Physical Modeling of Protein Folding

Stefan Wallin

Department of Theoretical Physics
Lund University, Sweden

Thesis for the degree of Doctor of Philosophy

Thesis Advisor: Anders Irbäck

Faculty Opponent: Cecilia Clementi
Rice University, Houston, USA

To be presented, with the permission of the Faculty of Natural Sciences of Lund University, for public criticism in Lecture Hall F of the Department of Theoretical Physics on Wednesday, the 4th of June 2003, at 10.15 a.m.

Organization LUND UNIVERSITY Department of Theoretical Physics Sölvegatan 14A SE-223 62 LUND	**Document Name** DOCTORAL DISSERTATION **Date of issue** May 2003 **CODEN:**
Author(s) Stefan Wallin	**Sponsoring organization**

Title and subtitle

Physical modeling of protein folding

Abstract

Sequence-based models for protein folding are developed and tested on peptides with both alpha- and beta-structure, and on small three-helix-bundle proteins. The interaction potentials of the models are minimalistic and based mainly on hydrogen bonding and effective hydrophobicity forces. By contrast, the geometric representation of the protein chain is detailed. We explore the thermodynamic behaviors of these models by using efficient Monte Carlo methods, and focus on obtaining a realistic physical description of the folding process. In particular, we investigate dynamical aspects of folding, such as 'two-state' behavior and secondary structure formation. In addition, the thesis includes a study on similarity measures for protein structures.

Summary in Swedish

Sekvensbaserade proteinveckningsmodeller utvecklas och testas på peptider med både betablad- och alfahelix-struktur, samt på små helix-proteiner. Modellernas potentialer är minimalistiska och baseras huvudsakligen på vätebindningar och effektiva hydrofobicitetskrafter. Den geometriska representationen av proteinkedjan är däremot detaljerad. Vi studerar det termodynamiska uppförandet med hjälp av effektiva Monte Carlo-metoder, och lägger tyngdpunkten på att uppnå en realistisk fysikalisk beskrivning av veckningsprocessen. Vi undersöker speciellt dynamiska aspekter av veckningen, som till exempel tvåtillståndsuppförande och bildandet av sekundärstruktur. I avhandlingen ingår dessutom ett arbete om likhetsmått för proteinstrukturer.

Key words

Protein folding, protein dynamics, two-state, three-helix bundle, similarity measure.

Classification system and/or index terms (if any)

Supplementary bibliographical information		**Language** English
ISSN and key title		**ISBN** 91-628-5671-5
Recipient's notes	**Number of pages** 160	**Price**
	Security classification	

Distribution by (name and address)

Stefan Wallin, Dept. of Theoretical Physics,
Sölveg. 14A, SE-223 62 Lund

Signature _____ **Date** 2003-05-09

This thesis is based on the following publications:

I Anders Irbäck, Fredrik Sjunnesson and Stefan Wallin,
Three-Helix-Bundle Protein in a Ramachandran Model,
Proceedings of the National Academy of Sciences USA **97**,
13614–13618 (2000).

II Giorgio Favrin, Anders Irbäck and Stefan Wallin,
Folding of a Small Helical Protein Using Hydrogen Bonds and Hydrophobicity Forces,
Proteins: Structure, Function, and Genetics **47**, 99–105 (2002).

III Stefan Wallin, Jochen Farwer and Ugo Bastolla,
Testing Similarity Measures with Continuous and Discrete Protein Models,
Proteins: Structure, Function, and Genetics **50**, 144–157 (2003).

IV Anders Irbäck, Björn Samuelsson, Fredrik Sjunnesson and Stefan Wallin,
Thermodynamics of α- and β-Structure Formation in Proteins,
LU TP 02-28, submitted to *Biophysical Journal.*

V Giorgio Favrin, Anders Irbäck, Björn Samuelsson and Stefan Wallin,
Two-State Folding over a Weak Free-Energy Barrier,
LU TP 03-07, submitted to *Biophysical Journal.*

VI Giorgio Favrin, Anders Irbäck and Stefan Wallin,
Sequence-Based Study of Two Related Proteins with Different Folding Behaviors,
LU TP 03-14, submitted to *Proteins: Structure, Function, and Genetics.*

Contents

Introduction

Proteins are biological macromolecules, shaped by millions of years of evolution into structures allowing them to perform specific functions. Thousands of different proteins exist in different organisms, and the functions they perform are diverse. They catalyze biochemical reactions, regulate gene expression, transport and store molecules, mediate cell signals, etc. [1]. Proteins can also act as structural building blocks, giving blood vessels and lungs their needed strength and elasticity, and they also contribute to the structure of bone and tendon.

The folding and dynamics of proteins is an area that is becoming increasingly exciting as comparisons between computer simulations and experiments are becoming feasible. In this thesis, we develop physical models for protein folding meaning that focus is on obtaining a realistic physical, or dynamical, description of the folding process. To study the dynamical behavior of these models, we use Monte Carlo methods. Included in the thesis is also a study of protein similarity measures which are important tools for analyzing computer simulations, but also widely used in other areas of protein science.

A protein is a heterogeneous polymer molecule [2]. It consists of tens to thousands of subunits which are linked together into a linear chain molecule. The subunits are the 20 naturally occurring amino acids and each protein is uniquely defined by its amino acid sequence. Individually, the amino acids are relatively uncomplicated small molecules that differ in size, shape and chemical properties. Proteins, however, display remarkably complex behaviors. One of them is the ability of protein chains to fold into a unique and well-defined structure, the so-called native structure. Furthermore, this process is relatively fast and it is possible to find examples of proteins that fold on the timescale of μs [3], although for some proteins it takes as long as a few seconds.

The folding process has been found to be reversible for many proteins [4]. This means that if a protein is unfolded by, for example, a change in temperature or pH, it will always return to its native structure once natural conditions are

restored. The fact that folding is reversible has far-reaching implications. It tells us that all information on the shape of the native structure (and how to reach it) is encoded in the amino acid sequence. It should therefore be possible to "decode" this information and predict the native structure of a protein given only its amino acid sequence. This is the famous "protein folding problem", which is, still, an open problem. It continues to attract great interest, as is manifested by the increasing number of research groups participating in the CASP competitions (almost 200 groups in the most recent CASP5 competition), which are blind tests for structure prediction organized every two years.

It has now been almost half a century since Anfinsen's experiments on the reversibility of the folding process. Proteins have since then been intensely studied and a thorough understanding of protein folding is one of the most important and sought-after goals in molecular biology. Naturally, a substantial knowledge of proteins has accumulated: large-scale sequencing projects, such as the Human Genome Project, has given the amino acid sequences of the large majority of proteins in several species; the three-dimensional structure of about 15 000 proteins have been experimentally determined and are accessible via the Protein Data Bank [5]; a wide range of experimental techniques have been invented to characterize the folding process [6]; a basic statistical-physics framework for protein folding has been developed (for recent reviews, see Refs. [7–9]) – the list could, of course, be greatly extended.

In spite of all these advances, developing theoretical models that are able to reproduce the folding of protein chains into their native states has proven a difficult task. Statistical methods, such as homology modeling and fold recognition, are being used for structure prediction purposes. However, in order to study the dynamics of the folding process it is necessary to use models based on physical and chemical principles. The models in this thesis are applied to peptides and small proteins, and we compare the results from our simulations with experimental data. Such comparisons are important in order to calibrate models in terms of their physical properties.

The thesis is based on six papers and organized as follows. In the next two sections, a brief introduction to protein chains and the interactions that govern them is given. Then follows a discussion of the different computational models and statistical-mechanical methods used to study them, including similarity measures. This introduction is ended with a summary of the papers. The thesis ends with the six papers.

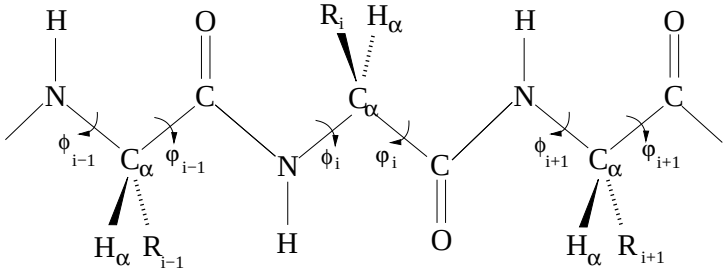
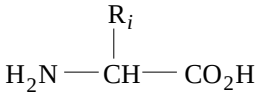


Figure 1: Schematic figure of the polypeptide chain.

Chain Geometry and Secondary Structure

As was mentioned above, proteins are constructed from the 20 naturally occurring amino acids. With one exception, proline, all amino acids have the same general form:



The amino acids are distinguished from each other only by their side chains R_i. Proline is special as its side chain is covalently attached also to the amide N atom. The size of the side chains vary from one hydrogen (glycine) up to 18 atoms (arginine and tryptophan). Two amino acids can be linked together by the formation of a peptide bond, produced by condensing the carboxyl group of one amino acid with the amino group of another and eliminating water in the process. During protein synthesis, amino acids are added sequentially to one end of the growing chain, each time forming a new peptide bond. The result is a polypeptide chain, which is schematically depicted in Fig. 1. It consists of a main chain, or backbone, (a repeating NH–$C_\alpha H_\alpha$–CO unit) from which the amino acid side chains R_i extend at regular intervals.

It is a well-known fact that the main degrees of freedom of the polypeptide chain are the Ramachandran torsion angles ϕ_i and ψ_i [10] (see Fig. 1). The torsion angle corresponding to the peptide bond is strongly restricted due to the partial double-bond character of this bond. Also the side chains are flexible and contain bonds with the same type of torsional degrees of freedom as the backbone chain.

Two types of local structures are very often found in the native states of pro-

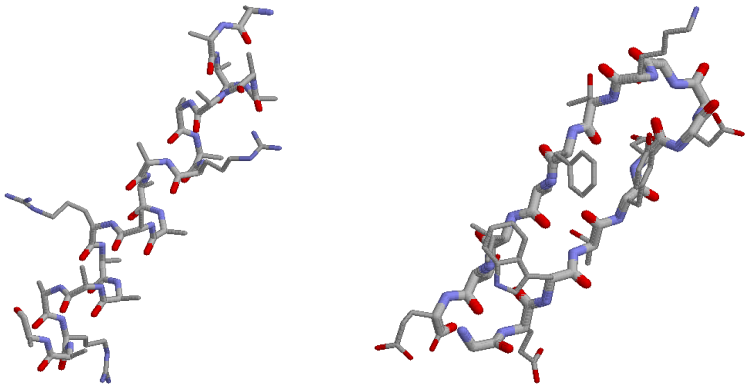


Figure 2: The α-helix (left) and the β-hairpin (right) studied in Paper IV.

teins, α-helices and β-sheets. These structural elements are referred to as the secondary structure of proteins, and Fig. 2 shows minimal secondary structure units. The arrangement of secondary structure into larger elements is called tertiary structure.

Physical Driving Forces

Although the protein chain itself is held together by covalent bonds, the folding and stability of globular proteins is governed by much weaker non-covalent interactions. An exception is disulfide bonds that sometimes occur between pairs of cysteine amino acids. There are several different types of non-covalent interactions: ionic bonds, hydrogen bonds and van der Waals bonds. Of these, hydrogen bonds play an especially important role, being a major stabilizing factor for both α-helices and β-sheets. Furthermore, hydrogen bonding between water molecules forms the basis of what is widely believed to be the main driving force for folding, the hydrophobic effect. Since hydrogen bonds and hydrophobicity play important roles in our models, we will discuss these effects in some more detail.

Hydrogen Bonds

A hydrogen bond is two electronegative atoms, a donor (D) and an acceptor (A), competing for the same hydrogen:

$$\text{—D—H} \cdots \text{A—}$$

The H atom is more closely bound to the donor atom. The favorable interaction arises because of the partial negative charges of D and A, and the partial positive charge of the H atom. This particular arrangement of charges also means that the interaction is strongest when the three atoms are aligned. In proteins, different types of hydrogen bonds can occur. The backbone NH- and CO-groups (see Fig. 1) can participate in hydrogen bonds, with the N and O atoms acting as donors and acceptors, respectively. Also hydrogen bonds between side chains, and between side chains and the backbone occur. Backbone-backbone hydrogen bonds are however most important in the sense that they help stabilize α-helices and β-sheets. In α-helices, the CO-group of amino acid i interact with the NH-group of amino acid $i + 4$. Helices with $i, i + 3$-bonds (3_{10}-helices) and $i, i + 5$-bonds (π-helices) do sometimes occur but are less frequent [11].

The Hydrophobic Effect

Perhaps the most well-known consequence of the hydrophobic effect is the poor solubility of oil in water. There appears to be a repulsive force between oil and water molecules expelling the oil from the water. This is, however, not strictly true. In fact, there are attractive van der Waals forces between oil and water molecules, but these are not nearly as strong as the attractions that occur between water molecules. In the liquid phase, water molecules form a loosely connected network of hydrogen bonds in which the O atoms act as both donors and acceptors. A hydrophobic (or "oil"-like) molecule introduced in water cannot participate in this hydrogen-bond network which is therefore disturbed. In response to this, water molecules surrounding the hydrophobic surface tend to become more organized, with a decrease in entropy as a result. Hence it is entropically favorable for hydrophobic molecules to cluster since it minimizes the contact surface between water and hydrophobic atoms.

This effective attraction between hydrophobic groups has large implications for globular proteins [12], which function in the water-soaked environment of the cell. The native structures of proteins are usually found to contain a core

dominated by hydrophobic amino acids without contact with water, and a surface that consists mainly of polar or charged amino acids.

The Folding Process

The usual formulation of the protein folding problem "predict the native structure given the amino acid sequence" implicitly implies a static picture of the folded states of proteins which is not very accurate. Proteins are flexible systems and understanding the dynamics of proteins, including *how* the native state is reached, is important for many reasons. It has recently become clear that it is not uncommon with proteins that are partly or entirely unstructured on their own, folding only upon binding to their target molecules [13]. This establishes a direct link between the folding process and protein function. Furthermore, understanding protein dynamics is also of utmost practical importance, as is underlined by the fact that an increasing number of severe diseases, such as Alzheimer's, are being linked to protein 'misfolding' and aggregation [14].

The folding process takes the protein from an unfolded state (U) to the folded, or native, state (N). The unfolded state is not unique but should be thought of as an ensemble of more or less disordered conformations. In 1991, the small protein chymotrypsin inhibitor 2 (CI2) was discovered to behave approximately as a two-state system [15], meaning that folding proceeds directly from the unfolded to the folded state, without significantly populating any meta-stable intermediate state. In terms of the free energy $F(X)$, where X is a suitably chosen reaction coordinate,[1] this has a particularly simple interpretation. $F(X)$ has only two minima corresponding to the folded and unfolded phases, respectively, and during folding a free-energy barrier (the transition state) separating U and N must be crossed for the protein to reach the folded state N. It has now become clear that CI2 is far from unique, and many such two-state proteins have been found [3].

The dynamics of folding are studied with a number of different experimental techniques. For example, fluorescence spectroscopy can sometimes be used because the amino acid side chains of tryptophan, tyrosine, and phenylalanine emit light when excited at the right wavelength. Information on the conformational states of proteins can then be inferred due to a shift in wavelength of the emitted light (Stokes shift) that occurs when the fluorescent amino acids are in contact with water; the reaction coordinate X mentioned above is here the

[1] $F(X) = -kT \ln P(X)$, where $P(X)$ is the probability distribution, k is Boltzmann's constant and T is the temperature.

hydrophobic burial of fluorescent side chains. Circular dichroism measurement are dependent on the fact that proteins are chiral molecules, and are used to probe the secondary structure content (most often the α-helix content). A less direct approach is taken by mutagenesis experiments (Φ-value analysis) [16]. Here selected single amino acid mutations are performed, and their effects on the folding rate and stability are measured. By comparing the results from the wild-type (unmutated) protein with that of the mutants, knowledge of the transition state can be obtained.

An interesting question on the folding of small two-state proteins is the order in which secondary structure formation and chain collapse occur during folding. Several different folding scenarios have been proposed: the framework (or diffusion-collision) model [17, 18], the nucleation-condensation model [19, 20], and the hydrophobic collapse model [21]. These scenarios all give different answers to the above posed question, with nucleation-condensation being midway between the other two and predicting the simultaneous formation of secondary structure and chain collapse. There is evidence of proteins (including CI2) that fold by a nucleation-condensation mechanism [19], but it cannot be excluded that there are proteins that fold in alternative ways [22]. For example, there are recent experiments on small helical proteins that have been interpreted as supporting a framework picture [23, 24].

Models and Methods

We now turn to the theoretical study of proteins, which is an area that has always relied heavily on computer simulations. Because of the size and complexity of proteins, calculations based on first principles are not computationally feasible. Instead, one has to postulate some effective form of the potential (describing the interactions between atoms or groups of atoms) in such a way that low-energy states resemble real protein structures. Finding the state of minimum energy, for a given form of the potential, is a difficult problem that has received a lot of attention. In order to study dynamics it is necessary to go beyond energy minimization and sample the full conformational space in a representative way.

Models for protein folding have been developed at widely different levels of resolution, ranging from simple lattice models to elaborate all-atom models with explicit water. Not the same questions, however, are addressed with all models. The standard programs [25–27] with elaborate force fields are becoming increasingly popular and are used for a variety of purposes such as refining experimental NMR data, studying functional mechanisms, etc. [28]. Lattice

models, where typically each amino acid is represented by a single site, are not meant to study specific proteins but rather the generic principles of folding. Many models "intermediate" in complexity exist. A particularly common choice is off-lattice C_α models where each amino acid is represented by its C_α atom.

A description of our approach to protein modeling is naturally divided into two parts.

1. The potential is kept as simple and transparent as possible, with a minimum of free parameters to tune.

2. The geometric representation of the protein chain is detailed.

A common way to determine model parameters is to estimate their values based on microscopic calculations. Our focus when optimizing parameters is rather the overall thermodynamic behavior of the model. To be able to carry out such an optimization, a small parameter space is necessary. The use of a detailed chain geometry has many advantages. It allows precise comparisons with experimental data (structure, stability, etc.) to be done, and a good local description of the protein chain is relatively easy to obtain. Last but not least, it makes the formulation of the potential easier.

Another central property of the models in this thesis is that they are sequence-based, meaning that the only information used when simulating a protein is its amino acid sequence. In contrast to sequence-based models are those of so-called Gō-type [29], in which interactions not favoring the native structure are ignored. The Gō prescription has received support from a statistical analysis showing that the contact order (CO) parameter [30], a quantity based solely on the native topology, is an important determinant for folding rates. There are also examples of proteins for which experimental folding pathways have been reproduced using Gō-type models [31]. Nevertheless, it is of practical and fundamental interest to try to go beyond Gō-type models.

Let us now turn to the statistical-mechanical description of protein folding. A protein model has a number of degrees of freedom and a state space consisting of all possible chain conformations C. The thermodynamic behavior is determined by the interaction potential, which is expressed as an energy function $E(C)$. $E(C)$ is an effective energy that generally depends on the temperature. This dependence is for simplicity neglected in our models. The probability to find the chain in a particular conformation C at temperature T, $p(C)$, is given by the Boltzmann distribution; that is $p(C) \propto \exp[-E(C)/k_BT]$, where k_B is Boltzmann's constant. The thermodynamic average of an observable O is then

given by

$$\langle O \rangle = \int p(C) O(C) dC \tag{1}$$

where the integration is over all state space. A direct evaluation of the integral in Eq. (1) could only be done for systems of miniscule sizes and is not an option for off-lattice protein models. Instead, one has to resort to some form of Monte Carlo method in order to estimate $\langle O \rangle$.

Monte Carlo Methods

The Metropolis Monte Carlo method [32] is a general scheme for generating a sequence of states weighted according to the Boltzmann distribution. This allows thermodynamic averages, such as $\langle O \rangle$ in Eq. (1), to be estimated as simple averages over Monte Carlo time. From a generated sequence of chain conformations $(C_1, ..., C_N)$, we have

$$\langle O \rangle \approx \frac{1}{N} \sum_{k=1}^{N} O(C_k) \tag{2}$$

for large N. The way the sequence of states is generated is by chain updates. In each step, the most recent chain conformation is updated producing a tentative state, which is then either accepted or rejected according to the Metropolis condition. The generation of states is a stochastic process in the sense that the chain updates and the outcome of the accept/reject question depend on random numbers. We use a few different types of chain updates. The simplest is the pivot move where a single torsion angle (picked at random) is turned. This move is non-local and generates large conformational changes. It would be helpful to have also an update that affected only a small local part of the chain, while keeping the rest of the chain structure rigid. Such a move would certainly help accelerate the conformational search at low temperatures. Unfortunately, it is somewhat cumbersome to design a strictly local move for a general chain such as the polypeptide backbone. We therefore include in our move set a semi-local move [33] in which 7 or 8 adjacent torsion angles are turned in a coordinated way, with a bias towards local deformations of the chain.

A general problem with the Metropolis method is that the generated states C_k tend to be strongly correlated, especially at low temperatures. The main reason for this is that the free-energy landscape of protein models are rugged, with many local minima where the system can easily get trapped. For the estimate in Eq. (2) to be accurate, the C_k's must represent a proper sampling of the entire conformational space, which means that N must be much longer

than the decorrelation time of the system. Many different solutions have been proposed to alleviate this problem: the multicanonical method [34, 35], simulated tempering [36–38], and parallel tempering [39, 40]. In this thesis, we use the simulated tempering scheme which has been shown to be comparable in efficiency with both parallel tempering [40] and the multicanonical method [41]. With simulated tempering it is easy to monitor convergence, and it also provides a convenient way to calculate free energies.

Similarity Measures

Measures of the similarity (or distance) between protein structures are used in many contexts. For example, similarity measures have been used to classify the structures in the PDB into different fold categories [42].

The most frequently used, and perhaps the most intuitive, distance measure is the root-mean-square deviation, D_{rmsd}. If $\mathbf{r}_i^{\text{a}}$ and $\mathbf{r}_i^{\text{b}}$ denote the atom positions in two structures a and b, respectively, then

$$D_{\text{rmsd}}^2 = \min \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{r}_i^{\text{a}} - \mathbf{r}_i^{\text{b}} \right)^2 . \tag{3}$$

The minimization is over all rotations and translations so that an optimal superposition of the two structures is obtained. It is, in fact, possible to perform this minimization analytically for any two given structures [43].

An alternative to the rmsd measure is to use the set of internal atomic distances r_{ij}^{a} and r_{ij}^{b} in the definition of a distance measure. Because of their construction, this class of measures have the property of being unable to discriminate between a structure and its mirror image. On the other hand, the minimization over rotations and translations that is required in the calculation of the rmsd can be avoided.

The Papers

Papers I and V

The three-helix bundle is one of the smallest existing protein folds. Yet it contains both secondary and tertiary structure, making three-helix-bundle proteins ideal as testbeds for protein models. Furthermore, α-helices are local structural elements and, as such, less complicated to model than β-sheets which

may involve chain segments that are globally separated in sequence. Another interesting aspect of the three-helix bundle is that it has two possible overall topologies; if the first two helices make a U, the third one can go either in front of or behind that U. In Papers I and V, we study a designed three-helix-bundle protein with 54 amino acids.

