *Protein Science.* **13** (3): 633-9. Copyright © 2004 The Protein Society.

2.2 Introduction

The hard sphere model (Richards 1977) has been an invaluable tool in characterizing fundamental aspects of protein molecules, including their accessible surface area (Lee and Richards 1971; Eriksson et al. 1992), packing (Richards 1977; Richards 1979), and fitting errors (Word et al. 1999). Clearly, atoms are not simply hard spheres; but, quoting Richards,

> "For chemically bonded atoms the distribution is not spherically symmetric nor are the properties of such atoms isotropic. In spite of all this, the use of the hard sphere model has a venerable history and an enviable record in explaining a variety of different observable properties" (Richards 1977).

Arguably, the most important application of the hard sphere model in biochemistry is the now famous ϕ,ψ-plot for a dipeptide, developed by Sasisekharan, Ramakrishnan and Ramachandran (Ramachandran et al. 1963; Ramachandran and Sasisekharan 1968). Recently, this simple idea has been applied to nucleic acids as well (Duarte and Pyle 1998; Murthy et al. 1999). In proteins, the hard sphere model identifies two major populated regions for an alanine dipeptide; backbone dihedral angles in these regions resemble those of an α-helix or a β-strand. Despite their remarkable structural diversity, protein molecules have main chain conformations that lie almost entirely within these two regions.

In this paper, we explore additional steric constraints on polypeptide chains beyond the dipeptide level. We find that an α-helix cannot be followed by a β-strand without an intervening linker. This restriction is a consequence of unavoidable collisions between backbone atoms, and it derives experimental support from the paucity of exceptions among high-resolution protein structures (Berman et al. 2000).

We use several conventions for describing regions of the ϕ,ψ-map. Figure 2.1 shows a ϕ,ψ-distribution of α-helix (yellow), β-strand (red), and polyproline II helix (blue), determined from structures in the PDB (Berman et al. 2000); the more populated the region, the darker the color. Throughout this chapter, β refers to the region given by $-135.0^{o} \leq \phi \leq -105.0^{o}$ and $120.0^{o} \leq \psi \leq 150.0^{o}$ and P_{II} refers to $-80.0^{o} \leq \phi \leq -55.0^{o}$ and $130.0^{o} \leq \psi \leq 155.0^{o}$. We define both a relaxed helical region α', where $-75.0^{o} \leq \phi \leq -45.0^{o}$ and $-60.0^{o} \leq \psi \leq -30.0^{o}$, and a strict helical region α, as a circle of radius 7.0^{o} centered about $\phi = -63.0^{o}$ and $\psi = -45.0^{o}$. Finally, κ represents the entirety of sterically-accessible ϕ,ψ-space for the alanine dipeptide.

An experimental observation raises a question

Hybrid segments consisting of an α-helix followed by a β-strand are rarely observed in the PDB. Instead, helices and strands are interconnected by a transition region – a turn, a loop, or some other linker. Only seven occurrences of three or more α' residues followed by a single β residue were found in a representative set of PDB structures (Hobohm and Sander 1994), but a pattern consisting of three or more α' residues followed by a non-α' residue was detected 37,563 times in this dataset. Why is the direct transition from α to β so rare?

2.3 Materials and Methods

Ramachandran Plots. Ramachandran plots were generated from ϕ,ψ-distributions by subdividing ϕ,ψ-space into a 5^{o} by 5^{o} grid; each of the 72 by 72 grid squares corresponds to a bin. These bins were ranked according to the number of ϕ,ψ-pairs they contain and then grouped into three categories representing the top (i) 33%, (ii) 66% and (iii) 90% of the data. The three groups are plotted in figure 2.1.

Mining the PDB. In all cases, the chains selected from the PDB were that subset of PDBSelect (Hobohm and Sander 1994) structures determined by X-ray diffraction. In all, 1455 chains at the 25% aligned sequence identity level and 5378 chains at the 90% level were included.

Idealized Secondary Structures. Secondary structure was assigned using PROSS (Srinivasan and Rose 1999), a method based solely on ϕ,ψ-angles. Unlike the more familiar DSSP (Kabsch and Sander 1983), PROSS does not include hydrogen bonding in its assignment criteria. Distributions of ϕ,ψ-values were grouped into three secondary structure categories: α-helix, β-strand, and polyproline II (figure 2.1A). The helical region was further subdivided into 1^{o} grid squares (figure 2.1B). Idealized ranges for α, β, and P_{II} were then defined, guided by those bins that represent the top 33% of the data. The α' region is a relaxed definition of α, similar in size to β and P_{II}. These definitions agree well with textbook classifications of secondary structure (Creighton 1984).

Simulations. Monte Carlo simulations of polyalanyl peptides were performed to determine how steric factors influence chain conformation. Polyalanine was chosen as a model for the peptide backbone. Simulated peptides had lengths ranging from 9 to 12 residues. Hydrogen atoms were not included. For each simulation, the ϕ,ψ-distributions of all sterically-allowed conformers were collected so as to accumulate up to 5000 clash-free structures from a maximum of 5 million attempts.

Generation of sterically allowed structures was accomplished by sampling backbone torsion angles at random from α, β, $P_{II,}$ or κ, as appropriate. ω-torsions were varied at random in the range [-175.0°, -185.0°] and assigned in conjunction with backbone torsions. Each attempt was checked for collisions; if none were found, the conformer was accepted and its torsion angles were retained. Otherwise, the conformer was rejected. Rejected structures with a single atomic collision occur at the boundaries between regions; these were cataloged by ϕ,ψ-angle and collision type for use in assembling a collision map. Structures with multiple atomic collisions are not localized at boundary regions and were ignored.

Hard sphere atomic radii from Word *et. al.* (Word et al. 1999) are among the most conservative in the literature and were adopted for this study (table 2.1). These radii were further scaled by a factor of 0.95, ensuring that the observed collisions are not methodological artifacts. The overall robustness of our results was tested extensively by determining the degree to which steric restrictions persist as radii diminish (described below). The collisions identified in this study do not include those with hydrogen atoms. Inclusion of hydrogens would have enlarged the effective radii and, consequently, imposed further restrictions on available conformational space.

2.4 Results

Flexibility of a central wild-card residue in an α-helical peptide: α_4-κ-α_4

A series of host-guest simulations was performed; each consisted of a wild-card κ-residue (the guest) in the middle of an 8-residue polyalanyl peptide that was constrained to be helical (the host). In every case, sterically disallowed patterns identified in simulations were validated against X-ray elucidated structures by searching for exceptions.

The resulting distribution of the guest residue (figure 2.2) plots 5000 allowed structures (from 418,213 attempts). Both raw (figure 2.2A) and binned (figure 2.2B) data exhibit a Y-shaped plot for the guest residue. A comparison of the two figures (2.2A and 2.2B) shows that the binning method captures the distribution successfully. Notably, the Y shape encompasses α-helix, but both β-strand and P_{II} are excluded. This conclusion is highlighted in figure 2.2B by superimposing the 66% contour for β-strand from figure 1A on the binned simulation data. *In short, a single β-guest residue cannot avoid a steric collision in an α-helical host.*

The collision maps in figure 2.3 rationalize this restriction, which is a consequence of a steric clash between the carbonyl oxygens of the guest *i*-residue (O_i) and the *i-3* α-residue (O_{i-3}). The distribution of points for this collision (figure 2.3, green) fits precisely into the void region of figure 2.2.

Additional collisions from this simulation are also shown in figure 2.3. Of particular note is the collision between O_{i-3} and C^{β}_{i+3} (in red) which is responsible for exclusion of the P_{II} region. These two atoms are brought into juxtaposition when the guest κ-residue, at *i*, samples the relevant region in ϕ,ψ-space.

Two different methods were used to test the robustness of these results. First, the atomic radii were reduced well beyond any plausible van der Waals limit by successively decrementing the scaling factor from 0.95 to 0.92, 0.90, 0.88, and 0.85 (figures 2.4A-D). Reduction of the atomic radii results in expansion of the Y-shaped boundaries of the guest residue. However, substantial strand exclusion survives even the most extreme reduction. Similar behavior was also observed when the α region was expanded to a radius of 14^{o} or 21^{o} (figures 2.4E and F). Again, the Y-shaped boundary expands, yet persists. Thus, steric exclusion of a β-residue in a sequence of α-residues is a robust finding, not an artifact of our helix definition or hard sphere radii.

As further validation, helices were excised from proteins of known structure and used as starting structures in simulations. Specifically, 40 12-residue helices were selected at random from X-ray elucidated structures in the PDB (table 2.2), and all side chain atoms beyond C_{β} were eliminated. Simulations were then performed as before, except that the definition of α was varied for each helical residue, using a radius of 7.0^{o} centered about its experimentally-determined ϕ,ψ-value: α_6-κ-$\alpha_{5.}$ With a radial scaling factor of 0.95, all but three helices were found to be sterically incompatible with β values for the central residue. The β region was largely, but not entirely, excluded in these three exceptions as well (figure 2.5); in each case, the ϕ,ψ-values of flanking residues were well outside the high-confidence α-region (figure 2.1B), sometimes extending into 3_{10} helix.

In all, these results culminate in a prediction that a β-residue cannot follow three or more consecutive α-residues, a testable hypothesis using the PDB. In the list of 5378 chains with sequence identity of 90% or less, a series of three or more α-residues was found 19,062 times; none was followed by a β-residue. However, as mentioned earlier, seven exceptions were found when α' is used instead of α (table 2.3). For two of these structures, a small but real overlap is indicated between carbonyl oxygens as assessed by either our unscaled radii or contact dots (Word et al. 1999). In a third case, backbone clash is avoided by an unusual progression of ω-torsions. The final four cases may be legitimate, albeit marginal, exceptions. One of the four cases involves a single β-residue that follows the α'-residues; the other three cases involve type III turns and do not represent an intermixing of helix and strand.

In sum, a direct transition from canonical α-helix to β-strand is disallowed: a single β-residue adjoined to a helical peptide results in a steric collision (figure 2.6). Further, this collision only affects residues N-terminal to the β-residue. Therefore, an N-terminal to C-terminal transition from helix to strand must pass through at least one “buffer” residue from the turn region. This finding rationalizes the familiar observation that many α-helices terminate in a 3_{10} helix (Richardson 1981), a progression that both satisfies helix capping requirements (Aurora and Rose 1998) and facilitates the transition from helix to strand, turn, or loop.

2.5 Discussion

More than four decades ago, Sasisekharan, Ramakrishnan and Ramachandran (Ramachandran et al. 1963; Ramachandran and Sasisekharan 1968) elucidated the steric map for an alanyl dipeptide (more precisely, the compound Cα-CO-NH–CαHR–CO-NH-Cα, which, has two degrees of backbone freedom like a dipeptide). Similar ideas about the importance of sterics as an organizing force in proteins were also implicit in space-filling models (Koltun 1965), developed during this same era. Such ideas have been validated repeatedly in proteins (Berman et al. 2000) and are now invoked routinely when assessing the quality of experimentally determined structures (Laskowski et al. 1993b).

Today, the restrictions that sterics impose on the conformation of a dipeptide are widely accepted. Yet, hard sphere models have played a comparatively small role in both protein structure prediction and analysis of the unfolded state. Why?

The perceived problem is one of scale. If each ϕ,ψ-pair is independent of its neighbors (Flory 1969), then conformational space grows exponentially, despite dipeptide restrictions. Accordingly, the conformations accessible to a peptide backbone – even a short one – can quickly overwhelm constraints imposed by dipeptide sterics. This view is often invoked by alluding to the "Levinthal paradox" (Levinthal 1969): how does a protein find its unique native conformation among the more-than-astronomical number of conformational possibilities? For Levinthal, this conundrum was a demonstration, not a paradox, indicating that additional conformational constraints must exist. But what additional constraints might have been overlooked in this well-cultivated field?

Earlier work using explicit counting showed that the size of conformational space is smaller than previously believed (Pappu et al. 2000) because local steric interactions exert influence beyond the dipeptide, winnowing the number of accessible conformations. Here, we focused specifically on steric restrictions in the α-helix.

Unfolded Proteins

It has been proposed that the *coil library* – defined as the set of all non-helical, non-strand structures in the PDB – can be used to model the unfolded state of proteins (Swindells et al. 1995; Avbelj and Baldwin 2003). Therefore, our steric rules, which were validated against the PDB, including the coil library, would also hold for this model of the unfolded state. Plausibly so, because van der Waals repulsive forces will be unaffected by whether or not the protein is folded or unfolded.

Repulsive forces can have an organizing influence on the folding reaction, N(ative) $\rightleftarrows$ U(nfolded), and related order-disorder transitions, such as helix-coil theory (Zimm and Bragg 1959; Van Holde et al. 1998), where each residue is characterized by initiation and propagation constants. For example, any "coil" conformation with ϕ,ψ-angles in the β-region would exert a cooperative influence on the helix-coil equilibrium, making it harder to initiate a helix from the coil state by constricting the size of conformational space accessible to a residue that follows a helix nucleation site. Conversely, once nucleated it would also be harder to melt a helix because the helical conformer would inhibit introduction of a central coil residue with ϕ,ψ-angles in either the β- or P_{II}-regions.

The Number of Protein Domains

Our analysis of short polyalanyl chains demonstrates that a β-conformer cannot be introduced into an α-helix without an accompanying steric clash. This restriction maintains the structural homogeneity of α-helices by excluding heterogeneous conformers consisting of a turn of α-helix followed by one or more β-residues. Exclusion of folds in which there is an immediate transition from helix to strand eliminates many conceivable protein domains.

In particular, when a protein folds, backbone polar groups removed from solvent will participate in compensatory intramolecular hydrogen bonds. To do so, they form segments of α-helix or strands of β-sheet, the only regular, repeating hydrogen-bonded protein structures that are sterically available (Aurora et al. 1997). Proteins are largely supramolecular complexes of helices and strands (Levitt and Chothia 1976), and their intramolecular recognition and self-assembly is facilitated by the sterically-imposed elimination of α-β hybrids.

2.6 Acknowledgments

We thank Rajgopal Srinivasan, Rohit Pappu, Teresa Przytycka, Patrick Fleming, Nicholas Panasik, Timothy Street, Haipeng Gong and Robert L. Baldwin for formative discussion and Ross Shiman for his critical reading of this manuscript. Support from the NIH and Mathers Foundation is gratefully acknowledged.

Table 2.1: Hard Sphere Radii Used in Simulations

Atom Type	Radius (Å)[†]	Scaled Radius (Å)[‡]
Carbon	1.75	1.66
Carbonyl Carbon	1.65	1.57
Nitrogen	1.55	1.47
Oxygen	1.40	1.33

[†] Atomic radii taken from (Word et al. 1999)

[‡] Radii shown use the scaling factor of 0.95

Table 2.2: Twelve-Residue Helices used in Testing Robustness

PDB ID*	Res (Å)	R factor	Residue Start	Residue End	PDB ID	Res (Å)	R factor	Residue Start	Residue End
1AIHA	2.5	0.21	176	187	1HBKA	2.0	0.20	51	62
1ALN	2.3	0.19	14	25	1HNNB	2.4	0.23	604	615
1B16A	1.4	0.18	109	120	1HQ6B	2.7	0.25	123	134
1CXQA	1.0	0.13	185	196	1IXH	1.0	0.12	298	309
1D2HA	3.0	0.20	247	258	1J8YF	2.0	0.23	4	15
1D6JA	2.0	0.21	110	121	1J9LA*	1.9	0.20	14	25
1DI2B	1.9	0.23	113	124	1JD22	3.0	0.25	167	178
1DSZA*	1.7	0.20	1153	1164	1JK7A	1.9	0.20	146	157
1EG9B	1.6	0.19	592	603	1JMVA	1.9	0.22	64	75
1EJ0A	1.5	0.20	34	45	1JN0A	3.0	0.21	252	263
1EJ3A	2.3	0.22	162	173	1KPGA	2.0	0.19	185	196
1EXJB	3.0	0.24	51	62	1MUN	1.2	0.12	30	41
1F0JB*	1.8	0.20	191	202	1POC	2.0	0.19	61	72
1F0JB	1.8	0.20	261	272	1POC	2.0	0.19	77	88
1F4LA	1.9	0.18	536	547	1QTWA	1.0	0.12	268	279
1FSGA	1.1	0.00	153	164	1XRC*	3.0	0.20	64	75
1FUIA	2.5	0.16	65	76	1YGE	1.4	0.20	159	170
1G9ZA	1.8	0.20	99	110	2ACY	1.8	0.17	22	33
1GAL	2.3	0.18	29	40	4HB1	2.9	0.23	11	22
1GRCB	3.0	0.19	12	23	8OHM	2.3	0.23	506	517

* The set of 40 12-residue helical segments simulated using experimentally determined ϕ,ψ-values (see text). Table entries marked with a star represent those in which the β-region was not excluded completely; even in these three cases, the distribution maintains a distinct Y-shape (see figure 5). The chain identifier is listed as the fifth character of the PDB ID.

Table 2.3: Violations of Relaxed Helical Complementarity in the PDB

PDB ID*	Res (Å)	R factor	Residue Start	Residue End	O-O Dist (Å)‡	Explanation
1CERO	2.5	0.20	43	46	2.8	Collision[1]
1PHK	2.2	0.21	197	200	2.9	Tight Packing[2]
1QCIA	2.0	0.23	177	180	2.8	Collision[1]
1RCD	2.0	0.19	128	131	3.1	Tight Packing[2]
1DSSG	1.9	0.17	43	46	3.1	Omega Angles[3]
1IFT	1.8	0.22	178	181	3.1	Tight Packing[2]
1HFUA	1.7	0.18	474	477	3.1	Tight Packing[2]

* The chain identifier is listed as the fifth character of the PDB ID.

‡ Distance between the carbonyl oxygens of the first and last residues.

[1] A collision between carbonyl oxygens is observed using unscaled radii (table 2.1).

[2] No collision observed; but packing is tight and perturbation of any torsion angle would lead to a collision.

[3] Violation occurs because of deviations from planarity in ω-torsions; deviations were greater than 5.0° for all four residues.

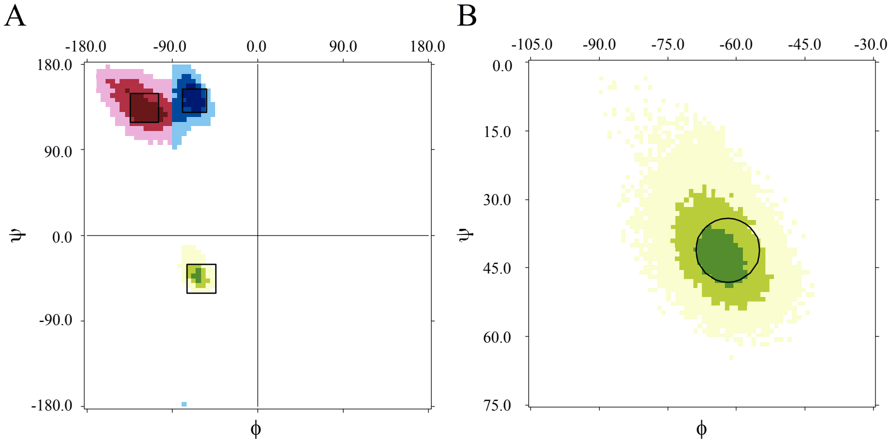


Figure 2.1. (A) The distribution of secondary structure from the PDB using 5^{o} by 5^{o} bins, as described in Methods. Color coding: β-strand: red, polyproline II: blue, and α-helix: green. The regions β, P_{II}, and α', defined in the text, are shown as black boxes embedded in the colored regions. (B) Smaller 1^{o} by 1^{o} bins were used to determine the size and location of the α region, shown as a black circle overlaid on the PDB distribution, in yellow.

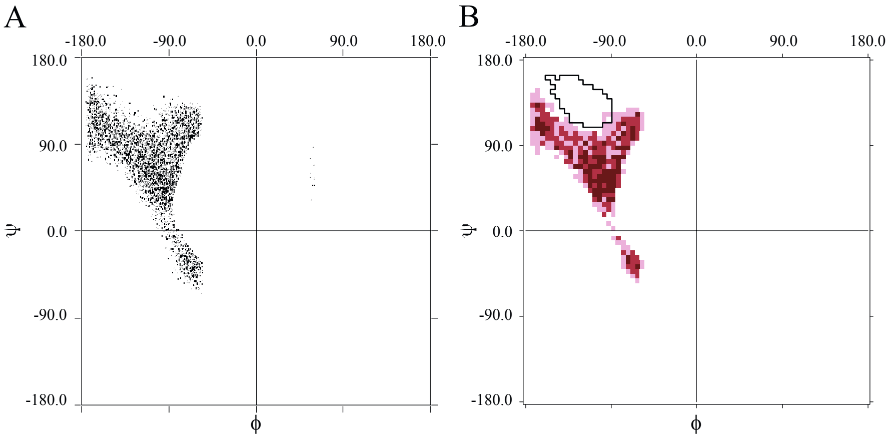


Figure 2.2. The φ,ψ-distribution in polyalanine for a central κ residue flanked on either side by four consecutive α residues. (A) Raw φ,ψ-values. Each point represents a sterically allowed structure when all residues were assigned random values of φ and ψ. (B) Same data as A, but grouped into 5° by 5° bins as described in Methods. Sterically disfavored regions fall outside the 90% boundaries; the most favorable regions are the most intensely colored. The 66% contour line of the observed β-strand distribution (from fig. 2.1A) is shown in black outline.

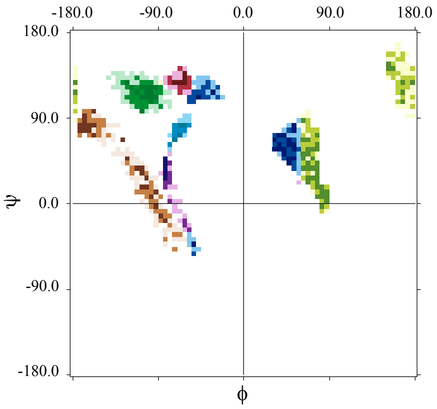


Figure 2.3. Collision map for a single κ residue in a sequence of α residues. Atom collisions responsible for the Y shape in figure 2.2 are color coded: $O_{i\text{-}3} – O_i$: green, $O_{i\text{-}3}$ – CB_{i+3}: red, $O_{i\text{-}1}$ – C_i: blue, $O_{i\text{-}3}$ – O_{i+1}: brown, $C_{i\text{-}1}$ – N_{i+1}: purple, $C_{i\text{-}1}$ – N_{i+1}: cyan, $O_{i\text{-}1}$ – CB_i: yellow. The most conspicuous collision, between O_i and $O_{i\text{-}3}$, is responsible for the void in the strand region, in figure 2.2.

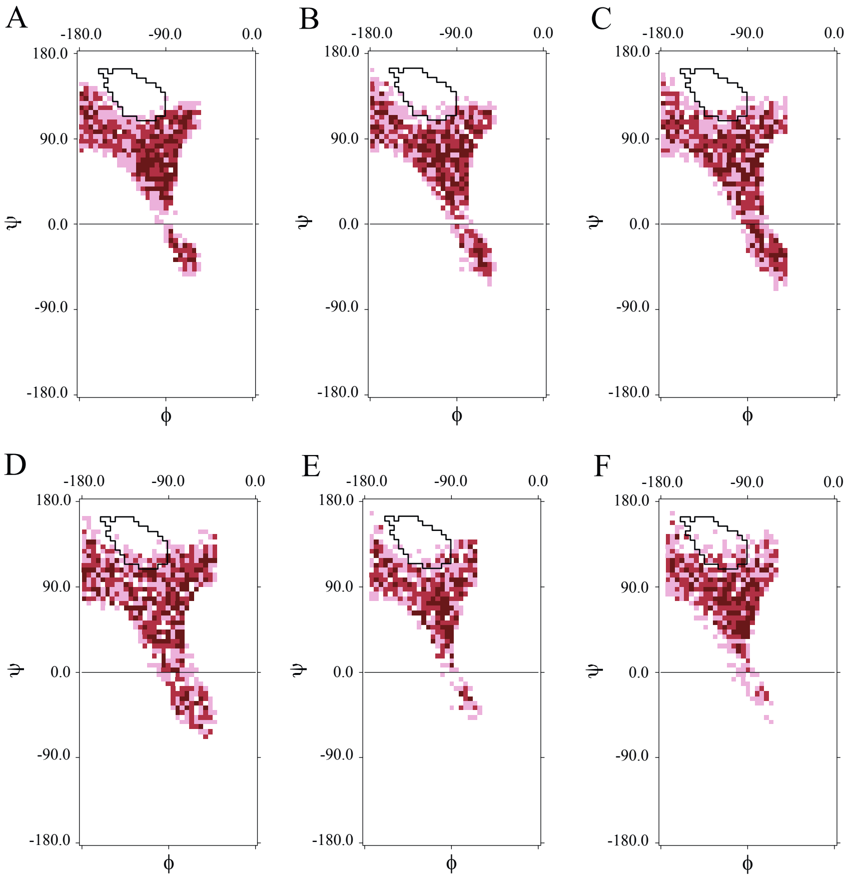


Figure 2.4. (A-D) Reducing the hard sphere scaling factor. Same experiment as figure 2.2B, but with scaling factors of 0.92, 0.90, 0.88, and 0.85, respectively. Even at the extreme of 0.85, a remnant of the original Y shape survives. (E and F) Relaxing the definition of α. A radius of 14° around ϕ = -63.0° and ψ = -45.0° (E) and 21° (F). On all plots, the 66% contour line from the observed β-strand distribution (in figure 2.1A) is overlaid in black.

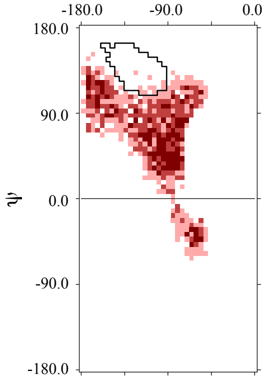


Figure 2.5. One of three extreme examples from the set of 40 12-residue helical segments (Table 2, PDB entry 1J9L, chain A, residues 14–25) in which simulations used experimentally determined ϕ's and ψ's to define α (see text). The resulting distribution still maintains a Y shape, but a slight overlap with the 66% β-strand contour is evident.

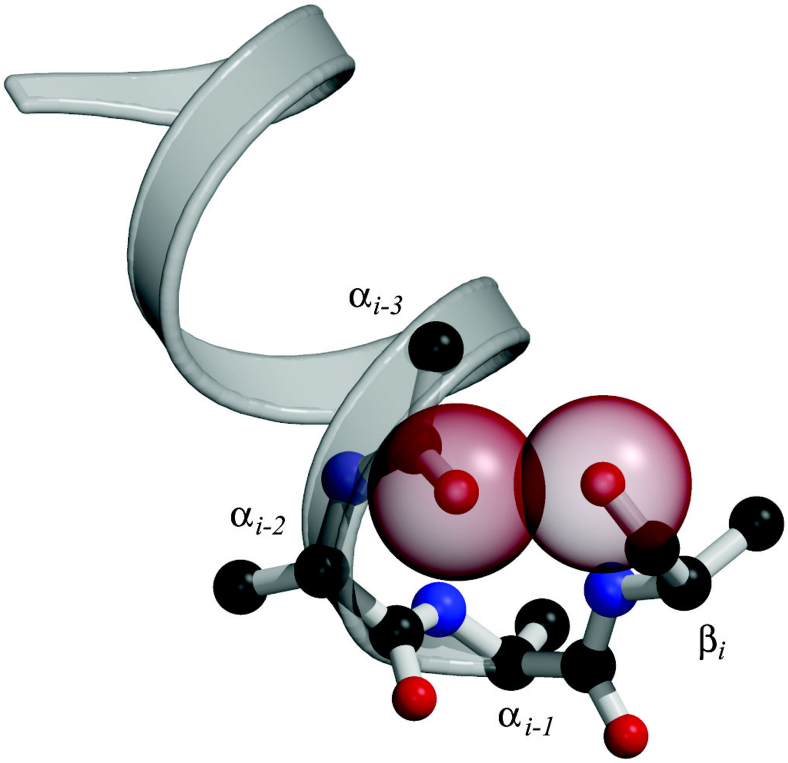


Figure 2.6. A β-residue cannot be added to three or more residues of α-helix without encountering a steric clash. Ball-and-stick backbone atoms for three residues of an α-helix (α_{i-3} - α_{i-1}) are shown superimposed on a longer helical ribbon, followed by a single β-residue (β_i). This conformation forces a substantial overlap between O_i and O_{i-3}, shown here as transparent van der Waals spheres. Atoms are rendered using conventional CPK colors, i.e. carbon:black, nitrogen: blue and oxygen:red.

CHAPTER 3

Reassessing Random Coil Statistics in Unfolded Proteins[*]

3.1 Abstract

The Gaussian-distributed random-coil has been the dominant model for denatured proteins since the 1950s, and it has long been interpreted to mean that proteins are featureless, statistical coils in 6M guanidinium chloride (GmHCl). Here, we demonstrate that random-coil statistics are not a unique signature of featureless polymers. The random-coil model does predict the experimentally determined coil dimensions of denatured proteins successfully. Yet, other equally convincing experiments have shown that denatured proteins are biased toward specific conformations, in apparent conflict with the random-coil model. We seek to resolve this paradox by introducing a contrived counterexample in which largely native protein ensembles nevertheless exhibit random-coil characteristics. Specifically, proteins of known structure were used to generate disordered conformers by varying backbone torsion angles at random for ~8% of the residues; the remaining ~92% of the residues remained fixed in their native conformation. Ensembles of these disordered structures were generated for 33 proteins using a torsion angle Monte Carlo algorithm with hard sphere sterics; bulk statistics were then calculated for each ensemble. Despite this extreme degree of imposed internal

[*] This chapter is reprinted with permission from the authors as allowed by the copyright agreement. Fitzkee, N. C. and Rose, G. D. (2004) *Proceedings of the National Academy of Sciences USA.* **101** (34): 12497-502. Copyright © 2004 The National Academy of Sciences of the USA.

structure, these ensembles have end-to-end distances and mean radii of gyration that agree well with random-coil expectations in all but two cases.

3.2 Introduction

The protein folding reaction, U(nfolded) ⇄ N(ative), is a reversible disorder ⇄ order transition. Typically, proteins are disordered (U) at high temperature, high pressure, extremes of pH, or in the presence of denaturing solvents, but they fold to uniquely ordered, biologically relevant conformers (N) under physiological conditions. With some exceptions (Dunker et al. 2001), the folded state is the biologically relevant form, and it can be characterized to atomic detail using X-ray crystallography and nuclear magnetic resonance spectroscopy (NMR). In contrast, our understanding of the unfolded state is based primarily on a statistical model – the random-coil model – which was developed largely by Flory (Flory 1969) and corroborated by Tanford (Tanford 1968) in the 1950s and 1960s.

In a *random-coil,* the energy differences among sterically accessible backbone conformers are of order ~kT. Consequently, there are no strongly preferred conformations, the energy landscape is essentially featureless, and a Boltzmann-weighted ensemble of such polymers would populate this landscape uniformly.

Our motivation here is to dispel the belief – widespread among protein chemists – that the presence of random coil statistics for denatured proteins confirms the absence of residual structure in these molecules. Indeed, it is well known to polymer chemists that rods of any stiffness – *e.g.*, steel I-beams – behave as Gaussian-distributed, temperature-dependent random coils if they are long enough. Chains in which the persistence length exceeds one physical link can be treated effectively by rewriting them as polymers of Kuhn segments (Flory 1969, pg. 12). Consequently, a protein chain can behave as a random coil even if it is comprised of non-random segments.

A denatured protein is a heteropolymer in which different amino acid residues will have differing average conformations, but in which an average backbone conformation is attained within a window of approximately ten residues. For such a heteropolymer, coil dimensions can be assessed using two related measures: the radius of gyration and the end-to-end distance. Flory showed (de Gennes 1979, pg. 43) that the radius of gyration, R_G, follows a simple scaling law:

$$R_G = R_0 N^{\nu} \tag{3.1}$$

where N is the number of residues, R_0 is a constant related to persistence length, and ν is the scaling factor of interest that depends on solvent quality. Values of ν range from 0.33 for a collapsed, spherical molecule in poor solvent through 0.5 for an ideal solvent to 0.6 in good solvent. The mean-squared end-to-end distance, $<L^2>$, for unfolded proteins is also expected to scale linearly with chain length:

$$\left\langle L^2 \right\rangle = L_0 N \tag{3.2}$$

with the L_0 prefactor obtained from experiment.

Tanford and coworkers (Tanford et al. 1966) corroborated these random-coil expectations for unfolded proteins using intrinsic viscosity measurements, which scale with chain length in a conformation-dependent way. From this relationship, they obtained values of $\nu = 0.67$ and $L_0 = 70 \pm 15$ Å^2. To a good approximation, end-to-end distances for random coils of sufficient length are Gaussian distributed (Chan and Dill

1991), and, in fact, this behavior has been observed in recent simulations (Goldenberg 2003).

Tanford emphasized that such measurements are meaningful only after eliminating all residual structure, requiring denaturation in 6M GmHCl (Aune et al. 1967). This is a crucial issue. Structure induced by peptide hydrogen bonds is abolished only under strongly denaturing conditions. As pointed out by Millet *et al.*, "Additional evidence that chemically or thermally denaturing conditions are typically good solvents for the unfolded state stems from the observation that R_G is generally fixed over a broad range of temperatures or denaturant conditions" (Millett et al. 2002, pg. 255 and ensuing discussion).

Today, the most reliable experimental values of R_0 and ν in equation 3.1 are obtained from small angle X-ray scattering (SAXS) (Millett et al. 2002). Using this approach for a series of 25 unfolded proteins, values of $R_0 = 2.08 \pm 0.19$ Å and $\nu = 0.581 \pm 0.017$ were obtained (Kohn et al. 2004). These results are a strong indicator of random-coil behavior. Additionally, SAXS data can be used to construct a Kratky plot, s versus $s^2 I(s)$, where s is the small angle scattering vector and $I(s)$ is the corresponding scattering intensity (Semisotnov et al. 1996; Doniach 2001). For random coils, the plot increases monotonically and approaches linearity in s (Pilz et al. 1979). This is the behavior observed for unfolded proteins, whereas folded proteins plotted in this way exhibit a notable maximum (figure 1 in (Millett et al. 2002)). Such plots have become the present-day standard for assessing random-coil behavior in unfolded proteins (Semisotnov et al. 1996; Doniach 2001).

The success of the random-coil model in fitting experimentally determined coil dimensions of unfolded proteins is undisputed. Accordingly, the field has grown accustomed to believing that unfolded proteins are featureless random coils. Here, we demonstrate that non-random coils can also exhibit random-coil statistics.

Tanford knew that denatured proteins need not be entirely random simply because they satisfy random-coil statistics, and he warned:

> "A cautionary word is in order regarding the use of the measurement of the radius of gyration of a particular protein as the sole criterion for random coil behavior. Other conformations can have similar radii of gyration. For example, an α-helical rod has a length of 1.50 Å per residue... There is a narrow range of N where essentially identical values of R_G are predicted for α-helices and random coils."
>
> (Tanford 1968)

In this paper, we introduce the *rigid-segment model*, a highly contrived, limiting model in which known protein structures are partitioned alternately into rigid segments linked by individual flexible residues. X-ray elucidated coordinates are retained for the rigid segments, but backbone torsions angles were allowed to vary freely for the flexible residues. The fraction of the chain allowed to vary, ~8%, was chosen to approximate one residue per peptide chain turn (Rose and Wetlaufer 1977). If this physically-unrealistic, extreme model still exhibits random-coil statistics, it follows that a lesser degree of pre-organization in the unfolded state need not violate random-coil expectations. In fact, we find that our limiting model still reproduces random-coil statistics when ~92% of the structure is held rigidly in its native conformation.

3.3 The Rigid-Segment Model

Our strategy is to devise an algorithm that operates on native protein structures and generates ensembles of highly structured, sterically allowed conformers. We then test these ensembles and determine the extent to which they exhibit random-coil statistics. A largely native ensemble that nevertheless appears random serves as a counterexample to the random-coil model.

The algorithm consists of several steps. First, each residue is examined in turn, and those with the maximum possible flexibility are identified. Flexibility is measured by evaluating the range of sterically allowed backbone torsion angles for each residue; the broader the range, the greater the flexibility. Next, using a biochemically-motivated rationale, a subset of these flexible residues is selected as links, transforming the polypeptide chain into rigid segments interconnected by flexible links. The links are then varied at random in concerted fashion to generate clash-free ensembles suitable for statistical analysis. These steps are now described in detail.

Identifying individual flexible residues

The first step quantifies the backbone flexibility of individual residues. For each residue, sterically allowed ϕ,ψ-space (Ramachandran et al. 1963) was explored using torsion angle Monte Carlo sampling with hard sphere sterics, with the acceptance ratio taken as the measure of flexibility. Steric clashes were evaluated in a window of 15 residues flanking the residue in question (but with diminishing window size nearing chain termini). A half-window of 15 residues was chosen to approximate the average size of a protein secondary structure element together with its adjoining turn (Rose and Wetlaufer 1977).

To construct a flexibility profile of acceptance ratio versus residue number, 10,000 backbone ϕ,ψ-pairs were sampled for each residue, as illustrated for lysozyme in figure 3.1. Generally, though not invariably, the most flexible residues correspond to turns; glycines also promote chain flexibility.

Selecting sets of flexible residues

Individual acceptance ratios were ranked by flexibility, and a set of suitable size was chosen based on the average length of a protein α-helix: 12 residues (Presta and Rose 1988). Accordingly, a flexible residue set, $\Re$, of size m = N/12 residues was chosen, having one flexible linker for every 12 residues in the protein. The value of m was rounded to the nearest integer, with a minimum value of one.

The most flexible residues were chosen for inclusion in $\Re$, with two minor qualifications: sites were chosen so as to be at least five residues apart, and those within five residues of chain termini were not included. These qualifications promote a uniform distribution of flexible links along the polypeptide chain and ensure that the chosen backbone torsion angles are independent of one another (Ohkubo and Brooks 2003).

An ensemble of structures was generated for each protein by concerted sampling of backbone torsions, chosen at random from all sterically allowed regions of ϕ,ψ-space. Random-coil statistical measures were then used to characterize this ensemble. Details are described next.

3.4 Materials and Methods

Thirty-three proteins of size 8 to 415 residues were selected from the protein data bank (Berman et al. 2000) based on structure quality, scientific interest, and size distribution (Table 3.2). Where possible, proteins studied previously by SAXS were included. All crystallographic waters, heteroatoms and non-biological chain terminators (Acetyl groups, N-Methylamide, etc.) were removed, and any disulfide bonds were broken.

Hard-sphere, torsion angle Monte Carlo simulations (Metropolis et al. 1953) were performed using a suite of freely available programs (http://roselab.jhu.edu/dist/index.html). Default van der Waals radii (Srinivasan and Rose 2002a) were used unless the experimentally reported distance between two atoms was smaller than the sum of their hard sphere radii, in which case the minimum inter-atomic distance was taken from PDB coordinates. At each Monte Carlo step, random values of backbone torsions, chosen from allowed regions on the dipeptide map, were assigned in concert to residues in $\Re$. In the event of a steric clash, the step was rejected.

Statistics of interest for each ensemble include the average radius of gyration and end-to-end distance. The geometric radius of gyration for a chain is given by:

$$R_G = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(\vec{r}_i - \vec{r}_c\right)^2} \qquad (3.3)$$

where M is the number of atoms in the protein structure, $\vec{r}_i$ is the position of atom i in three-dimensional space, and $\vec{r}_C$ is the geometric center of the molecule. Weighting by mass or atomic scattering factor does not change the radius of gyration significantly, and therefore the ensemble-averaged radius of gyration was computed simply by averaging R_G over all chains in the ensemble.

The mean squared end-to-end distance, $<L^2>$, is given by:

$$\left\langle L^2 \right\rangle = \frac{1}{n}\sum_{j=1}^{n} L_j^2 \tag{3.4}$$

where n is the number of conformers in the ensemble, and L_j is the end-to-end distance of conformer j, taken from the amino-terminal nitrogen to the carboxy-terminal oxygen. End-to-end distance histograms were generated using the R statistics package (R Development Core Team 2003).

For each protein in the dataset, an ensemble of at least 1,000 clash-free conformers was generated as described above, with flexible residues selected from the corresponding flexibility profile (*e.g.,* figure 3.1). This process was repeated five times. To assure convergence, standard deviations for both R_G and $<L^2>$ were calculated. As a further test, ensembles of 10,000 structures and 500 structures were examined; all have similar statistics.

The program CRYSOL (Svergun et al. 1995) was used to generate simulated SAXS scattering profiles for every conformer in each ensemble. In CRYSOL, the scattering vector s is defined as:

$$s = 4\pi\frac{\sin\theta}{\lambda} \tag{3.5}$$

where θ is the scattering angle and λ is the X-ray wavelength (in Ångstroms). Default options were used for all values. Scattering profiles of all conformers were averaged at every point, and errors were reckoned as the standard deviation of *I(s)* for that point over the entire ensemble. Simulated Kratky plots were produced by plotting *s* against $s^2 I(s)$ for every point.

3.5 Results

Detailed results for lysozyme (1HEL) using the rigid-segment model are described as an illustrative example. Almost all flexible residues are situated in turn and coil regions (figure 3.1), as identified from backbone torsion angles (Srinivasan and Rose 1999). The set of flexible linker residues, $\mathfrak{R}$, selected by our algorithm is (see Table 3.2):

$$\{16, 22, 41, 47, 55, 71, 79, 86, 102, 117, 123\}$$

and the resultant ensemble of segmentally-rigid chains was found to be consistent with random-coil expectations. In particular, the value of R_G for denatured lysozyme predicted by equation (3.1) is 35.0 ± 4.3 Å, and the average R_G from five rigid-segment simulations is 37.93 ± 0.14 Å, in good agreement. The experimentally determined R_G for trifluoroethanol (TFE)-denatured lysozyme is 35.8 ± 0.5 Å (Hoshino et al. 1997); this value may be especially relevant for comparison with the rigid-segmental model because TFE stabilizes helical segments (Nelson and Kallenbach 1986). Similarly, the value $<L^2>$ for denatured lysozyme predicted by equation (3.2) lies between 7,095 Å^2 and 10,965 Å^2, and $<L^2>$ from rigid-segment ensembles is 10,690 ± 160 Å^2, near the high end of the predicted Gaussian distribution (figure 3.2). Thus, highly structured lysozyme chains (figure 3.3), generated using the rigid-segment model, exhibit random-coil statistics.

The rigid-segment model was applied to 33 proteins in all, as summarized in table 3.3. In general, values of both R_G and $<L^2>$ are consistent with random-coil expectations, and histograms of the end-to-end distances fit well to a Gaussian curve with two exceptions: Angiotensin II (1N9V, 8 residues) and PKC δ-Cys2 Domain (1PTQ, 50 residues). Both outliers are small and deviate from the normal distribution expected for longer chains (more than ~100 residues), consistent with the systematic deviations from equations (3.1) and (3.2) that Tanford noted for short chains (Tanford 1968 figure 2; Cantor and Schimmel 1980, pg. 994). However, two other small proteins in our dataset (*e.g.*, 1VII, 36 residues; 2GB1, 56 residues) behave as expected for longer chains. The rigid-segment model, which tends to localize chain flexibility at peptide chain turns, is expected to be sensitive to differences in the average segment length between consecutive turns. This expectation is borne out: in comparison to the values predicted by equation (3.1), the rigid-segment model under-estimates R_G for α-helical proteins (1VII, 1LMB, 1HRC, 2HMQ, 1CM1, 1MBO and 1MUN) but over-estimates R_G for β-sheet proteins (1SHF, 1CSP, 2PCY and 1IFB), as shown in table 3.1.

Among the R_G's, one outlier warrants particular comment. The value of R_G for creatine kinase (1QK1) from rigid-segment calculations is 79.812 ± 0.078 Å, but the corresponding value predicted by equation (3.1) is only 66.5 ± 8.9 Å. It is noteworthy that both values substantially exceed the actual, experimentally determined value of 46.1 ± 1.5 Å observed using SAXS. We find no explanation for this anomalous behavior.

Data from all 33 proteins were fit to equations (3.1) and (3.2) and are displayed in figure 3.4. A nonlinear least squares best fit (R Development Core Team 2003) to equation (3.1) gives $R_0 = 1.98 \pm 0.37$ Å and $\nu = 0.602 \pm 0.035$, which are indistinguishable from recent experimentally-determined values (Kohn et al. 2004). The corresponding fit to equation (3.2) gives $L_0 = 81.8 \pm 3.4$ Å^2, similar to Tanford's value of

L_0 = 70 ± 15 Å^2 (Tanford et al. 1966). The standard deviations reported here for R_G and $<L^2>$ represent a convergence criterion, not the actual uncertainties of those values, and weights were not used during the fits.

Values of R_G derived from the rigid-segment and random-coil models are strongly correlated (r^2 = 0.916, figure 3.5). In all, characteristic statistics for the random-coil model resemble those for the rigid-segment model, despite the fact that in the latter, 92% of each chain is fixed in its native conformation.

SAXS and Kratky Plots

SAXS profiles monitor the correlation among inter-atomic distances. In our simulations, inter-atomic distances do not vary within each rigid segment, so it is conceivable that a segmentally rigid ensemble could have random-coil values of $\mathrm{R_G}$ and $<L^2>$ but yet appear structured in a Kratky plot. To test this possibility, a Kratky plot was calculated for random chains from the lysozyme ensemble (figure 3.6A). Although the simulated plot has a maximum at 0.275 Å^{-1}, it lacks the pronounced hump typical of Kratky plots for native proteins. A second test shows that side chain rigidity is a major factor contributing to this maximum. After removal of side chain atoms beyond $\mathrm{C_\beta}$, the corresponding plot now resembles that of a denatured protein (figure 3.6B).

3.6 Discussion

The random-coil model has a long and impressive record of successfully predicting the chain dimensions of denatured proteins (Tanford 1968; Millett et al. 2002; Kohn et al. 2004). However, two recent lines of evidence suggest that denatured protein chains may be far from random. First, experiments have identified native-like organization in unfolded proteins. Using residual dipolar couplings (RDCs) from NMR, Shortle and Ackerman showed that native-like topology persists under strongly denaturing conditions in a truncated staphylococcal nuclease (Shortle and Ackerman 2001). Contention about the origin of RDCs in unfolded proteins notwithstanding (Louhivuori et al. 2003), other NMR methods also detect structure in the unfolded state. Using triple-resonance NMR, native-like topology has been observed in protein L (Yi et al. 2000). A second line of evidence suggests that unfolded proteins are conformationally biased toward polyproline II (P_{II}) helical conformations. Both theory (Mu and Stock 2002; Pappu and Rose 2002; Avbelj and Baldwin 2004; Drozdov et al. 2004; Garcia 2004; Kentsis et al. 2004; Mezei et al. 2004; Vila et al. 2004) and experiment (Tiffany and Krimm 1968a; Woutersen and Hamm 2000; Rucker and Creamer 2002; Shi et al. 2002a; Ferreon and Hilser 2003) have investigated the preference for P_{II} in unfolded peptide ensembles. If the experimental results are correct and the ensemble is not random, why is the random-coil model so successful? This paradox has been dubbed *the reconciliation problem* by Plaxco and co-workers (Millett et al. 2002).

Our contrived counterexample was designed to address the reconciliation problem directly. Indeed, we find that the random-coil model is insensitive to a preponderance of stiff segments in an otherwise flexible chain.

In our simulations, chains of interest are comprised of rigid segments of native protein structure interconnected by flexible hinge residues. This approach is deliberately extreme in its neglect of physical reality, and we emphasize that *it is not intended as a model of the unfolded state*. With the exception of steric repulsion, all interatomic forces and temperature-dependent effects are ignored, together with resultant structural fluctuations. Yet, this physically absurd model – in which 92% of the native structure is retained – successfully reproduces random-coil statistics for R_G and $<L^2>$ in good solvent (*e.g.,* 6M GmHCl). Therefore, it is none too surprising that transient organization in denatured proteins could also give rise to the random-coil statistics observed in experiment (Kohn et al. 2004).

The presence of pre-organization in denatured proteins changes our perspective about the disorder $\rightleftarrows$ order transition that occurs during protein folding. In the prevailing view, denatured proteins are random coils, lacking in correlations beyond nearest chain neighbors. If so, there is a puzzling, time-dependent search problem as unfolded polypeptide chains negotiate self-avoiding Brownian excursions through this featureless landscape en route to their native conformation (Levinthal 1969). Concepts like folding funnels, kinetic traps, and frustration all arose as attempts to rationalize this process (Dill 1999). However, such conundrums are eliminated by the presence of sufficient conformational bias in the unfolded state (Zwanzig et al. 1992; Srinivasan and Rose 2002b). In fact, significant conformational bias is inescapable, and it originates from sterically imposed chain organization that extends beyond nearest sequential neighbors, such as those discussed in chapter 2 (Pappu et al. 2000; Fitzkee and Rose 2004b).