To this end, we develop in Paper I a model with a realistic backbone geometry (all non-hydrogen backbone atoms are included), and where each side chain is represented by a single atom, a large C_β. The degrees of freedom are the Ramachandran torsion angles ϕ_i and ψ_i, and the amino acids can be of three different types: hydrophobic (Hyd), polar (Pol) or glycine (Gly; this amino acid has no C_β atom). The sequence is designed in such a way that the three-helix-bundle fold will have a core of hydrophobic amino acids, and flexible glycines located in the turn regions between the helices. The energy function has the general form:

$$E = E_{\rm loc} + E_{\rm sa} + E_{\rm hb} + E_{\rm hp}\,. \tag{4}$$

The self-avoidance term $E_{\rm sa}$ and the local term $E_{\rm loc}$ enforce steric constraints on the chain and limits its flexibility; overlapping atom pairs are subject to a strong repulsive force. $E_{\rm hb}$ represent backbone-backbone hydrogen bonds; any NH- and CO-group can form a hydrogen bond in the model. Water molecules are not explicitly represented in the model. The effect of water is accounted for in a crude way by the hydrophobicity term, $E_{\rm hp}$, which represents pairwise forces between the C_β atoms of hydrophobic amino acids. $E_{\rm hb}$ and $E_{\rm hp}$ are responsible for the stability of the protein.

We study the thermodynamic behavior of this model and find that is has a number of non-trivial properties. First of all, as the temperature is lowered, the chain is indeed found to form a stable three-helix bundle, although the model does not discriminate between the two possible topologies. The transition from an expanded state at high temperatures to the folded state is found be cooperative (i.e. abrupt). In fact, by also studying the corresponding one- and two-helical segments, we obtain a size dependence consistent with that of an ordinary first-order phase transition. Second, the one- and two-helical segments are found to be less stable on their own than as parts of the full chain, as expected for a cooperative system. Third, we find that helix formation and chain collapse proceed in parallel in the model, which means that the behavior of this model lies closest to the nucleation-condensation folding scenario. A simultaneous chain collapse and helix formation is supported by recent experiments on small helical proteins [44].

In paper V, the thermodynamic analysis of the three-helix-bundle model is extended and a Monte Carlo-based kinetic study is performed. The goal is to investigate whether or not this system can be understood in terms of a simple

two-state picture, which seems to be a good description of so many small single-domain proteins [3]. The fact that the folding transition is first-order-like in our model indicates that this could be the case. Following what is often done in analyzing experimental data, we perform fits of the temperature dependence of a few observables (melting curves) to a two-state model. The fits are good and hence the model shows the same type of thermodynamic behavior as two-state proteins. The kinetic runs differ from our thermodynamic simulations in two ways. First, the temperature is held constant (at the melting temperature), and second, only the semi-local move [33] is used to update the chain, and not the non-local pivot move. Starting from high-temperature conformations, a large number of folding simulations are recorded. We find that the relaxation behavior of ensemble averages show relatively small deviations from a single exponential. This means that our model reproduces the behavior of small two-state proteins not only thermodynamically, but also kinetically.

In spite of this agreement with experimental data, we find that the two-state picture is far from perfect in the model, which is interesting. We see this most easily from the free energy $F(X)$, where the barrier separating the folded and unfolded phases is small, or absent, depending on which reaction coordinate X is used. One consequence of this is that the two-state fit parameters must be interpreted with great care.

Papers II and VI

In Paper II we extend the model in Paper I slightly and then confront it with experimental data on a real three-helix-bundle protein, a 46-amino acid fragment from the B domain of staphylococcal protein A. This protein has been extensively studied both experimentally, and theoretically by other groups. In particular, this means that we can compare our results with those obtained using Gō-type models. Our previous model is extended to include two more amino acid types. Proline (Pro) is given a special geometric representation which is needed in order to capture the helix breaking property of this amino acid. The B domain sequence contains two prolines, one in each of the two turn regions. Also, the amino acid alanine (Ala) is taken to form its own weaker hydrophobicity class. Thus the model has 5 amino acid types: Pro and Gly with their special geometries, and Hyd, Pol, and Ala which share the same geometry but differ in hydrophobicity.

We begin our study of the B domain sequence by looking at the collapse transition, and compare with sequences obtained by randomly reshuffling the amino acids of the original sequence (keeping the two prolines fixed at their turn positions). Naively, one would expect these sequences to show similar collapse

behaviors, since the amino acid composition is the same. However, it turns out that the B domain sequence collapses much more efficiently than the random sequences. Next, we turn to the structure of the collapsed phase and find that one of the topologies is thermodynamically favored. It is indeed the topology of the experimentally determined structure for this protein, although the suppression of the wrong topology is not strong in the model. If we restrict ourselves to the thermodynamically favored topology, the lowest-energy conformation is found very close the native structure; in terms of the rmsd measure, this conformation has $D_{\mathrm{rmsd}} = 1.8$ Å. We also study the three individual helix segments, and compare with experimental data for these short chains.

Monte Carlo-based kinetic runs are also performed for the B domain sequence. In spite of large variations, they show one common and stable trend, namely that the hydrophobic collapse is always at least as fast as helix formation. This means that our results are in disagreement with previous results for this sequence obtained with an off-lattice C_α model [45] of Gō-type. Here, fast folding was found to be associated with fast helix formation and a rate-limiting collapse step. There have been claims, however, that these simulations were effectively carried out at an unreasonably low temperature [46].

Recently, interesting experiments [47] was performed on the Z domain of protein A and on a related engineered protein, $Z_{\mathrm{SPA}-1}$, which differ from the Z domain in 13 amino acid positions. The folding behaviors of these two sequences are very different. The Z domain, which is almost identical to the B domain, adopts a three-helix-bundle fold in solution, while $Z_{\mathrm{SPA}-1}$ is rather a disordered molten globule. Together in solution, they form a complex in which both proteins adopt three-helix-bundle folds. Hence, this is an example of coupled folding and binding. In Paper VI, we study both sequences (individually) using exactly the same model, and model parameters, as for the B domain in Paper II. The aim of the study is twofold. First, we investigate whether or not the difference in folding behavior of these two sequences can be understood in terms of our simple model. Second, we use the model to make predictions on the properties of the collapsed phase of $Z_{\mathrm{SPA}-1}$. To decide whether or not these predictions are correct requires further experimental data.

Paper III

In this work, we investigate and compare the properties of a number of frequently used similarity measures. In particular, we compare the standard rmsd with measures based on intramolecular distances r_{ij}. We also propose a new distance measure of the latter type.

The analysis is carried out using two different models, one continuous and one discrete. The continuous model is the three-helix-bundle model in Paper I, which is used to examine the relationship between distance to the native structure and energy. It is well known that the rmsd correlates only weakly with energy. We show that a stronger energy correlation can be obtained if measures based on the r_{ij}'s are used. The continuous model is also used to examine to what extent the various measures can discriminate between the two different three-helix-bundle topologies. This task is particularly challenging for measures based on the r_{ij}'s given their inabilities to distinguish between a structure and its mirror image. With the discrete model, we perform fits of model structures to real protein structures. The properties of the fitted structures are investigated, and the quality of the fits are found to depend on the scale considered and on the distance measure used in obtaining the fits. We find that the rmsd is good at reproducing global structural properties, and that measures based on the r_{ij}'s are good at reproducing local, small-scale properties, for which the rmsd fail.

Combining the results obtained with both models, this work clearly shows that similarity measures must be carefully chosen depending on the application at hand.

Paper IV

As was mentioned early on in this introduction, simulating protein folding at atomic resolution is a difficult task. It is, however, no longer computationally impossible as has been demonstrated by recent all-atom Gō-type studies [31, 48]. Extending these calculations to purely sequence-based potentials remains, however, an open problem, due to well-known uncertainties about the form and relevance of different terms in the potential. In the process of calibrating potentials, short amino acid sequences with protein-like properties are likely to be very important. Breakthrough experiments in the last ten years have given examples of such sequences. In paper IV, we develop a minimalistic all-atom model and test it on two short chains with 21 and 16 amino acids making an α-helix and a β-hairpin, respectively (see Fig. 2). The melting curves for these peptides have been experimentally determined, and the β-hairpin has been shown to exhibit two-state character.

The model in Paper IV goes beyond our previous models in that full side chains are added. It is, in fact, geometrically more detailed than standard "all-atom" models, since all hydrogen atoms are explicitly represented. The general form of the energy function is as in Eq. (4), with the exception that E_{loc} is absent. To be able to remove this term, the inclusion of the hydrogen atoms

is crucial. The form of the E_{sa} and E_{hb} terms are not very different from the previous models, although E_{hb} is extended to include some side chain-side chain and side chain-backbone hydrogen bond types. The hydrophobicity term E_{hp} represents, as before, pairwise attractions between hydrophobic amino acids. Here, the relative strengths of these attractions are chosen according to the contact energies estimated by Miyazawa and Jernigan [49] by statistical analysis of PDB structures.

Despite the minimalistic potential of our model, it turns out that it is able to form both the α-helix and the β-hairpin structure (with all model parameters held fixed). By extensive Monte Carlo sampling, we calculate the melting curves for both sequences and find that simple two-state fits provide good descriptions of these curves, and that the fit parameters are in reasonable *quantitative* agreement with those obtained from experiments. Previous theoretical studies have not provided the correct temperature dependence [50]. As in Paper V, we find also here that the two-state picture is an oversimplification. The free energy $F(X)$ does not show any significant barrier separating the folded and unfolded phases, which renders the two-state fit parameters ambiguous. We also perform the same type of kinetic simulations as in Paper V, and find that the α-helix forms faster than the β-hairpin, in accord with experimental observations.

When calibrating the model, our goal was to ensure that, without resorting to parameter changes, our two sequences made an α-helix and a β-hairpin, respectively, which was not an easy task. Once this goal had been achieved, the thermodynamic calculations were carried out without any further fine-tuning of the potential. It is therefore hard to believe that the generally quite good agreement with experimental data is accidental. A more plausible explanation of the agreement is that the thermodynamics of these peptides are largely governed by backbone-backbone hydrogen bonds and hydrophobic collapse forces, as assumed by the model.

Acknowledgments

I would like to thank my supervisor Anders for excellent and generous guidance, and for good company on the running trails around Skryllegården. Thanks also to the rest of the "protein folding group", Björn, Fredrik and Giorgio for great team work, and to everybody else at the department for all the good times. A special thought goes to the many after-work "meetings" over a refreshment, with core participants Pierre "Jacques" Dhonte and Giorgio Favrin. It has been a pleasure to share an office with Fredrik for the past three years. Finally, thanks to Jan-Inge Wallin for proofreading this introduction and to Thomas Breslin for help with the cover page.

References

[1] Alberts, B., Bray, D.,Lewis, J., Raff, M., Roberts, K. & Watson, J.D. (1994) "Molecular biology of the cell", Garland Publishing, 3rd edition.

[2] Branden, C. & Tooze, J. (1999) "Introduction to protein structure", Garland Publishing, 2nd edition.

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

[4] Anfinsen, C.B. (1973) "Principles that govern the folding of protein chains", *Science* **181**, 223–230.

[5] Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) "The Protein Data Bank: A computer-based archival file for macromolecular structures", *J. Mol. Biol.* **112**, 535–542.

[6] Creighton, T.E. (1993) "Proteins: Structures and molecular properties", W.H. Freeman and Company, New Tork, 2nd edition.

[7] Plotkin, S.S. & Onuchic, J.N. (2002) "Understanding protein folding with energy landscape theory I: Basic concepts", *Q. Rev. Biophys.* **35**, 111–167.

[8] Plotkin, S.S. & Onuchic, J.N. (2002) "Understanding protein folding with energy landscape theory II: Quantitative aspects", *Q. Rev. Biophys.* **35**, 205–286.

[9] Chan, H.S., Kaya, H. & Shimizu, S. (2002) "Computational methods for protein folding: Scaling a hierarchy of complexities", In: Current topics in computational molecular biology, edited by Jiang, T., Xu, Y. & Zhang, M.Q. (MIT Press, Cambridge, USA), pp. 403–447.

[10] Ramachandran, G.N. & Sasisekharan, V. (1968) "Conformation of polypeptides and proteins", *Adv. Protein Chem.* **23**, 283–437.

[11] Fodje, M.N. & Al-Karadaghi, S. (2002) "Occurrence, conformational features and amino acid propensities for the pi-helix", *Protein Eng.* **15**, 353–358.

[12] Kauzmann, W. (1959) "Some factors in the interpretation of protein denaturation", *Adv. Protein Chem.* **14**, 1–63.

[13] Dyson, H.J. & Wright, P.E. (2002) "Coupling of folding and binding for unstructured proteins", *Curr. Opin. Struct. Biol.* **12**, 54–60.

[14] Dobson, C.M. (2001) "The structural basis of protein folding and its links with human disease", *Phil. Trans. R. Soc. Lond.* B **356**, 133–145.

[15] Jackson, S.E. & Fersht, A.R. (1991) "Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition", *Biochemistry* **30**, 10428–10435.

[16] Fersht, A.R., Matouschek, A. & Serrano, L. (1992) "The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding", *J. Mol. Biol.* **224**, 771–782.

[17] Kim, P.S. & Baldwin, R.L. (1982) "Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding", *Annu. Rev. Biochem.* **51**, 459–489.

[18] Karplus, M. & Weaver, D.L. (1976) "Protein-folding dynamics", *Nature* **260**, 404–406.

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

[20] Abkevich, V.I., Gutin, A.M. & Shakhnovich, E.I. (1994) "Specific nucleus as the transition state of protein folding: Evidence from the lattice model", *Biochemistry* **33**, 10026–10036.

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

[22] Daggett, V. & Fersht, A.R. (2003) "Is there a unifying mechanism for protein folding?", *Trends Biochem. Sci.* **28**, 18–25.

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

[24] Myers, J.K. & Oas, T.G. (2001) "Preorganized secondary structure as an important determinant of fast protein folding", *Nat. Struct. Biol.* **8**, 552–558.

[25] Brooks, B.R., Bruccoleri, R.E., Olafson. B.D., States, D.J., Swaminathan, S. & Karplus, M. (1983) "CHARMM: A program for macromolecular energy minimization, and dynamics calculations", *J. Comput. Chem.* **4**, 187–217.

[26] Weiner, P.W. & Kollman, P.A. (1981) "AMBER: Assisted model building with energy refinement. A general program for modelling molecules and their interactions", *J. Comput. Chem.* **2**, 287–303.

[27] Scott, W.R.P., Hünenberger, P.H., Tironi, I.G., Mark, A.E., Billeter, S.R., Fennen, J., Torda, A.E., Huber, T., Krüger, P. & van Gunsteren, W.F. (1999) "The GROMOS biomolecular simulation program package", *J. Phys. Chem.* A **103**, 3596–3607.

[28] Karplus, M. & McCammon, J.A. (2002) "Molecular dynamics simulations of biomolecules", *Nat. Struct. Biol.* **9**, 646–652.

[29] Gō, N. & Taketomi, H. (1978) "Respective roles of short- and long-range interactions in protein folding", *Proc. Natl. Acad. Sci. USA* **75**, 559–563.

[30] Plaxco, K.W., Simons, K.T. & Baker, D. (1998) "Contact order, transition state placement and the refolding rates of single domain proteins", *J. Mol. Biol.* **277**, 985–994.

[31] Clementi, C., García, A.E. & Onuchic, J.N. (2003) "Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: All-atom representation study of protein L", *J. Mol. Biol.* **326**, 933–954.

[32] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. (1953) "Equation of state calculations by fast computing machines", *J. Chem. Phys.* **21**, 1087–1092.

[33] Favrin, G., Irbäck, A. & Sjunnesson, F. (2001) "Monte Carlo update for chain molecules: Biased gaussian steps in torsional space", *J. Chem. Phys.* **114**, 8154–8158.

[34] Berg, B.A. & Neuhaus, T. (1991) "Multicanonical algorithms for 1st order phase-transitions", *Phys. Lett.* B **267**, 249–253.

[35] Hansmann, U.H.E & Okamoto, Y. (1993) "Prediction of peptide conformation by multicanonical algorithm: New approach to the multiple-minima problem", *J. Comput. Chem.* **14**, 1333–1338.

[36] Lyubartsev, A.P., Martsinovski, A.A., Shevkunov, S.V. & Vorontsov-Velyaminov, P.V. (1992) "New approach to Monte Carlo calculation of the free energy: Method of expanded ensembles", *J. Chem. Phys.* **96**, 1776–1783.

[37] Marinari, E. & Parisi, G. (1992) "Simulated Tempering: A new Monte Carlo scheme", *Europhys. Lett.* **19**, 451–458.

[38] Irbäck, A. & Potthast, F. (1995) "Studies of an off-lattice model for protein folding: Sequence dependence and improved sampling at finite temperature", *J. Chem. Phys.* **103**, 10298–10305.

[39] Swendsen, R.H. & Wang, J.-S. (1986) "Replica Monte Carlo simulation of spin-glasses", *Phys. Rev. Lett.* **57**, 2607–2609.

[40] Irbäck, A. & Sandelin, E. (1999) "Monte Carlo study of the phase structure of compact polymer chains", *J. Chem. Phys.* **110**, 12245–12262.

[41] Hansmann, U.H.E. & Okamoto, Y. (1997) "Numerical comparisons of three recently proposed algorithms in the protein folding problem", *J. Comput. Chem.* **18**, 920–933.

[42] Holm, L. & Sanders, C. (1994) "The FSSP database of structurally aligned protein fold families", *Nucleic Acids Res.* **22**, 3600–3609.

[43] von Neumann, J. (1937) "Some matrix-inequalities and metrization of matric-space", *Tomsk. Univ. Rev.* **1**, 286–300.

[44] Krantz, B.A., Srivastava, A.K, Nauli, S., Baker, D., Sauer, R.T. & Sosnick, T.R. (2002) "Understanding protein hydrogen bond formation with kinetic H/D amide isotope effects", *Nat. Struct. Biol.* **9**, 458–463.

[45] Zhou, Y. & Karplus, M. (1999) "Interpreting the folding kinetics of helical proteins", *Nature* **401**, 400–403.

[46] Mirny, L. & Shakhnovich, E.I. (2001) "Protein folding: Matching theory and experiment", In: Proceeding of the international school of physics 'Enrico Fermi': Protein folding, evolution and design, edited by Broglia, R.A, Shakhnovich, E.I. & Tiana, G. (IOS Press), pp. 37–68.

[47] Wahlberg, E., Lendel, C., Helgstrand, M., Allard, P., Dincbas-Renqvist, V., Hedqvist, A., Berglund, H., Nygren, P.-Å. & Härd, T. (2002) "An affibody in complex with a target protein: Structure and coupled folding", *Proc. Natl. Acad. Sci. USA* **100**, 3185–3190.

[48] Shimada, J. & Shakhnovich, E.I. (2002) "The ensemble folding kinetics of protein G from an all-atom Monte Carlo simulation", *Proc. Natl. Acad. Sci. USA* **99**, 11175–11180.

[49] Miyazawa, S. & Jernigan, R.L. (1996) "Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density, for simulation and threading", *J. Mol. Biol.* **256**, 623–644.

[50] Zhou, R., Berne, B.J. & Germain, R. (2001) "The free energy landscape for β-hairpin folding in explicit water", *Proc. Natl. Acad. Sci. USA* **98**, 14931–14936.

Three-Helix-Bundle Protein in a Ramachandran Model

Paper I

Three-Helix-Bundle Protein in a Ramachandran Model

Anders Irbäck, Fredrik Sjunnesson and Stefan Wallin

Complex Systems Division, Department of Theoretical Physics
Lund University, Sölvegatan 14A, SE-223 62 Lund, Sweden
http://www.thep.lu.se/complex/

Proceedings of the National Academy of Sciences USA, **97**,
13614–13618 (2000)

Abstract:
We study the thermodynamic behavior of a model protein with 54 amino acids that forms a three-helix bundle in its native state. The model contains three types of amino acids and five to six atoms per amino acid and has the Ramachandran torsional angles ϕ_i, ψ_i as its degrees of freedom. The force field is based on hydrogen bonds and effective hydrophobicity forces. For a suitable choice of the relative strength of these interactions, we find that the three-helix-bundle protein undergoes an abrupt folding transition from an expanded state to the native state. Also shown is that the corresponding one- and two-helix segments are less stable than the three-helix sequence.

1.1 Introduction

It is not yet possible to simulate the formation of proteins' native structures on the computer in a controlled way. This goal has been achieved in the context of simple lattice and off-lattice models, where typically each amino acid is represented by a single interaction site corresponding to the C_α atom, and such studies have provided valuable insights into the physical principles of protein folding [1–5] and the statistical properties of functional protein sequences [6,7]. However, these models have their obvious limitations. Therefore, the search for computationally feasible models with a more realistic chain geometry remains a highly relevant task.

In this paper, we discuss a model based on the well-known fact that the main degrees of freedom of the protein backbone are the Ramachandran torsional angles ϕ_i, ψ_i [8]. Each amino acid is represented by five or six atoms, which makes this model computationally slightly more demanding than C_α models. On the other hand, it also makes interactions such as hydrogen bonds easier to define. The formation of native structure is, in this model, driven by hydrogen-bond formation and effective hydrophobicity forces; hydrophobicity is widely held as the most important stability factor in proteins [9, 10], and hydrogen bonds are essential to properly model the formation of secondary structure.

In this model, we study in particular a three-helix-bundle protein with 54 amino acids, which represents a truncated and simplified version of the four-helix-bundle protein *de novo* designed by Regan and DeGrado [11]. This example was chosen partly because there have been earlier studies of similar-sized helical proteins using models at comparable levels of resolution [12–18]. The behavior of small fast-folding proteins is a current topic in both theoretical and experimental research, and a three-helix-bundle protein that has been extensively studied both experimentally [19,20] and theoretically [14,17,21,22] is fragment B of staphylococcal protein A.

In addition to the three-helix protein, to study size dependence, we also look at the behavior of the corresponding one- and two-helix segments. By using the method of simulated tempering [23–25], a careful study of the thermodynamic properties of these different chains is performed.

Not unexpectedly, it turns out that the behavior of the model depends strongly on the relative strength of the hydrogen-bond and hydrophobicity terms. In fact, the situation is somewhat reminiscent of what has been found for homopolymers with stiffness [26–29], with hydrogen bonds playing the role of the stiffness term. Throughout this paper, we focus on one specific empirical choice of these parameters.

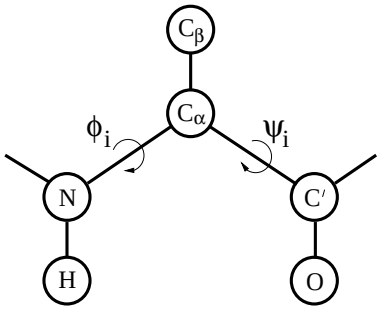


Figure 1.1: Schematic figure showing the representation of one amino acid.

For this choice of parameters, we find that the three-helix-bundle protein has the following three properties. First, it does form a stable three-helix bundle (except for a 2-fold topological degeneracy). Second, its folding transition is abrupt, from an expanded state to the native three-helix-bundle state. Third, compared to the one- and two-helix segments, it forms a more stable secondary structure. It should be stressed that these properties are found without resorting to the popular Gō approximation [30], in which interactions that do not favor the desired structure are ignored.

1.2 The Model

The model we study is a reduced off-lattice model. The chain representation is illustrated in Fig. 1.1. As mentioned in the introduction, each amino acid is represented by five or six atoms. The three backbone atoms N, C_α and C' are all included. Also included are the H and O atoms shown in Fig. 1.1, which we use to define hydrogen bonds. Finally, the side chain is represented by a single atom, C_β, which can be hydrophobic, polar or absent. This gives us the following three types of amino acids: A with hydrophobic C_β, B with polar C_β, and G (glycine) without C_β.

The H, O and C_β atoms are all attached to the backbone in a rigid way. Furthermore, in the backbone, all bond lengths, bond angles and peptide torsional angles (180°) are held fixed. This leaves us with two degrees of freedom per amino acid, the Ramachandran torsional angles ϕ_i and ψ_i (see Fig. 1.1). The parameters held fixed can be found in Table 1.1.

Bond lengths (Å)		Bond angles (°)	
NC_α	1.46	$C'NC_\alpha$	121.7
$C_\alpha C'$	1.52	$NC_\alpha C'$	111.0
$C'N$	1.33	$C_\alpha C'N$	116.6
NH	1.03	$NC_\alpha C_\beta$	110.0
$C_\alpha C_\beta$	1.53	$C'C_\alpha C_\beta$	110.0
$C'O$	1.23		

Table 1.1: Geometry parameters.

Our energy function

$$E = E_{\rm loc} + E_{\rm sa} + E_{\rm hb} + E_{\rm AA} \tag{1.1}$$

is composed of four terms. The local potential $E_{\rm loc}$ has a standard form with 3-fold symmetry,

$$E_{\rm loc} = \frac{\epsilon_\phi}{2}\sum_i (1+\cos 3\phi_i) + \frac{\epsilon_\psi}{2}\sum_i (1+\cos 3\psi_i)\,. \tag{1.2}$$

The self-avoidance term $E_{\rm sa}$ is given by a hard-sphere potential of the form

$$E_{\rm sa} = \epsilon_{\rm sa}\sum_{i<j}{}' \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}, \tag{1.3}$$

where the sum runs over all possible atom pairs except those consisting of two hydrophobic C_β. The hydrogen-bond term $E_{\rm hb}$ is given by

$$E_{\rm hb} = \epsilon_{\rm hb}\sum_{ij} u(r_{ij})v(\alpha_{ij},\beta_{ij})\,, \tag{1.4}$$

where

$$u(r_{ij}) = 5\left(\frac{\sigma_{\rm hb}}{r_{ij}}\right)^{12} - 6\left(\frac{\sigma_{\rm hb}}{r_{ij}}\right)^{10} \tag{1.5}$$

$$v(\alpha_{ij},\beta_{ij}) = \begin{cases} \cos^2\alpha_{ij}\cos^2\beta_{ij} & \alpha_{ij},\beta_{ij} > 90^\circ \\ 0 & \text{otherwise} \end{cases} \tag{1.6}$$

In Eq. 1.4 i and j represent H and O atoms, respectively, and r_{ij} denotes the HO distance, α_{ij} the NHO angle, and β_{ij} the HOC' angle. Any HO pair can form a hydrogen bond. The last term in Eq. 1.1, the hydrophobicity term $E_{\rm AA}$, has the form

$$E_{\rm AA} = \epsilon_{\rm AA}\sum_{i<j}\left[\left(\frac{\sigma_{\rm AA}}{r_{ij}}\right)^{12} - 2\left(\frac{\sigma_{\rm AA}}{r_{ij}}\right)^{6}\right], \tag{1.7}$$

σ_i(Å)					
N	C_α	C′	H	C_β	O
1.65	1.85	1.85	1.0	2.5	1.65

ϵ_ϕ	ϵ_ψ	$\epsilon_{\rm sa}$	$\epsilon_{\rm hb}$	$\epsilon_{\rm AA}$	$\sigma_{\rm hb}$ (Å)	$\sigma_{\rm AA}$ (Å)
1	1	0.0034	2.8	2.2	2.0	5.0

Table 1.2: Parameters of the energy function. Energies are in dimensionless units, in which the folding transition occurs at $kT \approx 0.65$ for the three-helix-bundle protein (see below).

where both i and j represent hydrophobic C_β. To speed up the simulations, a cutoff radius r_c is used,[1] which is 4.5Å for $E_{\rm sa}$ and $E_{\rm hb}$, and 8Å for $E_{\rm AA}$.

In this energy function, roughly speaking, the first two terms, $E_{\rm loc}$ and $E_{\rm sa}$, enforce steric constraints, whereas the last two terms, $E_{\rm hb}$ and $E_{\rm AA}$, are the ones responsible for stability. Force fields similar in spirit, emphasizing hydrogen bonding and hydrophobicity, have been used with some success to predict structures of peptides [31] and small helical proteins [15].

The parameters of our energy function were determined largely by trial and error. The final parameters are listed in Table 1.2. The parameters σ_{ij} of Eq. 1.3 are given by

$$\sigma_{ij} = \sigma_i + \sigma_j + \Delta\sigma_{ij} \,,$$

where σ_i, σ_j can be found in Table 1.2, and $\Delta\sigma_{ij}$ is zero except for C_βC′, C_βN and C_βO pairs that are connected by three covalent bonds. In these three cases, we put $\Delta\sigma_{ij} = 0.625$Å. This could equivalently be described as a change of the local ϕ_i and ψ_i potentials. In Fig. 1.2, we show ϕ_i, ψ_i scatter plots for nonglycine (A and B) and glycine for our final parameters, which are in good qualitative agreement with the ϕ_i, ψ_i distributions of real proteins [8, 32].

Finally, we determined the strengths of the hydrogen-bond and hydrophobicity terms on the basis of the resulting overall thermodynamic behavior of the three-helix sequence. For this purpose, we performed a set of trial runs for fixed values of the other parameters. An alternative would have been to use the method of Shea *et al.* [33]. The result of our empirical determination of $\epsilon_{\rm hb}$ and $\epsilon_{\rm AA}$ does not seem unreasonable; at the folding temperature of the three-helix sequence (see below), we get $\epsilon_{\rm hb}/kT \approx 4.3$ and $\epsilon_{\rm AA}/kT \approx 3.4$.

In this model, we study the three sequences shown in Table 1.3, which contain

[1] The cutoff procedure is $f(r) \mapsto \tilde{f}(r)$ where $\tilde{f}(r) = f(r) - f(r_c) - (r - r_c)f'(r_c)$ if $r < r_c$ and $\tilde{f}(r) = 0$ otherwise.

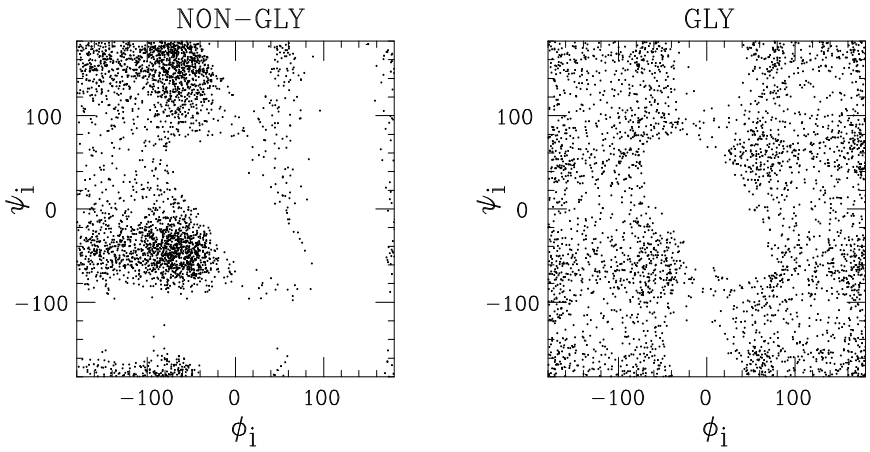


Figure 1.2: ϕ_i, ψ_i scatter plots for nonglycine and glycine, as obtained by simulations of the chains GXG for X=A/B and X=G, respectively, at $kT = 0.625$ (shown is ϕ_i, ψ_i for X).

1H:	BBABBAABBABBAABB
2H:	1H–GGG–1H
3H:	1H–GGG–1H–GGG–1H

Table 1.3: The sequences studied.

16, 35 and 54 amino acids, respectively. Following the strategy of Regan and DeGrado [11], the A and B amino acids are distributed along the sequence 1H in such a way that this segment can form a helix with all hydrophobic amino acids on the same side. The sequence 3H, consisting of three such stretches of As and Bs plus two GGG segments, is meant to form a three-helix bundle. This particular sequence was recently studied by Takada *et al.* [18], who used a more elaborate model with nonadditive forces.

1.3 Results

To study the thermodynamic behavior of the chains described in the previous section, we use the method of simulated tempering. This means that we first select a set of allowed temperatures and then perform simulations in which the temperature is a dynamical variable. This is done to speed up low-temperature

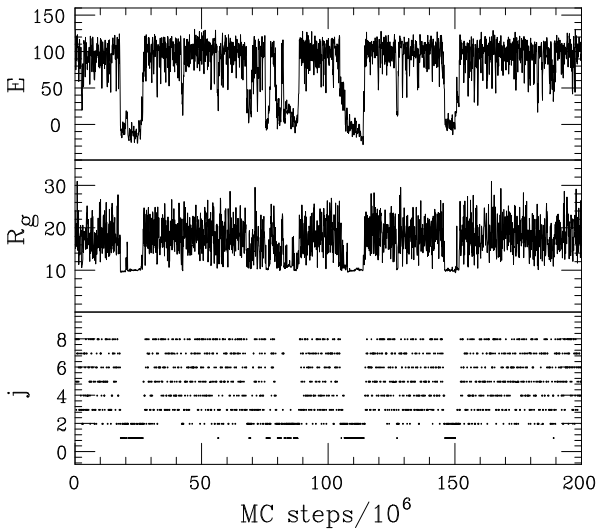


Figure 1.3: Monte Carlo evolution of the energy and radius of gyration in a typical simulation of the three-helix sequence. The bottom panel shows how the system jumps between the allowed temperatures T_j, which are given by $T_j = T_{\min}(T_{\max}/T_{\min})^{(j-1)/(J-1)}$ [34] with $kT_{\min} = 0.625$, $kT_{\max} = 0.9$ and $J = 8$. The temperature $T_{\min}$ is chosen to lie just below the collapse transition, whereas $T_{\max}$ is well into the coil phase (see Fig. 1.4).

simulations. In addition, it provides a convenient method for calculating free energies.

An example of a simulated-tempering run is given in Fig. 1.3, which shows the Monte Carlo evolution of the energy E and radius of gyration $R_{\rm g}$ (calculated over all backbone atoms) in a simulation of the three-helix sequence. Also shown, bottom panel, is how the system jumps between the different temperatures. Two distinct types of behavior can be seen. In one case, E is high, fluctuations in size are large, and the temperatures visited are high. In the other case, E is low, the size is small and almost frozen, and the temperatures visited are low. Interesting to note is that there is one temperature, the next-lowest one, which is visited in both cases. Apparently, both types of behavior are possible at this temperature.

In Fig. 1.4a we show the specific heat as a function of temperature for the one-, two- and three-helix sequences. A pronounced peak can be seen that gets stronger with increasing chain length. In fact, the increase in height is not inconsistent with a linear dependence on chain length, which is what one

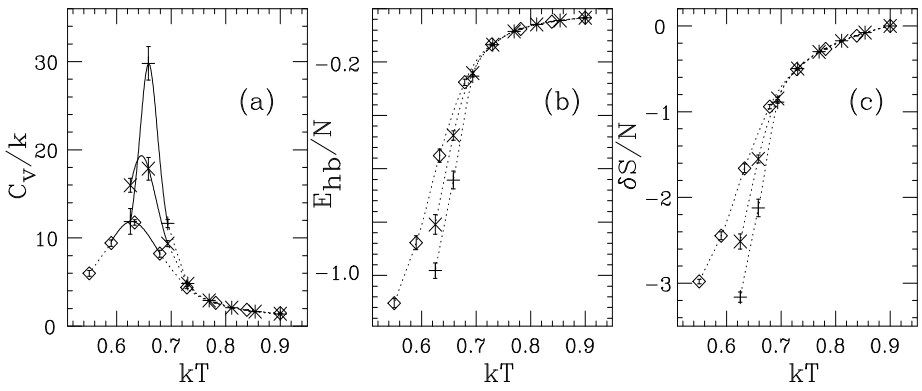


Figure 1.4: Thermodynamic functions against temperature for the sequences 1H ($\diamond$), 2H ($\times$) and 3H ($+$) in Table 1.3. (a) Specific heat $C_v = (\langle E^2\rangle - \langle E\rangle^2)/NkT^2$, N being the number of amino acids. (b) Hydrogen-bond energy per amino acid, E_{hb}/N. (c) Chain entropy per amino acid, $\delta S/N = [S - S(kT = 0.9)]/N$. The full lines in (a) represent single-histogram extrapolations [35]. Dotted lines are drawn to guide the eye.

would have expected if it had been a conventional first-order phase transition with a latent heat.

Our results for the radius of gyration (not displayed) show that the specific heat maximum can be viewed as the collapse temperature. The specific heat maximum is also where hydrogen-bond formation occurs, as can be seen from Fig. 1.4b. Important to note in this figure is that the decrease in hydrogen-bond energy *per amino acid* with decreasing temperature is most rapid for the three-helix sequence, which implies that, compared to the shorter ones, this sequence forms more stable secondary structure. The results for the chain entropy shown in Fig. 1.4c provide further support for this; the entropy loss per amino acid with decreasing temperature is largest for the three-helix sequence.

It should be stressed that the character of the collapse transition depends strongly on the relative strength of the hydrogen-bond and hydrophobicity terms. Figure 1.4 shows that the transition is very abrupt or "first-order-like" for our choice $(\epsilon_{hb}, \epsilon_{AA}) = (2.8, 2.2)$. A fairly small decrease of $\epsilon_{hb}/\epsilon_{AA}$ is sufficient to get a very different behavior with, for example, a much weaker peak in the specific heat. In this case, the chain collapses to a molten globule without specific structure rather than to a three-helix bundle. A substantially weakened transition was observed for $\epsilon_{hb} = \epsilon_{AA} = 2.5$. If, on the other hand,

Figure 1.5: Representative low-temperature structures, FU and BU, respectively. Drawn with RasMol [36].

$\epsilon_{\rm hb}/\epsilon_{\rm AA}$ is too large, then it is evident that the chain will form one long helix instead of a helical bundle.

We now turn to the three-dimensional structure of the three-helix sequence in the collapsed phase. It turns out that it does form a three-helix bundle. This bundle can have two distinct topologies: if we let the first two helices form a U, then the third helix can be either in front of or behind that U. The model is, not unexpectedly, unable to discriminate between these two possibilities. To characterize low-temperature conformations, we therefore determined two representative structures, one for each topology, which, following [18], are referred to as FU and BU, respectively. These structures are shown in Fig. 1.5. They were generated by quenching a large number of low-T structures to zero temperature, and we feel convinced that they provide good approximations of the energy minima for the respective topologies. Given an arbitrary conformation, we then measure the root-mean-square distances δ_i ($i = {\rm FU}, {\rm BU}$) to these two structures (calculated over all backbone atoms). These distances are converted into similarity parameters Q_i by using

$$Q_i = \exp(-\delta_i^2/100\text{Å}^2)\,. \tag{1.8}$$

At temperatures above the specific heat maximum, both Q_i tend to be small. At temperatures below this point, the system is found to spend most of its time close to one or the other of the representative structures; either $Q_{\rm FU}$ or $Q_{\rm BU}$ is close to 1. Finally, at the peak, all three of these regions in the $Q_{\rm FU}, Q_{\rm BU}$ plane are populated, as can be seen from Fig. 1.6a. In particular, this implies that the folding transition coincides with the specific heat maximum.

The folding transition can be described in terms of a single "order parameter" by taking $Q = \max(Q_{\rm FU}, Q_{\rm BU})$ as a measure of nativeness. Correspondingly, we put $\delta = \min(\delta_{\rm FU}, \delta_{\rm BU})$. In Fig. 1.6b, we show the free-energy profile $F(Q)$ at the folding temperature. The free energy has a relatively sharp minimum at

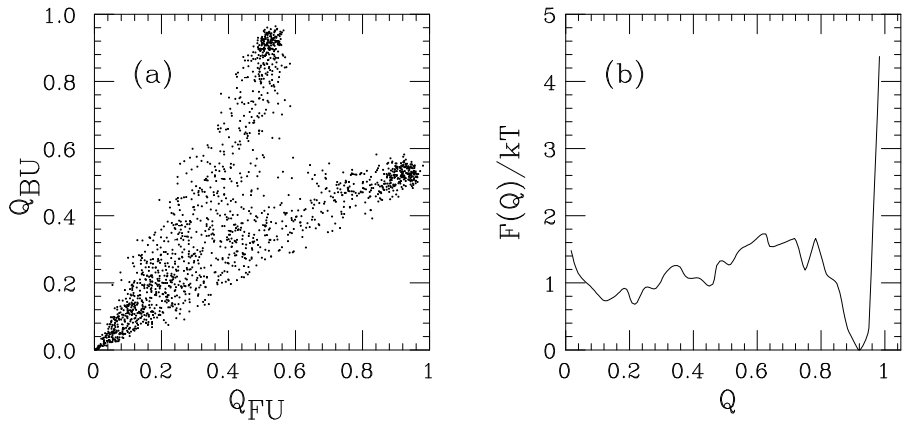


Figure 1.6: (a) $Q_{\rm FU}, Q_{\rm BU}$ (see Eq. 1.8) scatter plot at the specific heat maximum ($kT = 0.658$). (b) Free energy $F(Q)$ as a function of $Q = \max(Q_{\rm FU}, Q_{\rm BU})$ at the same temperature.

$Q \approx 0.9$, corresponding to $\delta \approx 3$Å. This is followed by a weak barrier around $Q = 0.7$, corresponding to $\delta \approx 6$Å. Finally, there is a broad minimum at small Q, where $Q = 0.2$ corresponds to $\delta \approx 13$Å.

What does the nonnative population at the folding temperature correspond to in terms of $R_{\rm g}$ and $E_{\rm hb}$? This can be seen from the $Q, R_{\rm g}$ and $Q, E_{\rm hb}$ scatter plots in Fig. 1.7. These plots show that the low-Q minimum of $F(Q)$ corresponds to expanded structures with a varying but not high secondary-structure content. Although a detailed kinetic study is beyond the scope of this paper, we furthermore note that the free-energy surfaces corresponding to the distributions in Fig. 1.7 are relatively smooth. Consistent with that, we found that standard fixed-temperature Monte Carlo simulations were able to reach the native state, starting from random coils.

Let us finally mention that we also performed simulations of some random sequences with the same length and composition as the three-helix sequence. The random sequences did not form stable structures and collapsed more slowly with decreasing temperature than the designed three-helix sequence.

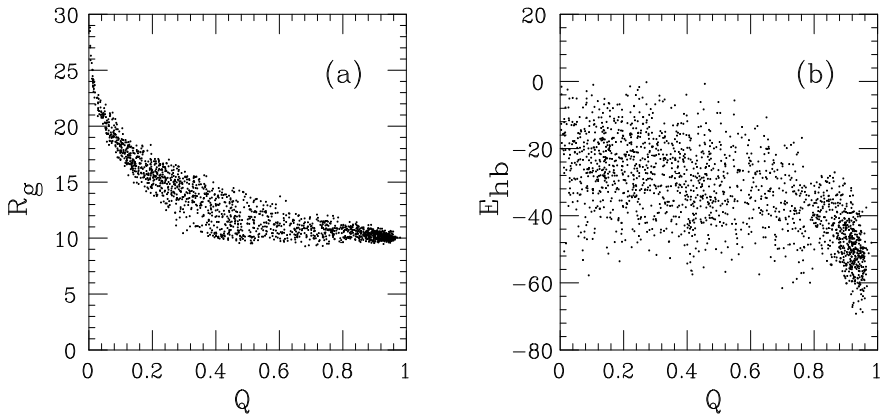


Figure 1.7: (a) $Q, R_{\rm g}$ and (b) $Q, E_{\rm hb}$ scatter plots at the folding temperature ($kT = 0.658$).

1.4 Summary and Outlook

We have studied a reduced protein model where the formation of native structure is driven by a competition between hydrogen bonds and effective hydrophobicity forces. Using this force field, we find that the three-helix-bundle protein studied has the following properties:

- It does form a stable three-helix-bundle state, except for a 2-fold topological degeneracy.
- It undergoes an abrupt folding transition from an expanded state to the native state.
- It forms more stable secondary structure than the corresponding one- and two-helix segments.

An obvious question that remains to be addressed is what is needed to lift the topological degeneracy. Not obvious, however, is whether this question should be addressed at the present level of modeling, before including full side chains.

A first-order-like folding transition that takes the system directly from the unfolded state to the native one is what one expects for small fast-folding proteins. For the model to show this behavior, careful tuning of the relative strength of the hydrogen-bond and hydrophobicity terms, $\epsilon_{\rm hb}/\epsilon_{\rm AA}$, is required. This $\epsilon_{\rm hb}/\epsilon_{\rm AA}$ dependence may at first glance seem unwanted but is not physically unreasonable; $\epsilon_{\rm hb}$ can be thought of partly as a stiffness parameter, and chain

stiffness has important implications for the phase structure, as shown by recent work on homopolymers [26–29]. Note also that incorporation of full side chains makes the chains intrinsically stiffer, which might lead to a weaker $\epsilon_{\rm hb}/\epsilon_{\rm AA}$ dependence.

Our three-helix sequence has previously been studied by Takada *et al.* [18], who used a more elaborate force field. It was suggested that it is essential to use context-dependent hydrogen bonds for the three-helix-bundle protein to make more stable secondary structure than its one-helix fragments. Our model shows this behavior, although its hydrogen bonds are context-independent.

Let us finally stress that we find a first-order-like folding transition without using the Gō approximation. Evidence for first-order-like folding transitions has been found for proteins with similar lengths in some C_α models [5,14,17,33], but these studies use this approximation.

Acknowledgments

This work was in part supported by the Swedish Foundation for Strategic Research.

References

[1] Săli, A., Shakhnovich, E. & Karplus, M. (1994) "Kinetics of Protein Folding: A Lattice Model Study of the Requirements for Folding to the Native State", *J. Mol. Biol.* **235**, 1614–1636.