The random coil model has been construed to imply that denatured proteins lack organization, an interpretation that has become a mainstay in protein folding studies. Against this backdrop, there was no motivation to seek out organizing steric interactions beyond the linked alanyl dipeptide (Ramachandran et al. 1963). Nonetheless, such interactions do exist (Fitzkee and Rose 2004b) and are easy to detect. Our rigid-segment counterexample was developed to challenge this conventional interpretation of the random-coil model and to remove a conceptual obstacle that has impeded alternative explanations.

3.7 Acknowledgments

We thank Kevin Plaxco for insights and unpublished data and Patrick Fleming, Rajgopal Srinivasan, Ross Shiman, Gary Pielak, Nicholas Panasik, Timothy Street and Haipeng Gong for many helpful discussions. Funding from the Mathers Foundation is gratefully acknowledged.

Table 3.1: Proteins Used in Rigid Segment Simulations

			Resolution	Refinement	Chain
Protein	**PDB ID**	**Chain**	**(Å)**	**Factor**	**Length**
Angiotensin II	1N9V	A	(NMR)	(NMR)	8
Chicken Villin Headpice	1VII		(NMR)	(NMR)	36
PKC delta Cys2 Domain	1PTQ		1.95	0.196	50
Protein G	2GB1		(NMR)	(NMR)	56
Fyn SH3	1SHF	A	1.90	0.180	59
CspB	1CSP		2.50	0.195	67
Ubiquitin	1UBQ		1.80	0.176	76
Lambda Repressor	1LMB	3	1.80	0.189	87
Barstar	1A19	A	2.76	0.203	89
CT Acylphosphatase (ctAcP)	2ACY		1.80	0.170	98
Plastocyanin	2PCY		1.80	0.160	99
Horse Cytochrome c	1HRC		1.90	0.179	104
pI3K SH2 (rat)	1FU6	A	(NMR)	(NMR)	111
Myohemerythrin	2HMQ	A	1.66	0.189	113
Bovine α-Lactalbumin	1F6S	A	2.20	0.216	122
Bovine Ribonuclease A	1XPT	A	1.90	0.162	124
CheY	1EHC		2.26	0.143	128
Lysozyme	1HEL		1.70	0.152	129
Intestinal FA Binding Protein	1IFB		1.96	0.188	131
Staphylococcal Nuclease	2SNS		1.50	N/A	141
Calmodulin	1CM1	A	2.00	0.234	143
Myoglobin	1MBO		1.50	0.159	153
Ribonuclease H	2RN2		1.48	0.196	155
ASV Integrase Core	1ASU		1.70	0.152	162
T4 Phage Lysozyme	2LZM		1.70	0.193	164
DHFR	1AI9	A	2.76	0.203	192
MutY Catalyic Domain	1MUN		1.20	N/A	225
Triosephosphate Isomerase	5TIM	A	1.83	0.183	249
Human Glyoxase II	1QH3	A	1.90	0.185	260
EcoRI Endonuclease	1ERI	A	2.70	0.170	261
UDP-Galactose 4-Epimerase	1NAH		1.80	0.165	338
Creatine Kinase	1QK1	A	2.70	0.195	379
Yeast PGK	3PGK		2.50	N/A	415

Table 3.2: Flexibility Set Selection in Lysozyme

Residue Number	SS Type[1]	Residue Type	Flexibility[2]	Included in Set?
102	C	GLY	0.694	Yes
16	C	GLY	0.645	Yes
126	T	GLY	0.640	No[†]
86	C	SER	0.635	Yes
71	T	GLY	0.630	Yes
129	C	LEU	0.592	No[†]
4	P	GLY	0.570	No[†]
22	T	GLY	0.546	Yes
117	T	GLY	0.542	Yes
128	P	ARG	0.375	No[‡]
47	T	THR	0.368	Yes
41	T	GLN	0.366	Yes
1	C	LYS	0.349	No[†]
127	P	CYS	0.327	No[†]
84	T	LEU	0.321	No[‡]
123	T	TRP	0.285	Yes
26	H	GLY	0.282	No[‡]
101	H	ASP	0.277	No[‡]
21	T	ARG	0.264	No[‡]
103	C	ASN	0.250	No[‡]
100	H	SER	0.207	No[‡]
79	P	PRO	0.196	Yes
55	T	ILE	0.191	Yes

[1] Secondary structure types were determined as in (Srinivasan and Rose 1999). C=coil, T=turn, P=polyproline II helix, and H=α-helix.

[2] Flexibility values, in rank order, correspond to those plotted in fig. 2.

[†] Not included owing to its proximity to the N- or C-terminus.

[‡] Not included owing to its proximity to a previously selected residue.

Table 3.3: Summary of Simulations and Comparison to the Random Coil Model and SAXS (Proteins 1-20)

			Radius of Gyration (Å)			**Mean-Squared End-to-End Distance (Å^2)**	
PDB ID	**Chain Length**	**Flexible Residues**	**SAXS**[1]	**Random Coil Model**[2]	**Segment Simulations**[3]	**Random Coil Model**[4]	**Segment Simulations**
1N9V	8	1	9.1±0.3	6.96±0.68	6.8790±0.0086	560±120	346.3±3.5
1VII	36	3		16.7±1.8	16.044±0.019	2,520±540	2,015±13
1PTQ	50	4		20.2±2.3	16.988±0.012	3,500±750	2,313±13
2GB1	56	5	23±1	21.6±2.5	25.396±0.039	3,920±840	5,407±57
1SHF	59	5		22.2±2.5	23.269±0.037	4,130±890	3,580±71
1CSP	67	6		23.9±2.8	29.047±0.066	4,700±1,000	4,261±77
1UBQ	76	6	25.2±0.2	25.8±3.0	25.176±0.048	5,300±1,100	4,290±120
1LMB	87	7		27.9±3.3	24.244±0.048	6,100±1,300	4,420±140
1A19	89	7		28.2±3.4	28.628±0.060	6,200±1,300	6,372±74
2ACY	98	8	30.5±0.4	29.9±3.6	34.945±0.095	6,900±1,500	7,430±270
2PCY	99	8		30.0±3.6	40.439±0.075	6,900±1,500	11,690±110
1HRC	104	9		30.9±3.7	28.06±0.10	7,300±1,600	5,200±180
1FU6	111	9	30.3±0.3	32.1±3.9	29.87±0.10	7,800±1,700	5,990±180
2HMQ	113	9		32.4±3.9	30.07±0.10	7,900±1,700	6,200±120
1F6S	122	10		33.9±4.2	36.04±0.17	8,500±1,800	8,650±240
1XPT	124	10	33.2±1.0	34.2±4.2	36.777±0.077	8,700±1,900	8,420±130
1EHC	128	11	38.0±1.0	34.9±4.3	36.613±0.049	9,000±1,900	8,270±200
1HEL	129	11	35.8±0.5	35.0±4.3	37.93±0.14	9,000±1,900	10,690±160
1IFB	131	11		35.3±4.4	47.61±0.15	9,200±2,000	15,260±370
2SNS	141	12	37.2±1.2	36.9±4.6	41.10±0.14	9,900±2,100	10,660±240

[1] SAXS data from Millett *et al.* (Millett et al. 2002) and Kohn *et al.* (Kohn et al. 2004).

[2] Random-coil radii of gyration calculated from equation 1 using constants from (Millett et al. 2002; Kohn et al. 2004). Error is calculated using standard propagation of error formulae.

[3] Segment simulation error was calculated as the error on the mean from five simulations.

[4] Random-coil mean-squared end-to-end distance values calculated from equation (3.2) (Tanford et al. 1966). Error is propagated from the initial constant.

Table 3.3: Summary of Simulations and Comparison to the Random Coil Model and SAXS (Proteins 21-33)

			Radius of Gyration (Å)			**Mean-Squared End-to-End Distance (Å^2)**	
PDB ID	**Chain Length**	**Flexible Residues**	**SAXS[1]**	**Random Coil Model[2]**	**PDB ID**	**Chain Length**	**Flexible Residues**
1CM1	143	12		37.2±4.6	33.76±0.25	10,000±2,100	7,920±320
1MBO	153	13	40±2	38.7±4.8	40.084±0.083	10,700±2,300	13,140±270
2RN2	155	13		39.0±4.9	39.50±0.21	10,900±2,300	11,850±200
1ASU	162	14		40.0±5.0	42.94±0.19	11,300±2,400	11,160±320
2LZM	164	14		40.3±5.1	36.83±0.19	11,500±2,500	9,730±300
1AI9	192	16	44±2	44.1±5.6	51.71±0.13	13,400±2,900	21,370±330
1MUN	225	19		48.4±6.3	47.12±0.21	15,800±3,400	16,200±710
5TIM	249	21		51.3±6.7	49.88±0.24	17,400±3,700	15,910±340
1QH3	260	22		52.6±6.9	61.34±0.54	18,200±3,900	21,240±810
1ERI	261	22		52.7±6.9	62.78±0.10	18,300±3,900	24,900±1,100
1NAH	338	28		61.3±8.3	62.67±0.61	23,700±5,100	23,700±680
1QK1	379	32	46.1±1.5	65.5±8.9	79.812±0.078	26,500±5,700	43,500±2,400
3PGK	415	35	71±1	69.0±9.5	67.58±0.41	29,100±6,200	32,200±1,100

[1] SAXS data from Millett *et al.* (Millett et al. 2002) and Kohn *et al.* (Kohn et al. 2004).

[2] Random coil radii of gyration calculated from equation (3.1) using constants from (Millett et al. 2002; Kohn et al. 2004). Error is calculated using standard propagation of error formulae.

[3] Segment simulation error was calculated as the error on the mean from five simulations.

[4] Random-coil mean-squared end-to-end distance values calculated from equation (3.2) (Tanford et al. 1966). Error is propagated from the initial constant.

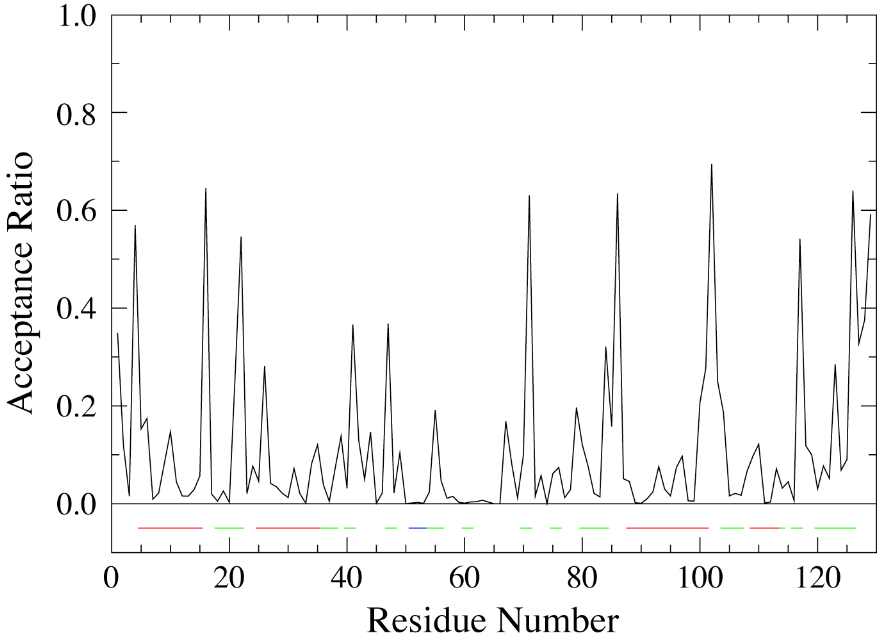


Figure 3.1. Flexibility profile for lysozyme (PDB 1HEL). Secondary structure is indicated by colored bars beneath the plot: red = α-helices, blue = β-strands and green = turns. Secondary structure determinations are based on backbone torsions, as described in (Srinivasan and Rose 1999).

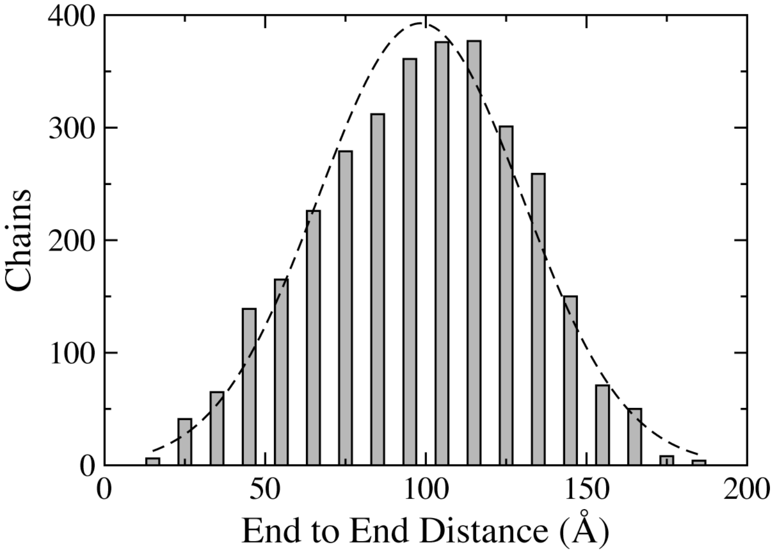


Figure 3.2. End-to-end distance histogram for lysozyme using 5000 chains generated from the rigid segment model. Chains were grouped into 10 Å-bins based on the distance from the N-terminal nitrogen to the C-terminal oxygen. For comparison, a Gaussian curve having the same mean and standard deviation as the actual distribution is also shown (dotted line).

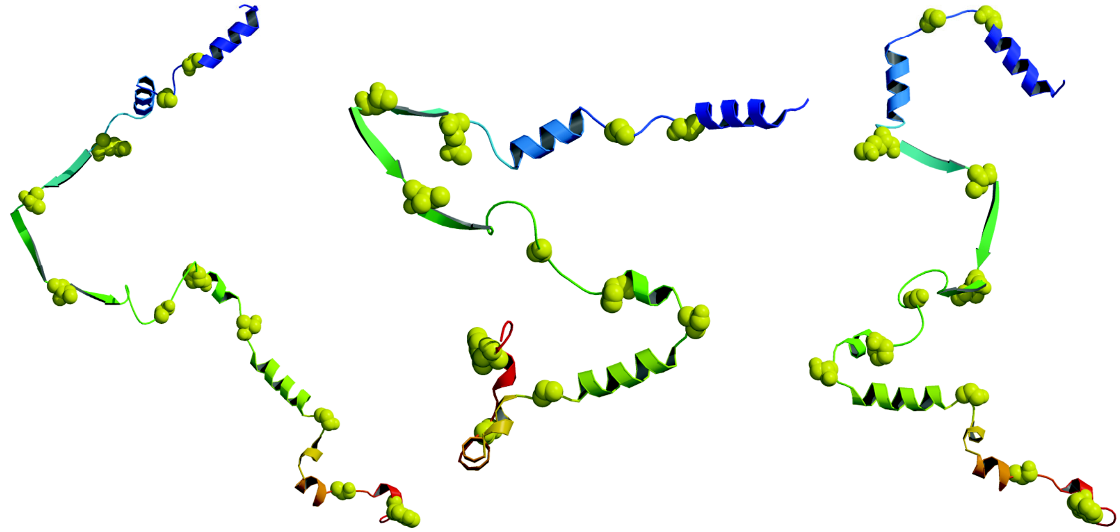


Figure 3.3. Representative lysozyme structures from rigid segment simulations. The entire chain was held fixed in its X-ray-determined conformation except for 11 flexible hinge residues (shown as yellow space-filling spheres). Ribbon diagram depict elements of secondary structure, defined here from the PDB header records and generated using MOLSCRIPT (Kraulis 1991) and Raster3D (Merritt and Bacon 1997). Termini are color-coded: blue=N-termini; red=C-termini.

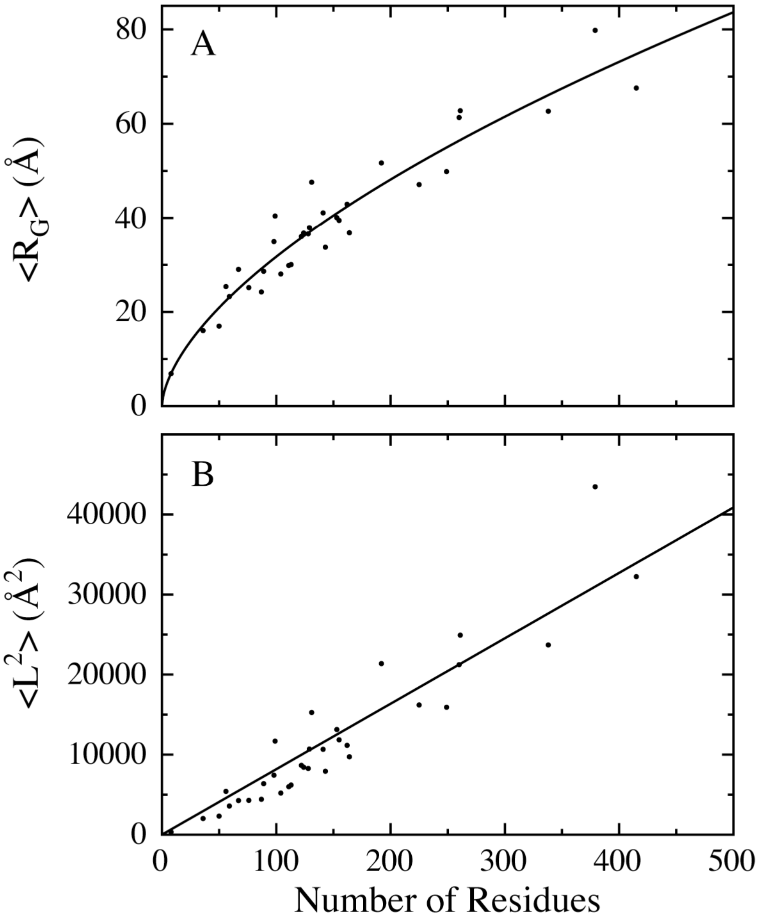


Figure 3.4. (A) Radius of gyration (<R_G>) versus chain length, in residues, for 33 ensembles from rigid segment simulations. The curve is well fit by equation (3.1), with $R_0 = 1.98 \pm 0.37$ Å and $\nu = 0.602 \pm 0.035$. (B) Mean squared end-to-end distance (<L^2>) versus chain length, in residues, for the same 33 ensembles. The best-fit value of L_0, the slope of the line, is 81.8 ± 3.4 Å^2. These fitted parameters are in close agreement with accepted random-coil values.

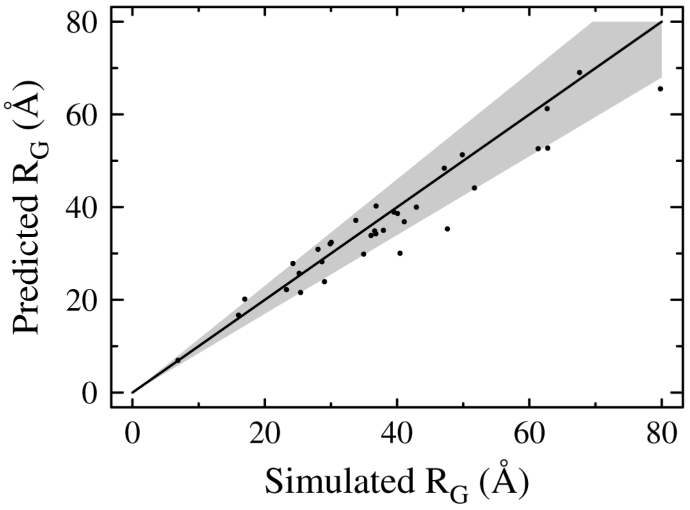


Figure 3.5. Comparison between our values of R_G from the rigid segment model and corresponding values of R_G from random-coil expectations using equation (3.1). All data points fall near the diagonal line. To aid in visualization, a shaded region marks the ±15% boundary, ranging between y = 1.15x and y = 0.85x.

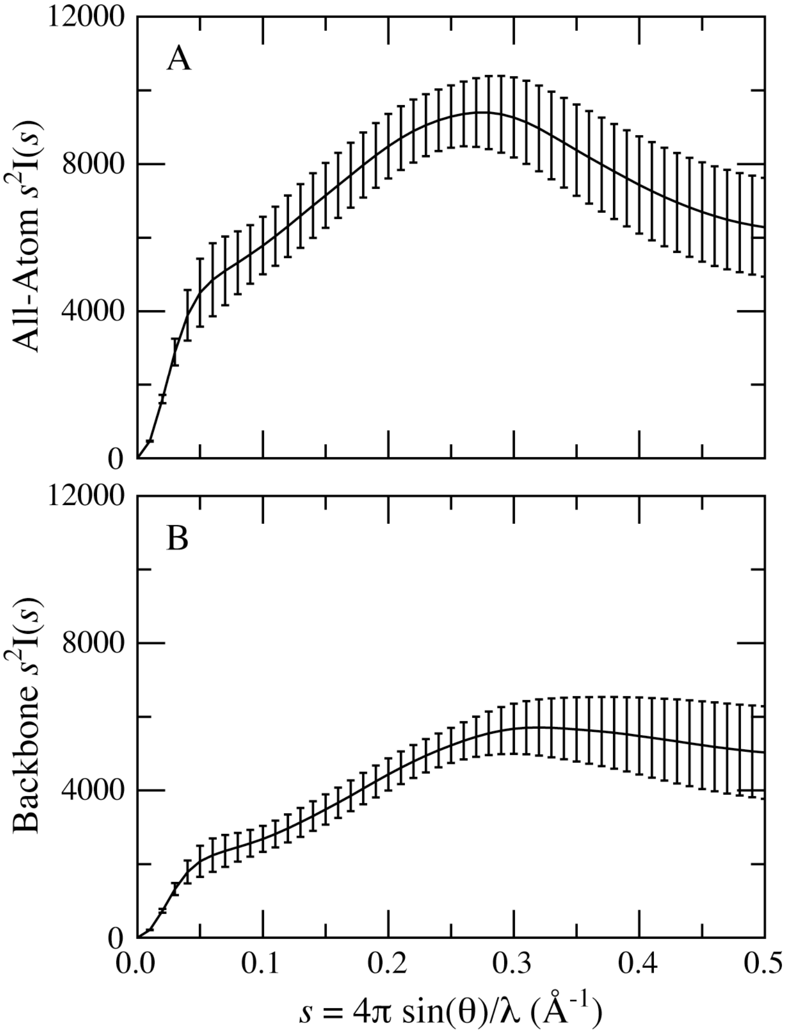


Figure 3.6. Kratky plots of rigid segment simulations. (A) calculated Kratky plot for 1,296 structures chosen at random from the lysozyme ensemble. (B) calculated Kratky plot for the same structures after removal of side chain atoms beyond C_β. The maximum in (A) is suggestive of a native protein while (B) resembles a denatured protein, suggestive of the fact that the hump in (A) is caused by side chain rigidity, not by lack of backbone flexibility.

CHAPTER 4

The Protein Coil Library:

A Structural Database of Non-helix, Non-strand Fragments Derived from the PDB[*]

4.1 Abstract

Approximately half the structure of folded proteins is either α-helix or β-strand. We have developed a convenient repository of all remaining structure after these two regular secondary structure elements are removed. The Protein Coil Library (http://roselab.jhu.edu/coil/) allows rapid and comprehensive access to non-α-helix and non-β-strand fragments contained in the Protein Data Bank (PDB). The library contains both sequence and structure information together with calculated torsion angles for both the backbone and side chains. Several search options are implemented, including a query function that uses output from popular PDB-culling servers directly. Additionally, several popular searches are stored and updated for immediate access. The library is a useful tool for exploring conformational propensities, turn motifs, and a recent model of the unfolded state.

[*] This chapter is reprinted with permission from the authors as allowed by the copyright agreement. Fitzkee, N. C., Fleming, P. J., and Rose, G. D. (2005) *Proteins: Structure, Function and Bioinformatics.* **58** (4): 852-4. Copyright © 2005 Wiley-Liss, Inc.