[2] Bryngelson, J.D., Onuchic, J.N., Socci, N.D. & Wolynes, P.G. (1995) "Funnels, Pathways, and the Energy Landscape of Protein Folding: A Synthesis", *Proteins Struct. Funct. Genet.* **21**, 167–195.

[3] Dill, K.A. & Chan, H.S. (1997) "From Levinthal to Pathways to Funnels", *Nat. Struct. Biol.* **4**, 10–19.

[4] Klimov, D.K. & Thirumalai, D. (1998) "Linking Rates of Folding in Lattice Models of Proteins with Underlying Thermodynamic Characteristics", *J. Chem. Phys.* **109**, 4119–4125.

[5] Nymeyer, H., García, A.E. & Onuchic, J.N. (1998) "Folding Funnels and Frustration in Off-lattice Minimalist Protein Landscapes", *Proc. Natl. Acad. Sci. USA* **95**, 5921–5928.

[6] Pande, V.S., Grosberg, A.Y. & Tanaka, T. (1994) "Thermodynamic Procedure to Synthesize Heteropolymers that Can Renature to Recognize a Given Target Molecule", *Proc. Natl. Acad. Sci. USA* **91**, 12976–12979.

[7] Irbäck, A., Peterson, C. and Potthast, F. (1996) "Evidence for Nonrandom Hydrophobicity Structures in Protein Chains", *Proc. Natl. Acad. Sci. USA* **93**, 9533–9538.

[8] Ramachandran, G.N. & Sasisekharan, V. (1968) "Conformation of Polypeptides and Proteins", *Adv. Protein Chem.* **23**, 283–437.

[9] Dill, K.A. (1990) "Dominant Forces in Protein Folding", *Biochemistry* **29**, 7133–7155.

[10] Privalov, P.L. (1992) "Physical Basis of the Stability of the Folded Conformations of Proteins", in *Protein Folding*, ed. Creighton, T.E. (Freeman, New York), pp. 83–126.

[11] Regan, L. & DeGrado, W.F. (1988) "Characterization of a Helical Protein Designed from First Principles", *Science* **241**, 976–978.

[12] Rey, A. & Skolnick, J. (1993) "Computer Modeling and Folding of Four-helix Bundles", *Proteins Struct. Funct. Genet.* **16**, 8–28.

[13] Guo, Z. & Thirumalai, D. (1996) "Kinetics and Thermodynamics of Folding of a *de Novo* Designed Four-helix Bundle Protein", *J. Mol. Biol.* **263**, 323–343.

[14] Zhou, Z. & Karplus, M. (1997) "Folding Thermodynamics of a Model Three-helix-bundle Protein", *Proc. Natl. Acad. Sci. USA* **94**, 14429–14432.

[15] Koretke, K.K., Luthey-Schulten, Z. & Wolynes, P.G. (1998) "Self-consistently Optimized Energy Functions for Protein Structure Prediction by Molecular Dynamics", *Proc. Natl. Acad. Sci. USA* **95**, 2932–2937.

[16] Hardin, C., Luthey-Schulten, Z. & Wolynes, P.G. (1999) "Backbone Dynamics, Fast Folding, and Secondary Structure Formation in Helical Proteins and Peptides", *Proteins Struct. Funct. Genet.* **34**, 281–294.

[17] Shea, J.-E., Onuchic, J.N. & Brooks, C.L., III (1999) "Exploring the Origins of Topological Frustration: Design of a Minimally Frustrated Model of Fragment B of Protein A", *Proc. Natl. Acad. Sci. USA* **96**, 12512–12517.

[18] Takada, S., Luthey-Schulten, Z. & Wolynes, P.G. (1999) "Folding Dynamics with Nonadditive Forces: A Simulation Study of a Designed Helical Protein and a Random Heteropolymer", *J. Chem. Phys.* **110**, 11616–11629.

[19] Bottomley, S.P., Popplewell, A.G., Scawen, M., Wan, T., Sutton, B.J. & Gore, M.G. (1994) "The Stability and Unfolding of an IgG Binding Protein Based upon the B Domain of Protein A from *Staphylococcus Aureus* Probed by Tryptophan Substitution and Fluorescence Spectroscopy", *Protein Eng.* **7**, 1463–1470.

[20] Bai, Y., Karimi, A., Dyson, H.J. & Wright, P.E. (1997) "Absence of a Stable Intermediate on the Folding Pathway of Protein A" *Protein Sci.* **6**, 1449–1457.

[21] Guo, Z., Brooks, C.L., III & Boczko, E.M. (1997) "Exploring the Folding Free Energy Surface of a Three-helix Bundle Protein", *Proc. Natl. Acad. Sci. USA* **94**, 10161-10166.

[22] Kolinski, A., Galazka, W. & Skolnick, J. (1998) "Monte Carlo Studies of the Thermodynamics and Kinetics of Reduced Protein Models: Application to Small helical, β and α/β Proteins", *J. Chem. Phys.* **108**, 2608–2617.

[23] Lyubartsev, A.P., Martsinovski, A.A., Shevkunov, S.V. & Vorontsov-Velyaminov, P.V. (1992) "New Approach to Monte Carlo Calculation of the Free Energy: Method of Expanded Ensembles", *J. Chem. Phys.* **96**, 1776–1783.

[24] Marinari, E. & Parisi, G. (1992) "Simulated Tempering: A New Monte Carlo Scheme", *Europhys. Lett.* **19**, 451–458.

[25] Irbäck, A. & Potthast, F. (1995) "Studies of an Off-lattice Model for Protein Folding: Sequence Dependence and Improved Sampling at Finite Temperature", *J. Chem. Phys.* **103**, 10298–10305.

[26] Kolinski, A., Skolnick, J. & Yaris, R. (1986) "Monte Carlo Simulations on an Equilibrium Globular Protein Folding Model", *Proc. Natl. Acad. Sci. USA* **83**, 7267–7271.

[27] Doniach, S., Garel, T. & Orland, H. (1996) "Phase Diagram of a Semiflexible Polymer Chain in a θ Solvent: Application to Protein Folding", *J. Chem. Phys.* **105**, 1601–1608.

[28] Bastolla, U. & Grassberger, P. (1997) "Phase Transitions of Single Semistiff Polymer Chains", *J. Stat. Phys.* **89**, 1061–1078.

[29] Doye, J.P.K., Sear, R.P. & Frenkel, D. (1998) "The Effect of Chain Stiffness on the Phase Behaviour of Isolated Homopolymers", *J. Chem. Phys.* **108**, 2134–2142.

[30] Gō, N. & Taketomi, H. (1978) "Respective Roles of Short- and Long-range Interactions in Protein Folding", *Proc. Natl. Acad. Sci. USA* **75**, 559–563.

[31] Ishikawa, K., Yue, K. & Dill, K.A. (1999) "Predicting the Structures of 18 Peptides Using Geocore", *Protein Sci.* **8**, 716–721.

[32] Zimmerman, S.S., Pottle, M.S., Némethy, G. & Scheraga, H.A. (1977) "Conformational Analysis of the 20 Naturally Occurring Amino Acid Residues Using ECEPP", *Macromolecules* **10**, 1–9.

[33] Shea, J.-E., Nochomovitz, Y.D., Guo, Z. & Brooks, C.L., III (1998) "Exploring the Space of Protein Folding Hamiltonians: The Balance of Forces in a Minimalist β-barrel Model", *J. Chem. Phys.* **109**, 2895–2903.

[34] Hansmann, U.H.E. & Okamoto, Y. (1997) "Numerical Comparisons of Three Recently Proposed Algorithms in the Protein Folding Problem", *J. Comput. Chem.* **18**, 920–933.

[35] Ferrenberg, A.M. & Swendsen, R.H. (1988) "New Monte Carlo for Studying Phase Transitions" *Phys. Rev. Lett.* **61**, 2635–2638, and erratum (1989) **63**, 1658, and references given in the erratum.

[36] Sayle, R. & Milner-White, E.J. (1995) "RasMol: Biomolecular Graphics for All", *Trends Biochem. Sci.* **20**, 374–376.

Folding of a Small Helical Protein Using Hydrogen Bonds and Hydrophobicity Forces

Paper II

Folding of a Small Helical Protein Using Hydrogen Bonds and Hydrophobicity Forces

Giorgio Favrin, Anders Irbäck and Stefan Wallin

Complex Systems Division, Department of Theoretical Physics
Lund University, Sölvegatan 14A, SE-223 62 Lund, Sweden
http://www.thep.lu.se/complex/

Proteins: Structure, Function, and Genetics **47**, 99-105 (2002)

Abstract:
A reduced protein model with five to six atoms per amino acid and five amino acid types is developed and tested on a three-helix-bundle protein, a 46-amino acid fragment from staphylococcal protein A. The model does not rely on the widely used Gō approximation where non-native interactions are ignored. We find that the collapse transition is considerably more abrupt for the protein A sequence than for random sequences with the same composition. The chain collapse is found to be at least as fast as helix formation. Energy minimization restricted to the thermodynamically favored topology gives a structure that has a root-mean-square deviation of 1.8 Å from the native structure. The sequence-dependent part of our potential is pairwise additive. Our calculations suggest that fine-tuning this potential by parameter optimization is of limited use.

2.1 Introduction

In recent years, several important insights have been gained into the physical principles of protein folding [1–6]. Still, in terms of quantitative predictions, it is clear that it would be extremely useful to be able to perform more realistic folding simulations than what is currently possible. In fact, most models that have been used so far for statistical-mechanical simulations of folding rely on one or both of two quite drastic approximations, the lattice and Gō [7] approximations.

The reason that lattice models have been used to study basics of protein folding is partly computational, but also physical — on the lattice, it is known what potential to use in order for stable and fast-folding sequences to exist (a simple contact potential is sufficient). How to satisfy these criteria for off-lattice chains is, by contrast, largely unknown, and therefore many current off-lattice models [5, 8–14] use Gō-type potentials [7] where non-native interactions are ignored. The use of the Gō approximation has some support from the finding that the native structure is a determinant for folding kinetics [15,16]. However, it is an uncontrolled approximation, and it is, of course, useless when it comes to structure prediction, as it requires prior knowledge of the native structure.

In this paper, we discuss an off-lattice model that does not follow the Gō prescription. Using this model, we perform extensive folding simulations for a small helical protein. The force field of the model is simple and based on hydrogen bonds and effective hydrophobicity forces (no explicit water). There exist other non Gō-like models with more elaborate force fields that have been used for structure prediction with some success [17–19]. However, it is unclear what the dynamical properties of these models are.

The original version of our model was presented in Ref. [20] and has three types of amino acids: hydrophobic, polar and glycine. This version was applied to a designed three-helix-bundle protein with 54 amino acids [20]. For a suitable relative strength of the hydrogen bonds and hydrophobicity forces, it was found that this sequence does form a stable three-helix bundle, except for a twofold topological degeneracy, and that its folding transition is first-order-like and coincides with the collapse transition (the parameter σ of Ref. [4] is zero).

Here, we extend this model from three to five amino acid types, by taking alanine to be intermediate in hydrophobicity between the previous two hydrophobic and polar classes, and by introducing a special geometric representation for proline, which is needed to be able to mimic the helix-breaking property of this amino acid. Otherwise, the model is the same as before. The modified model is tested on a real three-helix-bundle protein, the 10–55-amino acid fragment

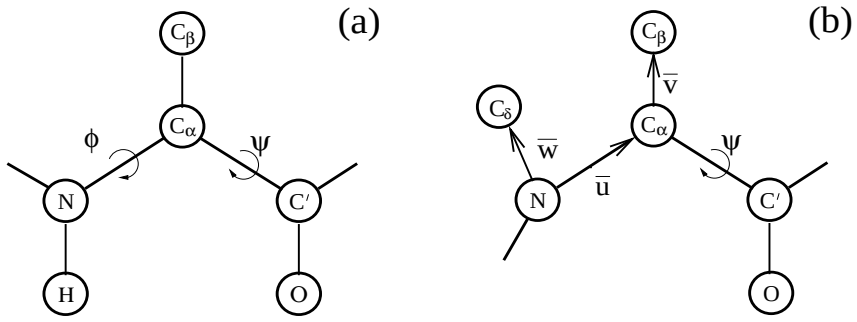


Figure 2.1: (a) Schematic figure showing the common geometric representation for all amino acids except glycine and proline. (b) The representation of proline. The C_δ atom is assumed to lie in the plane of the N, C_α and C_β atoms. The N-C_δ bond vector $\bar{w}$ is given by $\bar{w} = -0.596\bar{u} + 0.910\bar{v}$, where the vectors $\bar{u}$ and $\bar{v}$ are defined in the figure. The numerical factors were obtained by an analysis of structures from the Protein Data Bank (PDB) [27].

of the B domain of staphylococcal protein A. The structure of this protein has been determined by NMR [21], and an energy-based structure prediction method has been tested on the sequence [17]. The folding properties have been studied too, both experimentally [22, 23] and theoretically [8, 10, 11, 24–26]. In particular, this means that we can compare the behavior of previous Gō-like models to that of our more realistic model.

2.2 Materials and Methods

2.2.1 Geometry

Our model is an extension of that introduced in Ref. [20]. It uses three different amino acid representations: one for glycine, one for proline and one for the rest. The non-glycine, non-proline representation is illustrated in Fig. 2.1a, and is identical to that of hydrophobic and polar amino acids in the original model. The three backbone atoms N, C_α and C′ are all included, whereas the side chain is represented by a single atom, a large C_β. The remaining two atoms, H and O, are used to define hydrogen bonds. The representation of glycine is the same except that C_β is missing.

The representation of proline is new compared to the original model. The side

chain of proline is attached to the backbone not only at C_α, but also at N. A well-known consequence of this is that proline can act as a helix breaker. For the model to be able to capture this important property, we introduce a special representation for proline, which is illustrated in Fig. 2.1b. It differs from that in Fig. 2.1a in two ways: first, the Ramachandran angle ϕ is held constant, at $-65°$; and second, the H atom is replaced by a side-chain atom, C_δ. This more realistic representation of proline is needed when studying the protein A fragment which has one proline at each of the two turns.

All amino acids except proline have the Ramachandran torsion angles ϕ and ψ (see Fig. 2.1a) as their degrees of freedom, whereas ψ is the only degree of freedom for proline. All bond lengths, bond angles and peptide torsion angles ($180°$) are held fixed. Numerical values of the bond lengths and bond angles can be found in Ref. [20] and Fig. 2.1b.

The helix-breaking property of proline manifests itself clearly in the shape of the ψ distribution for amino acids that are followed by a proline in the sequence (with the proline on their C′ side). Helical values of ψ are suppressed for such amino acids. This is illustrated in Fig. 2.2a, where the peak on the left corresponds to α-helix. From Fig. 2.2b, it can be seen that the model shows a qualitatively similar behavior.

2.2.2 Force Field

Our energy function

$$E = E_{\rm loc} + E_{\rm sa} + E_{\rm hb} + E_{\rm col} \tag{2.1}$$

is composed of four terms. The first two terms $E_{\rm loc}$ and $E_{\rm sa}$ are local ϕ, ψ and self-avoidance potentials, respectively (see Ref. [20]). The third term is the hydrogen-bond energy $E_{\rm hb}$, which is given by

$$E_{\rm hb} = \epsilon_{\rm hb} \sum_{ij} \left[5 \left(\frac{\sigma_{\rm hb}}{r_{ij}} \right)^{12} - 6 \left(\frac{\sigma_{\rm hb}}{r_{ij}} \right)^{10} \right] v(\alpha_{ij}, \beta_{ij}) \tag{2.2}$$

$$v(\alpha_{ij}, \beta_{ij}) = \begin{cases} \cos^2 \alpha_{ij} \cos^2 \beta_{ij} & \alpha_{ij}, \beta_{ij} > 90° \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

where i and j represent H and O atoms, respectively, and where r_{ij} denotes the HO distance, α_{ij} the NHO angle, and β_{ij} the HOC′ angle.

The last term in Eq. (2.1), the hydrophobicity or collapse energy $E_{\rm col}$, has the

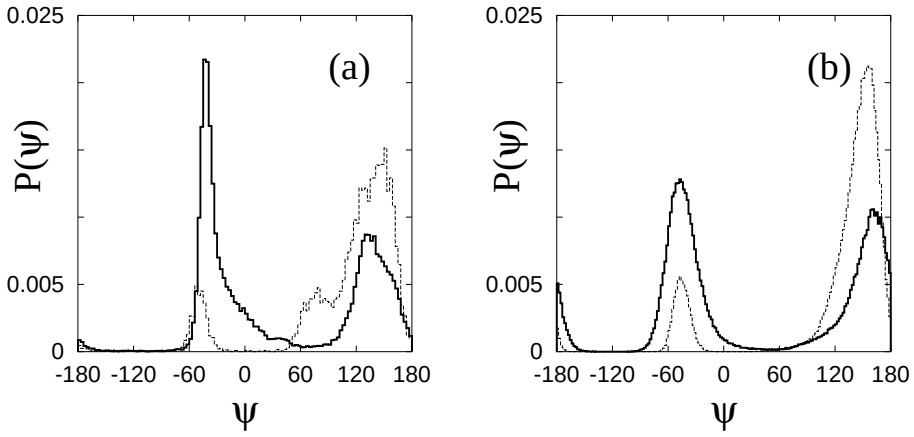


Figure 2.2: (a) Distributions of the Ramachandran angle ψ, based on PDB data. The full (dashed) line represents non-glycine, non-proline amino acids that are followed by a non-proline (proline) in the sequence. (b) The corresponding histograms for the model, as obtained by simulations of Gly-X-X (full line) and Gly-X-Pro (dashed line) at $kT = 0.55$, where X denotes polar amino acids (shown is the ψ distribution for the middle of the three amino acids).

form

$$E_{\rm col} = \epsilon_{\rm col} \sum_{i<j} \Delta(s_i, s_j) \left[\left(\frac{\sigma_{\rm col}}{r_{ij}} \right)^{12} - 2 \left(\frac{\sigma_{\rm col}}{r_{ij}} \right)^{6} \right] , \tag{2.4}$$

where the sum runs over all possible $C_\beta C_\beta$ pairs and s_i denotes amino acid type. To define $\Delta(s_i, s_j)$, we divide the amino acids into three classes: hydrophobic (H; Leu, Ile, Phe), alanine (A; Ala) and polar (P; Arg, Asn, Asp, Gln, Glu, His, Lys, Pro, Ser, Tyr).[1] There are then six kinds of $C_\beta C_\beta$ pairs, and the corresponding $\Delta(s_i, s_j)$ values are taken to be

$$\Delta(s_i, s_j) = \begin{cases} 1 & \text{for HH and HA pairs} \\ 0 & \text{for HP, AA, AP and PP pairs} \end{cases} \tag{2.5}$$

The main change in the force field compared to Ref. [20] is that alanine forms its own hydrophobicity class, besides the previous two hydrophobic and polar classes. Alanine is taken as intermediate in hydrophobicity, meaning that there is a hydrophobic interaction between HA pairs but not between AA pairs. In

[1] Cys, Met, Thr, Trp and Val do not occur in the sequence studied.

addition, the interaction strength $\epsilon_{\rm col}$ is increased slightly, from 2.2 to 2.3.[2] Finally, in the self-avoidance potential, the C_δ atom of proline is assigned the same size as C_β atoms. Otherwise, the entire force field, including parameter values, is exactly the same as in Ref. [20].

With these changes in geometry and force field, we end up with five different amino acid types in the new model. First, we have hydrophobic, alanine and polar which share the same geometric representation but differ in hydrophobicity, and then glycine and proline with their special geometries.

In this paper, we test this model on the 10–55-amino acid fragment of the B domain of staphylococcal protein A. Calculated structures are compared to the minimized average NMR structure [21] with PDB code 1bdd. Throughout the paper, this structure is referred to as the native structure.

As a first test of our model, two different fits to the native structure were made. The first fit is purely geometrical. Here, we simply minimized the root-mean-square deviation (rmsd) from the native structure, δ (calculated over all backbone atoms). This was done by using simulated annealing, and the best result was $\delta = 0.14$ Å. In the second fit, we took into account the limitations imposed by the first three terms of the potential, by minimizing the function

$$\tilde{E} = E_{\rm loc} + E_{\rm sa} + E_{\rm hb} + \kappa \sum_i (\mathbf{r}_i - \mathbf{r}_i^0)^2 \,, \tag{2.6}$$

where $\kappa = 1$ Å^{-2} and $\{\mathbf{r}_i^0\}$ denotes the structure obtained from the first fit. The minimum-$\tilde{E}$ structure had $\delta = 0.32$ Å. These results show that our model, in spite of relatively few degrees of freedom, permits a quite accurate description of the real structure.

2.2.3 Numerical Methods

To simulate the thermodynamic behavior of this model, we use simulated tempering [28–30], which means that the temperature is a dynamical variable (for details, see Refs. [28–30]). The temperature update is a standard Metropolis step. Our conformation updates are of two different types: the simple non-local pivot move where a single torsion angle is turned, and the semi-local biased Gaussian step proposed in Ref. [31]. The latter method works with the Ramachandran angles of four adjacent amino acids. These are turned with a bias toward local rearrangements of the chain. The degree of bias is governed by

[2] The energy unit is dimensionless and such that $kT_{\rm c} = 0.62$, $T_{\rm c}$ being the collapse temperature (see Sec. 2.3).

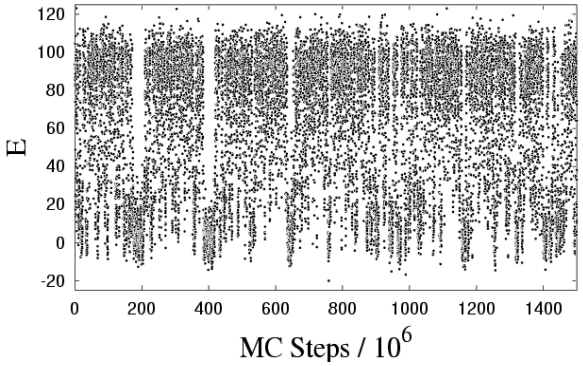


Figure 2.3: Monte Carlo evolution of the energy in a simulated-tempering run.

a parameter b. In our thermodynamic simulations, we take $b = 10$ (rad/Å)^2, which gives a strong bias toward deformations that are approximately local [31].

Figure 2.3 shows the evolution of the energy in a simulated-tempering run that took about two weeks on an 800 MHz processor. Data corresponding to all the different temperatures are shown (eight temperatures, ranging from $kT = 0.54$ to $kT = 0.90$). We see that there are many independent visits to low-energy states, which is necessary in order to get a reliable estimate of the relative populations of the folded and unfolded states. To test the usefulness of the semi-local update, we repeated the same calculation using pivot moves only. The difference in performance was not quantified, but it was clear that the sampling of low energies was less efficient in the run relying solely on pivot moves.

For our kinetic simulations, we do not use the pivot update but only the semi-local method. The parameter b is taken to be 1 (rad/Å)^2 in the kinetic runs, which turned out to give an average change in the end-to-end vector squared of about 0.5 Å^2.

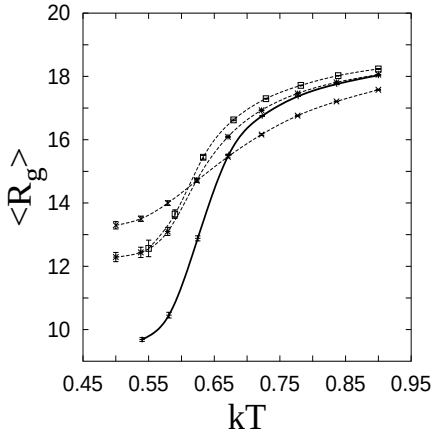


Figure 2.4: The radius of gyration (in Å) against temperature. Full and dashed lines represent the protein A sequence and the three random sequences (see the text), respectively.

2.3 Results and Discussion

2.3.1 Thermodynamics

We begin our study of the model defined in Sec. 2.2 by locating the collapse transition. In Fig. 2.4, we show the radius of gyration (calculated over all backbone atoms) against temperature for both the protein A sequence and three random sequences with the same length and composition. The random sequences were generated keeping the two prolines of the protein A sequence fixed at their positions, one at each turn. The remaining 44 amino acids were randomly reshuffled.

Naively, one may expect these sequences to show similar collapse behaviors, since the composition is the same. However, the protein A sequence turns out to collapse much more efficiently than the random sequences (see Fig. 2.4). The native structure has a radius of gyration of 9.25 Å, which is significantly smaller than one finds for the random sequences in this temperature range. The specific heat (data not shown) has a pronounced peak in the region where the collapse occurs. Taking the maximum as the collapse temperature $T_{\rm c}$, we obtain $kT_{\rm c} = 0.62$ for the protein A sequence.

The chain collapse is not as abrupt for the protein A sequence as for the de-

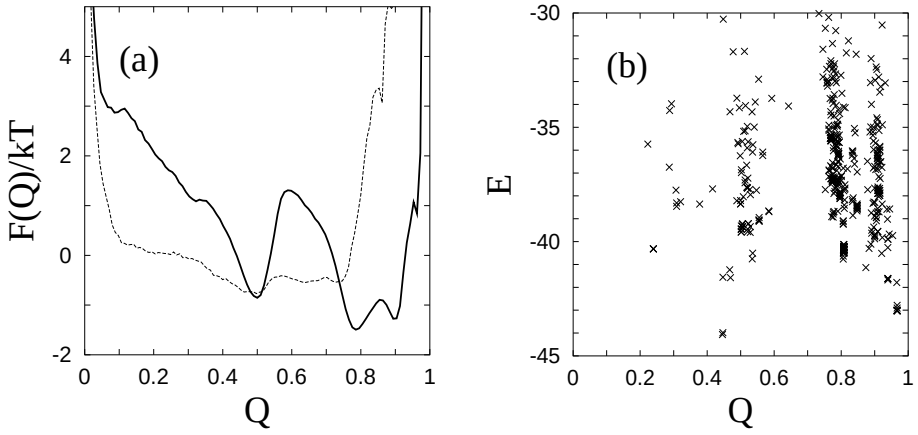


Figure 2.5: (a) Free-energy profile $F(Q) = -kT \ln P(Q)$ at $kT = 0.54$ (full line), where $P(Q)$ is the probability distribution of Q. Also shown (dashed line) is the result for one of the random sequences at $kT = 0.50$. (b) Q, E scatter plot for quenched conformations with low energy.

signed sequence studied in Ref. [20]. This is not surprising, as that sequence has a hydrophobicity pattern that fits its native structure perfectly. The protein A sequence does not have a fully perfect hydrophobicity pattern, but still the collapse behavior is highly cooperative, as can be seen from the comparison with the random sequences.

Next, we turn to the structure of the collapsed state. As a measure of similarity with the native structure, we use

$$Q = \exp(-\delta^2/100 \text{ Å}^2)\,, \tag{2.7}$$

where δ, as before, denotes rmsd. An alternative would be to base the similarity measure on the number of native contacts present, rather than rmsd. The problem with such a definition is that it does not provide an efficient discrimination between the two possible topologies of a three-helix bundle [32] — the third helix can be either in front of or behind the U formed by the first two helices. This problem is avoided by using rmsd.

In Fig. 2.5a, we show the free-energy profile $F(Q)$ in the collapsed phase at $kT = 0.54$. We see that there is a broad minimum at $Q \approx 0.8$–0.9, with two distinct local minima at $Q = 0.78$ and $Q = 0.90$, respectively. Both these minima correspond to the native overall topology. There is also a minimum at $Q = 0.50$, which corresponds to the wrong topology. The $Q = 0.50$ minimum

is more narrow and slightly higher, so the native topology is the favored one. However, it should be stressed that it is difficult to discriminate between the two topologies using a pairwise additive potential (see Sec. 2.3.4). To be able to do that in a proper way, it is likely that one has to include multibody terms and/or more side-chain atoms in the model.

The main difference between the two minima at $Q = 0.78$ and $Q = 0.90$ lies in the shape and orientation of helix III, which comprises amino acids 41–55 in the native structure. At the $Q = 0.78$ minimum, there tends to be a sharp bend in this segment, and the amino acids before the bend, 41–44, are disordered rather than helical. The remaining amino acids, 45–55, tend to make a helix, but its orientation differs from that in the native structure. Relative to the $Q = 0.90$ minimum, where helix III is much more native-like, we find that the $Q = 0.78$ minimum is entropically favored but energetically disfavored. The separation in energy between these minima is probably underestimated by our model. There is, for example, a stabilizing electrostatic interaction between helices I and III in the native structure (Glu16-Lys50), which should favor the $Q = 0.90$ minimum but is missing in our model.

Also shown in Fig. 2.5a is the result for one of the random sequences. The probability of finding this sequence in the vicinity of the native structure is, not unexpectedly, very low. The same holds true for the other two random sequences too (data not shown).

To extract representative conformations for the collapsed state, we used simulated annealing followed by a conjugate-gradient minimization. Using this procedure, a large set of low-temperature Monte Carlo conformations were quenched to zero temperature. In Fig. 2.5b, we show the quenched conformations with lowest energy in a Q, E scatter plot. Our minimum-energy structure is found at $Q = 0.44$, corresponding to $\delta = 9.1$ Å. However, our thermodynamic calculations show that this conformation is not very relevant, in spite of its low energy. If we restrict ourselves to conformations with the native-like and thermodynamically most relevant topology, then the lowest energy is at $Q = 0.97$, corresponding to $\delta = 1.8$ Å. This conformation is shown in Fig. 2.6 along with the native structure. It is worth noting that the $Q = 0.44$ and $Q = 0.97$ minima both were revisited in independent runs.

These results can be compared with those of Scheraga and coworkers [17], who tested an energy-based structure prediction method on the same sequence. With their energy function, the global minimum was found to have an rmsd of 3.8 Å from the native structure (calculated over C_α atoms).

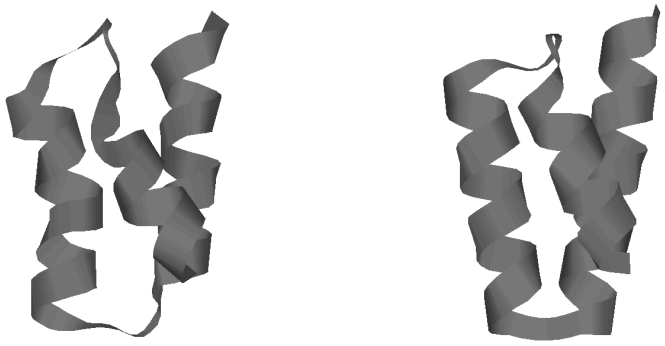


Figure 2.6: Schematic illustrations of the native structure (left) and our minimum-energy structure for the native topology (right). Drawn with RasMol [33].

Segment	Sequence	Amino acids
I	QQNAFYEILHL	10–20
II	NEEQRNGFIQSLKDD	24–38
III	QSANLLAEAKKLNDA	41–55

Table 2.1: The one-helix fragments studied.

2.3.2 Helix Stability

Having discussed the overall thermodynamic behavior, we now take a closer look at the stability of the secondary structure and how it varies along the chain. To this end, we monitored the hydrogen-bond energy between the CO group of amino acid i and the NH group of amino acid $i+4$ [see Eqs. (2.2,2.3)], $e_{\mathrm{hb}}(i)$, as a function of i. This was done not only for the protein A sequence, but also for the corresponding three one-helix segments, which are listed in Table 2.1. An experimental study [23] of essentially the same three segments found segment III to be the only one that shows some stability on its own.

The results of our calculations are shown in Fig. 2.7, from which we see that the difference between the full sequence and the one-helix segments is not large in the model. However, the segments I and II definitely make less stable helices on their own than as interacting parts of the full system; they are stabilized by interhelical interactions. Furthermore, among the three one-helix segments, the model correctly predicts segment III to be the most stable one. That this segment does not get more stable as part of the full system is probably related

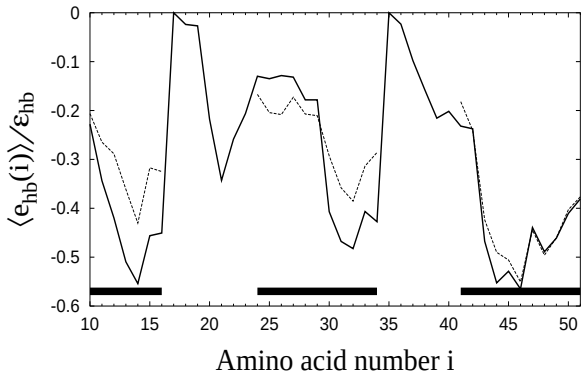


Figure 2.7: Hydrogen-bond profile showing the normalized average energy of α-helical hydrogen bonds, $\langle e_{\rm hb}(i)\rangle/\epsilon_{\rm hb}$, against amino acid number i, at $kT = 0.58$. The full line represents the protein A sequence, whereas the dashed lines represent the corresponding three one-helix segments (see Table 2.1). The thick horizontal lines indicate hydrogen bonds present in the native structure.

to the observation above that helix III is distorted at the $Q = 0.78$ minimum.

A striking detail in Fig. 2.7 is that the beginning of segment II is quite unstable. This can be easily understood. This segment has a flexible glycine at position 30, and the amino acids before the glycine, 24–29, are all polar, so there are no hydrophobic interactions that can help to stabilize this part.

2.3.3 Kinetics

Using the semi-local update [31], we performed a set of 30 kinetic simulations at $kT = 0.54$. The runs were started from random coils. There are big differences between these runs, partly because the system, as it should, sometimes spent a significant amount of time in the wrong topology. Nevertheless, the data show one stable and interesting trend, namely, that the formation of helices was never faster than the collapse. This is illustrated in Fig. 2.8, which shows the evolution of the similarity parameter Q_0, the hydrogen-bond energy $E_{\rm hb}$ and the radius of gyration, $R_{\rm g}$, in one of the runs. Q_0 is defined as Q in Eq. (2.7), except that it measures similarity to the optimized model structure in Fig. 2.6 rather than the native structure. In Fig. 2.8, we see that $E_{\rm hb}$ converges slowly, whereas the collapse occurs relatively early.

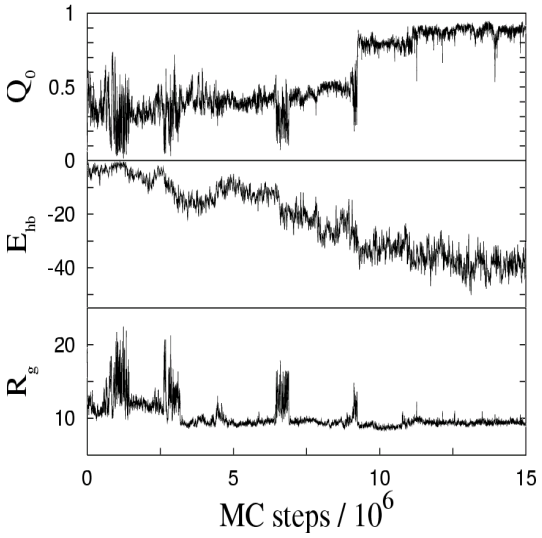


Figure 2.8: Monte Carlo evolution of the similarity parameter Q_0 (top), the hydrogen-bond energy E_{hb} (middle) and the radius of gyration R_{g} (bottom) in a kinetic simulation at $kT = 0.54$.

Now, at a first glance, it may seem easy to make the helix formation faster by simply increasing the strength of the hydrogen bonds. Therefore, it is important to note that the hydrogen bonds cannot be made much stronger without making the ground state non-compact and thus destroying the three-helix bundle [34]. This means that the conclusion that the collapse is at least as fast as helix formation holds for any reasonable choice of parameters in this model.

It is interesting to compare these results to those of Zhou and Karplus [10], who studied the same protein using a Gō-type potential and observed fast folding when the Gō forces were strong. Under these conditions, the helix formation was found to be fast, whereas the collapse was the rate-limiting step.

However, a Gō-like model ignores a large fraction of the interactions that drive the collapse, which can make the collapse artificially slow. In a recent Gō model study [14], this problem was addressed by eliminating backbone terms from the potential until a reasonable helix stability was achieved. No such calibration was carried out in Ref. [10]. This may explain why these authors find a behavior that our model cannot reproduce.

Let us finally mention that we also performed the same type of kinetic simulations for the designed sequence studied in Ref. [20] which, as discussed earlier, has a very abrupt collapse transition. It turns out that $E_{\rm hb}$ and $R_{\rm g}$ evolve in a strongly correlated manner in this case. So, the helix formation and collapse occur simultaneously for this sequence.

2.3.4 Fine-Tuning?

In Sec. 2.3.1, we discussed the relative weights of the two possible overall topologies, which is a delicate issue. What changes are needed in order for the model to more strongly suppress the wrong topology? Is it necessary to change the form of the energy function, or would it be sufficient to fine-tune the interaction matrix $\Delta(s_i, s_j)$ in Eq. (2.4)?

One way to do such a fine-tuning of $\Delta(s_i, s_j)$ would be to maximize $\langle Q\rangle'$, where Q is the similarity parameter and $\langle\cdot\rangle'$ denotes a thermodynamic average restricted to compact conformations ($R_{\rm g} < 10$ Å say). This is essentially the overlap method of Ref. [35]. The gradient of the quantity $\langle Q\rangle'$ can be written as

$$\frac{\partial\langle Q\rangle'}{\partial\Delta(s_i, s_j)} = -\frac{\epsilon_{\rm col}}{kT}\left(\langle QX\rangle' - \langle Q\rangle'\langle X\rangle'\right) , \tag{2.8}$$

where X is a sum of Lennard-Jones terms, $(\sigma_{\rm col}/r_{ij})^{12} - 2(\sigma_{\rm col}/r_{ij})^6$, over all possible $C_\beta C_\beta$ pairs of type s_i, s_j.

We calculated the Q, X correlation in Eq. (2.8) for all pairs s_i, s_j with $\Delta(s_i, s_j) = 1$ at $kT = 0.54$, and found that $|\partial\langle Q\rangle'/\partial\Delta(s_i, s_j)|$ was small (≤ 0.15) for all these pairs. Hence, there is no sign that a significant increase in $\langle Q\rangle'$ can be achieved by fine-tuning $\Delta(s_i, s_j)$; the contact patterns seem to be too similar in the two topologies. To include more side-chain atoms and/or multibody terms in the model is likely to be a more fruitful approach.

2.4 Conclusion

We have explored a five-letter protein model with five to six atoms per amino acid, where the formation of native structure is driven by hydrogen bonding and effective hydrophobicity forces. This model, which does not follow the Gō prescription, was tested on a small but real sequence, a three-helix-bundle fragment from protein A.

Using this model, the protein A sequence was found to collapse much more

efficiently than random sequences with the same composition. In the collapsed phase, we found that the native topology dominates, although the suppression of the wrong three-helix-bundle topology is not strong. Energy minimization constrained to the thermodynamically favored topology gave a structure with an rmsd of 1.8 Å from the native structure.

In our kinetic simulations, the collapse was always at least as fast as helix formation, which is in sharp contrast with previous results for the same protein that were obtained using a Gō-like C_α model [10]. A possible explanation for the conflicting conclusions is that the Gō approximation makes the collapse artificially slow by ignoring a large fraction of the interactions driving the collapse. In our model, the conclusion that the helix formation is not faster than collapse seems unavoidable; if one tries to speed up the helix formation by increasing the strength of the hydrogen bonds, then the chain does not fold into a compact helical bundle.

The force field of our model was deliberately kept simple. In particular, the hydrophobicity potential was taken to be pairwise additive, with a simple structure for the interaction matrix $\Delta(s_i, s_j)$ [see Eq. (2.5)]. In the future, it would be very interesting to look into the behavior of the model in the presence of multibody terms. A simpler alternative is to stick to the pairwise additive potential and fine-tune the parameters $\Delta(s_i, s_j)$. However, the calculations in this paper give no indication that there is much to be gained from such a fine-tuning.

Acknowledgments

This work was in part supported by the Swedish Foundation for Strategic Research.

References

[1] Săli A, Shakhnovich E, Karplus M. Kinetics of protein folding: A lattice model study of the requirements for folding to the native state. *J. Mol. Biol.* 1994 ; 235 : 1614–1636.

[2] Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins Struct. Funct. Genet.* 1995 ; 21 : 167–195.

[3] Dill KA, Chan HS. From Levinthal to pathways to funnels. *Nat. Struct. Biol.* 1997 ; 4 : 10–19.

[4] Klimov DK, Thirumalai D. Linking rates of folding in lattice models of proteins with underlying thermodynamic characteristics. *J. Chem. Phys.* 1998 ; 109 : 4119–4125.

[5] Nymeyer H, García AE, Onuchic JN. Folding funnels and frustration in off-lattice minimalist protein landscapes. *Proc. Natl. Acad. Sci. USA* 1998 ; 95 : 5921–5928.

[6] Hao M-H, Scheraga HA. Theory of two-state cooperative folding of proteins. *Acc. Chem. Res.* 1998 ; 31 : 433–440.

[7] Gō N, Taketomi H. Respective roles of short- and long-range interactions in protein folding. *Proc. Natl. Acad. Sci. USA* 1978 ; 75 : 559–563.

[8] Zhou Y, Karplus M. Folding thermodynamics of a model three-helix-bundle protein. *Proc. Natl. Acad. Sci. USA* 1997 ; 94 : 14429–14432.

[9] Shea J-E, Nochomovitz YD, Guo Z, Brooks CL III. Exploring the space of protein folding Hamiltonians: The balance of forces in a minimalist β-barrel model. *J. Chem. Phys.* 1998 ; 109 : 2895–2903.

[10] Zhou Y, Karplus M. Interpreting the folding kinetics of helical proteins. *Nature* 1999 ; 401 : 400–403.

[11] Shea J-E, Onuchic JN, Brooks CL III. Exploring the origins of topological frustration: Design of a minimally frustrated model of fragment B of protein A. *Proc. Natl. Acad. Sci. USA* 1999 ; 96 : 12512–12517.

[12] Clementi C, Nymeyer H, Onuchic JN. Topological and energetic factors: What determines the structural details of the transition state ensemble and 'en-route' intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* 2000 ; 298 : 937–953.

[13] Clementi C, Jennings PA, Onuchic JN. How native-state topology affects the folding of dihydrofolate reductase and interleukin-1β. *Proc. Natl. Acad. Sci. USA* 2000 ; 97 : 5871–5876.

[14] Shimada J, Kussell EL, Shakhnovich EI. The folding thermodynamics and kinetics of crambin using an all-atom Monte Carlo simulation. *J. Mol. Biol.* 2001 ; 308 : 79–95.

[15] Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* 1998 ; 277 : 985–994.

[16] Baker D. A surprising simplicity to protein folding. *Nature* 2000 ; 405 : 39–42.

[17] Lee J, Liwo A, Scheraga HA. Energy-based *de novo* protein folding by conformational space annealing and an off-lattice united-residue force field: Application to the 10–55 fragment of staphylococcal protein A and to apo calbindin D9K. *Proc. Natl. Acad. Sci. USA* 1999 ; 96 : 2025–2030.

[18] Pillardy J, Czaplewski C, Liwo A, Lee J, Ripoll DR, Kaźmierkiewicz R, Ołdziej S, Wedemeyer WJ, Gibson KD, Arnautova YA, Saunders J, Ye Y-J, Scheraga HA. Recent improvements in prediction of protein structure by global optimization of a potential energy function. *Proc. Natl. Acad. Sci. USA* 2000 ; 98 : 2329–2333.

[19] Hardin C, Eastwood MP, Luthey-Schulten Z, Wolynes PG. Associative memory Hamiltonians for structure prediction without homology: alpha-helical proteins. *Proc. Natl. Acad. Sci. USA* 2000 ; 97 : 14235–14240.

[20] Irbäck A, Sjunnesson F, Wallin S. Three-helix-bundle protein in a Ramachandran model. *Proc. Natl. Acad. Sci. USA* 2000 ; 97 : 13614–13618.

[21] Gouda H, Torigoe H, Saito A, Sato M, Arata Y, Shimada I. Three-dimensional solution structure of the B domain of staphylococcal protein A: comparisons of the solution and crystal structures. *Biochemistry* 1992 ; 31 : 9665–9672.

[22] Bottomley SP, Popplewell AG, Scawen M, Wan T, Sutton BJ, Gore MG. The stability and unfolding of an IgG binding protein based upon the B domain of protein A from *Staphylococcus Aureus* probed by tryptophan substitution and fluorescence spectroscopy. *Protein Eng.* 1994 ; 7 : 1463–1470.

[23] Bai Y, Karimi A, Dyson HJ, Wright PE. Absence of a stable intermediate on the folding pathway of protein A. *Protein Sci.* 1997 ; 6 : 1449–1457.

[24] Boczko EM, Brooks CL III. First-principles calculation of the folding free energy of a three-helix bundle protein. *Science* 1995 ; 269 : 393–396.

[25] Guo Z, Brooks CL III, Boczko EM. Exploring the folding free energy surface of a three-helix bundle protein. *Proc. Natl. Acad. Sci. USA* 1997 ; 94 : 10161-10166.

[26] Kolinski A, Galazka W, Skolnick J. Monte Carlo studies of the thermodynamics and kinetics of reduced protein models: Application to small helical, β and α/β proteins. *J. Chem. Phys.* 1998 ; 108 : 2608–2617.

[27] Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* 1977 ; 112 : 535–542.

[28] Lyubartsev AP, Martsinovski AA, Shevkunov SV, Vorontsov-Velyaminov PV. New approach to Monte Carlo calculation of the free energy: Method of expanded ensembles. *J. Chem. Phys.* 1992 ; 96 : 1776–1783.

[29] Marinari E, Parisi G. Simulated tempering: A new Monte Carlo scheme. *Europhys. Lett.* 1992 ; 19 : 451–458.