4.2 Introduction

The structures of folded proteins are inherently complex, and many cognitive schemes have been developed to simplify and organize protein substructure. Cartoon illustrations that reduce α-helices and β-strands to visual icons (Kraulis 1991) have been especially useful tools because approximately half of any given folded protein adopts either or both of these two regular secondary structure motifs. Here, we focus on the other half of the protein, *i.e.* the "coil" regions.

The intriguing hypothesis that coil regions are apt models for the unfolded state of proteins has motivated several important studies. Swindells *et. al.* distinguished between α-helices, β-strands, polyproline-II helices and coil (everything else) when calculating conformational propensities for amino acids (Swindells et al. 1995). Serrano compared the ϕ torsion angle propensities found in the coil conformation to NMR measurements of the unfolded state (Serrano 1995), an approach that has been pursued in very recent work (Avbelj and Baldwin 2004; Fleming et al. 2005). On the whole, however, comparatively few investigators have capitalized on the wealth of structural information stored in coil fragments.

The Protein Coil Library (PCL) is designed to address this issue. It classifies protein structure using a torsion-angle based standard and stores non-helix, non-strand fragments in an online database. The library includes molecular coordinates, dihedral angles, and sequence information for each fragment, and users can browse this information using a convenient web interface. Data can also be accessed via FTP. Versatile search tools are provided via a queued system, and the output from several online PDB-culling servers can be used to select the list of proteins to be included in a search. Additionally, the library provides basic utility programs to assist users in analyzing their search results.

4.3 Implementation

Secondary Structure Classification

The method used to classify secondary structure in the PCL, similar to that described by Srinivasan and Rose (Srinivasan and Rose 1999), tiles Ramachandran dihedral space (Ramachandran and Sasisekharan 1968) into a course-grained 30° x 30° ϕ, ψ-grid. We refer to these grid squares as *mesostates*; each is assigned a unique identifier. Any protein backbone conformation can be approximated by its linear sequence of mesostate identifiers, and regular expressions of mesostate sequences can be used to define α-helices, β-strands, and turns. Hydrogen bonds are not included in our method, but, nevertheless, the results are in close agreement with those of other secondary structure classification programs (*e.g.* DSSP (Kabsch and Sander 1983)) that do utilize hydrogen bonds. Mesostate bins are illustrated in figure 4.1, overlaid on to a contour plot of Ramachandran dihedral angles calculated by Hovmöller *et. al* (Hovmöller et al. 2002). The regular expressions used to define secondary structures (α-helix, β-strand, polyproline-II helix, turns, and coil) are given on the PCL web page and in table 4.1.

Coil Fragment Excision

Using the secondary structure classification algorithm described above, non-helix and non-strand fragments were extracted from the Protein Data Bank (Berman et al. 2000). Each fragment was inspected for chain breaks. Residues lacking any backbone atom (N, CA, C, or O) and single-residue fragments were excluded from the library. As a result, all fragments in the PCL are continuous and include at least two residues. Where possible, up to two flanking residues at both the N- and C-termi were also extracted to provide the context of the fragment. The resulting coordinates were stored in standard PDB format.

Torsion angle calculations

Accompanying every fragment is a data file that includes the sequence of the fragment along with the ϕ, ψ, ω, τ, and χ_n torsion angle values for each residue, according to the IUPAC-IUB standard (IUPAC-IUB 1970). The file also includes the per-residue mesostate identifiers and secondary structure classifications. The file format is designed to ease high-capacity analysis and is described in detail on the PCL website.

File naming and organization

Data from the coil library are stored in a collection of compressed text files that can be accessed via the web or anonymous FTP. Filenames reflect the origin and details of each fragment: the PDB identifier, chain, fragment length and start residue are all reported within each file name. All files are organized hierarchically by PDB ID and fragment length to minimize strain on the server file system. File naming conventions are described in detail on the website.

4.4 Interface and Usage

The Protein Coil Library can be accessed at http://roselab.jhu.edu/coil/. For simple searches, single chains from a PDB identifier may be browsed interactively. For each chain, coil fragments are listed and ranked according to size. The browsing functionality allows the user to download molecular coordinates directly or to view dihedral angle and secondary structure data in an HTML document. A cross-reference link to the Protein Data Bank site is associated with each coil fragment.

For more complex queries, a batch search form is provided that allows users to specify fragment sizes in addition to PDB and chain identifiers. In addition to a simple text file containing PDB ID's, PDBSelect (Hobohm and Sander 1994) and PISCES (Wang and Dunbrack 2003) formatted lists may be uploaded that specify which chains to include in the search. Using a PDBSelect or PISCES list allows the user to filter fragments based on sequence identity, resolution, and refinement quality (R-value). Once submitted, batch searches are queued, and when the results have been calculated, the user is notified that the search results are available on the server. Results are returned as a list of fragments stored on the server as well as a compressed archive of the dihedral angle data for all matched fragments. Coordinates for search results must be downloaded separately or extracted from a local copy of the PDB using one of the included utilities. Search results are removed from the server after two weeks.

Given the popularity of PISCES, two lists are generated automatically to ease resource consumption. The first list contains fragments extracted from PDB entries with a 90% sequence identity cutoff, a resolution of 2.0 Å or better, and an R-value of 25% or better. The second list contains fragments with a 20% sequence identity cutoff, a resolution of 1.5 Å or better, and an R-value of 25% or better. The results from these searches are always available as precompiled lists, and as demand arises other searches can be scheduled automatically as well. While the coil library itself is updated nightly from the PDB, these lists are only updated weekly, in coordination with the distribution of new PISCES lists.

Finally, a repository of analysis tools is provided on the website. In addition to a utility that will extract structural coordinates given a dihedral angle file, a tool is provided that can catalog the number of times different structural motifs appear in a dataset. As additional tools are implemented or contributed, they will be posted at this location.

4.5 Statistics

There are presently 784,257 coil fragments contained in the PCL, representing 55,111 chains in 25,392 unique PDB identifiers. The culled list containing fragments having less than 90% sequence identity cutoff, resolution of 2.0 Å or better, and an R-value better than 25% currently has 57,402 fragments representing 3,959 chains in 3,652 unique PDB identifiers. The distributions of fragment sizes for both lists are markedly skewed toward short fragments (figure 4.2). This is not surprising in light of hydrogen bonding considerations: α-helix and β-strand are the only regular structures that can satisfy hydrogen bonds for long chain segments, and the PCL lacks these structures. However, hydrogen-bonded structures are also abundant in short chain fragments. Indeed, using the least stringent hydrogen bond definitions outlined by Kortemme *et. al.* (Kortemme et al. 2003)*,* approximately 40% of the residues in the PCL are involved in an *i* to *i+3* hydrogen-bonded turn.

4.6 Acknowledgments

We thank Nick Panasik, Timothy Street, and Haipeng Gong for their assistance in critiquing and debugging this work. Financial support from the Mathers Foundation is gratefully acknowledged.

Table 4.1: Secondary Structure Mesostate Definitions

Secondary Structure	SS Code	Mesostates	Description
α-helix	H	De, Df, Ed, Ee, Ef, Fe	A region is identified as α-helix if there are five or more contiguous residues in this mesostate set.
β-strand	E	Bj, Bk, Bl, Cj, Ck, Cl, Dj, Dk, Dl	A region is identified as β-strand if there are three or more contiguous residues in this mesostate set.
Turn[†]	T	EfDf, EeEf, EfEf, EfDg, EeDg, EeEe, EfCg, EeDf, EkJf, EkIg, EfEe, EkJg, EeCg, DfDf, EfCf, DgDf, DfDg, IhIg, EfDe, EkIh, DgCg, DfCg, IbDg, DfEe, FeEf, IbEf, DfEf, IhJf, IhJg, IgIg, EfCh, DgEe, DgEf, EeEg, IhIh, EeDe, IgJg, EkKf, EeCh, IbDf, DgDg, EgDf, FeDg, ElIg, IgIh, DfDe, EjIg, EeCf, DfCh, DgCf, DfCf, DeEe, DkIh, FeDf, EkIf, EeDh, DgCh, IgJf, EjJg, FeEe, DlIh, EgCg, ElIh, EjJf, FeCg, DlIg, IbCg, EfEg, EkJe, FkJf, ElJg, DgDe, DlJg, EgCf, IaEf, FkIg, JaEf, EjIh, EgEf, DkJg, DeEf, EeCi, JgIh, IcEf, EkKe, DkIg, IbEe, EgDg, EeFe, EjKf, IaDf, HhIg, HbDg, ElJf, EfDh, IcDf, EfBh, IcDg, IcCg, FkJg, FeCh, IgKf, FdDg, EkHh, DfDh, DgBh, DfBh, DeDf, DfFe, EfFe, EgEe, EgDe, DkJf, JgJg, IbEg, IbCh, EfBg, DgCe, JlEf, CgCg, HhJf, EeBi, DfBi, IhIf, FeEg, FdEf, EdEf, DlJf, DhCg, JgIg, IeBg, FjIg, FdCh, EdEe, JfIh, JaEe, HhJg, HbEf, HbCh, FkIh, FjJf, ElJe, DhDf, CgDf	All dipeptide pairs which match a combination in this mesostate set are identified as turn.[‡]
P_{II}[†]	P	Dk, Dl, Ek, El	Residues in this mesostate set that are left over after β-strand has been classified are identified as polyproline-II.
Coil[†]	C	All	After all other classifications have been made, unclassified residues are identified as coil.

[†] Residues identified as turns, polyproline II helix, and coil are included in the library.

[‡] Definitions adapted from Rose, *et. al.*(Rose et al. 1985)

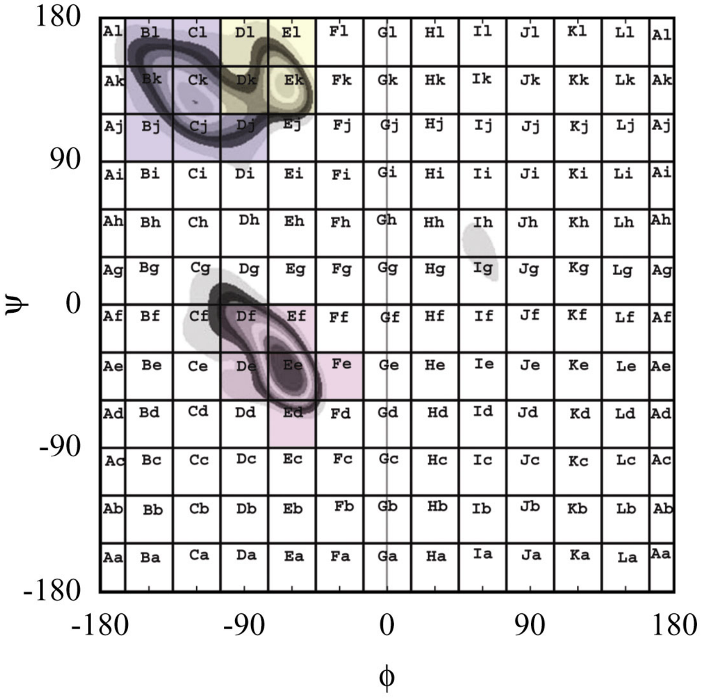


Figure 4.1. Contour plot of Ramachandran dihedral space(Hovmöller et al. 2002) overlaid with mesostate tile definitions. Mesostate identifiers are two character strings (*e.g.* Ae): the first character indicates a region along the ϕ axis and the second character indicates a region along the ψ axis. Mesostates used to identify β-strands are shaded blue, those used for identifying α-helices are shaded pink, and those used for identifying polyproline-II are yellow. Mesostates Dl and Dk, while normally used to identify polyproline-II, will be classified as β-strand if adjoining residues are also β-strand. See table 4.1 for further detail.

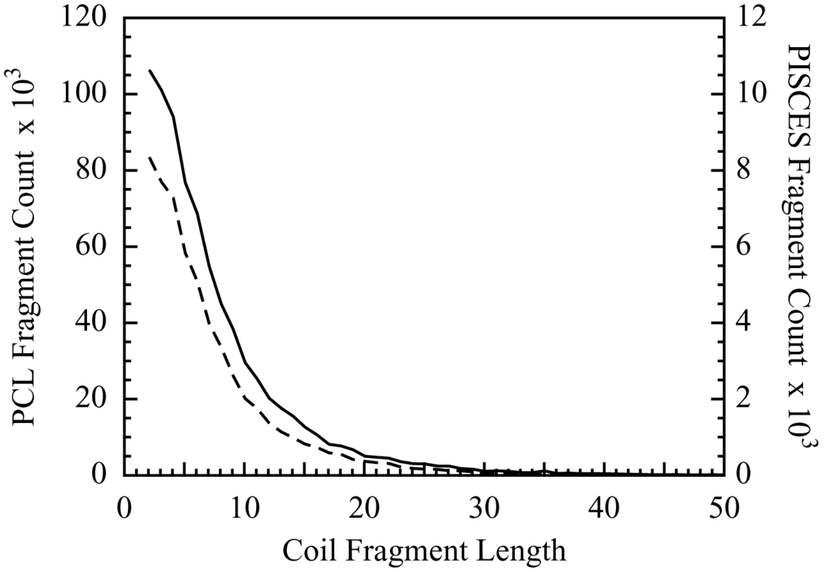


Figure 4.2. Comparing the coil fragment length in the PCL to the number of times that length occurs. The solid line (left axis) shows the distribution for the entire coil library, and the dashed line (right axis) shows the distribution for a culled list of chains (90% sequence identity cutoff, 2.0 Å resolution or better, and 25% or better R-value). Both graphs indicate a decline in frequency as fragment size increases.

CHAPTER 5

Sterics and Solvation Winnow Accessible Conformational Space for Unfolded Proteins[*]

5.1 Abstract

The magnitude of protein conformational space is over-estimated by the traditional random-coil model, in which local steric restrictions arise exclusively from interactions between adjacent chain neighbors. Using a five-state model, we assessed the extent to which steric hindrance and hydrogen bond satisfaction – energetically significant factors – impose additional conformational restrictions on polypeptide chains, beyond adjacent residues. Steric hindrance is repulsive: the distance of closest approach between any two atoms cannot be less than the sum of their van der Waals radii. Hydrogen bond satisfaction is attractive: polar backbone atoms must form hydrogen bonds, either intramolecularly or to solvent water. To gauge the impact of these two factors on the magnitude of conformational space, we systematically enumerated and classified the disfavored conformations that restrict short polyalanyl backbone chains. Applying such restrictions to longer chains, we derived a scaling law to estimate conformational restriction as a function of chain length. Disfavored conformations predicted by the model were tested against experimentally determined structures in the coil library, a non-helix, non-strand subset of the PDB. These disfavored conformations are usually absent from the coil library, and exceptions can be uniformly rationalized.

[*] This chapter has been submitted for publication in *The Journal of Molecular Biology.*

5.2 Introduction

Protein folding, the transition from unfolded to folded ensembles, is an inherently complex process (Dill 1990; MacKerell et al. 1998), and the field is divided as to whether the essential features of this process can be captured using simplified models. The hard-sphere model for noncovalent atomic contacts (Lee and Richards 1971; Richards 1977) is a case in point. Despite its oversimplification of the underlying quantum mechanics, the model has provided fundamental insights into conformational preferences in proteins (Ramachandran et al. 1963; Ramachandran and Sasisekharan 1968; Fitzkee and Rose 2004b) and polymer statistics (Flory 1969; Fitzkee and Rose 2004a). Indeed, the model is now used routinely to assess the validity of X-ray elucidated protein structures (Laskowski et al. 1993a).

Solvation also plays a significant role in protein structure (Chellgren and Creamer 2004; Kentsis et al. 2004; Mezei et al. 2004; Fleming et al. 2005) and is another complex topic that invites simplification. Essentially all polar groups in proteins are hydrogen bonded, either to other protein groups or to solvent water (Panasik et al.; Fleming and Rose 2005). Each unsatisfied hydrogen bond in a folded protein comes at a cost of ~5kcal/mol, an energetic penalty that rivals the free energy difference between the folded and unfolded forms of the molecule (Fleming et al. 2005). This stiff energy cost can be exploited in the form of a simple screening algorithm in which polar atoms that fail to participate in an intramolecular hydrogen bond (Stickle et al. 1992) are assessed for solvent-accessibility (Lee and Richards 1971). Further, the solvation energy can be quantified by the detailed extent to which such atoms can be hydrated (Petukhov et al. 2004; Fleming et al. 2005). This simple strategy avoids the incompletely-understood complexity of explicit water but still identifies solvent-inaccessible, disfavored conformations successfully.

Steric exclusion and solvation are thought to be the dominant forces in the unfolded state of proteins (Dill and Shortle 1991; Pappu and Rose 2002; Chellgren and Creamer 2004), and we investigate their effects using the two simplifying approximations described above. It is noteworthy that much earlier work, stemming from Flory, was grounded in the assumption that hard-sphere sterics will have little effect on the energy landscape of the unfolded state. This assumption was based on the isolated pair hypothesis (IPH) (Flory 1969), which posits that the only systematic local steric constraints on a residue are exerted by its adjacent chain neighbors and are described by the well-known Ramachandran diagram for the alanyl dipeptide (Ramachandran and Sasisekharan 1968). However, on re-examination, Pappu *et al.* found that local steric constraints do extend beyond the alanyl dipeptide (Pappu et al. 2000) and do influence the unfolded population. Although other workers confirmed this result (Ohkubo and Brooks 2003), they concluded that it is limited to peptides of six residues or less, with only a small effect on the total conformational entropy of the unfolded population (Zaman et al. 2003).

Today, the extent to which steric restrictions sculpt the conformational landscape remains unclear. Protein secondary structure can be predicted largely on the basis of sterics and hydrogen bonding considerations (Srinivasan and Rose 1999) and may be sufficient to determine tertiary structure as well (Przytycka et al. 1999; Gong and Rose 2005). An undifferentiated tube of finite thickness with a hydrogen bonding potential can reproduce the entire repertoire of small, single-domain protein folds (Hoang et al. 2004), indicating that finite chain thickness is the principal organizing factor for polypeptide chains (Banavar et al. 2004).

Levinthal deduced that undisclosed organizing interactions must bias a protein as it negotiates the folding process because a random search for the native state could not be accomplished on a biologically realistic timescale (Levinthal 1969). But what are these organizing interactions? Are steric restrictions sufficient to justify Levinthal's conclusion by reducing the accessible energy landscape of an otherwise vast unfolded state?

Steric repulsion is an appealing explanation, and especially so when hydrogen bonds to solvent are also included in the analysis. But other models exist as well. For example, in the foldon model, proteins organize via step-wise assembly of small, quasi-independent subunits (Maity et al. 2005). In the contrasting nucleation-collapse model, the search is reduced by an overall cooperative collapse around an expanded native-like nucleus (Daggett and Fersht 2003). Alternatively, the chain might collapse around a hydrophobic core and then subsequently self-organize in a highly confined space (Dill and Stigter 1995). Another model proposes that evolution has selected sequences that can fold with minimal frustration, avoiding non-productive conformational excursions (Go 1984; Onuchic and Wolynes 2004). Some of these models are not mutually exclusive, of course.

Here, we explore the hypothesis that steric restriction and protein:solvent hydrogen bonding reduce conformational complexity in the unfolded state. Work done in chapter two identified a single sterically disfavored conformation: three consecutive residues in α-helical conformation followed by a residue in β-strand will encounter an unavoidable steric clash (Fitzkee and Rose 2004b). This *steric restriction* organizes protein structure by constraining the interaction between the two fundamental secondary structures, α-helix and β-strand, which must be separated by intervening residues in turn or coil. Hybrid segments will be suppressed.

The goal of this chapter is to enumerate such restrictions systematically and to quantify their significance in organizing the unfolded state. Our methodology combines simulation with analysis of known structure. In simulations, protein conformation was represented using five discrete states, drawn from assiduously chosen regions of conformational space for a dipeptide. The five-state model was validated by showing that it is sufficient to represent proteins of known structure satisfactorily. Using this model, conformations of short peptides were generated exhaustively and tested for restrictions imposed by either sterics or solvation requirements. The probability of occurrence for a given conformation was measured by its *acceptance ratio*, the fraction of sterically-acceptable, solvation-available conformations encountered in a statistically significant number of randomly-generated attempts. A *restriction* is then defined as a conformation with an acceptance ratio of less than e^{-1}, corresponding to an ambient-temperature fluctuation in our statistical energy function, as described in Methods.

The five-state model was then applied to proteins of known structure to determine whether model-based restrictions are correspondingly disfavored in experimental structures. In general, such conformations are usually absent altogether, and exceptions to this trend can be explained readily. This conclusion was examined in detail for tetramer fragments, including an atomic-level description of ten illustrative, highly disfavored examples. Finally, restrictions derived from short peptides were extrapolated to longer chains to derive a scaling law. These results are now presented in detail.

5.3 Results

Rebuilding Proteins from Five States

To simplify the inherent complexity of protein structure, we developed a five-state model that is intended to capture the fundamental backbone conformations (figure 5.1 and table 5.1). The model was validated by rebuilding the backbones of six arbitrarily chosen proteins using backbone torsion angles from these five conformational states (see Methods). Rebuilding proteins using ideal bond lengths and angles remains a difficult challenge, even when exact backbone torsions are used (Holmes and Tsai 2004). Using the five-state model, good results were obtained for five of the six proteins (table 5.2 and figure 5.2), with an RMSD < 3.0 Å from the native structure in all five cases. The sixth protein, hen egg lysozyme (1HEL), is also well represented in large part, but with a hinge-like opening of the structural core around residues 38-45. Still, the RMSD for this case, 4.47 Å, is well below the value expected for a random conformation (Cohen and Sternberg 1980). From a structural standpoint, artificially limiting the conformations of proteins to five discrete states is surely an oversimplification. However, the five-state model is based on more than mere convenience; previous analysis indicates that a limited number of energy basins is sufficient to account for the majority of the equilibrium thermodynamic population, both for short peptides (Pappu and Rose 2002) and for proteins (Srinivasan et al. 2004).

Summary of Five-State Simulations

Hard sphere simulations were performed on blocked polyalanyl peptides, N-acetyl-(Ala)$_n$-N-methylamide, n = 1-6 as described in Methods. Every combination of the five conformational states at each chain length was simulated. Resultant clash-free structures were assessed for hydrogen bond satisfaction to ensure that all backbone polar groups could participate in a hydrogen bond. The probability of occurrence for a given conformation was measured by its acceptance ratio, the fraction of successful attempts. Conformations with a low acceptance ratio ($\leq e^{-1} \sim 38\%$, see Methods) were further investigated for steric clash and/or lack of solvent access. A complete list of these disfavored conformations and their interactions can be found as at http://roselab.jhu.edu/fivestate/.

Systematic conformational strain was not observed for polyalanine chains of length n = 1-2, but unfavorable conformations were observed for n = 3-6 (table 5.3). Given five states, there are 5^n possible conformational strings for each polymer of length n. The fraction of conformational space that is disallowed can be estimated as the number of conformations that are disfavored normalized by the total number of conformations. Beyond three residues, observed trends indicate that many of the disfavored conformations for an $n+1$-residue peptide can be predicted from the corresponding conformations in an n-residue peptide. For example, if HHE is disfavored in trialanine, then HHHE will be disfavored in tetraalanine. Exceptions to this extrapolation pertain to conformations that are only marginally disfavored in the n-residue case and probably result from statistical fluctuations in the data, not systematic problems with the analysis. As anticipated by Ohkubo and Brooks (Ohkubo and Brooks 2003), the number of unique disfavored conformations dies away beyond n = 6.

However, at this peptide length, approximately 50% of conformational space has already been eliminated owing to steric clash and/or lack of solvent access.

Within the range of interest (n = 3-6), the number of highly improbable conformations increases with peptide length, reinforcing the proposition that residues do not behave as independent ϕ,ψ-pairs (Pappu et al. 2000). Referring to the histograms in figure 5.3, the trimer distribution of low acceptance ratios ($\leq e^{-1}$) is sparse, with the lowest ratio at 8.0%. The number of conformations with acceptance ratios near 0% increases dramatically with increasing chain length, suggesting that disfavored interactions are cooperative. As atoms are added to a peptide, the number of opportunities for disfavored interactions increases, a continuing trend over the range of interest. The persisting peak in acceptance ratio at 8% is noteworthy and may be a result of decorating a constant trimer core with new, structurally allowed conformations. However, this explanation fails to account for the disappearance of the non-persisting peak at 24%.