[30] Irbäck A, Potthast F. Studies of an off-lattice model for protein folding: Sequence dependence and improved sampling at finite temperature. *J. Chem. Phys.* 1995 ; 103 : 10298–10305.

[31] Favrin G, Irbäck A, Sjunnesson F. Monte Carlo update for chain molecules: Biased Gaussian steps in torsional space. *J. Chem. Phys.* 2001 ; 114 : 8154–8158.

[32] Bastolla U., Farwer J, Wallin S. On distance measures for protein structures. Manuscript in preparation.

[33] Sayle R, Milner-White EJ. RasMol: Biomolecular graphics for all. *Trends Biochem. Sci.* 1995 ; 20 : 374–376.

[34] Irbäck A, Sjunnesson F, Wallin S. Hydrogen bonds, hydrophobicity forces and the character of the collapse transition. e-print cond-mat/0107177 (to appear in *J. Biol. Phys.*).

[35] Bastolla U, Vendruscolo M, Knapp E-W. A statistical mechanical method to optimize energy functions for protein folding. *Proc. Natl. Acad. Sci. USA* 2000 ; 97 : 3977–3981.

Testing Similarity Measures with Continuous and Discrete Protein Models

Paper III

Testing Similarity Measures with Continuous and Discrete Protein Models

Stefan Wallin

Complex Systems Division, Dept. of Theoretical Physics
Lund University, Sölvegatan 14A, SE-223 62 Lund, Sweden

Jochen Farwer

Free University of Berlin, Dept. of Biology,
Chemistry and Pharmacy,
Institute of Chemistry, Takustr. 6, D-14195 Berlin, Germany

Ugo Bastolla

Centro de Astrobiología (CSIC-INTA),
Ctra. de Ajalvir km. 4, 28850 Torrejón de Ardoz (Madrid), Spain

Proteins: Structure, Function, and Genetics **50**, 144-157 (2003)

Abstract

There are many ways to define the distance between two protein structures, thus assessing their similarity. Here, we investigate and compare the properties of five different distance measures, including the standard root-mean-square deviation (cRMSD) and a new one that we propose, called the power distance. The performance of these measures is studied from different perspectives with two different protein models, one continuous and the other discrete. Using the continuous model, we examine the correlation between energy and native distance, and the ability of the different measures to discriminate between the two possible topologies of a three-helix bundle. Using the discrete model, we perform fits to real protein structures by minimizing different distance measures. The properties of the fitted structures are found to depend strongly on the distance measure used and the scale considered. We find that the cRMSD measure effectively describes long-range features but is less effective with short-range features, and it correlates weakly with energy. A stronger correlation with energy and a better description of short-range properties is obtained when we use measures based on intramolecular distances, such as the power distance.

3.1 Introduction

Protein structures are complex three-dimensional objects, characterized by a large number of linked atoms and a hierarchy of description levels, from primary to quaternary structure. Several distance measures may be defined for the task of comparing protein structures and assessing their similarity. The root-mean-square deviation (cRMSD) is the measure most commonly used, but several others are frequently applied in the literature. Here, we set up to compare four of the most widely used distance measures, plus a new one that we propose.

Measuring protein similarity is becoming increasingly important in the field of bioinformatics, both for the sake of evaluating protein structure prediction methods [1] and for the sake of classifying proteins and investigating distant evolutionary relationships [2–4]. To this end, several similarity scores have been developed. In a recent review [5], merits and drawbacks of some of these measures are pointed out. Similarity measures have been assessed for their ability to generate robust and accurate clusters in hierarchical classification of proteins [6], with the conclusion that the problem of protein structure comparison does not have a unique answer [7–9]. In this work, we choose another approach and assess protein structure similarity measures by using statistical mechanical models of protein folding.

A similarity measure is appropriate if objects scored as similar share similar properties. In the present context, the objects to be compared are reduced representations of protein structures, and their most important property is their effective energy, which can be thought of as the free energy of the reduced structure as obtained by integrating out degrees of freedom which are not represented in the model (depending on the model, these can be solvent atoms, side chains, etc.). The effective energy depends on solvent properties like temperature, pH, denaturant concentration and so on. In practice, it is impossible to compute the effective energy from first principles, and one has to postulate some energy function in such a way that the low-energy states of the model resemble real protein structures.

According to the thermodynamic view of protein folding [10], the native state of a protein is the state of minimal free energy of the protein plus solvent system. Since the configurational entropy of the protein chain in its native state is very small, it is customary to identify the native state of the protein as the reduced state of minimal effective energy plus its thermal fluctuations. Theoretical considerations based on spin glass theory [11] and comparisons between minimal models of random heteropolymers [12,13] and models of well designed sequences [14–16] have shown that a necessary condition for a simplified model to be a viable model of protein folding is that the energy landscape is well

correlated. Qualitatively, this means that structures very different from the ground state should have high effective energy, so that they have a negligible weight in the thermodynamic ensemble. A well correlated energy landscape is a prerequisite for fast folding [11, 14, 17, 18], thermodynamic stability with respect to changes in the solvent, and stability with respect to mutations [19, 20] of the model protein. In assessing the shape of an energy landscape, it is important that the effective energy function and the similarity measure used are well correlated.

There are some reasons why one can expect that the cRMSD measure does not correlate very well with the energy (see also Ref. [21]). The cRMSD compares atoms at the same position in two chains. All positions have the same weight in this comparison, but not all positions contribute in the same way to the effective energy. Residues in a loop can sometimes be moved without changing the effective energy much, thus generating structures with a high cRMSD from each other but very similar energies. By contrast, even a small displacement of a residue in the hydrophobic core of the native state is likely to produce an atomic collision and thereby a drastic increase in energy. Thus we can find structures with high similarity but very different energies.

A natural way to try to increase the correlation with energy is to replace the cRMSD by a distance measure based on intramolecular atomic distances. In fact, most energy terms are functions of atom-atom distances. It seems reasonable to expect that such a measure can provide a stronger energy correlation, especially if atom pairs at short distances are given a higher weight.

To investigate this issue quantitatively, we perform two kinds of tests. In the first, we evaluate how different distance measures correlate with the effective energy function of a continuous model for protein folding [22–24]. For this purpose we use configurations generated through Monte Carlo sampling at the folding temperature. At this temperature, folded as well as unfolded structures exist, which makes it possible to evaluate the correlation between energy and distance over a wide range of distances.

In the second test, we investigate discrete protein models. These models maintain the simplicity of lattice models in that they have a finite state space, and can, at the same time, reproduce native structures of real proteins very closely, as showed by Park and Levitt [25]. Following their approach, we work with a set of six directions; each residue along the chain selects one of these directions. We find that structures obtained this way can fit all protein structures in a very large database with an average cRMSD of less than 1.6 Å. However, a high similarity in terms of cRMSD does not mean that the fitted structures are similar to the real ones in all respects. In particular, we find that the dis-

tribution of C_α-C_α distances is quite different for fitted and real structures, respectively. The discrete structures that are most similar to the native ones in terms of cRMSD tend to exhibit a large number of atomic collisions, which are of course prohibited in real proteins. Such unphysical collisions can be avoided by using, for example, the power distance introduced in this work.

The third issue that we address is the ability of different distance measures to distinguish structures that are locally similar but globally different. We address this issue using the continuous protein model, which possesses two states that are very similar in terms of energy, entropy and secondary structure, but different in overall topology. In this test, the cRMSD outperforms distance measures based on intramolecular distances which, in the presence of thermal noise, have difficulties in discriminating between the two topologies.

On the other hand, when it comes to energy correlations, our analysis shows that three of the four distance measures that we compare to the cRMSD are indeed better than this measure. In particular, this holds for a contact-based distance measure. Hence, the choice of distance measure depends on the type of problem addressed. Some distance measures perform rather poorly for certain tasks, and it is advisable not to use them in such situations.

The paper is organized as follows. In the next section we introduce the distance measures studied and the continuous and discrete models used to test them. In section 3.3 the results of the tests are presented. The paper is closed with a short discussion and summary of the results.

3.2 Materials and Methods

3.2.1 Distance Measures

From a mathematical point of view, a distance is a function $D(a, b)$ associating to any pair of elements, a and b, of a metric space a non-negative real number. It has the distinctive properties: (1) symmetry, $D(a, b) = D(b, a)$; (2) positivity, $D(a, b) \geq 0$, where equality holds if and only if a and b are identical; and (3) the triangular inequality, that is $D(a, b) \leq D(a, c) + D(b, c)$ for any c.

In this work, we represent a structure Γ as an ordered set of coordinates of N atoms, $\Gamma = \{\mathbf{r}_1 \cdots \mathbf{r}_N\}$. Each amino acid will be represented either by a single atom, C_α or C_β, or by the three backbone atoms N, C_α and C'. If the two structures to be compared would represent different proteins, one would need a criterion to define an alignment, i.e. a correspondence between the

residues of the first protein and those in the other one. Alignment methods aim at putting in correspondence evolutionarily related amino acids (sequence alignment) or structurally related amino acids (structural alignment). In what follows, we assume that the two proteins have already been aligned so that there is a one-to-one correspondence between the atoms of the two structures Γ.

There are several ways to define distance measures in the $3N$-dimensional space of all possible structures Γ. We consider five of them, which are listed below.

3.2.2 Root-Mean-Square Deviation (cRMSD)

The most natural and most frequently used distance measure is the coordinate root-mean-square deviation (cRMSD), which is the ordinary Euclidean distance in $3N$-dimensional space:

$$D_{\mathrm{cRMSD}}(\Gamma^a, \Gamma^b) = \min \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\mathbf{r}_i^a - \mathbf{r}_i^b)^2} \,. \tag{3.1}$$

Structures related through rigid rotations or translations of all the N atoms are considered equivalent. The cRMSD between two equivalence classes Γ^a and Γ^b is defined as the minimal cRMSD with respect to all possible translations and rotations of the two structures. This minimization can be performed analytically [26, 27].

It has been noted that the cRMSD tends to increase with the number of atoms, N. Maiorov and Crippen proposed a normalization meant to make the cRMSD effectively independent of N [28]. This correction is important if one compares distances corresponding to different N. However, in the present paper we deal with fixed values of N and will adopt the usual definition, Eq. (3.1).

The cRMSD measure gives the same weight to all atoms, whatever their structural role. As discussed in the Introduction, this typically leads to a poor correlation with energy, which makes it interesting to look at alternative distance measures.

3.2.3 Distance RMSD (dRMSD)

A protein consisting of N atoms has a set of $N(N-1)/2$ (not independent) internal atomic distances r_{ij}. Using distance measures based on the r_{ij}'s is

appealing partly because the effective energy usually is a function of these distances. Another advantage is that the minimization needed to obtain the cRMSD can be avoided.

The simplest way to construct a distance measure based on the r_{ij}'s is to use a Euclidean distance. This gives the so-called dRMSD measure [29–31], defined as

$$D_{\mathrm{dRMSD}}(\Gamma^a, \Gamma^b) = \sqrt{\frac{1}{N_{\mathrm{pair}}} \sum_{i<j} \left(r_{ij}^a - r_{ij}^b\right)^2}\,, \tag{3.2}$$

where N_{pair} is the number of distances compared. $D_{\mathrm{dRMSD}}(\Gamma^a, \Gamma^b)$ is the root-mean-square deviation between two sets of distances, r_{ij}^a and r_{ij}^b.

An unwanted feature of $D_{\mathrm{dRMSD}}(\Gamma^a, \Gamma^b)$ is that the pairs of atoms contributing most are those with largest r_{ij}, which are those contributing least to the energy. As a result, one may expect the correlation with energy to be even worse than for cRMSD.

3.2.4 Contact Distance

A simple way to remove this unwanted feature is to introduce the binary quantities C_{ij} which, for a given structure with atomic distances r_{ij}, are defined as

$$C_{ij} = \begin{cases} 1 & \text{if } r_{ij} < r_c \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

where r_c is a cutoff distance. The matrix C is called the contact map and represents an equivalence class of structures whose configurational entropy is extensive (of order N), as has been seen using lattice models to represent protein structures [32, 33].

Despite the gross simplification implied by Eq. (3.3), models based on the contact map have been intensively studied in the last 15 years (for a recent review, see Ref. [34]). Such models have an energy function of the form $E(\Gamma, \{B_{ij}\})/k_BT = \sum_{ij} C_{ij}(\Gamma) B_{ij}$, where the quantities B_{ij} are effective contact interactions in units of k_BT. For these models, the natural measure of similarity between two structures Γ^a and Γ^b is their contact overlap, defined as

$$q(\Gamma^a, \Gamma^b) = \frac{\sum_{i<j} C_{ij}^a C_{ij}^b}{\max\left(\sum_{i<j} C_{ij}^a\,, \sum_{i<j} C_{ij}^b\right)}\,, \tag{3.4}$$

where C^a and C^b are the contact maps of the structures Γ^a and Γ^b, respectively. The numerator is the number of common contacts in the two structures and

the denominator is chosen such that the overlap takes the maximum value $q = 1$ if and only if the two contact maps coincide. Alternatively, we could use a denominator of the form $\sqrt{\sum_{i<j} C_{ij}^a \sum_{i<j} C_{ij}^b}$, which is qualitatively very similar and has the advantage that the contact overlap could be interpreted as the "cosine" between the two contact maps. A contact-based distance measure is most easily defined as

$$D_{\mathrm{cont}}(\Gamma^a, \Gamma^b) = 1 - q(\Gamma^a, \Gamma^b)\,. \tag{3.5}$$

We prove in the Appendix that this distance satisfies the triangular inequality.

3.2.5 The Holm and Sander Score

In the context of structure alignment, Holm and Sander [35, 36] proposed a similarity score based on internal atomic distances, which does not have the drawback of the dRMSD of strongly weighting large values of r_{ij}, and which, in contrast to the contact distance, does not require a discretization. The corresponding distance measure can be defined as

$$D_{\mathrm{HS}}^*(\Gamma^a, \Gamma^b) = \sum_{i<j} \frac{|r_{ij}^a - r_{ij}^b|}{r_{ij}^a + r_{ij}^b}\, \mathrm{e}^{-(r_{ij}^a + r_{ij}^b)^2/4r_0^2}\,. \tag{3.6}$$

Only features on scales of order r_0 and smaller contribute significantly to this score (in the original paper, the value $r_0 = 20\,\text{Å}$ was chosen). The exponential weight function is qualitatively similar to a softening of the condition (3.3). For the task of comparing protein structures, we found it more convenient to use a normalized version of the Holm and Sander score, given by

$$D_{\mathrm{HS}}(\Gamma^a, \Gamma^b) = \frac{D_{\mathrm{HS}}^*(\Gamma^a, \Gamma^b)}{\sum_{i<j} \mathrm{e}^{-(r_{ij}^a + r_{ij}^b)^2/4r_0^2}}\,. \tag{3.7}$$

This score is restricted to the interval $[0, 1]$ regardless of protein length and has the advantage of being less sensitive to details than the original one.

Neither the Holm and Sander score nor its normalized version is a real distance, since they do not strictly satisfy the triangular inequality (see Appendix). However, an advantage of the normalized version D_{HS} is that violations of the triangular inequality seem to become extremely rare. In fact, no violation was found for D_{HS} for the set of protein-like structures obtained with the continuous protein model. For D_{HS}^*, violations were observed for values of r_0 less than $11\,\text{Å}$, but not for the original value $r_0 = 20\,\text{Å}$ used in Ref. [35].

3.2.6 Power Distance

Finally, we propose another distance measure that gives a large weight to pairs of atoms with small r_{ij}. This measure is defined as

$$D^{(0)}_{\mathrm{pow}}(\Gamma^a, \Gamma^b) = \sum_{i<j} \left| (r^a_{ij})^{-m} - (r^b_{ij})^{-m} \right| , \tag{3.8}$$

where m is a parameter that effectively controls the relative weight of small distances r_{ij}. If m is too small, the correlation between energy and distance to the native state becomes weak. If, on the other hand, m is too large, then local features are weighted too strongly, so that structures that are similar in terms of $D^{(0)}_{\mathrm{pow}}$ can have rather different global features. We tested different integer values of m, finding the best overall results for m equal to 2 or 3.

The function $D^{(0)}_{\mathrm{pow}}$ satisfies all properties of a distance, but it is expressed in bizarre units ($\mathrm{\AA}^{-m}$), and it depends very strongly on the length and compactness of the structures compared. These problems can be alleviated by normalizing the measure so that it takes values between 0 and 1. One possible normalization is given by

$$D^{(1)}_{\mathrm{pow}}(\Gamma^a, \Gamma^b) = \frac{1}{N_{\mathrm{pair}}} \sum_{i<j} \frac{\left| (r^a_{ij})^{-m} - (r^b_{ij})^{-m} \right|}{(r^a_{ij})^{-m} + (r^b_{ij})^{-m}} , \tag{3.9}$$

where N_{pair} is the number of distances compared. The function $D^{(1)}_{\mathrm{pow}}$ coincides with the Holm and Sander score if $m = -1$ and $r_0 = \infty$ in Eq. (3.7). Unlike the general Holm and Sander score, $D^{(1)}_{\mathrm{pow}}$ fulfills the triangular inequality (see Appendix). The definition (3.9) has the drawback, however, that it is not necessarily true that small r_{ij}'s are given a higher weight. This can be seen by rescaling r^a_{ij} and r^b_{ij} for one particular pair ij by a common factor λ. Then, as $\lambda \to 0$, the term ij becomes dominant in Eq. (3.8), whereas it remains unchanged in Eq. (3.9). An alternative normalization is to use

$$D_{\mathrm{pow}}(\Gamma^a, \Gamma^b) = \frac{\sum_{i<j} \left| (r^a_{ij})^{-m} - (r^b_{ij})^{-m} \right|}{\sum_{i<j} \left[(r^a_{ij})^{-m} + (r^b_{ij})^{-m} \right]} . \tag{3.10}$$

This distance measure does not strictly fulfill the triangular inequality. However, we found very few violations of this inequality for randomly generated coordinates and none for protein-like structures. In the following, we will use D_{pow}, defined by Eq. (3.10), as the power distance.

3.2.7 Sequence Cutoff

Giving a large weight to short distances r_{ij}, as some of the discussed distances do, has the side-effect that amino acid pairs kl with short sequence separation $|k-l|$ get a high weight. This is an unwanted property because the relative position of such pairs provide little information about the overall structure of the chain. To overcome this problem, it is useful to define a sequence cutoff s and only consider pairs such that $|k-l| > s$ in the evaluation of the distances. Since in alpha helices amino acids at separation $|k-l| = 4$ form hydrogen bonds, $s = 2$ seems to be a good choice to get rid of local details without losing important energetic contributions.

3.2.8 Continuous Protein Model

The first part of our analysis is performed using a continuous protein model, introduced and studied in Refs. [22] and [23]. The model describes a protein with 54 amino acids which are of three different types (hydrophobic, polar and glycine). Each amino acid is represented by the backbone atoms N, C_α, C', H and O, as well as the C_β atom (except for glycine), which is treated as either hydrophobic or polar, depending on the amino acid type. The degrees of freedom are the Ramachandran torsion angles ϕ_i and ψ_i. The energy function is composed of four terms:

$$E = E_{\rm loc} + E_{\rm sa} + E_{\rm hb} + E_{\rm AA}\,. \tag{3.11}$$

The local potential $E_{\rm loc}$ has a standard form with 3-fold symmetry,

$$E_{\rm loc} = \frac{\epsilon_\phi}{2}\sum_i(1+\cos 3\phi_i) + \frac{\epsilon_\psi}{2}\sum_i(1+\cos 3\psi_i)\,. \tag{3.12}$$

The self-avoidance term $E_{\rm sa}$ is given by a hard-sphere potential of the form

$$E_{\rm sa} = \epsilon_{\rm sa}{\sum_{i<j}}'\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}, \tag{3.13}$$

where the sum runs over all possible atom pairs except those consisting of two hydrophobic C_β. The hydrogen-bond term $E_{\rm hb}$ is given by

$$E_{\rm hb} = \epsilon_{\rm hb}\sum_{ij} u(r_{ij})v(\alpha_{ij},\beta_{ij})\,, \tag{3.14}$$

where i and j represent H and O atoms respectively, and

$$u(r_{ij}) \;=\; 5\left(\frac{\sigma_{\rm hb}}{r_{ij}}\right)^{12} - 6\left(\frac{\sigma_{\rm hb}}{r_{ij}}\right)^{10} \tag{3.15}$$

$$v(\alpha_{ij},\beta_{ij}) \quad = \quad \begin{cases} \cos^2\alpha_{ij}\cos^2\beta_{ij} & \alpha_{ij},\beta_{ij} > 90^\circ \\ 0 & \text{otherwise} \end{cases} \tag{3.16}$$

In these equations, r_{ij} denotes the HO distance, α_{ij} the NHO angle, and β_{ij} the HOC$'$ angle. Finally, the hydrophobicity term E_{AA} has the form

$$E_{\mathrm{AA}} = \epsilon_{\mathrm{AA}} \sum_{i<j} \left[\left(\frac{\sigma_{\mathrm{AA}}}{r_{ij}} \right)^{12} - 2 \left(\frac{\sigma_{\mathrm{AA}}}{r_{ij}} \right)^{6} \right], \tag{3.17}$$

where both i and j represent hydrophobic C_β. Further details of the model, including numerical values of all the parameters, can be found in Ref. [22].

The model exhibits a first-order-like folding transition from an extended phase to the folded one, where the chain forms a three-helix bundle. A three-helix bundle has two possible topologies; if the first two helices form a U, the third helix can go either in front of or behind this U. The model is unable to distinguish between these two possible ways of arranging the helices. Hence, the model exhibits a twofold topological degeneracy.

The thermodynamic behavior of this model protein has been studied in detail before [22] through extensive Monte Carlo simulations at different temperatures. A set of 5000 configurations, separated by at least 10^5 elementary Monte Carlo steps, were recorded at each temperature. In our analysis, we use ensembles of configurations sampled at the folding temperature T_{f}, where the folded and unfolded phases coexist, and at a lower temperature $T_{\mathrm{low}} = 0.95T_{\mathrm{f}}$, where the chain is folded.

The previous study also determined representative conformations for the two topologies, through a quenching procedure [22]. These two structures, having minimum energy within their respective topologies, are referred to as FU and BU. In Fig. 3.1 we compare the contact maps of these two conformations. The contacts can be divided into interhelical and intrahelical ones, the latter ones being located near the diagonal. The two intrahelical contacts sets are found to be essentially the same, as expected. The two interhelical contact sets resemble each other, but many of the contacts are not exactly the same in the two conformations.

In our analysis of the model, all distances except the contact distance are calculated over the backbone atoms N, C_α and C$'$. In the contact distance, we use the C_β atoms to define contacts.

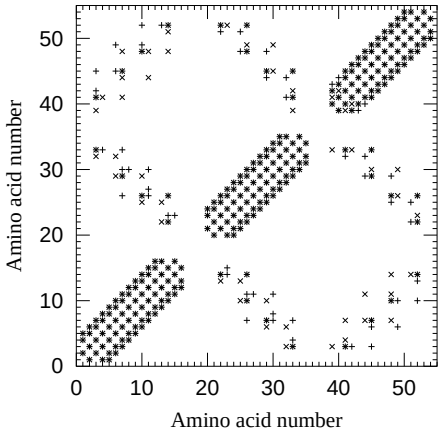


Figure 3.1: Contact maps of the two minimum-energy conformations BU (+) and FU (×); shared contacts appear as '∗'. Two amino acids with C_β atoms closer than 7 Å are regarded in contact.

3.2.9 Discrete Protein Models

Continuous models are much more complex to simulate than lattice models where the conformation space is much simpler. Most minimal models of random heteropolymers and designed sequences have been studied on the cubic lattice where each new residue that is added to the chain has at most five possible directions (going backwards is forbidden by hard core repulsion). Although these models have provided valuable qualitative insights into the principles of folding, their application to real proteins is severely limited. Several groups have then studied models based on more complex lattices (reviewed in Ref. [25]) and discrete off-lattice models where each new residue can choose between a finite number of predetermined dihedral angles [25, 37].

Park and Levitt [25] evaluated the accuracy of discrete models measuring the minimal cRMSD between model structures and protein native states, for a set of 149 proteins. They showed that, for four-state models, it is possible to optimize the set of allowed directions in such a way that the average of the minimal cRMSD is 2.2 Å. Moreover, two six-state models with angles chosen according to the distribution of torsion angles in the PDB by Rooman *et al.* [37] and Park and Levitt [25] can fit the same proteins up to 1.74 Å and 1.90 Å, respectively, on the average. These results suggest that optimized discrete models can reproduce protein structures very accurately, despite their low complexity.

In this work, following Park and Levitt, we optimize sets of six directions. We choose a C_α representation of protein structures, parameterizing directions in terms of the pseudo-bond angle α (the angle formed by three consecutive C_α atoms) and the pseudo-torsion angle τ (the torsion angle formed by four consecutive C_α atoms). Our scoring function is based on distances to a training set of native protein structures. The same analysis is repeated using different distance measures.

3.2.10 Modified Build-Up Algorithm

The task to determine the discrete structure that minimizes the distance to a given target protein structure is computationally very hard. One approach to this problem is to apply a Monte Carlo algorithm at low temperature, or some form of simulated annealing. However, Park and Levitt [25] showed that a simple deterministic algorithm provides good approximate solutions to the problem.

One such algorithm is the build-up algorithm, where new residues are attached to the growing chain one at a time. Since each residue can be attached in k different configurations, where $k = 6$ for six-state models, the number of possible chain configurations increases exponentially with the number of residues. At every growth stage n, configurations are ranked according to their score and the N_{keep} configurations with the best score are selected. These configurations are then used as building blocks to build the kN_{keep} configurations at stage $n+1$. The procedure is iterated until the chain is completed. Surprisingly, this very simple algorithm, which is guaranteed to find the lowest energy configuration for $N_{\text{keep}} = k^N$, where N is the chain length, converges close to the optimal value already at N_{keep} of the order of a few hundreds, apparently independent of N [25].

We found out that the performance of this build-up algorithm can be improved by adding a little bit of randomness. This is accomplished by selecting the N_{keep} conformations at stage n using the scoring function

$$\text{Score} = D(\Gamma^n, \Gamma^{\text{nat}}) + T_r \epsilon_n \left(1 - \frac{n}{N}\right) , \tag{3.18}$$

where $D(\Gamma^n, \Gamma^{\text{nat}})$ is the distance between the discrete structure at stage n and the first n residues of the native structure, T_r is a parameter measuring the amount of randomness in the score, and ϵ_n is a random variable uniformly distributed between -1 and 1. For $T_r = 0$, the original, deterministic build-up algorithm is recovered. The factor $1 - n/N$ is chosen such that the random part of the score vanishes for the completed chain, since $D(\Gamma^N, \Gamma^{\text{nat}})$ is to

be minimized. For moderate but non-vanishing T_r and sizeable N_{keep}, the modified algorithm typically finds structures of lower distance than for $T_r = 0$. In fact, for $T_r = 0$, all the selected configurations are rather correlated, and the algorithm explores a relatively small number of directions in a high-dimensional space. After adding some randomness, the algorithm explores a broader spectrum of directions and has a better chance to find low-distance states.

3.3 Results and Discussion

3.3.1 Relationships Between Distances

The first question we address is how the different distance measures are related to each other. Obviously, we expect that they are correlated, but the extent of the correlation will depend on factors such as the relative weight given to local versus global features. We study this question using the conformations obtained by sampling of the continuous system described above at its folding temperature T_{f}.

Since the native state of this chain has a twofold degeneracy, we define native distance as

$$D^{\text{n}}(\Gamma) \equiv \min\left[D(\text{FU},\Gamma), D(\text{BU},\Gamma)\right] , \tag{3.19}$$

where FU and BU are the representative conformations of the two topologies, and D denotes any one of the distances previously introduced. We consider five different native distances: $D^{\text{n}}_{\text{cRMSD}}$, $D^{\text{n}}_{\text{dRMSD}}$, $D^{\text{n}}_{\text{cont}}$, D^{n}_{HS} and $D^{\text{n}}_{\text{pow}}$. Throughout our analysis of the continuous model, the parameters of $D^{\text{n}}_{\text{cont}}$, D^{n}_{HS} and $D^{\text{n}}_{\text{pow}}$ are taken as $r_c = 7\,\text{Å}$, $r_0 = 13\,\text{Å}$ and $m = 3$, respectively. These values were chosen so as to maximize the correlation with energy (see section 3.3.2).

Table 3.1 shows correlation coefficients between different pairs of native distances. The three measures $D^{\text{n}}_{\text{cont}}$, D^{n}_{HS} and $D^{\text{n}}_{\text{pow}}$ are found to be strongly correlated with each other, the correlation being strongest between D^{n}_{HS} and $D^{\text{n}}_{\text{pow}}$. Figure 3.2a shows the correlation between D^{n}_{HS} and $D^{\text{n}}_{\text{cont}}$.

These three measures turn out to be approximately exponentially related to the two others, $D^{\text{n}}_{\text{cRMSD}}$ and $D^{\text{n}}_{\text{dRMSD}}$. This is illustrated in Fig. 3.2b, which shows the $D^{\text{n}}_{\text{pow}}, D^{\text{n}}_{\text{dRMSD}}$ distribution. An exponential relation has been observed previously between $D^{\text{n}}_{\text{cont}}$ and $D^{\text{n}}_{\text{cRMSD}}$, in a study of database structures [38]. A fit of our data to the form $D^{\text{n}}_{\text{cRMSD}} \approx a \exp(b * D^{\text{n}}_{\text{cont}})$ yields an exponent $b = 3.4$. The correlation coefficient is 0.84 between $D^{\text{n}}_{\text{cont}}$ and $\log(D^{\text{n}}_{\text{cRMSD}}+1)$

	$\log(D^{\mathrm{n}}_{\mathrm{dRMSD}}+1)$	$D^{\mathrm{n}}_{\mathrm{cont}}$	$D^{\mathrm{n}}_{\mathrm{HS}}$	$D^{\mathrm{n}}_{\mathrm{pow}}$
$\log(D^{\mathrm{n}}_{\mathrm{cRMSD}}+1)$	0.97	0.84	0.87	0.90
$\log(D^{\mathrm{n}}_{\mathrm{dRMSD}}+1)$		0.81	0.85	0.89
$D^{\mathrm{n}}_{\mathrm{cont}}$			0.95	0.97
$D^{\mathrm{n}}_{\mathrm{HS}}$				0.98

Table 3.1: Correlation coefficients between pairs of native distances.

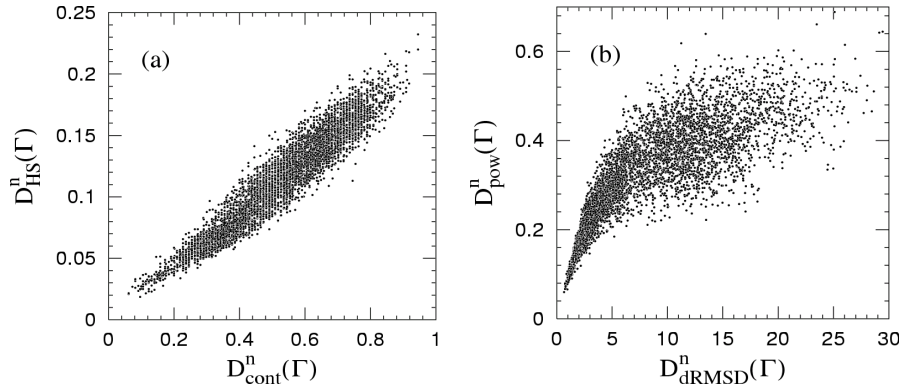


Figure 3.2: (a) $D^{\mathrm{n}}_{\mathrm{HS}}$, $D^{\mathrm{n}}_{\mathrm{cont}}$ and (b) $D^{\mathrm{n}}_{\mathrm{pow}}$, $D^{\mathrm{n}}_{\mathrm{dRMSD}}$ scatter plots for conformations at T_{f}.

(see Table 3.1). We note that the logarithm of a distance can be used to define another distance through the formula $D_{\log}(a,b) = \log((D(a,b)+\epsilon)/\epsilon)$ for any $\epsilon > 0$; it is easy to verify that $D_{\log}$ satisfies the three distance properties if D does so.

The two Euclidean distances $D^{\mathrm{n}}_{\mathrm{cRMSD}}$ and $D^{\mathrm{n}}_{\mathrm{dRMSD}}$ correlate well and linearly with each other. Consistent with a previous study [30] we find that $D^{\mathrm{n}}_{\mathrm{cRMSD}}$ tends to be slightly larger than $D^{\mathrm{n}}_{\mathrm{dRMSD}}$ at low values ($< 10\,\text{Å}$), and that the situation is the opposite at high values.

3.3.2 Energy Correlation

We now turn to the correlation between native distance, defined in Eq. (3.19), and energy. Before describing our results, we should remark that they are of course dependent on the specific choice of energy function and its parameters. In particular, we note that the folding properties of the model used here depend strongly on the hydrogen bond and hydrophobicity strengths, ϵ_{hb} and ϵ_{AA}, respectively [see Eqs. (3.14,3.17)]. For our choice of these parameters, it turns out that folding and collapse occur at the same temperature, and that the transition is first-order-like [22]. Even a moderate change of the relative strength of these parameters leads to a very different behavior. If $\epsilon_{\mathrm{hb}}/\epsilon_{\mathrm{AA}}$ is made too large, the ground state becomes one long helix rather than a helical bundle. If, on the other hand, $\epsilon_{\mathrm{hb}}/\epsilon_{\mathrm{AA}}$ is made too small, chain collapse occurs before folding [23]. In these situations, we expect the effective energy gap between the native state and other states to be smaller, which may result in a weaker correlation between native distance and energy.

The main justification for the present form of the model is that it does show two-state folding, as observed for many small proteins. Furthermore, the fact that collapse and helix formation occur simultaneously is in accord with recent experiments [39] on small helical proteins, which found that hydrophobic association and helix formation cannot be separated in the transition state.

Let us first look at the behavior of the cRMSD measure. Figure 3.3 shows a scatter plot of the energy against the native distance $D^{\mathrm{n}}_{\mathrm{cRMSD}}$, for conformations taken at the folding temperature T_{f}. Low $D^{\mathrm{n}}_{\mathrm{cRMSD}}$ conformations have a fairly wide range of energies so that, as expected, the correlation is not very strong.

Figure 3.4 shows scatter plots of the energy against native distance for the other four distance measures. The correlation coefficient R between energy and all the five different distance measures can be found in Table 3.2. From this table we see that the energy correlation is strongest for $D^{\mathrm{n}}_{\mathrm{cont}}$ and $D^{\mathrm{n}}_{\mathrm{pow}}$, strong also for $D^{\mathrm{n}}_{\mathrm{HS}}$, and significantly weaker for $D^{\mathrm{n}}_{\mathrm{cRMSD}}$ and $D^{\mathrm{n}}_{\mathrm{dRMSD}}$. From Table 3.2 we also see that $\log(D^{\mathrm{n}}_{\mathrm{cRMSD}}+1)$ and $\log(D^{\mathrm{n}}_{\mathrm{dRMSD}}+1)$ are more strongly correlated with energy than $D^{\mathrm{n}}_{\mathrm{cRMSD}}$ and $D^{\mathrm{n}}_{\mathrm{dRMSD}}$, respectively. This is not surprising since we have seen that $D^{\mathrm{n}}_{\mathrm{cRMSD}}$ is approximately exponentially related to $D^{\mathrm{n}}_{\mathrm{cont}}$, and that $D^{\mathrm{n}}_{\mathrm{cont}}$ is very strongly correlated with energy.

The parameters r_c, r_0 and m for $D^{\mathrm{n}}_{\mathrm{cont}}$, $D^{\mathrm{n}}_{\mathrm{HS}}$ and $D^{\mathrm{n}}_{\mathrm{pow}}$, respectively, are important for the properties of these measures and therefore also for the relationship between energy and distance. The plots in Fig. 3.4b–d were obtained for the choices $r_c = 7\,\text{Å}$, $r_0 = 13\,\text{Å}$ and $m = 3$, respectively, for which the correlation coefficient R is close to maximal. It turns out, however, that the energy

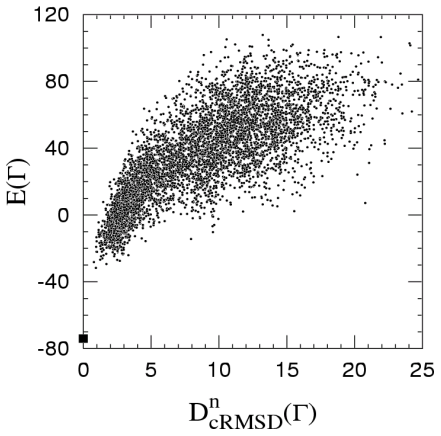


Figure 3.3: Energy $E(\Gamma)$ (in dimensionless units) against $D^{\mathrm{n}}_{\mathrm{cRMSD}}$ (in Å), as obtained at the folding temperature T_{f}. Also indicated are the representative conformations FU and BU (filled box); both $E(\mathrm{FU})$ and $E(\mathrm{BU})$ are found within the size of the plot symbol.

correlation remains strong for fairly wide ranges of these parameters. For example, for the correlation between $D^{\mathrm{n}}_{\mathrm{cont}}$ and energy we find that $R > 0.90$ if $6\,\text{Å} \leq r_c \leq 11\,\text{Å}$. The corresponding parameter intervals for $D^{\mathrm{n}}_{\mathrm{HS}}$ and $D^{\mathrm{n}}_{\mathrm{pow}}$ are $4\,\text{Å} \leq r_0 \leq 18\,\text{Å}$ and $2 \leq m \leq 11$, respectively. Finally, we note that $R = 0.83$ for $D^{\mathrm{n}}_{\mathrm{HS}}$ and $r_0 = \infty$, and that $R = 0.79$ for $D^{\mathrm{n}}_{\mathrm{pow}}$ and $m = -1$.

It should be remembered that the above results were obtained for a model whose native state is a three-helix bundle, and an important question is, of course, to what extent they hold for general proteins. In particular, it would be very interesting to perform the same analysis for proteins with a large beta sheet content. We did repeat the analysis for a 16-amino acid beta hairpin. In this case, the energy correlations were somewhat lower (between 0.73 and 0.78 for all the five distance measures). This is not unexpected because the beta hairpin is smaller and less stable than the three-helix-bundle protein studied. To see whether there is a systematic difference between alpha and beta proteins, it is clear that data for larger beta proteins are needed. Performing such simulations is, however, a very difficult problem. In fact, it takes a lot of effort to develop continuous protein models with suitable folding properties, and it is very hard to obtain models for beta proteins.

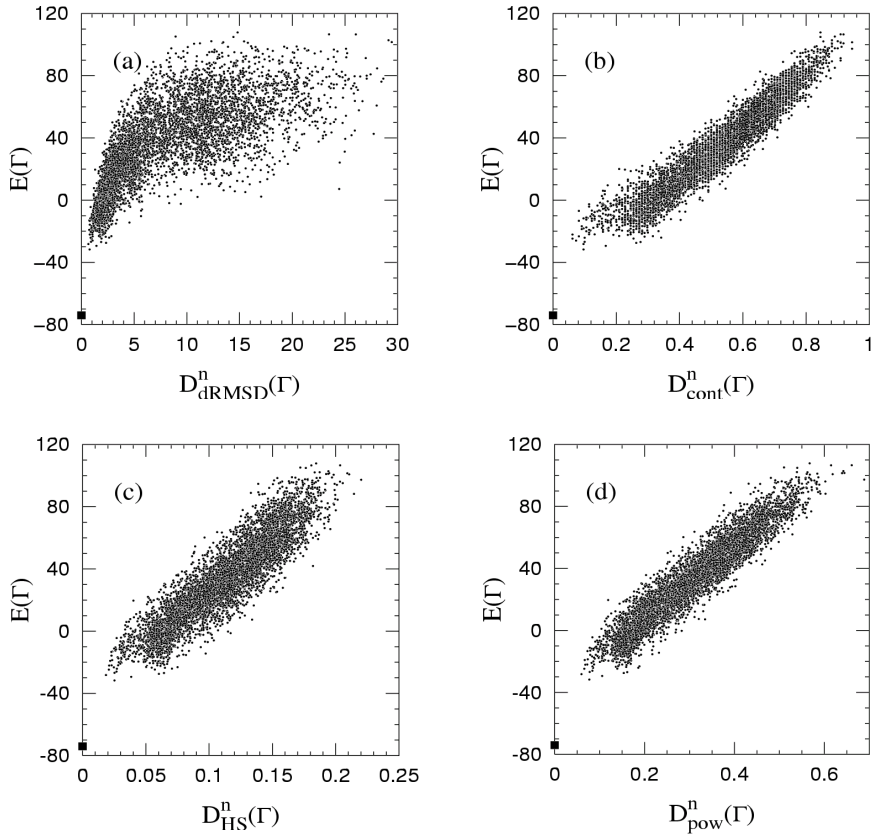


Figure 3.4: Scatter plots of the energy against different native distances, as measured by (a) $D^{\mathrm{n}}_{\mathrm{dRMSD}}$ (in Å), (b) $D^{\mathrm{n}}_{\mathrm{cont}}$ with $r_c = 7\,\text{Å}$ (see Eq. (3.5)), (c) $D^{\mathrm{n}}_{\mathrm{HS}}$ with $r_0 = 13\,\text{Å}$ (see Eq. (3.7)) and (d) $D^{\mathrm{n}}_{\mathrm{pow}}$ with $m = 3$ (see Eq. (3.10)). The plots are based on conformations taken at the folding temperature T_{f}. FU and BU are indicated using a filled box, and native distance is defined through Eq. (3.19).

3.3.3 Topology Discrimination

An interesting issue is the ability of different distance measures to discriminate structures that are locally similar but globally different. Our three-helix-bundle protein is challenging in this respect, since the FU and BU conformations, in a global sense, are related by an approximate mirror symmetry. It is easy

Distance	R	parameter
$D^{\mathrm{n}}_{\mathrm{cRMSD}}$	0.78	
$\log(D^{\mathrm{n}}_{\mathrm{cRMSD}}+1)$	0.81	
$D^{\mathrm{n}}_{\mathrm{dRMSD}}$	0.73	
$\log(D^{\mathrm{n}}_{\mathrm{dRMSD}}+1)$	0.80	
$D^{\mathrm{n}}_{\mathrm{HS}}$	0.91	$r_0 = 13\,\text{Å}$
$D^{\mathrm{n}}_{\mathrm{pow}}$	0.95	$m = 3$
$D^{\mathrm{n}}_{\mathrm{cont}}$	0.95	$r_c = 7\,\text{Å}$

Table 3.2: Correlation coefficient R between the energy and different native distances.

to see that any distance measure based on intramolecular distances is unable to distinguish between a structure and its mirror image. On the other hand, FU and BU are not exact mirror images of each other since there are no left-handed helices, so one may still hope that some of the four measures based on intramolecular distances are able to solve this task.

A relevant parameter for monitoring the ability of a distance measure to discriminate between the two topologies is

$$\Delta(\Gamma) = \frac{D(\mathrm{FU},\Gamma) - D(\mathrm{BU},\Gamma)}{D(\mathrm{FU},\mathrm{BU})}\,. \tag{3.20}$$

Using the triangular inequality, it is easy to see that $-1 \le \Delta(\Gamma) \le 1$, with equality only if Γ coincides with either FU or BU. The behavior of the parameter $\Delta(\Gamma)$ is studied at the temperature $T_{\mathrm{low}} = 0.95T_{\mathrm{f}}$, where the unfolded population is very small.

Figure 3.5 shows the probability distribution of Δ_{cRMSD}. We see that most conformations indeed belong to the basin of attraction of either FU or BU. These two basins of attraction are separated by a large energy barrier, since changing topology requires one of the end helices to cross the U formed by the other two. The fact that the Δ_{cRMSD} distribution has a bimodal shape implies that the cRMSD measure is able to discriminate efficiently between the two topologies.

In the corresponding analysis of the four measures based on intramolecular distances, we make the measures more sensitive to global differences by increasing the sequence cutoff to $s = 4$ (see section 3.2.7). This is useful because the FU and BU conformations are very similar locally, as we saw in Fig. 3.1. The D_{dRMSD}, D_{cont}, D_{HS} and D_{pow} measures obtained this way will be denoted by D'_{dRMSD}, D'_{cont}, D'_{HS} and D'_{pow}, respectively.

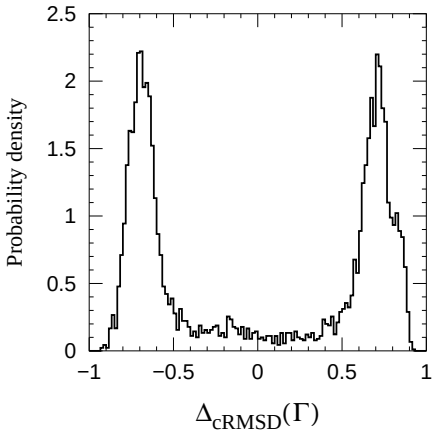


Figure 3.5: Probability distribution of Δ_{cRMSD} at T_{low}, showing that the cRMSD measure clearly discriminates between the two topologies of the three-helix-bundle protein.

Figure 3.6 shows the Δ distributions obtained using the D'_{dRMSD}, D'_{cont}, D'_{HS} and D'_{pow} measures. For D'_{dRMSD}, we see that the distribution is unimodal, in sharp contrast to the clear bimodal shape seen in Fig. 3.5. As opposed to D_{cRMSD}, the D'_{dRMSD} measure is unable to discriminate between the two topologies of the three-helix bundle. The situation is less clear for D'_{HS}, D'_{pow} and D'_{cont}. These distributions show signs of bimodality, but the separation of the two peaks is much weaker than in Fig. 3.5.