Of the 17 disfavored conformations for trialanine (table 5.3), 15 are a consequence of solvation effects. In these 15 cases, the NH group is unsolvated, and in some conformations the C=O is buried as well. In particular, a backbone N-H at position *i+2* following an *i+1* residue with backbone dihedrals in the bridge region (B) is shielded from hydrogen-bonding and inaccessible to solvent whenever the i^{th} residue adopts an extended conformation, either polyproline II (P) or β-strand (E). This same effect also prevails, though to a lesser extent, when the i^{th} residue is in left-handed (L) conformation. In other words, in a polyalanyl peptide, a conformation like $P_{i-1}P_iB_{i+1}X_{i+2}$ has a low acceptance ratio (8%) because, for all X, the N-H of residue X cannot participate in a hydrogen bond. Specifically, it cannot hydrogen bond to the backbone because the two PP residues at *i* and *i-1* direct possible backbone acceptors away from the N-H(*i+2*) group. Neither can it hydrogen bond readily to water because the C_β atoms of polyalanine inhibit solvent-access. There are only two ways to satisfy the hydrogen bond in this example: the backbone might be adjusted so as to compensate an unfavorable geometry by the energetically favorable hydrogen bond. Alternatively, a side chain to main chain hydrogen bond could satisfy this N-H group without encountering a steric clash because the covalent radius of the side chain acceptor would be smaller than the corresponding hard sphere radius of either a non-local acceptor or a water oxygen. Authentic proteins with segments in PPB conformation have most often utilized this latter strategy (see below).

Analysis of Real Structures

Our entire dataset, cross-referenced with interactions, is available online at http://roselab.jhu.edu/fivestate/. The tetraalanine trends described below also apply to trialanine and, where statistics are significant, to penta- and hexaalanine. Tetramers represent a good compromise between a population that is large enough to be statistically significant but small enough to be individually analyzed.

In simulations using the five-state model, there are $5^4 = 625$ tetramers; 174 have e^{-1}-level restrictions. We validated these simulations by comparing them to the conformational trends seen in the protein coil library, a database of non-helix, non-strand segments from the PDB (see Methods for justification and chapter four for a detailed description). The polyalanine model, which uses fixed bond lengths and scalar angles, has disfavored conformers that can be compensated in authentic proteins. For example, a backbone to side chain hydrogen bond could rescue a peptide with PPB conformation, as described above, whereas that conformation would be highly disfavored in an all-alanine model. Table 5.4 lists common reasons for discrepancies between simulations and experimental structures. The listed rationalization classes may expose limitations of the polyalanine model or the accuracy of the experimental structure. For example, steric clashes (type I) or buried backbone hydrophilic atoms (type IIIb) are thought to be disfavored in proteins (Laskowski et al. 1993a) and may signal a problem structure when found in the coil library. On the other hand, a *cis*-proline residue (type IIa) is simply beyond the scope of our all-*trans*-polyalanine model. Given that fragments in the structural dataset were selected to conform to the five-state model and do not contain glycine, rationalization classes are not relevant for those cases.

Tetramer conformations. In the coil-library dataset of 3,864 tetrameric fragments, 497 were identified as disfavored based on the 174 e^{-1}-level model-based restrictions. Classifying these fragments according to the four rationalization classes yields the distribution shown in fig. 5.4. The most frequent class is type IIIa, cases in which a backbone-side chain hydrogen bond satisfies an otherwise inaccessible backbone N-H. The other predominant class is type IIa, *cis*-peptide bonds, where conformations are not well described by our model. Remaining classes include 19 instances of tetramers where backbone hydrophilic atoms appeared to be genuinely desolvated according to our model. However, it should be noted that the model fails to take into account hydrogen bonds to electron-rich aromatic rings, a conformation that was observed at least once. Finally, 36 of these 497 fragments could not be rationalized; all occurred among conformations with higher acceptance ratios. A similar situation was observed for the 12,731 fragments in the trialanine dataset: 845 were disfavored and 116 could not be rationalized.

Tetramer statistics. It is informative to compare the number of times a given conformation is observed to the number of times it is expected, under the assumption that the individual conformational states assort independently. This comparison is explored in figure 5.5, where all 174 model-based tetramer restrictions are ranked and the log ratio of expected to actual occurrences is plotted after rationalizations are taken into account. Only one conformation, BLBE, is observed more frequently than would be expected by chance. The plot indicates that model-based predictions of disfavored conformations are pertinent to experimental structures from the PDB: structures predicted to be disfavored by the model are less prevalent in the coil library, and they can be rationalized easily when they do occur (table 5.4), most often by a backbone to side chain hydrogen bond or by a *cis*-peptide bond.

Examples of Disfavored Conformations

In this section, we describe ten examples of disfavored tetramers and pentamers. An exhaustive description of all such conformations would be prohibitive, but full simulation data are available in at http://www.roselab.jhu.edu/fivestate/. The conformations presented here were selected based on acceptance ratio, disparity between expected and observed occurrences in the coil library, and biophysical interest.

Table 5.5 and figure 5.6 summarize the conformations described below. Only sterically allowed conformations were tested for hydrogen bond satisfaction. For each interaction in table 5.5 (column 5), the listed frequency is the number of times that interaction occurs divided by the total number of disallowed conformations, *i.e.* the fraction disallowed. This fraction may be misleading for compound conformations involving both steric clash and unsolvated polar groups: if a steric clash occurs in 80% of the conformations, hydrogen bond satisfaction would only be tested for the remaining 20%, and the calculated frequencies will be bounded accordingly. Similarly, if a given backbone nitrogen is unsolvated 90% of the time but a steric clash is present as well, the clash will be the dominant effect, and the calculated frequency will reflect this. For non-compound cases – conformations with unsolvated atoms but no steric clash, or conformations with steric clash but no unsolvated atoms – the indicated frequencies are a valid measure of that particular interaction. In fact, this is the case for most disallowed conformations in our simulations, which involve either steric clashes or unsolvated atoms, but not both.

Helix-Strand Transitions: HHEB and HHHE. Trialanine simulations indicate that the HHE conformation is disfavored for steric reasons, with the eleventh lowest acceptance ratio (0.20 ± 0.01) of the 125 possible conformations. Thus, it follows that the two tetrameric conformations HHEB and HHHE also have correspondingly low acceptance ratios: $2.66 \pm 0.29 \times 10^{-2}$ for HHEB and $1.20 \pm 0.04 \times 10^{-2}$ for HHHE. The primary interaction in both cases is an O_{i-1} to O_{i+2} steric clash, described previously in chapter two (Fitzkee and Rose 2004b). The two conformations are distinguished by an additional interaction that arises when a bridge residue immediately follows an extended P or E residue: both EBX and PBX bias peptides toward conformations that shield the next residue from hydrogen bond access, regardless of X, resulting in an overall acceptance ratio of ~3% (table 5.5).

Occurrences of HHEB and HHHE in the coil library resemble their expected frequencies. Given independent assortment, HHEB is expected three times and HHHE twice. In fact, HHEB occurs twice and HHHE once. All three observed occurrences are rationalized easily: both occurrences of HHEB have a *cis*-peptide bond which relaxes conformational strain and exposes the otherwise shielded backbone nitrogen to solvent. The sole occurrence of HHHE has several atypical bond lengths and angles, as identified by PROCHECK (Laskowski et al. 1993a), which relieve steric clash. Exceptions like this one are expected to be minor or non-existent in the denatured state, which lacks persisting contacts that can compensate for distortion of equilibrium bond lengths and angles. But even in the native state, the fact that peptide geometry is strained in these conformations lends support to the polyalanine model.

Extended Residues and the Bridge Region: HBEB. As described above, a residue in an extended conformation followed by one in the bridge region (*i.e.* EB) tends to shield the following residue from either water or peptide hydrogen bonds. Many of the restrictions described here are a consequence of this tendency, as illustrated by HBEB. Similar to HHEB, in that a steric clash between two carbonyl oxygens restricts conformational space, the trimeric HBE conformation is more permissive than HHE, with an acceptance ratio of 31% (vs. 19% for HHE). The acceptance ratio for an HBEB tetramer, $3.54 \pm 0.14 \times 10^{-2}$, is substantially reduced relative to the trimer, largely because of the additional solvation requirement for the amino nitrogen at position *i+4* (shown with its associated virtual water in figure 5.6B). As seen in the illustration, the β-carbon of the adjacent extended residue inhibits access to the amino nitrogen. This situation would be obviated for a glycine residue. Similarly, an adjacent serine or threonine, with its side chain hydrogen bond acceptor, could satisfy the otherwise occluded amino nitrogen. In fact, both HBEB structures in the coil library are rationalized by a side chain-backbone hydrogen bond from a serine hydroxyl oxygen. Significantly, a structure like the one in figure 5.6B, but with an occluding side chain that is hydrophobic, is never observed in our dataset.

Mixing P_{II} helix with bridge residues: EPBH, HPBB, PBPB, PPPB. Conformations with mixtures of P and B are often disfavored owing to a solvent-inaccessible N-H group, as illustrated in figures 5.6C-F. In each of these cases – EPBH, HPBB, PBPB, PPPB – a similar interaction blocks access to the amino nitrogen of the residue immediately following the PB combination. Given the relatively large fraction of P and B residues in our dataset (table 5.1), PB-mixtures would be frequent in the coil library if these two conformations assorted independently. However, as shown in figure 5.5, all such structures occur less frequently than predicted, and almost every occurrence can be rationalized, typically by a backbone-side chain hydrogen bond.

Compact conformations: HBLB. The conformation HBLB was chosen for discussion because of the large disparity between its expected and observed frequencies: three times vs. 38 times, respectively. Superficially, this disparity seems to expose a model deficiency, but, in fact, each of the 38 occurrences can be rationalized by a local backbone-side chain hydrogen bond that satisfies an otherwise inaccessible amino nitrogen (figure 5.6H). In this conformation, the side chain of residue *i-1* is poised to serve as a hydrogen bond acceptor for residue *i+4*. Indeed, this is the arrangement most often observed in the coil library, where, typically, aspartic acid or threonine are preferred at the *i-1* position.

Five-residue conformations: EHELL and HPLLP. Finally, we include two disfavored pentamer conformations: EHELL and HPLLP (figures 5.6I-J). Both are similar in causing the peptide chain to wrap back upon itself. Neither conformation is observed in the coil library, but the statistical distribution of pentamer fragments is too sparse to draw reliable conclusions from this fact. Nevertheless, these conformations are expected to be rare because the steric clash is severe.

Summary of Examples. For the tetrameric conformations described above, as well as those structures not described here, it is almost always true that a low acceptance ratio corresponds to a population in the coil library that is less than expected based on independent assortment. When these conformations do appear, they can usually be rationalized by the limitations of our simple model. It is problematic to draw such conclusions for pentamers and hexamers where data are more sparse, but the success for tetramers and trimers bolsters confidence that our procedures can be reliably extended to longer peptides.

A Scaling Law: String Simulations

Local restrictions from sterics and solvation winnow conformational space, as described previously. To estimate the magnitude of these effects on longer peptides, a series of simulations was performed in which strings were generated at random from the five-state model, using the weights observed in the coil library. Each string was then accepted or rejected based on the acceptance ratios for six-residue restrictions (see Methods). For example, acceptance of the seven-residue string, HHEBEEE, would be based on the subsumed six-residue substring, HHEBEE, and its acceptance ratio of

2.99%. Substrings that were accepted based on the acceptance ratio of one six-residue restriction could not result in rejection when later compared with another six-residue restriction. String simulations were performed for strings of length 4 to 60, and their acceptance ratios were regarded as a statistical energy (equation 5.2 in Methods) and plotted on a log scale (figure 5.7). The log of the acceptance ratios falls on a straight line, and these data were fit to the equation:

$$log(r) = mN + b \tag{5.1}$$

where r is the acceptance ratio and N is the string length (*i.e.* number of residues). Using nonlinear least squares fitting with R (R Development Core Team 2003), the parameters m and b are -0.19801 ± 0.00039 and 0.504 ± 0.014, respectively ($R > 0.99$). There is no indication that this trend will deviate from linearity when extrapolated to longer peptides, although care must be taken to calculate the uncertainties for extrapolated values (Bevington and Robinson 1992).

Using this fit, the acceptance ratio for a random string was compared to the acceptance ratio for authentic proteins (table 5.2). In every case, the authentic protein's acceptance ratio is greater than that expected for a random sequence of the same length, typically by several orders of magnitude. Protein acceptance ratios were obtained by converting the structure to the five-state model and using the resultant conformation string in lieu of a random string (see Methods). This procedure underestimates the intrinsic protein acceptance ratios because the five-state representations lack side chains and would be additionally filtered using the rationalizations described previously. Although authentic proteins are not constrained to five discrete states, the data in table 5.2 demonstrate that estimates derived from our all-polyalanine model can provide a useful lower bound on the fraction of conformational space that is eliminated by local conformational restrictions. From the random-string acceptance ratio for a 100-residue protein (*e.g.,* urease from *B. pasteurii* (1UBP), table 5.2), that fraction is $4.2 \pm 3.7 \times 10^{-9}$, approximately nine orders of magnitude.

5.4 Discussion

The goal of this work is to study the local conformational constraints on the peptide backbone that are imposed by sterics and hydrogen bonding. Similar to our previous study in chapter two (Fitzkee and Rose 2004b), we used a computational approach involving both simulation and analysis of known structure. The earlier study identified a single constraint in proteins that limits conjunctions between an α-helix and a β-strand. Here, we seek to detect the full range of such constraints, to estimate their impact on the size of allowed conformational space, and to catalog some of the more important examples in atomic detail. Our results document specific interactions that lead to the failure of the Flory isolated pair hypothesis (Pappu et al. 2000), and they provide an estimate of the degree to which local backbone interactions contribute to resolution of the Levinthal paradox (Levinthal 1969).

The simulations presented here assume idealized bond lengths and scalar angles, presumably a modest assumption for the unfolded state, where there is a deficit of interactions that could compensate for locally strained conformations. Yet, even the folded state appears to be largely free of significant conformational strain, as indicated by methyl-rotors and side chains, which are found preferentially in staggered configurations (Kossiakoff et al. 1990; Butterfoss and Hermans 2003). Accordingly, many restrictions identified here are likely to be relevant to folded proteins, despite the use of idealized geometries in their identification. Still, it is important to bear in mind that our approach is based on equilibrium thermodynamics, where highly disfavored conformations can nevertheless occur.

Sterics and Solvation in Protein Folding

Hydrogen bond satisfaction plays a central role in organizing the denatured state and limiting conformations in the folded state. In our tetramer simulations, 148 of the 174 highly disfavored conformations (85%) involve solvation alone. In hexamers, this fraction decreases to 72%, with solvation still the dominant effect. Equivalently, it is clear that sterics alone play a lesser role in organizing these short peptides, although excluded volume effects become highly significant at longer length scales (Dill 1985), to be sure.

To further investigate the impact of peptide-water hydrogen bonding, we simulated a blocked alanine dipeptide with inclusion of the solvation criteria described in Methods, akin to a classical ϕ,ψ-plot (Ramachandran and Sasisekharan 1968) but with conformations rejected either for steric clash or for solvent-shielding. The resultant diagram (figure 5.8) departs from the iconic Ramachandran plot (dashed lines), differing significantly in the bridge region, and with the emergence of a distinct peninsula below the polyproline II region (below ϕ,ψ = -90°,60°). This peninsula is observed in proteins of known structure (Hovmöller et al. 2002; Ho et al. 2003), and our simple solvation model can account for its existence.

On the other hand, depletion of the bridge region, as seen in figure 5.8, is not observed in experimental structures (Hovmöller et al. 2002). Rather, many residues in the B region are involved in type I turns (Rose et al. 1985). We note, however, that the conformation of a four-residue β-turn (*i* to *i+3*) is established by the backbone dihedral angles of its two inner residues (*i+1, i+2*), which reside in the H and B regions, respectively, in a type I turn. When isolated residues from the bridge region are adjacent to extended residues (*e.g.,* PPPB) or to left-handed helical residues (*e.g.,* HBLB), instead of a turn-forming residue (*e.g.,* HHHB), the C-terminal N-H is sequestered from solvent access. This is not an issue in a β-turn, of course, which has an intrapeptide hydrogen bond (Rose et al. 1985). Depletion of isolated residues in the bridge region would serve to rarify the remaining population of turn residues, and consistent with this inference, removing turns from the coil library depletes the B region significantly (Panasik et al.). We conclude that hydrogen bond satisfaction both organizes accessible conformational space in unfolded proteins and shapes the observed ϕ,ψ-distribution in folded proteins.

The Levinthal Paradox

As reported in Results, local sterics and solvation reduce conformational space by at least nine orders of magnitude. This number is a likely underestimate for several reasons. First, the addition of side chains would result in further reduction (Bromberg and Dill 1994). Observed correlations between side chain rotamers and backbone conformations provide evidence that side chains restrict more conformational space than they allow (Dunbrack and Karplus 1994). Secondly, our parameters were designed to be conservative. All hard sphere radii, including the water radius, were scaled to 90% of their accepted values, a highly permissive strategy (Fitzkee and Rose 2004b). Additionally, our hydrogen bond criteria were chosen to be the maximally permissive values reported in Kortemme *et. al.* (Kortemme et al. 2003). Increasing hard sphere radii and using less permissive hydrogen bond criteria would have increased the number of disfavored conformations identified in these simulations. Finally, the 38% threshold for disfavored conformations is a permissive choice. Of course, higher acceptance ratios are even more permissive, but incorporating more relaxed acceptance ratios into string simulations increases the likelihood that longer strings will be rejected. For these three reasons, the actual reduction of conformational space may be several orders of magnitude greater than our conservative estimate.

It might be thought that a more realistic model, with a larger number of states, would increase the apparent size of conformational space. However, size does not scale with the number of conformational states in a straightforward manner because the number of disfavored conformations also increases with the number of states. Our five-state model can be likened to the discrete states in lattice models, and possibly such models could provide a convenient strategy for computing scaling laws of interest (Dill and Stigter 1995).

We hasten to add that nine orders of magnitude loses significance in a background of 5^{100}~10^{70} conformations for a 100-residue protein. Clearly, other forces are at work as well. Excluded volume constraints are thought to eliminate ~44 further orders of magnitude (Dill 1985). In contrast to this long-range excluded volume reduction, the restrictions described here are essentially short-range, with little overlap between the two types of contributions. Therefore, estimated conservatively, the two values account for a reduction of at least 53 orders of magnitude.

The Denatured State and Protein Folding

Today, three views dominate thinking about the unfolded state of proteins. The traditional view regards unfolded proteins as statistical coils, with little or no persisting structure (Brant and Flory 1965b; a; Tanford 1968). A more recent proposal, based on NMR experiments (Yi et al. 2000; Shortle and Ackerman 2001; Shortle 2002), holds that the denatured state retains native-like topology, although this view is not without controversy (Louhivuori et al. 2003; Jha et al. 2005). The third view regards unfolded proteins as fluctuating ensembles of polyproline II helix (Tiffany and Krimm 1968a; Pappu and Rose 2002; Shi et al. 2002a; Shi et al. 2002b; Mezei et al. 2004; Tran et al. 2005).

The work presented here is pertinent to all three views. Clearly, the statistical coil model cannot be rigorously correct in light of our evidence for structural correlations arising from local sterics and solvation. These local interactions may represent only a minor perturbation from the statistical coil denatured state. Alternatively, inclusion of side chains, together with the restraints documented here, may bias the backbone toward native-like secondary structure (Baldwin and Rose 1999b; a). Regarding the second view, sterics and solvation could explain a bias toward native-like structure by extensive depletion of other alternatives (Baldwin and Zimm 2000). Finally, the restrictions are consistent with a prevalence of polyproline II helix in the denatured state. No disfavored conformation involves E or P exclusively; consequently, disfavored interactions in other states would serve to shift the equilibrium population toward the northwest region of the ϕ,ψ-map, where further preference for the P region is exerted via favorable solvation (Mezei et al. 2004).

Summary

Two simple principles – hard sphere sterics and hydrogen bond satisfaction – were shown to restrict the local conformational space of proteins substantially. Using a five-state model, the effects of sterics and hydrogen bonding on the conformation of short peptides were investigated by simulation and analysis of known structures. Disfavored conformations in simulations were found to be depleted in the coil library. When present at all, those conformations were usually rationalized by the presence of a *cis*-peptide bond or a side chain-backbone hydrogen bond, neither of which are included in our simplified model. Highly disfavored conformations identified in this study reduce conformational space for a 100-residue chain by approximately nine orders of magnitude, and at least 53 orders of magnitude when long-range excluded volume effects are included as well. Finally, contracted conformations provide increased opportunities for steric clash and unfavorable solvent shielding of polar groups, a realization that sheds light on current models of the unfolded state.

5.5 Materials and Methods

Peptide Structures

Alanine was chosen as a model for the peptide backbone (Hummer et al. 2001; Margulis et al. 2002). All simulations were performed using blocked alanine polymers, N-acetyl-(Ala)$_n$-N-methylamide, for n = 1-6; local systematic interactions are known to die out beyond n = 6 (Ohkubo and Brooks 2003). Bond lengths and scalar angles were taken from the LINUS simulation package (available at http://roselab.jhu.edu/dist/) (Srinivasan and Rose 1995; Srinivasan et al. 2004) and held fixed throughout all simulations. Backbone torsion angles ϕ, ψ, and ω were allowed to vary as described below. Backbone amino hydrogens were included and used in reproducing the Ramachandran plot, shown as a dashed line in figure 5.1; other hydrogens were omitted.

Five-State Conformational Model

Protein conformation was represented using five discrete states: α-helix (H), β-strand (E), left-handed α-helix (L), polyproline II helix (P), and the bridge region (B). Each state included all ϕ,ψ values within a 30^{o} by 30^{o} box (figure 5.1) around its central position (table 5.1), which was chosen to represent typical examples of each respective secondary structure type. The B state corresponds to the *i+1* position of a type I β-turn (Rose et al. 1985).

The adequacy of the model was validated by testing how well these five states can represent the fold of arbitrarily chosen proteins (figure 5.2 and table 5.2), using a straightforward protocol. For six test proteins of known structure, side chain atoms beyond C_β were stripped away and each non-glycine residue was assigned to the state that best approximates its experimental backbone dihedral angles. Starting from the first residue, the all-atom root-mean-square positional difference (RMSD) from the experimental structure was minimized in 1,000 Monte Carlo trials, with ϕ,ψ sampling constrained to be within the box surrounding each residue's respective conformational state. Glycines falling within one of the five states were treated like non-glycine residues; otherwise they were sampled within ±15^{o} of their original ϕ,ψ values. All ω torsions were sampled within ±5^{o} of their original values. The approximate structure determined in this way was then subjected to successive rounds of steepest descent and conjugate gradient minimization (Press et al. 1992) to further minimize the RMSD and eliminate hard sphere bumps. Remaining bumps in the five-state structure were small (generally $\leq$ 0.2 Å) and comparable in number to those in the experimentally determined starting structure.

Weights for the five states were taken from the observed distributions in proteins of known structure (Berman et al. 2000). In detail, a dataset was extracted from the coil library as described in chapter four (Fitzkee et al. 2005), a subset of non-helix, non-strand fragments in the PDB. The coil library – postulated to model unfolded and disordered protein systems (Serrano 1995; Swindells et al. 1995; Avbelj and Baldwin 2004; Jha et al. 2005) – was culled from the PISCES list (Wang and Dunbrack 2003) dated February 13, 2005. All are X-ray elucidated structures, with aligned sequence identity of 90% or less and resolution and refinement values better than 2.0 Å and 0.25, respectively. From a total of 63,798 fragments, only glycine-free fragments consistent with the five-state model were used: 12,731 fragments for $n = 3$; 3,864 for $n = 4$; 1,127 for $n = 5$; and 348 for $n = 6$. The distribution and relative fraction of residues falling within the five states are shown in table 5.1. With five states, there are 5^{N} possible conformational strings for a fragment of length N. The weights in table 5.1, together with the number of fragments, were used to calculate an expectation value for strings of varying length, under the assumption that string elements assort independently.

Simulations

Disfavored conformers were identified by low acceptance ratios in hard sphere simulations, performed as follows. Atomic radii were described previously (Fitzkee and Rose 2004b); water was modeled as a sphere of radius 1.4 Å. All radii were further scaled by a factor of 0.90, chosen to minimize the possibility of hard sphere artifacts (Fitzkee and Rose 2004b). Clash-free structures were further tested for hydrogen bond satisfaction using the least stringent criteria described in Kortemme *et. al.*(Kortemme et al. 2003), which maximize the number of potential hydrogen bonds. Unsatisfied backbone polar groups were probed for access to solvent using five virtual waters as described in Fleming *et. al.* (Fleming et al. 2005). Structures inaccessible to solvent were rejected. Surviving structures were guaranteed to be clash-free, with hydrogen bond partners for all backbone polar groups.

A conformational string is a sequence of letters from the five-state alphabet. For any given conformational string, a round of simulation consisted of 5,000 concerted attempts to sample ϕ, ψ and ω at random, subject to relevant five-state constraints (described above). Each sterically allowed attempt was further tested for hydrogen bond satisfaction by appending one residue to either end of the original conformation and sampling an additional 1,000 randomly-chosen ϕ,ψ angles for the two appended residues. The two single-residue extensions increased the opportunities for polar groups to be satisfied by a non-local backbone hydrogen bond. Structures were rejected if they had a steric clash in the first tier of the simulations or unsatisfied hydrogen bonds in the second tier. Each round of 5,000 attempts was repeated five times to assure convergence.

The probability of occurrence for a given conformation was measured by its *acceptance ratio*, the fraction of successful attempts. A *restriction* was defined as a conformational string for which the specific sequence was found to have an acceptance ratio of less than e^{-1} (~ 38%). At this threshold, the statistical energy function

$$E = RT \ln(\text{acceptance ratio}) \qquad (5.2)$$

has a value of *RT*, approximately one ambient-temperature energy fluctuation. This choice of threshold established an upper bound for the number of restrictions, but the same trends would have been observed were the restrictions defined by a smaller acceptance ratio.

Analysis of Structures

To validate these simulations, the coil library was screened to determine whether disfavored conformers are also suppressed in experimental structures. It is possible that a restriction is salvaged by compensating interactions of a kind that exceed the limited scope of our polyalanyl model. Accordingly, we sought to rationalize disfavored conformations observed in the coil library, placing them, when possible, into one of four classes (table 5.4).

Class I: Structures that exhibit a hard sphere steric clash (with radii scaled to 90%).

Class II: Structures with geometric anomalies. A broad range of nonstandard bond lengths and angles were permitted, but structures that systematically violated standard geometric constraints were excluded. For all subclasses except IIe, histograms were generated from structures contained in the coil library to establish reasonable cutoffs. For class IIe, the program PROCHECK (Laskowski et al. 1993a) was used to identify anomalous geometry.

Class III: Structures lacking hydrogen bond satisfaction. The program HBPLUS (McDonald and Thornton 1994) with Kortemme criteria (Kortemme et al. 2003) was used to identify hydrogen bonds in both simulations and in experimental structures. Structures with a local side chain to backbone hydrogen bond were classified as type IIIa. Such conformations are possible in peptides with side chains that can participate in hydrogen bonds, but not in polyalanine. Solvent-inaccessible structures lacking a backbone hydrogen bond are unlikely (Fleming and Rose 2005) and were classified as type IIIb. Finally, proline imino nitrogens, which cannot be hydrogen bond donors, were classified as type IIIc.

Class IV: Structures lacking electron density. For structures that could not otherwise be rationalized, electron density maps were downloaded from the electron density server (Kleywegt et al. 2004), normalized using MAPMAN (Kleywegt and Jones 1996), and visualized with O (Jones et al. 1991) or PyMOL (DeLano 2002). Structures lacking density at the 1.0 sigma level were classified as type IV. When structure factors were unavailable, the PDB headers were interrogated for a crystallographer's note about poor density.

String Simulations

To estimate the way string acceptance ratios scale with chain length, random strings were generated over the five-state alphabet, with relative weights for each state that reflect its frequency of occurrence in the coil library. Strings of length six and greater were assessed using the list of 8,654 disfavored conformations for polyalanine hexamers at the 38% acceptance ratio threshold. Four- and five-residue strings used the corresponding tetramer and pentamer lists of 174 and 1,322 disfavored conformations, respectively. When a disfavored substring was identified, it was accepted or rejected according to its acceptance ratio. If accepted, the substring was flagged and then exempted from application of other conformational restrictions. For each round of simulation, 10^6 strings of length *n* were generated, with five repetitions to assess convergence of the acceptance ratio. A similar method was applied to authentic proteins that were rebuilt from the five-state model, as described above. In this case, however, the protein's conformational string was used in lieu of a randomly generated string.

5.6 Acknowledgments

The authors thank Pat Fleming, Timothy Street, Nick Panasik, and Haipeng Gong for much helpful discussion. Support from the Mathers foundation is gratefully acknowledged.

Table 5.1: The Five-State Model

State	ϕ (^{o})	ψ (^{o})	Observed Residues[1]	Fraction	Relative Fraction[2]
H	-60	-45	28,732	0.071	0.188
E	-120	135	14,410	0.036	0.094
P	-70	140	62,296	0.154	0.407
B	-90	0	35,894	0.089	0.234
L	60	35	11,750	0.029	0.077
Others	N/A	N/A	250,331	0.621	N/A

[1] Observed residues in the coil library, filtered by PISCES, as described in Methods.

[2] Relative fraction for each state with respect to the other states. Other residues are excluded.

Table 5.2: Modeling Real Proteins to the Five-State Model

PDB	Residues	RMSD (Å)[1]	Observed Frequency[2]	Expected Frequency[3]
1VII	36	2.99	7.637×10^{-3}	$(1.3 \pm 1.2) \times 10^{-3}$
2GB1	56	2.43	8.769×10^{-4}	$(2.5 \pm 2.3) \times 10^{-5}$
1UBQ	76	2.15	1.019×10^{-1}	$(4.8 \pm 4.3) \times 10^{-7}$
1LMB	87	1.89	1.079×10^{-4}	$(5.5 \pm 4.9) \times 10^{-8}$
2UBP	100	2.70	1.255×10^{-3}	$(4.2 \pm 3.7) \times 10^{-9}$
1HEL	129	4.47	6.384×10^{-8}	$(1.3 \pm 1.2) \times 10^{-11}$

[1] Backbone atom (N, C_α, C, O, C_β) RMSD of the final protein structure when constrained to the five-state model. RMSD to the native structure was minimized with soft-sphere steric and ϕ, ψ torsion angle restraints as described in Methods.

[2] The calculated acceptance ratio for the protein given its five-state string.

[3] Expected acceptance ratio for a randomly sampled five-state string of comparable size. Errors are calculated using standard propagation of error formulas on equation (5.1).

Table 5.3: Simulation Statistics and Peptide Length

Size (N)	3	4	5	6
Possible Conformations	125	625	3,125	15,625
Unfavorable Conformations				
Total[†]	17 (13.6%)	174 (27.8%)	1,322 (42.3%)	8,654 (55.4%)
Predicted from N-1[‡]	0 (0.0%)	168 (96.6%)	1,253 (94.8%)	8,472 (97.9%)
Predicted but not observed[‡]	0 (0.0%)	5 (2.9%)	49 (3.7%)	243 (2.8%)
Observed but not predicted[‡]	17 (100.0%)	11 (6.3%)	118 (8.9%)	425 (4.9%)

[†] Percentages calculated with respect to the total number of conformations.

[‡] Percentages calculated with respect to the number of unfavorable conformations.

Table 5.4: Rationalization Classes

Type	Description
I	Fragment contains a backbone steric clash at the 90% hard sphere scaling level.
IIa	One or more omega (ω) torsions deviate more than 10 degrees from planarity. Typically a *cis*-peptide bond.
IIb	Two or more omega (ω) torsions deviate more than 5 degrees from planarity.
IIc	One or more tau (τ) scalar angles lie outside of 111 ± 10 degrees (4 standard deviations).
IId	Two or more tau (τ) scalar angles lie outside of 111 ± 5 degrees (2 standard deviations).
IIe	PROCHECK program reports three or more geometric parameters that differ by two or more standard deviations from the ideal values.
IIIa	Local side chain satisfies an otherwise inaccessible backbone hydrogen bond donor or acceptor.
IIIb	Backbone hydrophilic atom is totally masked from solvent and protein hydrogen bond partners.
IIIc	Proline residue at an otherwise unsatisfiable N-H bond donor.
IV	No electron density is observed for the backbone at the 1.0 sigma level.

Table 5.5: A Sample of Unfavorable Conformations in Proteins

Conformational			Violation		Times	Times	Figure 5.6
String	Acceptance Ratio[1]	Rank[2]	Description	Frequency[3]	Observed[4]	Rationalized[5]	Label[6]
HHEB	$(2.66 \pm 0.29) \times 10^{-2}$	1	Clash between O_{i-1} and O_{i+2}	82.1%	2	2	A
			Unsolvated N_{i+4}	17.0%			
HBEB	$(3.54 \pm 0.14) \times 10^{-2}$	2	Clash between O_{i-1} and O_{i+2}	66.5%	2	2	B
			Unsolvated N_{i+4}	30.2%			
EPBH	$(3.65 \pm 0.29) \times 10^{-2}$	3	Unsolvated N_{i+3}	94.4%	5	4	C
HPBB	$(6.79 \pm 0.41) \times 10^{-2}$	17	Unsolvated N_{i+3}	95.6%	0	N/A	D
PBPB	$(7.64 \pm 0.28) \times 10^{-2}$	23	Unsolvated N_{i+4}	98.5%	2	2	E
			Unsolvated N_{i+2}	90.3%			
PPPB	$(7.97 \pm 0.34) \times 10^{-2}$	30	Unsolvated N_{i+4}	99.8%	23	22	F
HHHE	$(1.197 \pm 0.035) \times 10^{-1}$	102	Clash between O_{i-1} and O_{i+2}	90.1%	1	1	G
HBLB	$(1.546 \pm 0.057) \times 10^{-1}$	113	Unsolvated N_{i+4}	91.6%	38	38	H
EHELL	$(3.96 \pm 0.57) \times 10^{-3}$	1	Clash between CB_i and CA_{i+5}	87.9%	0	N/A	I
			Clash between CB_i and N_{i+5}	86.3%			
HPLLP	$(3.26 \pm 0.34) \times 10^{-2}$	94	Clash between O_{i-1} and CB_{i+4}	90.7%	0	N/A	J
			Clash between O_{i-1} and CA_{i+4}	89.1%			

[1] Simulation acceptance ratio averaged over five trials.

[2] Conformational rank when ordered by increasing acceptance ratio for four or five residue fragments.

[3] For rejected structures, the frequency of occurrence for each violation. For highly strained conformations, multiple violations may occur at the same time, though, at most, two are listed. More details are available in the supplemental data. Any steric clash preempts further investigation of solvation violations.

[4] The number of times this conformation is observed in the PISCES-filtered coil library.

[5] As described in the text, the number of observations in the coil library that can be rationalized as being outside the scope of the model.

[6] For the structural examples in figure 5.6, the label for this particular conformation.

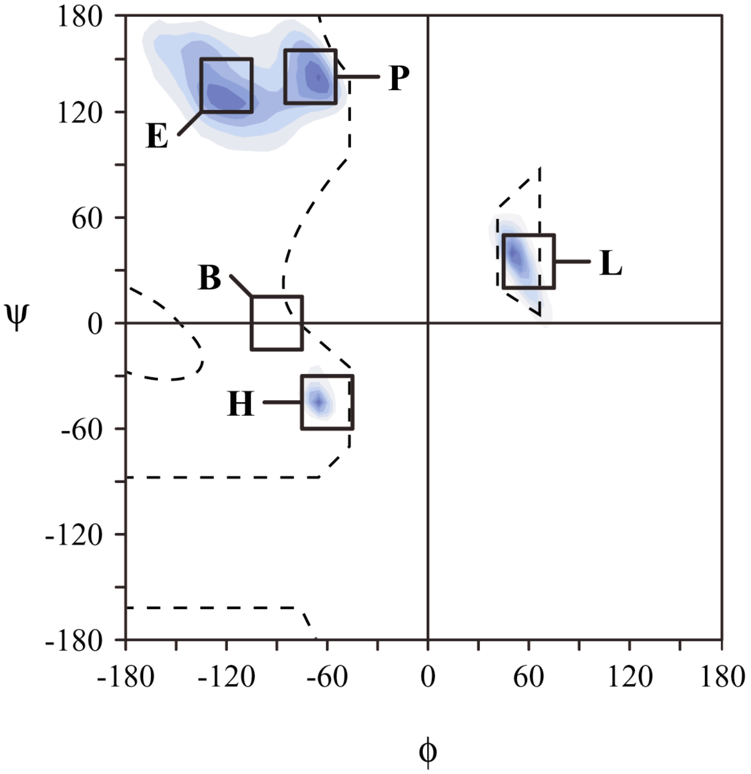


Figure 5.1. Labeled ϕ,ψ bins used in the five-state model, overlaid on contour plots of the extended, helix, and left-handed helical regions, using data from the coil library. Each bin is 30^{o} x 30^{o}, centered on the coordinate position listed in table 5.1. The dashed outline represents the conventional Ramachandran plot for an alanine dipeptide (Ramachandran and Sasisekharan 1968). In our simulations, hard sphere radii were smaller than those used in the original Ramachandran plot, resulting in an expansion of the sterically allowed region. All five bins are fully allowed.

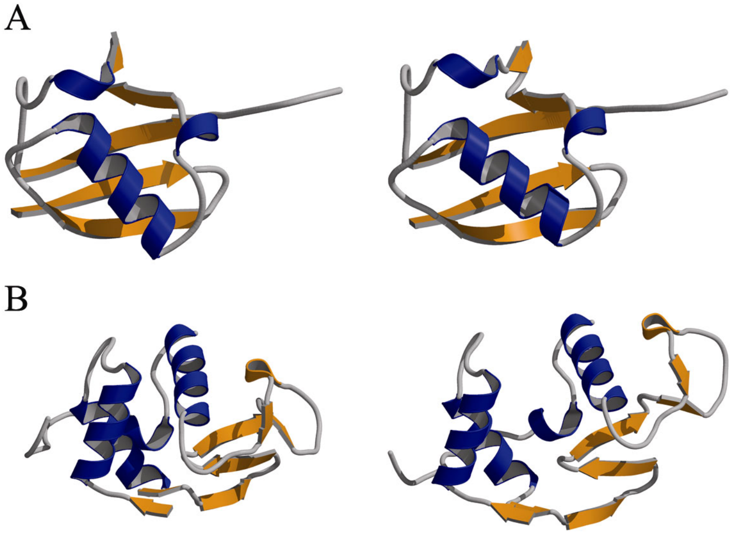


Figure 5.2. X-ray (left) vs. five-state (right) structures of (A) ubiquitin (1UBQ) and (B) lysozyme (1HEL). Five-state structures were obtained as described in Methods. RMS differences between the experimental structure and its five-state model were small (table 5.2). Even for lysozyme – a worst case – the experimental structure is largely captured by its five-state model, except for a two-residue segment that is responsible for the hinge-like opening of the structural core.

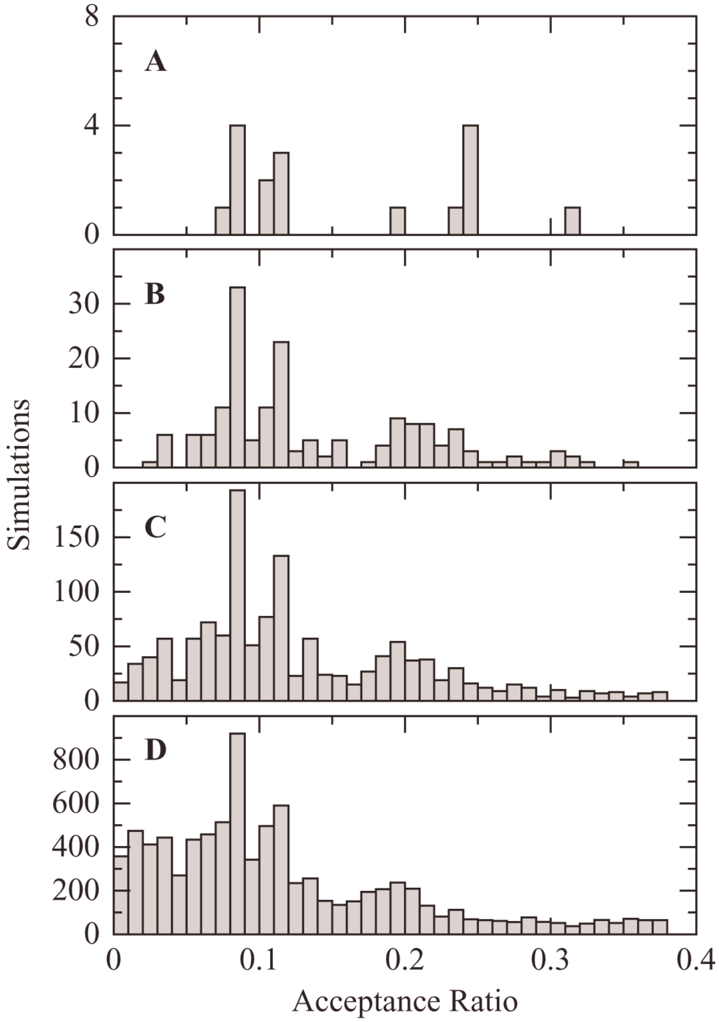


Figure 5.3. Histograms of acceptance ratios for disfavored conformations, ranging from trimers to hexamers: (A) *N = 3* (B) *N = 4* (C) *N = 5* (D) *N = 6*. As peptide length increases, the number of highly disfavored conformations (acceptance ratio approaching zero) also increases.

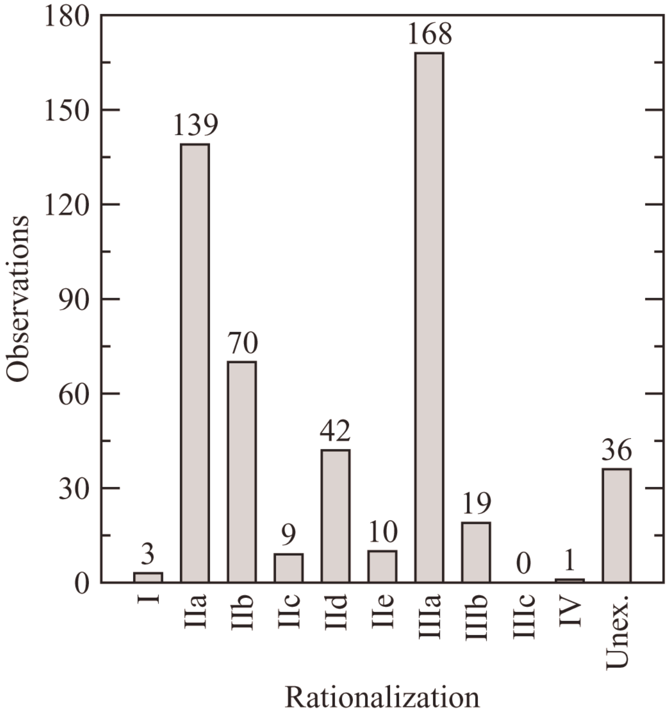


Figure 5.4. Bar graph showing the distribution of the 497 disfavored tetrameric fragments across the 10 rationalization classes (table 5.4) and an 11th, unexplained category. The two predominant reasons why a conformation is disfavored in simulations but found in the coil library are the presence of a *cis*-peptide bond (IIa) or a backbone-side chain hydrogen bond (IIIa), neither of which are included in our simplified model.

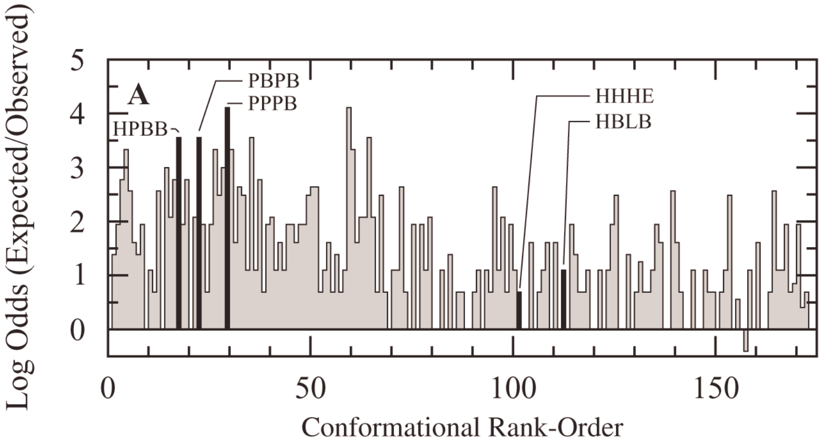
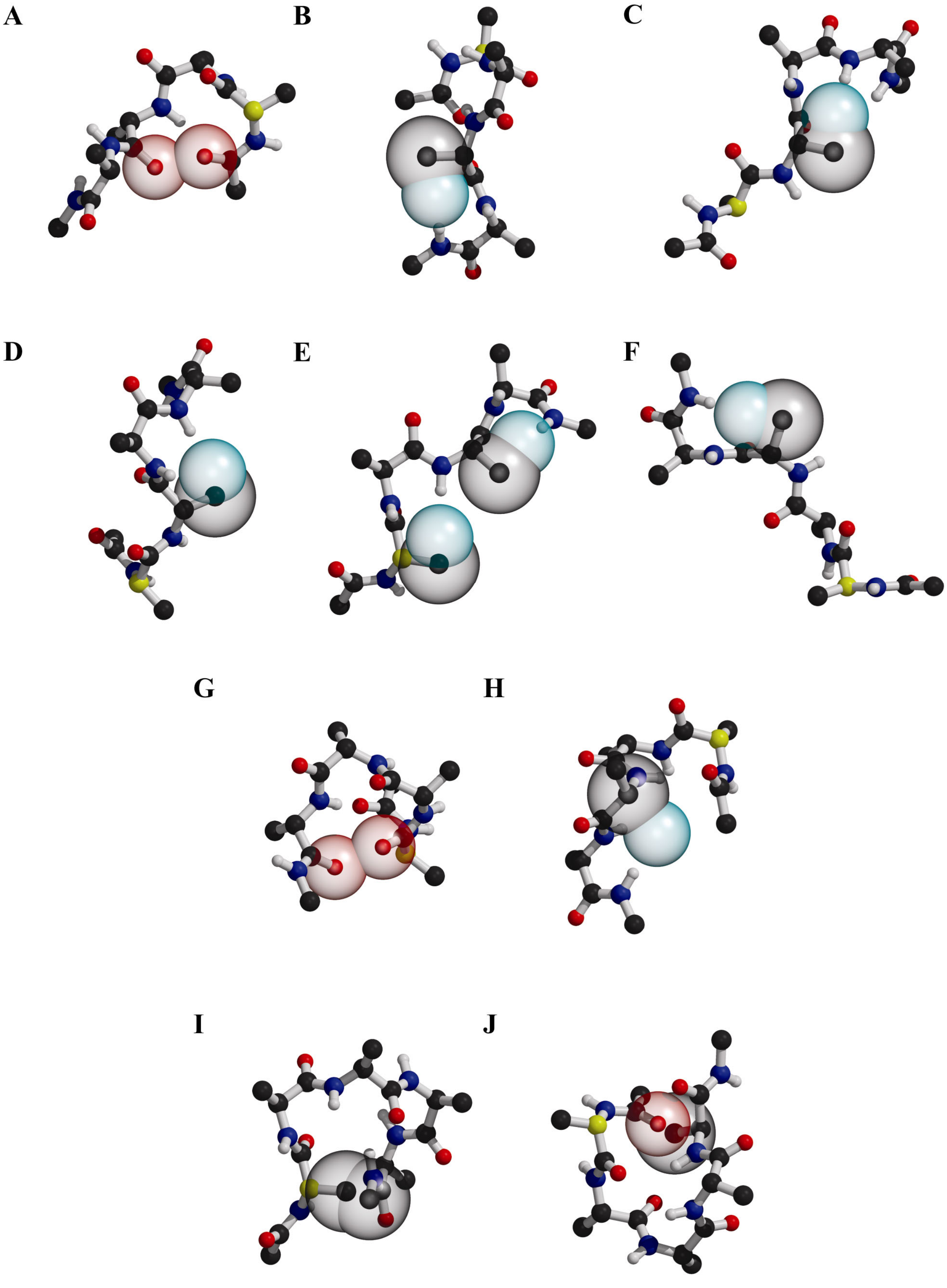


Figure 5.5. Log Odds ratios – log(expected/observed) – of the 174 restricted tetramer conformations, ordered by acceptance ratio. The expected number of structures was calculated from its frequency of occurrence in the coil library, assuming that each residue assorts independently. The log odds ratio is positive when the number of expected structures exceeds the number actually observed. Observed conformations were counted after removing structures with *cis*-peptide bonds, backbone-side chain hydrogen bonds, and other rationalizations from table 5.4. Conformations used as examples in figure 5.6 are shown as annotated, dark bars. To avoid log(0), conformations with a frequency of zero were assigned a value of unity. Only one conformation, BLBE, occurs more frequently than expected by chance (*i.e.* negative log odds ratio).

Figure 5.6. Examples of disfavored conformations identified in simulations (key given in table 5.5). For each structure, hard sphere collisions are displayed as overlapping, semitransparent CPK spheres. Virtual waters hydrogen-bonded to backbone N-H atoms are displayed in cyan; the i^{th} C_α carbon is shown in yellow. Images were generated with MOLSCRIPT (Kraulis 1991) and Raster3D (Merritt and Bacon 1997).


A
B
C
D
E
F
G
H
I
J

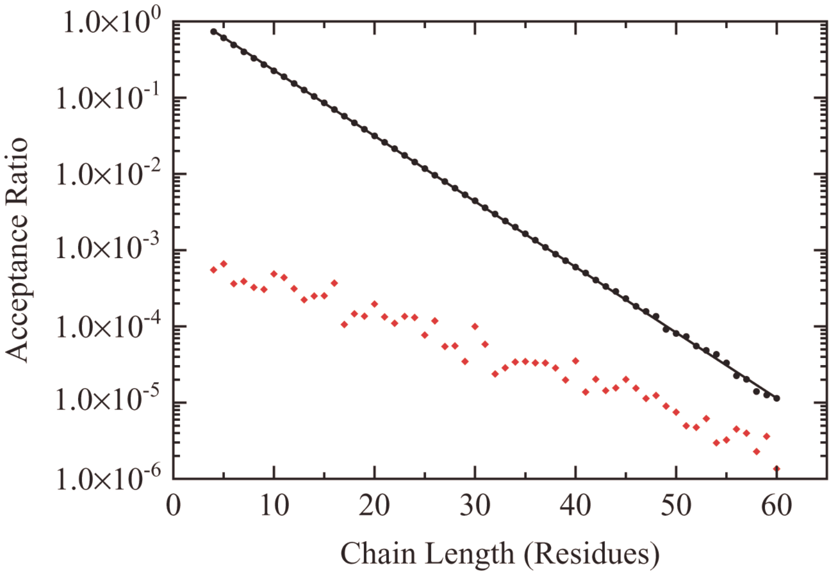


Figure 5.7. Acceptance ratios from string simulations for strings ranging from *N = 4,60*. Acceptance ratios for strings are based on local steric and hydrogen bond interactions derived from five-state, hard sphere simulations. Black dots plot the average acceptance ratio from five separate simulations; red dots represent the standard deviation of this average. Convergence for all the points is at least an order of magnitude smaller than the values themselves, and often several orders of magnitude smaller.

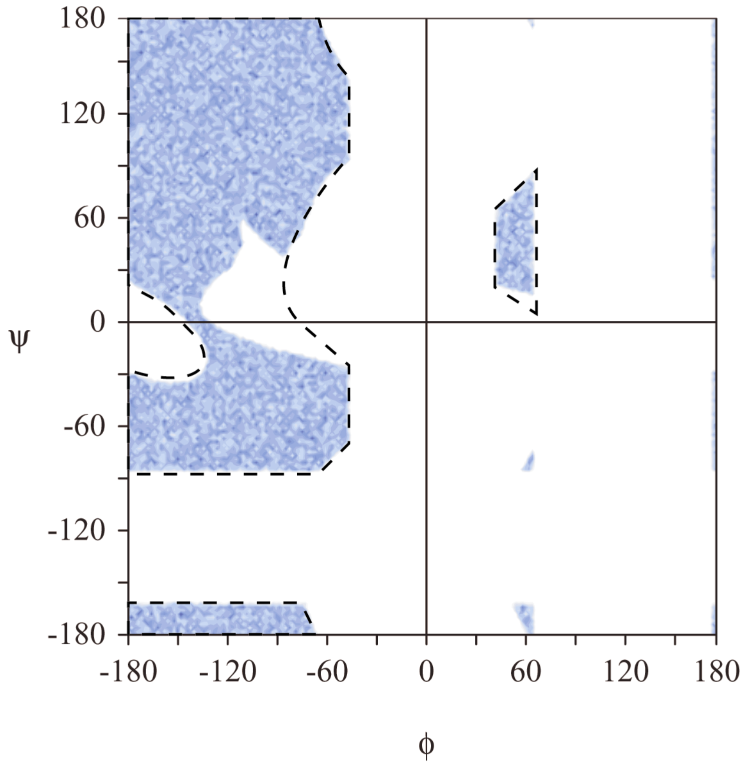


Figure 5.8. ϕ,ψ-plot of the blocked alanyl dipeptide, N-acetyl-Ala-N-methylamide. Allowed conformations (in blue) were derived from simulations (see Methods) and are based on hard sphere sterics and hydrogen bond satisfaction, with radii scaled to 95%. The conventional Ramachandran plot (Ramachandran and Sasisekharan 1968), shown as a dashed line, is based solely on hard sphere sterics. In comparison, this plot, which rejects conformations with solvent-inaccessible polar groups, has a large missing section in the bridge region and a distinct peninsula below the β-strand/P_{II} region.

REFERENCES

Ackerman, M.S., and Shortle, D. 2002a. Molecular alignment of denatured states of staphylococcal nuclease with strained polyacrylamide gels and surfactant liquid crystalline phases. *Biochemistry* **41:** 3089-3095.

Ackerman, M.S., and Shortle, D. 2002b. Robustness of the long-range structure in denatured staphylococcal nuclease to changes in amino acid sequence. *Biochemistry* **41:** 13791-13797.

Ahmad, F., and Bigelow, C.C. 1982. Estimation of the free energy of stabilization of ribonuclease A, lysozyme, alpha-lactalbumin, and myoglobin. *J Biol Chem* **257:** 12935-12938.

Alexandrescu, A.T., Abeygunawardana, C., and Shortle, D. 1994. Structure and dynamics of a denatured 131-residue fragment of staphylococcal nuclease: a heteronuclear NMR study. *Biochemistry* **33:** 1063-1072.

Anderson, A.G., and Hermans, J. 1988. Microfolding: conformational probability map for the alanine dipeptide in water from molecular dynamics simulations. *Proteins* **3:** 262-265.

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

Aune, K.C., Salahuddin, A., Zarlengo, M.H., and Tanford, C. 1967. Evidence for residual structure in acid- and heat-denatured proteins. *J Biol Chem* **242:** 4486-4489.

Aurora, R., Creamer, T.P., Srinivasan, R., and Rose, G.D. 1997. Local interactions in protein folding: Lessons from the α-helix. *J Biol Chem* **272:** 1413-1416.

Aurora, R., and Rose, G.D. 1998. Helix Capping. *Protein Science* **7:** 21-38.

Auton, M., and Bolen, D.W. 2004. Additive transfer free energies of the peptide backbone unit that are independent of the model compound and the choice of concentration scale. *Biochemistry* **43:** 1329-1342.

Avbelj, F., and Baldwin, R.L. 2003. Role of backbone solvation and electrostatics in generating preferred peptide backbone conformations: distributions of phi. *Proc Natl Acad Sci U S A* **100:** 5742-5747.

Avbelj, F., and Baldwin, R.L. 2004. Origin of the neighboring residue effect on peptide backbone conformation. *Proc Natl Acad Sci U S A* **101:** 10967-10972.

Baldwin, R.L., and Rose, G.D. 1999a. Is protein folding hierarchic? I. Local structure and peptide folding. *Trends Biochem Sci* **24:** 26-33.

Baldwin, R.L., and Rose, G.D. 1999b. Is protein folding hierarchic? II. Folding intermediates and transition states. *Trends Biochem Sci* **24:** 77-83.

Baldwin, R.L., and Zimm, B.H. 2000. Are denatured proteins ever random coils? *Proc Natl Acad Sci U S A* **97:** 12391-12392.

Banavar, J.R., Hoang, T.X., Maritan, A., Seno, F., and Trovato, A. 2004. Unified perspective on proteins: a physics approach. *Phys Rev E Stat Nonlin Soft Matter Phys* **70:** 041905.

Barrick, D., and Baldwin, R.L. 1993. Three-state analysis of sperm whale apomyoglobin folding. *Biochemistry* **32:** 3790-3796.

Becktel, W.J., and Schellman, J.A. 1987. Protein stability curves. *Biopolymers* **26:** 1859-1877.

Ben-Tal, N., Sitkoff, D., Topol, I.A., Yang, A.S., Burt, S.K., and Honig, B. 1997. Free energy of amide hydrogen bond formation in vacuum, in water, and in liquid alkane solution. *J Phys Chem B* **101:** 450-457.

Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., and Hermans, J. 1981. Interaction models for water in relation to protein hydration. In *Intermolecular Forces*. (ed. B. Pullman), pp. 331-342. D. Reidel Publishing Company, Boston, MA.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res* **28:** 235-242.

Bevington, P.R., and Robinson, D.K. 1992. *Data Reduction and Error Analysis for the Physical Sciences*, 2nd ed. WCB/McGraw Hill, New York.

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

Bower, M.J., Cohen, F.E., and Dunbrack, R.L., Jr. 1997. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *J Mol Biol* **267:** 1268-1282.

Brant, D.A., and Flory, P.J. 1965a. Configuration of Random Polypeptide Chains. I. Experimental Results. *Journal of the American Chemical Society* **87:** 2788-&.

Brant, D.A., and Flory, P.J. 1965b. Configuration of Random Polypeptide Chains. II. Theory. *Journal of the American Chemical Society* **87:** 2791-&.

Bromberg, S., and Dill, K.A. 1994. Side-chain entropy and packing in proteins. *Protein Sci* **3:** 997-1009.

Burton, R.E., Myers, J.K., and Oas, T.G. 1998. Protein folding dynamics: quantitative comparison between theory and experiment. *Biochemistry* **37:** 5337-5343.

Butterfoss, G.L., and Hermans, J. 2003. Boltzmann-type distribution of side-chain conformation in proteins. *Protein Sci* **12:** 2719-2731.

Calmettes, P., Roux, B., Durand, D., Desmadril, M., and Smith, J.C. 1993. Configurational distribution of denatured phosphoglycerate kinase. *J Mol Biol* **231:** 840-848.

Cantor, C.R., and Schimmel, P.R. 1980. *Biophysical Chemistry. Part III The Behavior of Biological Macromolecules*. Freeman, New York.

Chan, H.S., and Dill, K.A. 1991. Polymer principles in protein structure and stability. *Annu Rev Biophys Biophys Chem* **20:** 447-490.

Chellgren, B.W., and Creamer, T.P. 2004. Effects of H2O and D2O on polyproline II helical structure. *J Am Chem Soc* **126:** 14734-14735.

Chen, K., Liu, Z., and Kallenbach, N.R. 2004. The polyproline II conformation in short alanine peptides is noncooperative. *Proc Natl Acad Sci U S A* **101:** 15352-15357.

Chothia, C. 1974. Hydrophobic bonding and accessible surface area in proteins. *Nature* **248:** 338-339.

Cohen, F.E., and Sternberg, M.J. 1980. On the prediction of protein structure: The significance of the root-mean-square deviation. *J Mol Biol* **138:** 321-333.

Creamer, T.P., and Campbell, M.N. 2002. Determinants of the polyproline II helix from modeling studies. *Adv Protein Chem* **62:** 263-282.

Creamer, T.P., Srinivasan, R., and Rose, G.D. 1995. Modeling unfolded states of peptides and proteins. *Biochemistry* **34:** 16245-16250.

Creamer, T.P., Srinivasan, R., and Rose, G.D. 1997. Modeling unfolded states of proteins and peptides. II. Backbone solvent accessibility. *Biochemistry* **36:** 2832-2835.

Creighton, T.E. 1984. *Proteins: Structures and Molecular Principles*. W. H. Freeman and Company, New York, pp. 515.

Daggett, V., and Fersht, A. 2003. The present view of the mechanism of protein folding. *Nat Rev Mol Cell Biol* **4:** 497-502.

de Gennes, P.-G. 1979. *Scaling concepts in polymer physics*. Cornell University Press, Ithaca.

DeLano, W.L. 2002. Thy PyMOL Molecular Graphics System.

Dill, K.A. 1985. Theory for the folding and stability of globular proteins. *Biochemistry* **24:** 1501-1509.

Dill, K.A. 1990. Dominant forces in protein folding. *Biochemistry* **29:** 7133-7155.

Dill, K.A. 1999. Polymer principles and protein folding. *Protein Sci* **8:** 1166-1180.

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

Dill, K.A., and Shortle, D. 1991. Denatured states of proteins. *Annu Rev Biochem* **60:** 795-825.

Dill, K.A., and Stigter, D. 1995. Modeling protein stability as heteropolymer collapse. *Adv Protein Chem* **46:** 59-104.

Dinner, A.R., and Karplus, M. 2001. Comment on the Communication "The Key to Solving the Protein-Folding Problem Lies in an Accurate Description of the Denatured State" by van Gunsteren et al. *Angew Chem Int Ed Engl* **40:** 4615-4616.

Doniach, S. 2001. Changes in biomolecular conformation seen by small angle X-ray scattering. *Chem Rev* **101:** 1763-1778.

Drozdov, A.N., Grossfield, A., and Pappu, R.V. 2004. Role of solvent in determining conformational preferences of alanine dipeptide in water. *J Am Chem Soc* **126:** 2574-2581.

Duarte, C.M., and Pyle, A.M. 1998. Stepping through an RNA structre: a novel approach to conformational analysis. *J. Mol. Biol.* **284:** 1465-1478.

Dunbrack, R.L., Jr., and Karplus, M. 1994. Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. *Nat Struct Biol* **1:** 334-340.

Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.M., Hipps, K.W., et al. 2001. Intrinsically disordered protein. *J Mol Graph Model* **19:** 26-59.

Eker, F., Griebenow, K., Cao, X., Nafie, L.A., and Schweitzer-Stenner, R. 2004. Preferred peptide backbone conformations in the unfolded state revealed by the structure analysis of alanine-based (AXA) tripeptides in aqueous solution. *Proc Natl Acad Sci U S A* **101:** 10054-10059.

Eriksson, A.E., Baase, W.A., Zhang, X.J., Heinz, D.W., Blaber, M., Baldwin, E.P., and Matthews, B.W. 1992. Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science* **255:** 178-183.

Ferguson, N., Schartau, P.J., Sharpe, T.D., Sato, S., and Fersht, A.R. 2004. One-state downhill versus conventional protein folding. *J Mol Biol* **344:** 295-301.

Fernandez, D.P., Mulev, Y., Goodwin, A.R.H., and Sengers, J.M.H.L. 1995. A Database for the Static Dielectric-Constant of Water and Steam. *J Phys Chem Ref Data* **24:** 33-69.

Ferreon, J.C., and Hilser, V.J. 2003. The effect of the polyproline II (PPII) conformation on the denatured state entropy. *Protein Sci* **12:** 447-457.

Fersht, A.R. 2000. 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 U S A* **97:** 1525-1529.

Fitzkee, N.C., Fleming, P.J., and Rose, G.D. 2005. The Protein Coil Library: a structural database of nonhelix, nonstrand fragments derived from the PDB. *Proteins* **58:** 852-854.

Fitzkee, N.C., and Rose, G.D. 2004a. Reassessing random-coil statistics in unfolded proteins. *Proc Natl Acad Sci U S A* **101:** 12497-12502.

Fitzkee, N.C., and Rose, G.D. 2004b. Steric restrictions in protein folding: an alpha-helix cannot be followed by a contiguous beta-strand. *Protein Sci* **13:** 633-639.

Fleming, P.J., Fitzkee, N.C., Mezei, M., Srinivasan, R., and Rose, G.D. 2005. A novel method reveals that solvent water favors polyproline II over beta-strand conformation in peptides and unfolded proteins: conditional hydrophobic accessible surface area (CHASA). *Protein Sci* **14:** 111-118.

Fleming, P.J., and Rose, G.D. 2005. Do all backbone polar groups in proteins form hydrogen bonds? *Protein Sci* **14:** 1911-1917.

Flory, P.J. 1953. *Principles of Polymer Chemistry*. Cornell University Press, Ithaca, NY, pp. 672.

Flory, P.J. 1969. *Statistical Mechanics of Chain Molecules*. John Wiley & Sons, Inc., New York.

Frieden, C. 2003. The kinetics of side chain stabilization during protein folding. *Biochemistry* **42:** 12439-12446.

Garcia, A.E. 2004. Characterization of non-alpha helical conformations in Ala peptides. *Polymer* **45:** 669-676.

Garcia-Mira, M.M., Sadqi, M., Fischer, N., Sanchez-Ruiz, J.M., and Muñoz, V. 2002. Experimental identification of downhill protein folding. *Science* **298:** 2191-2195.

Gianni, S., Guydosh, N.R., Khan, F., Caldas, T.D., Mayor, U., White, G.W., DeMarco, M.L., Daggett, V., and Fersht, A.R. 2003. Unifying features in protein-folding mechanisms. *Proc Natl Acad Sci U S A* **100:** 13286-13291.

Gillespie, J.R., and Shortle, D. 1997a. Characterization of long-range structure in the denatured state of staphylococcal nuclease. I. Paramagnetic relaxation enhancement by nitroxide spin labels. *J Mol Biol* **268:** 158-169.

Gillespie, J.R., and Shortle, D. 1997b. Characterization of long-range structure in the denatured state of staphylococcal nuclease. II. Distance restraints from paramagnetic relaxation and calculation of an ensemble of structures. *J Mol Biol* **268:** 170-184.

Go, N. 1984. The consistency principle in protein structure and pathways of folding. *Adv Biophys* **18:** 149-164.

Goldenberg, D.P. 2003. Computational simulation of the statistical properties of unfolded proteins. *J Mol Biol* **326:** 1615-1633.

Gong, H., Isom, D.G., Srinivasan, R., and Rose, G.D. 2003. Local secondary structure content predicts folding rates for simple, two-state proteins. *J Mol Biol* **327:** 1149-1154.

Gong, H., and Rose, G.D. 2005. Does secondary structure determine tertiary structure in proteins? *Proteins, Structure, Function and Bioinformatics***:** In Press.

Greene, R.F., Jr., and Pace, C.N. 1974. Urea and guanidine hydrochloride denaturation of ribonuclease, lysozyme, alpha-chymotrypsin, and beta-lactoglobulin. *J Biol Chem* **249:** 5388-5393.

Guillot, B. 2002. A reappraisal of what we have learnt during three decades of computer simulations on water. *J Mol Liq* **101:** 219-260.

Gunasekaran, K., Tsai, C.J., and Nussinov, R. 2004. Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. *J Mol Biol* **341:** 1327-1341.

Hill, T.L. 1960. *An Introduction to Statistical Thermodynamics*. Dover Publications, Inc., New York, pp. 508.

Ho, B.K., Thomas, A., and Brasseur, R. 2003. Revisiting the Ramachandran plot: hard-sphere repulsion, electrostatics, and H-bonding in the alpha-helix. *Protein Sci* **12:** 2508-2522.

Hoang, T.X., Trovato, A., Seno, F., Banavar, J.R., and Maritan, A. 2004. Geometry and symmetry presculpt the free-energy landscape of proteins. *Proc Natl Acad Sci U S A* **101:** 7960-7964.

Hobohm, U., and Sander, C. 1994. Enlarged representative set of protein structures. *Protein Science* **3:** 522-524.

Holmes, J.B., and Tsai, J. 2004. Some fundamental aspects of building protein structures from fragment libraries. *Protein Sci* **13:** 1636-1650.

Hopfinger, A.J. 1973. *Conformational properties of macromolecules*. Academic Press, New York.

Hoshino, M., Hagihara, Y., Hamada, D., Kataoka, M., and Goto, Y. 1997. Trifluoroethanol-induced conformational transition of hen egg-white lysozyme studied by small-angle X-ray scattering. *FEBS Lett* **416:** 72-76.

Hovmöller, S., Zhou, T., and Ohlson, T. 2002. Conformations of amino acids in proteins. *Acta Crystallogr D Biol Crystallogr* **58:** 768-776.

Hu, H., Elstner, M., and Hermans, J. 2003. Comparison of a QM/MM force field and molecular mechanics force fields in simulations of alanine and glycine "dipeptides" (Ace-Ala-Nme and Ace-Gly-Nme) in water in relation to the problem of modeling the unfolded peptide backbone in solution. *Proteins* **50:** 451-463.

Hummer, G., Garcia, A.E., and Garde, S. 2001. Helix nucleation kinetics from molecular simulations in explicit solvent. *Proteins* **42:** 77-84.

IUPAC-IUB. 1970. IUPAC-IUB Commission on Biochemical Nomenclature. Abbreviations and symbols for the description of the conformation of polypeptide chains. *J Mol Biol* **52:** 1-17.

Ivankov, D.N., and Finkelstein, A.V. 2004. Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. *Proc Natl Acad Sci U S A* **101:** 8942-8944.

Jackson, S.E., and Fersht, A.R. 1991. Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition. *Biochemistry* **30:** 10428-10435.

Jha, A., Colubri, A., Freed, K.F., and Sosnick, T.R. 2005. Statistical coil model for the unfolded state: Resolving the reconciliation problem. *Proc Natl Acad Sci U S A*: In Press.

Jones, T.A., Zou, J.Y., Cowan, S.W., and Kjeldgaard. 1991. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr A* **47 (Pt 2):** 110-119.

Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., and Klein, M.L. 1983. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **79:** 926-935.

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

Karplus, M., and Weaver, D.L. 1976. Protein-folding dynamics. *Nature* **260:** 404-406.

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

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

Kauzmann, W. 1959. Some factors in the interpretation of protein denaturation. *Adv Protein Chem* **14:** 1-63.

Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H., and Phillips, D.C. 1958. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **181:** 662-666.

Kentsis, A., Mezei, M., Gindin, T., and Osman, R. 2004. Unfolded state of polyalanine is a segmented polyproline II helix. *Proteins* **55:** 493-501.

Khorasanizadeh, S., Peters, I.D., Butt, T.R., and Roder, H. 1993. Folding and stability of a tryptophan-containing mutant of ubiquitin. *Biochemistry* **32:** 7054-7063.

Kleywegt, G.J., Harris, M.R., Zou, J.Y., Taylor, T.C., Wahlby, A., and Jones, T.A. 2004. The Uppsala Electron-Density Server. *Acta Crystallogr D Biol Crystallogr* **60:** 2240-2249.

Kleywegt, G.J., and Jones, T.A. 1996. xdlMAPMAN and xdlDATAMAN - programs for reformatting, analysis and manipulation of biomacromolecular electron-density maps and reflection data sets. *Acta Crystallogr D Biol Crystallogr* **52:** 826-828.

Klotz, I.M., and Franzen, J.S. 1962. Hydrogen bonds between model peptide groups in solution. *J. Am. Chem. Soc.* **84:** 3461-3466.

Kohn, J.E., Millett, I.S., Jacob, J., Zagrovic, B., Dillon, T.M., Cingel, N., Dothager, R.S., Seifert, S., Thiyagarajan, P., Sosnick, T.R., et al. 2004. Random-coil behavior and the dimensions of chemically unfolded proteins. *Proc Natl Acad Sci U S A* **101:** 12491-12496.

Koltun, W.L. 1965. Precision space-filling atomic models. *Biopolymers* **3:** 665-679.

Kortemme, T., Morozov, A.V., and Baker, D. 2003. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J Mol Biol* **326:** 1239-1259.

Kossiakoff, A.A., Shpungin, J., and Sintchak, M.D. 1990. Hydroxyl hydrogen conformations in trypsin determined by the neutron diffraction solvent difference map method: relative importance of steric and electrostatic factors in defining hydrogen-bonding geometries. *Proc Natl Acad Sci U S A* **87:** 4468-4472.

Kraulis, P.J. 1991. Molscript - a Program to Produce Both Detailed and Schematic Plots of Protein Structures. *J Appl Crystallogr* **24:** 946-950.

Krishna, M.M., and Englander, S.W. 2005. The N-terminal to C-terminal motif in protein folding and function. *Proc Natl Acad Sci U S A* **102:** 1053-1058.

Laskowski, R.A., Macarthur, M.W., Moss, D.S., and Thornton, J.M. 1993a. Procheck - a Program to Check the Stereochemical Quality of Protein Structures. *J Appl Crystallogr* **26:** 283-291.

Laskowski, R.A., Macarthur, M.W., Moss, D.S., and Thornton, J.M. 1993b. PROCHECK: A Program to Check the Stereochemical Quality of Protein Structures. *J Appl Crystallogr* **26:** 283-291.

Lazaridis, T., and Karplus, M. 1997. "New view" of protein folding reconciled with the old through multiple unfolding simulations. *Science* **278:** 1928-1931.

Le Guillou, J.C., and Zinn-Justin, J. 1977. Critical Exponents for the N-Vector Model in 3 Dimensions from Field Theory. *Phys Rev Lett* **39:** 95-98.

Lee, B. 1991. Solvent reorganization contribution to the transfer thermodynamics of small nonpolar molecules. *Biopolymers* **31:** 993-1008.

Lee, B., and Richards, F.M. 1971. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol* **55:** 379-400.

Levinthal, C. 1969. How to fold graciously. In *Mössbauer Spectroscopy in Biological Sytems*. (eds. P. Debrunner, J.C.M. Tsibris, and E. Münck), pp. 22-24. Univ. of Illinois Press, Urbana.

Levitt, M., and Chothia, C. 1976. Structural patterns in globular proteins. *Nature* **261:** 552-558.

Levitt, M.H. 2001. *Spin Dynamics*. John Wiley & Sons, Ltd., New York, pp. 686.

Li, A.J., and Nussinov, R. 1998. A set of van der Waals and coulombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking. *Proteins* **32:** 111-127.

Lifson, S., and Roig, A. 1961. On the theory of helix-coil transition in polypeptides. *J. Chem. Phys.* **34:** 1963-1974.

Lindorff-Larsen, K., Kristjansdottir, S., Teilum, K., Fieber, W., Dobson, C.M., Poulsen, F.M., and Vendruscolo, M. 2004. Determination of an ensemble of structures representing the denatured state of the bovine acyl-coenzyme a binding protein. *J Am Chem Soc* **126:** 3291-3299.

Liu, Z., Chen, K., Ng, A., Shi, Z., Woody, R.W., and Kallenbach, N.R. 2004. Solvent dependence of PII conformation in model alanine peptides. *J Am Chem Soc* **126:** 15141-15150.

Louhivuori, M., Fredriksson, K., Paakkonen, K., Permi, P., and Annila, A. 2004. Alignment of chain-like molecules. *J Biomol NMR* **29:** 517-524.

Louhivuori, M., Paakkonen, K., Fredriksson, K., Permi, P., Lounila, J., and Annila, A. 2003. On the origin of residual dipolar couplings from denatured proteins. *J Am Chem Soc* **125:** 15647-15650.

MacKerell, A.D., Bashford, D., Bellott, M., Dunbrack, R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., et al. 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* **102:** 3586-3616.

Maity, H., Maity, M., Krishna, M.M., Mayne, L., and Englander, S.W. 2005. Protein folding: the stepwise assembly of foldon units. *Proc Natl Acad Sci U S A* **102:** 4741-4746.

Makarov, D.E., and Plaxco, K.W. 2003. The topomer search model: A simple, quantitative theory of two-state protein folding kinetics. *Protein Sci* **12:** 17-26.

Margulis, C.J., Stern, H.A., and Berne, B.J. 2002. Helix unfolding and intramolecular hydrogen bond dynamics in small alpha-helices in explicit solvent. *J. Phys. Chem. B* **106:** 10748-10752.

Matouschek, A. 2003. Protein unfolding--an important process in vivo? *Curr Opin Struct Biol* **13:** 98-109.

Matthews, C.R. 1987. Effect of point mutations on the folding of globular proteins. *Methods Enzymol* **154:** 498-511.

Maxwell, K.L., Wildes, D., Zarrine-Afsar, A., De Los Rios, M.A., Brown, A.G., Friel, C.T., Hedberg, L., Horng, J.C., Bona, D., Miller, E.J., et al. 2005. Protein folding: defining a "standard" set of experimental conditions and a preliminary kinetic data set of two-state proteins. *Protein Sci* **14:** 602-616.

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

Merritt, E.A., and Bacon, D.J. 1997. Raster3D: Photorealistic molecular graphics. *Method Enzymol* **277:** 505-524.

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

Mezei, M., Fleming, P.J., Srinivasan, R., and Rose, G.D. 2004. Polyproline II helix is the preferred conformation for unfolded polyalanine in water. *Proteins* **55:** 502-507.

Millett, I.S., Doniach, S., and Plaxco, K.W. 2002. Toward a taxonomy of the denatured state: small angle scattering studies of unfolded proteins. *Adv Protein Chem* **62:** 241-262.

Mirsky, A.E., and Pauling, L. 1936. On the structure of native, denatured, and coagulated proteins. *Proc. Natl. Acad. Sci. USA* **22:** 439-447.

Mohana-Borges, R., Goto, N.K., Kroon, G.J., Dyson, H.J., and Wright, P.E. 2004. Structural characterization of unfolded states of apomyoglobin using residual dipolar couplings. *J Mol Biol* **340:** 1131-1142.

Mu, Y.G., and Stock, G. 2002. Conformational dynamics of trialanine in water: A molecular dynamics study. *J Phys Chem B* **106:** 5294-5301.

Muñoz, V., and Sanchez-Ruiz, J.M. 2004. Exploring protein-folding ensembles: a variable-barrier model for the analysis of equilibrium unfolding experiments. *Proc Natl Acad Sci U S A* **101:** 17646-17651.

Murthy, V.L., Srinivasan, R., Draper, D.E., and Rose, G.D. 1999. A complete conformational map for RNA. *J Mol Biol* **291:** 313-327.

Myers, J.K., Pace, C.N., and Scholtz, J.M. 1995. Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. *Protein Sci* **4:** 2138-2148.

Naganathan, A.N., and Munoz, V. 2005. Scaling of folding times with protein size. *J Am Chem Soc* **127:** 480-481.

Naganathan, A.N., Perez-Jimenez, R., Sanchez-Ruiz, J.M., and Munoz, V. 2005. Robustness of downhill folding: guidelines for the analysis of equilibrium folding experiments on small proteins. *Biochemistry* **44:** 7435-7449.

Nelson, J.W., and Kallenbach, N.R. 1986. Stabilization of the ribonuclease S-peptide alpha-helix by trifluoroethanol. *Proteins: Structure, Function, and Genetics* **1:** 211-217.

Nölting, B. 1999. *Protein Folding Kinetics: Biophysical Methods*. Springer, New York, pp. 190.

Ohkubo, Y.Z., and Brooks, C.L., 3rd. 2003. Exploring Flory's isolated-pair hypothesis: statistical mechanics of helix-coil transitions in polyalanine and the C-peptide from RNase A. *Proc Natl Acad Sci U S A* **100:** 13916-13921.

Ohnishi, S., Lee, A.L., Edgell, M.H., and Shortle, D. 2004. Direct demonstration of structural similarity between native and denatured eglin C. *Biochemistry* **43:** 4064-4070.

Onuchic, J.N., and Wolynes, P.G. 2004. Theory of protein folding. *Curr Opin Struct Biol* **14:** 70-75.

Pace, C.N., Hebert, E.J., Shaw, K.L., Schell, D., Both, V., Krajcikova, D., Sevcik, J., Wilson, K.S., Dauter, Z., Hartley, R.W., et al. 1998. Conformational stability and thermodynamics of folding of ribonucleases Sa, Sa2 and Sa3. *J Mol Biol* **279:** 271-286.

Pace, C.N., and Shaw, K.L. 2000. Linear extrapolation method of analyzing solvent denaturation curves. *Proteins* **Suppl 4:** 1-7.

Panasik, N., Jr., Fleming, P.J., and Rose, G.D. Hydrogen-bonded turns in proteins: the case for a recount. Submitted.

Pappu, R.V., and Rose, G.D. 2002. A simple model for polyproline II structure in unfolded states of alanine-based peptides. *Protein Sci* **11:** 2437-2455.

Pappu, R.V., Srinivasan, R., and Rose, G.D. 2000. The Flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding. *Proc Natl Acad Sci U S A* **97:** 12565-12570.

Pauling, L., and Corey, R.B. 1951. The pleated sheet, a new layer configuration of polypeptide chains. *Proc Natl Acad Sci U S A* **37:** 251-256.

Pauling, L., Corey, R.B., and Branson, H.R. 1951. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci U S A* **37:** 205-211.

Petukhov, M., Rychkov, G., Firsov, L., and Serrano, L. 2004. H-bonding in protein hydration revisited. *Protein Sci* **13:** 2120-2129.

Pilz, I., Glatter, O., and Kratky, O. 1979. Small-angle X-ray scattering. *Methods Enzymol* **61:** 148-249.

Plaxco, K.W., Simons, K.T., and Baker, D. 1998. Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol* **277:** 985-994.

Pletneva, E.V., Gray, H.B., and Winkler, J.R. 2005. Many faces of the unfolded state: conformational heterogeneity in denatured yeast cytochrome C. *J Mol Biol* **345:** 855-867.

Poland, D., and Scheraga, H.A. 1970. *Theory of Helix-Coil Transitions in Biopolymers*. Academic Press, New York, pp. 797.

Press, W.H., Teukolsky, S.A., Vetterline, W.T., and Flannery, B.P. 1992. Chapter 10. Minimization or Maximization of Functions. In *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed, pp. 394-455. Cambridge University Press, New York.

Presta, L.G., and Rose, G.D. 1988. Helix signals in proteins. *Science* **240:** 1632-1641.

Prestegard, J.H., al-Hashimi, H.M., and Tolman, J.R. 2000. NMR structures of biomolecules using field oriented media and residual dipolar couplings. *Q Rev Biophys* **33:** 371-424.

Privalov, P.L. 1979. Stability of proteins: small globular proteins. *Adv Protein Chem* **33:** 167-241.

Privalov, P.L., and Gill, S.J. 1988. Stability of protein structure and hydrophobic interaction. *Adv. Prot. Chem.* **39:** 191-234.

Privalov, P.L., and Khechinashvili, N.N. 1974. A thermodynamic approach to the problem of stabilization of globular protein structure: a calorimetric study. *J Mol Biol* **86:** 665-684.

Przytycka, T., Aurora, R., and Rose, G.D. 1999. A protein taxonomy based on secondary structure. *Nat Struct Biol* **6:** 672-682.

R Development Core Team. 2003. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramachandran, G.N., Ramakrishnan, C., and Sasisekharan, V. 1963. Stereochemistry of polypeptide chain configurations. *J Mol Biol* **7:** 95-99.

Ramachandran, G.N., and Sasisekharan, V. 1968. Conformation of polypeptides and proteins. *Adv Prot Chem* **23:** 283-438.

Ramakrishnan, V., Ranbhor, R., and Durani, S. 2004. Existence of specific "folds" in polyproline II ensembles of an "unfolded"alanine peptide detected by molecular dynamics. *J Am Chem Soc* **126:** 16332-16333.

Richards, F.M. 1977. Areas, volumes, packing and protein structure. *Annu Rev Biophys Bioeng* **6:** 151-176.

Richards, F.M. 1979. Packing defects, cavities, volume fluctuations, and access to the interior of proteins. Including some general comments on surface area and protein structure. *Carlsberg Res. Commun.* **44:** 47-63.

Richardson, J.M., Lopez, M.M., and Makhatadze, G.I. 2005. Enthalpy of helix-coil transition: missing link in rationalizing the thermodynamics of helix-forming propensities of the amino acid residues. *Proc Natl Acad Sci U S A* **102:** 1413-1418.

Richardson, J.M., and Makhatadze, G.I. 2004. Temperature dependence of the thermodynamics of helix-coil transition. *J Mol Biol* **335:** 1029-1037.

Richardson, J.S. 1981. The anatomy and taxonomy of protein structure. *Adv. Prot. Chem.* **34:** 168-340.

Riddiford, L.M. 1966. Solvent perturbation and ultraviolet optical rotatory dispersion studies of paramyosin. *J Biol Chem* **241:** 2792-2802.

Ripoll, D.R., Vila, J.A., and Scheraga, H.A. 2005. On the orientation of the backbone dipoles in native folds. *Proc Natl Acad Sci U S A* **102:** 7559-7564.

Ropson, I.J., Gordon, J.I., and Frieden, C. 1990. Folding of a predominantly beta-structure protein: rat intestinal fatty acid binding protein. *Biochemistry* **29:** 9591-9599.

Rose, G.D. 1979. Hierarchic organization of domains in globular proteins. *J Mol Biol* **134:** 447-470.

Rose, G.D., Gierasch, L.M., and Smith, J.A. 1985. Turns in peptides and proteins. *Adv Protein Chem* **37:** 1-109.

Rose, G.D., and Wetlaufer, D.B. 1977. The number of turns in globular proteins. *Nature* **268:** 769-770.

Rucker, A.L., and Creamer, T.P. 2002. Polyproline II helical structure in protein unfolded states: lysine peptides revisited. *Protein Sci* **11:** 980-985.

Sallum, C.O., Martel, D.M., Fournier, R.S., Matousek, W.M., and Alexandrescu, A.T. 2005. Sensitivity of NMR residual dipolar couplings to perturbations in folded and denatured staphylococcal nuclease. *Biochemistry* **44:** 6392-6403.

Santoro, M.M., and Bolen, D.W. 1988. Unfolding free energy changes determined by the linear extrapolation method. 1. Unfolding of phenylmethanesulfonyl alpha-chymotrypsin using different denaturants. *Biochemistry* **27:** 8063-8068.

Scalley, M.L., Yi, Q., Gu, H., McCormack, A., Yates, J.R., 3rd, and Baker, D. 1997. Kinetics of folding of the IgG binding domain of peptostreptococcal protein L. *Biochemistry* **36:** 3373-3382.

Schellman, J.A. 1955. The stability of hydrogen-bonded peptide structures in aqueous solution. *C R Trav Lab Carlsberg [Chim]* **29:** 230-259.

Schellman, J.A., and Schellman, C. 1964. The Conformation of Polypeptide Chains in Proteins. In *The Proteins: Composition, Structure, and Function*. (ed. H. Neurath), pp. 1-137. Academic Press, New York.

Scholtz, J.M., Marqusee, S., Baldwin, R.L., York, E.J., Stewart, J.M., Santoro, M., and Bolen, D.W. 1991. Calorimetric determination of the enthalpy change for the alpha-helix to coil transition of an alanine peptide in water. *Proc Natl Acad Sci U S A* **88:** 2854-2858.

Semisotnov, G.V., Kihara, H., Kotova, N.V., Kimura, K., Amemiya, Y., Wakabayashi, K., Serdyuk, I.N., Timchenko, A.A., Chiba, K., Nikaido, K., et al. 1996. Protein globularization during folding. A study by synchrotron small-angle X-ray scattering. *J Mol Biol* **262:** 559-574.

Serrano, L. 1995. Comparison between the phi distribution of the amino acids in the protein database and NMR data indicates that amino acids have various phi propensities in the random coil conformation. *J Mol Biol* **254:** 322-333.

Shi, Z., Olson, C.A., Rose, G.D., Baldwin, R.L., and Kallenbach, N.R. 2002a. Polyproline II structure in a sequence of seven alanine residues. *Proc Natl Acad Sci U S A* **99:** 9190-9195.

Shi, Z., Woody, R.W., and Kallenbach, N.R. 2002b. Is polyproline II a major backbone conformation in unfolded proteins? *Adv Protein Chem* **62:** 163-240.

Shortle, D. 2002. The expanded denatured state: an ensemble of conformations trapped in a locally encoded topological space. *Adv Protein Chem* **62:** 1-23.

Shortle, D., and Ackerman, M.S. 2001. Persistence of native-like topology in a denatured protein in 8 M urea. *Science* **293:** 487-489.

Shortle, D., Stites, W.E., and Meeker, A.K. 1990. Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. *Biochemistry* **29:** 8033-8041.

Shrake, A., and Rupley, J.A. 1973. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.* **79:** 351-371.

Sims, G.E., Choi, I.G., and Kim, S.H. 2005. Protein conformational space in higher order phi-Psi maps. *Proc Natl Acad Sci U S A* **102:** 618-621.

Sitkoff, D., Sharp, K.A., and Honig, B. 1994. Accurate Calculation of Hydration Free-Energies Using Macroscopic Solvent Models. *J Phys Chem-Us* **98:** 1978-1988.

Srinivasan, R., Fleming, P.J., and Rose, G.D. 2004. Ab initio protein folding using LINUS. *Methods Enzymol* **383:** 48-66.

Srinivasan, R., and Rose, G.D. 1995. LINUS: a hierarchic procedure to predict the fold of a protein. *Proteins* **22:** 81-99.

Srinivasan, R., and Rose, G.D. 1999. A physical basis for protein secondary structure. *Proc Natl Acad Sci U S A* **96:** 14258-14263.

Srinivasan, R., and Rose, G.D. 2002a. Ab initio prediction of protein structure using LINUS. *Proteins* **47:** 489-495.

Srinivasan, R., and Rose, G.D. 2002b. Methinks it is like a folding curve. *Biophys Chem* **101-102:** 167-171.

Stapley, B.J., and Creamer, T.P. 1999. A survey of left-handed polyproline II helices. *Protein Sci* **8:** 587-595.

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

Svergun, D., Barberato, C., and Koch, M.H.J. 1995. CRYSOL - A program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. *J Appl Crystallogr* **28:** 768-773.

Swindells, M.B., MacArthur, M.W., and Thornton, J.M. 1995. Intrinsic phi, psi propensities of amino acids, derived from the coil regions of known structures. *Nat Struct Biol* **2:** 596-603.

Tanford, C. 1968. Protein denaturation. *Adv Protein Chem* **23:** 121-282.

Tanford, C. 1970. Protein denaturation. C. Theoretical models for the mechanism of denaturation. *Adv Protein Chem* **24:** 1-95.

Tanford, C., Kawahara, K., and Lapanje, S. 1966. Proteins in 6-M guanidine hydrochloride. Demonstration of random coil behavior. *J Biol Chem* **241:** 1921-1923.

Tiffany, M.L., and Krimm, S. 1968a. Circular dichroism of poly-L-proline in an unordered conformation. *Biopolymers* **6:** 1767-1770.

Tiffany, M.L., and Krimm, S. 1968b. New chain conformations of poly(glutamic acid) and polylysine. *Biopolymers* **6:** 1379-1382.

Tobias, D.J., and Brooks, C.L. 1992. Conformational Equilibrium in the Alanine Dipeptide in the Gas-Phase and Aqueous-Solution - a Comparison of Theoretical Results. *J Phys Chem-Us* **96:** 3864-3870.

Tran, H.T., Wang, X., and Pappu, R.V. 2005. Reconciling observations of sequence specific conformational propensities with the generic polymeric behavior of denatured proteins. *Biochemistry***:** In Press.

van Gunsteren, W.F., Burgi, R., Peter, C., and Daura, X. 2001a. The Key to Solving the Protein-Folding Problem Lies in an Accurate Description of the Denatured State. *Angew Chem Int Ed Engl* **40:** 351-355.

van Gunsteren, W.F., Burgi, R., Peter, C., and Daura, X. 2001b. Reply. *Angew Chem Int Ed Engl* **40:** 4616-4618.

Van Holde, K.E., Johnson, W.C., and Ho, P.S. 1998. *Principles of Physical Biochemistry*. Prentice Hall, Upper Saddle River, NJ.

Vila, J.A., Baldoni, H.A., Ripoli, D.R., Ghosh, A., and Scheraga, H.A. 2004. Polyproline II helix conformation in a proline-rich environment: a theoretical study. *Biophysical J.* **86:** 731-742.

Wallin, S., and Chan, H.S. 2005. A critical assessment of the topomer search model of protein folding using a continuum explicit-chain model with extensive conformational sampling. *Protein Sci* **14:** 1643-1660.

Wang, G., and Dunbrack, R.L., Jr. 2003. PISCES: a protein sequence culling server. *Bioinformatics* **19:** 1589-1591.

Whitten, S.T., and Garcia-Moreno, E.B. 2000. pH dependence of stability of staphylococcal nuclease: evidence of substantial electrostatic interactions in the denatured state. *Biochemistry* **39:** 14292-14304.

Wojcik, J., Altmann, K.H., and Scheraga, H.A. 1990. Helix Coil Stability-Constants for the Naturally-Occurring Amino-Acids in Water .24. Half-Cystine Parameters from Random Poly(Hydroxybutylglutamine-Co-S-Methylthio-L-Cysteine). *Biopolymers* **30:** 121-134.

Wolynes, P.G., Onuchic, J.N., and Thirumalai, D. 1995. Navigating the folding routes. *Science* **267:** 1619-1620.

Wong, K.B., Clarke, J., Bond, C.J., Neira, J.L., Freund, S.M., Fersht, A.R., and Daggett, V. 2000. Towards a complete description of the structural and dynamic properties of the denatured state of barnase and the role of residual structure in folding. *J Mol Biol* **296:** 1257-1282.

Word, J.M., Lovell, S.C., LaBean, T.H., Taylor, H.C., Zalis, M.E., Presley, B.K., Richardson, J.S., and Richardson, D.C. 1999. Visualizing and Quantifying Molecular Goodness-of-Fit: Small-probe Contact Dots with Explicit Hydrogen Atoms. *Journal of Molecular Biology* **285:** 1711-1733.

Woutersen, S., and Hamm, P. 2000. Structure determination of trialanine in water using polarization sensitive two-dimensional vibrational spectroscopy. *J Phys Chem B* **104:** 11316-11320.

Wu, H. 1931. Studies on denaturation of proteins. XIII. A theory of denaturation. *Chinese Journal of Physiology* **V:** 321-344.

Yi, Q., Scalley-Kim, M.L., Alm, E.J., and Baker, D. 2000. NMR characterization of residual structure in the denatured state of protein L. *J Mol Biol* **299:** 1341-1351.

Zagrovic, B., and Pande, V.S. 2003. Structural correspondence between the alpha-helix and the random-flight chain resolves how unfolded proteins can have native-like properties. *Nat Struct Biol* **10:** 955-961.

Zagrovic, B., and Pande, V.S. 2004. How does averaging affect protein structure comparison on the ensemble level? *Biophys J* **87:** 2240-2246.

Zagrovic, B., Snow, C.D., Khaliq, S., Shirts, M.R., and Pande, V.S. 2002. Native-like mean structure in the unfolded ensemble of small proteins. *J Mol Biol* **323:** 153-164.

Zaman, M.H., Shen, M.Y., Berry, R.S., Freed, K.F., and Sosnick, T.R. 2003. Investigations into sequence and conformational dependence of backbone entropy, inter-basin dynamics and the Flory isolated-pair hypothesis for peptides. *J Mol Biol* **331:** 693-711.

Zimm, B.H., and Bragg, J.K. 1959. Theory of the phase transition between helix and random coil in polypeptide chains. *J. Chem. Phys.* **31:** 526-535.

Zwanzig, R., Szabo, A., and Bagchi, B. 1992. Levinthal's paradox. *Proc. Natl. Acad. Sci. USA* **89:** 20-22.

VITA

Nicholas Charles Fitzkee was born on October 6, 1978 in York, Pennsylvania. The son of an architect and an elementary school librarian, Nicholas attended York Suburban High School, where he pursued interests in science as well as in music. After graduating in 1997, he enrolled at Carnegie Mellon University in Pittsburgh. During his freshman year, Nicholas decided to major in physics after seeking advice from Hugh Young, one of his physics professors. It was at this time that Nicholas also responded to the Gospel call and professed faith in Jesus Christ.

Throughout his undergraduate career, Nicholas was fortunate to receive instruction from many gifted professors in both the physics as well as the computer science departments. Drs. Stephen Garoff and Richard Edelstein taught the modern physics laboratory, where Nicholas had his first real instruction in giving scientific presentations. Nicholas also had the privilege to work with Dr. John Rosenberg on the crystallization of the restriction endonuclease EcoRI. Nicholas was inducted into Phi Beta Kappa as a junior and also was awarded an Andrew Carnegie Scholarship from the university. In addition, the physics department awarded him with the Richard Cutkosky award. Nicholas graduated from Carnegie Mellon in 2001 with honors in physics with a computational emphasis and a biology minor.

Nicholas attended Johns Hopkins University for graduate school, where he worked with Dr. George Rose on the unfolded state of proteins. Shortly after joining George's lab, Nicholas began courting Jennifer Davis, and they married in October of 2003. He plans to continue his studies with Dr. Bertrand Garcia-Moreno E., also at Johns Hopkins.