It is interesting to note that the distance between the "ideal" conformations FU and BU, as measured by the contact measure, is quite large. More precisely, one finds that $D'_{\mathrm{cont}}(\mathrm{FU}, \mathrm{BU}) = 0.62$. The difficulty in discriminating between the two topologies arises when thermal fluctuations are added.

To quantify the ability of the different distance measures to discriminate between the two topologies, we introduce three structural classes: (I) close to FU, (II) close to BU and (III) far from both FU and BU. Given a measure $D(\Gamma^a, \Gamma^b)$, we assign a conformation Γ to class I if

$$D(\mathrm{FU}, \Gamma) < \kappa D(\mathrm{FU}, \mathrm{BU}) \tag{3.21}$$

where κ is a parameter, and analogously for class II. Choosing $\kappa < 0.5$ assures that the classes I and II are disjoint, as can be seen with use of the triangular inequality. Conformations Γ such that $D(*, \Gamma) > \gamma D(\mathrm{FU}, \mathrm{BU})$ for both $*$=FU and $*$=BU, where we choose $\gamma = 1.0$, are assigned to class III. Conforma-

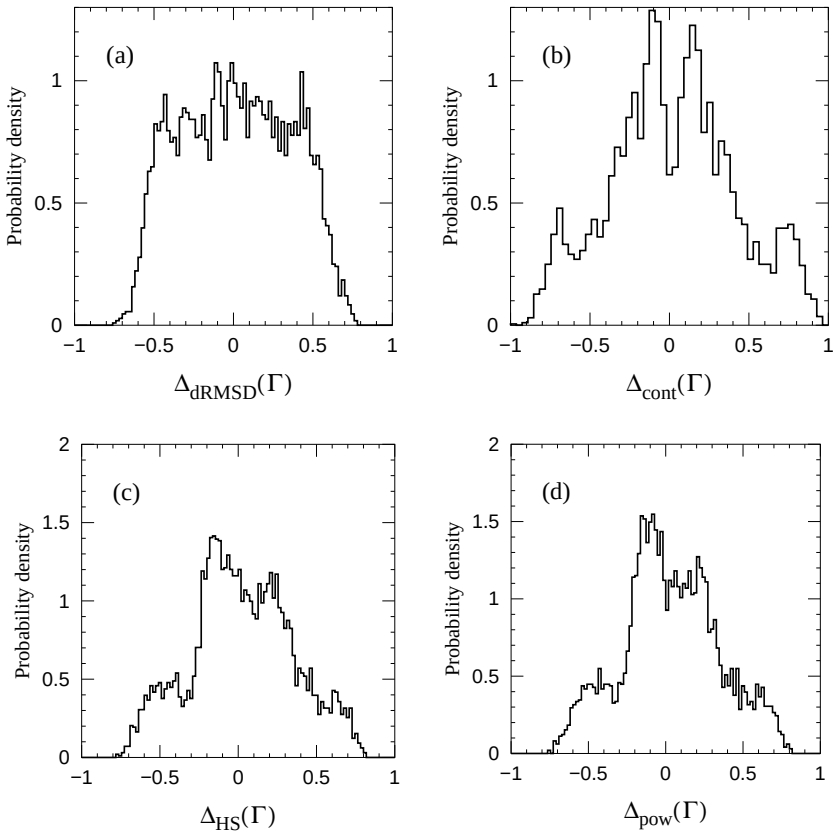


Figure 3.6: Probability distributions of the Δ parameter (see Eq. (3.20)), as constructed using the distance measures (a) D'_{dRMSD}, (b) D'_{cont}, (c) D'_{HS} and (d) D'_{pow}, at T_{low}. The parameters $r_0 = 13\,\text{Å}$, $m = 3$ and $r_c = 7\,\text{Å}$ of D'_{HS}, D'_{pow} and D'_{cont} respectively, are the same as in Fig. 3.4.

tions with native distances in between $\kappa D(\mathrm{FU}, \mathrm{BU})$ and $\gamma D(\mathrm{FU}, \mathrm{BU})$ are not assigned to any class.

Table 3.3 shows the result of this classification for D_{cRMSD}, D'_{HS}, D'_{pow}, D'_{cont} and D'_{dRMSD}. We see that the D_{cRMSD} measure classifies a large fraction of the conformations as belonging to either class I or II. By contrast, the other measures classify a large fraction as class III, and leave many conformations unclassified. This tendency is least strong for the contact measure, which does identify significant populations of FU- and BU-like conformations.

Class	I	II	III
D_{cRMSD}	0.39	0.39	0.07
D'_{dRMSD}	0.00	0.01	0.69
D'_{cont}	0.07	0.09	0.33
D'_{HS}	0.01	0.02	0.62
D'_{pow}	0.01	0.02	0.61

Table 3.3: Observed frequencies of class I, II and III at T_{low}, as obtained for the different distance measures, with $\kappa = 0.45$ and $\gamma = 1.0$ (see text).

From Table 3.3 it can be seen that there must be many conformations belonging to class I or II with the cRMSD measure that belong to class III with other measures. Notice that for a conformation in class III, the distances to both FU and BU are larger than the distance between FU and BU.

3.3.4 Discrete Models for Real Protein Structures

We now turn to the problem of approximating real protein structures. For this purpose, we use discrete C_α models where each virtual C_α-C_α bond has six possible directions, parameterized by the pseudo-bond angle α and the pseudo-torsion angle τ.

The discrete values of (α, τ) are determined by a Monte Carlo minimization procedure in the twelve-dimensional space of all possible (α, τ) angles. We follow Park and Levitt [25] and minimize a score which, for a given set of (α, τ) angles, is the average distance between a training set of 21–38 protein structures and the corresponding best model fits. The model fits are not strictly optimal since we use a stochastic algorithm (see section 3.2.10), but the distance found this way approximate well the minimal distance.

The optimized set of angles is then tested on a much larger set of proteins, consisting of 774 non-redundant protein structures without gaps in the crystal structure. The same analysis is carried out for three different distance measures, namely the cRMSD, the power distance with $m = 3$ (see Eq. (3.10)) and the contact distance with $r_c = 11\,\text{Å}$ (see Eq. (3.3)). The sequence cutoff s (see section 3.2.7) is taken to be $s = 2$ for the power and the contact distances.

Model fits obtained minimizing the cRMSD and the contact distance may exhibit C_α-C_α distances that are very small. Such structures are unphysical, since they contain atomic collisions. Therefore, we also optimized discrete models with hard core repulsion, rejecting all structures with two atoms closer than

	D_{cRMSD}	D_{cont}	D_{pow}
(α,τ) angles	$71.5, 71.2$	$83.3, 72.4$	$85.5, -62.4$
	$87.9, 55.6$	$92.0, 39.8$	$94.9, 96.0$
	$104.2, -111.0$	$115.0, -163.4$	$103.6, 163.0$
	$104.6, 36.6$	$118.0, -64.6$	$115.8, -152.2$
	$124.0, -160.0$	$128.5, 111.5$	$119.6, -22.0$
	$129.5, 128.8$	$129.7, -119.2$	$125.2, 126.8$
average score	1.57 Å	0.23	0.19

Table 3.4: Optimized (α,τ) angles, obtained without hard core repulsion.

a cutoff distance R_c. The problem of atomic collisions does not arise for the power distance, since very short atomic distances are heavily penalized by this distance measure (see Eq. (3.8)). Below we present results obtained both with and without hard core repulsion.

3.3.5 Models Without Hard Core Repulsion

The optimized discrete model obtained using cRMSD as distance measure can fit every structure up to a cRMSD of 2.1 Å (it is larger than 2.0 Å for two structures only), with an average cRMSD of 1.57 Å. This is at least 10% better than previous results with the same number of angles. The minimal cRMSD tends to increase with protein length up to around 150 residues and then stay constant, as observed in a previous study [25]. Thus our results show that the similarity between real protein structures and structures built using only six allowed C_α directions is very high in terms of cRMSD.

We performed the same analysis for the contact and the power distances. The contact distance between real and fitted model structures was found to be 0.46 at most and 0.23 on average, the latter value corresponding to 77% common contacts. For the power distance, the maximal and average distances are 0.33 and 0.19, respectively. The optimal sets of (α, τ) angles obtained with the three similarity measures are reported in Table 3.4.

These fits to real structures are based on distance minimization. In the following, we investigate how well model structures that are similar to real structures in terms of the distance measures can reproduce other geometrical features of real proteins. We focus our attention on the distribution of distances between C_α atoms. In Fig. 3.7 we plot the normalized number of pairs of C_α atoms at

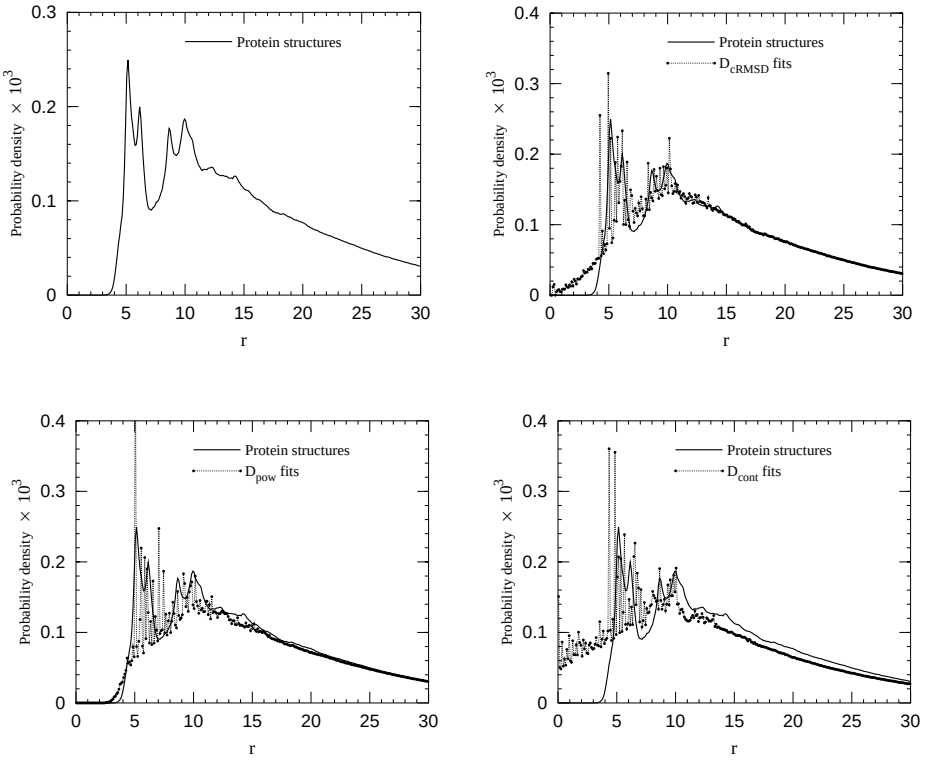


Figure 3.7: Distributions of C_α-C_α distances (in Å) for real proteins (upper left panel and solid lines in the other panels) and the corresponding models fits (dots). The upper right panel is for cRMSD and the lower panels are for the power distance (left) and the contact distance (right). In the lower left panel, the highest peak is outside the figure. Its value is $f(r) \times 10^3 = 0.53$ at $r = 5.0$.

distance r, $f(r)$, defined as

$$f(r) = \frac{N(r)}{r^2 \Delta r}\,, \tag{3.22}$$

where $N(r)$ is the fraction of pairs with C_α-C_α distance in the interval $[r, r + \Delta r]$, and Δr is taken to be 0.1 Å. The figure shows this distribution for real structures and for the three sets of fitted structures discussed above, corresponding to different distance measures. All these distributions are generated using a sequence cutoff of $s = 2$.

The upper left panel of Fig. 3.7 shows the C_α-C_α distance distribution for crystal structures of real proteins. Several details are worth noting. First, distances smaller than 3.5 Å are absent, which is due to atomic collisions. Second, there is a double-peak at 5.2–6.2 Å, which can be attributed to favorable inter-residue interactions. Third, there is a deep valley at 7.5 Å, probably due to the excluded volume effect of residues in contact with the observed residue. The tail of the distribution decays slowly with r for small r, and then approximately exponentially. In fact, the large-r behavior, $r > 13$ Å, is well fitted by an exponential function $f(r) \propto \exp(-r/\xi)$, with $\xi = 8.30 \pm 0.01$ Å. The average C_α-C_α distance is 28.75 Å.

To what extent the fitted structures reproduce these features depends on the distance measure used. For the cRMSD measure (upper right panel in Fig. 3.7), it can be seen that there is a significant population of atomic pairs with distance $r < 3.5$ Å. Such distances lead to atomic collisions and are therefore unphysical. On the other hand, the large-r behavior of this distribution is in good agreement with the results for real structures. An exponential fit for large r yields $\xi = 8.30 \pm 0.01$ Å and the average C_α-C_α distance is 28.72 Å. These numbers are very close to the corresponding ones for real proteins.

For the power distance (lower left panel in Fig. 3.7), the situation is the opposite. Small distances are in this case absent, as they should, due to a strong penalty for atomic collisions. However, this repulsion has the disadvantage that the model structures become less compact than the real ones. This is reflected in the values of the fitted exponent ξ, 8.50 ± 0.01 Å, and the average C_α-C_α distance, 29.82 Å, which are larger than for real proteins. Thus short range features are well reproduced but long range features are not.

Finally, for the contact distance (lower right panel of Fig. 3.7), we find that neither short range features nor long range ones are well reproduced. The density at $r < 3.5$ Å is even larger than in the cRMSD case; small distances are in fact favored if they increase the number of common contacts. As for long range features, one finds $\xi = 10.69 \pm 0.01$ Å and an average C_α-C_α distance of 32.8 Å. This implies that the contact distance leads to an effective long-range repulsion even stronger than for the power distance.

This effective long-range repulsion is at first sight surprising, but it can be explained by considering the definition of the contact distance. This distance measures the number of common contacts between two structures divided by the maximal number of contacts of the two structures. In order to minimize this distance, one first has to maximize the number of common contacts. Having maximized this number, the contact distance can still be further decreased if the total number of contacts in the model structure is made smaller than

the corresponding number for the native structure. Because of this, one finds that the model fits tend to have fewer contacts and be less compact than real structures.

Summarizing, we have seen that discrete structures obtained minimizing the cRMSD reproduce global but not local features of real protein structures, and that those obtained minimizing the power distance, by contrast, reproduce local but not global features. For the contact overlap measure, neither local nor global features are well reproduced.

3.3.6 Models With Hard Core Repulsion

We now turn to the calculations where atomic collisions are avoided by rejecting all model structures having two atoms closer than a cutoff distance $R_c = 2.6\,\text{Å}$. This value was chosen because it is the minimal C_α–C_α distance among the 774 native structures in our dataset. New optimal angles for the cRMSD and the contact distance were determined, and are shown in Table 3.5. The model fits obtained with the power distance satisfy the hard core repulsion constraint without rejecting any structures.

The discrete model obtained minimizing the cRMSD with hard core repulsion performs very similarly to the one reported in the previous section: every protein structure can be fitted to less than 2.1 Å cRMSD and the average cRMSD is 1.54 Å. The distribution of the C_α-C_α distances for these fits is shown in Fig. 3.8 and is very similar to the one obtained without repulsion, except that no distance smaller than R_c is present. The average distance is 28.76 Å and the large-r exponential fit to the probability distribution yields a characteristic length $\xi = 8.29 \pm 0.01\,\text{Å}$.

The inclusion of hard core repulsion does not modify the performance much for the contact distance either. The average value of the contact distance is 0.23, the same as before, and the largest contact distance is 0.42, slightly better than before. The average C_α-C_α distance for the model fits are 32.50 Å, and the fitted characteristic length is $\xi = 10.62 \pm 0.01\,\text{Å}$. Overall, the distribution is very similar to the one without repulsion, except for the zero probability density for $r < R_c$.

In conclusion, atomic collisions can be removed from the discrete models by introducing explicit hard core repulsion, without significantly changing the average quality of the fits. This hard core repulsion does, of course, affect the small-r part of the distribution of C_α-C_α distances for the fitted structures. However, the overall shape of this distribution, above R_c, changes very little

	D_{cRMSD}	D_{cont}
(α,τ) angles	$83.7, 62.0$	$82.7, 63.9$
	$95.9, 41.6$	$104.8, 169.4$
	$109.7, -151.9$	$106.9, 25.9$
	$110.8, -104.4$	$113.3, -137.3$
	$129.3, 186.7$	$113.8, -70.6$
	$134.0, 120.9$	$128.0, 109.9$
average score	1.54 Å	0.23

Table 3.5: Optimized (α,τ) angles, obtained with hard core repulsion.

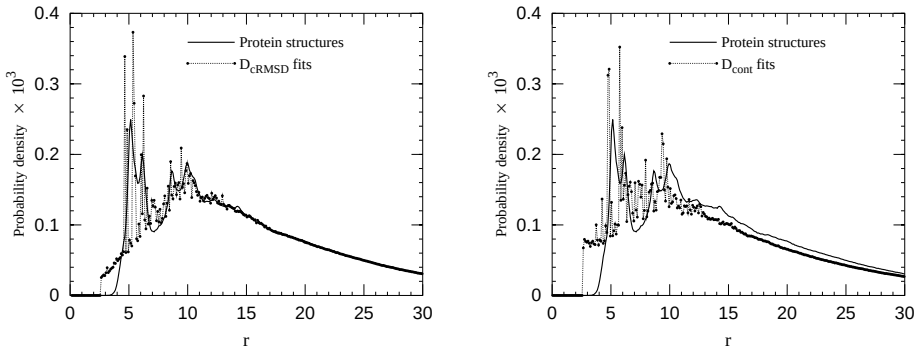


Figure 3.8: Distributions of C_α-C_α distances (in Å) for fits to real proteins (dots) obtained with hard core repulsion. Left and right panels are for cRMSD and the contact distance, respectively.

when introducing the hard core repulsion.

3.4 Conclusions

We have investigated the properties of five different distance measures: the standard cRMSD measure and four measures based on intramolecular distances. We recall that the Holm and Sander measure is not strictly speaking a distance, since it does not fulfill the triangular inequality, but violations of this inequality are very rare. In fact, with our normalized version of this measure, D_{HS}, there were no such violations in our data set.

Using a continuous three-helix-bundle model, the correlation between native distance and energy was studied. We find that this correlation is significantly stronger for the measures D_{HS}, D_{pow} and D_{cont} than for D_{cRMSD} and D_{dRMSD}. This suggests that the former distance measures are more suitable to investigate the shape of the energy landscape. It must be remembered, however, that these results were obtained for one particular protein. This three-helix-bundle protein was chosen because the model has a relatively realistic chain representation and exhibits two-state folding. How general our conclusions are remains to be seen.

On the other hand, the ability to discriminate between the two different topologies of our three-helix-bundle protein is found to be quite limited for all the four measures based on intramolecular distances, while this task is easily solved by the cRMSD measure. This topology problem concerns structures in which not all contacts have been formed, while the two "ideal" minimum-energy structures can be distinguished without difficulties, at least with the contact distance. Therefore, one can still expect that native protein structures determined by X-ray crystallography can be assigned the correct topology. However, our results suggest that caution should be taken with the distance measures based on intramolecular distances. Among these, the best discrimination ability is exhibited by the contact distance and the worst one by the dRMSD.

Using different similarity measures as scoring functions, we performed fits of discrete C_α models with six directions per amino acid to real protein structures. The properties of the fitted structures turn out to depend quite strongly on the similarity measure used. Structures obtained using the cRMSD measure provide a good representation of large scale properties, but fail to reproduce small scale properties. The problem here is that the fitted structures tend to contain unphysical atomic collisions, which are not penalized by the cRMSD measure. Such collisions can be avoided by introducing a hard core repulsion that explicitly rejects structures containing atomic distances below a cutoff value. For the power distance, we find the opposite behavior. Here, the fitted structures reproduce local features very well and global features much worse. The contact measure is bad in both respects; the fits contain atomic collisions and are, at the same time, not as compact as they should. The latter problem reflects the fact that the fits have fewer contacts than the real structures. These results imply that a straightforward application of standard energy functions to fitted discrete structures can be misleading, due to artifacts at small or large scale.

Comparing these results to those obtained using the continuous model, a consistent picture seems to emerge. The cRMSD is by far the best for reproducing long range properties of protein structures, but fails to reproduce short range properties, as reflected in a poor energy correlation and the appearance of

unphysical C_α-C_α collisions in the fitted structures. The power distance introduced here and the Holm and Sander measure are, by contrast, good at reproducing small scale properties, but the overall size of the fitted structures is too large and the ability to discriminate between the three-helix-bundle topologies is quite poor. For the contact distance the picture is less simple. This measure gives a good energy correlation, a short range property, and is not too bad at topology discrimination, a long range property. However, the fitted structures are relatively poor at both small and large scale. To conclude, it seems that there is no distance measure good for all purposes; a distance measure to be used in the complex space of protein conformations should be chosen in consideration of the application in question.

Acknowledgments

We are indebted to Anders Irbäck and Ernst-Walter Knapp for valuable discussions and suggestions and to an anonymous referee for suggesting the inclusion of hard core repulsion in the discrete models.

References

[1] Venclovas C, Zemla A, Fidelis K, Moult J. Comparison of performance in successive CASP Experiments. *Proteins Struct Funct Genetics, Suppl.* 2001; 5: 163-170.

[2] Holm L, Sander C. The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res* 1994; 22: 3600-3609.

[3] Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH: A hierarchic classification of protein domain structures. *Structure* 1997; 5: 1093-1108.

[4] Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995; 247: 536-540.

[5] Koehl P. Protein Structure similarities. *Curr Opin Struct Biol* 2001; 11: 348-53.

[6] May ACW. Towards more meaningful hierarchical classification of amino acid scoring matrices. *Protein Eng* 1999; 12: 707-712.

[7] Orengo CA, Swindells MB, Michie AD, Zvelebil MJ, Driscoll PC, Waterfield MD, Thornton JM. Structural similarity between the pleckstrin homology domain and verotoxin: the problem of measuring and evaluating structural similarity. *Protein Sci* 1995; 4: 1977-1983.

[8] Feng ZK, Sippl MJ. Optimum superimposition of protein structures: ambiguities and implications. *Fold Des* 1996; 1: 123-132.

[9] Godzik A. The structural alignment between two proteins: is there a unique answer? *Protein Sci* 1996; 5: 1325-1338.

[10] Anfinsen CB. Principles that govern the folding of protein chains. *Science* 1973; 181: 223–230.

[11] Bryngelson JD, Wolynes PG. Spin-glasses and the statistical-mechanics of protein folding. *Proc Natl Acad Sci USA* 1987; 84: 7524-7528.

[12] Garel T, Orland H. Mean-field model for protein folding. *Europhys Lett* 1988; 6: 307–310.

[13] Shakhnovich EI, Gutin AM. Formation of unique structure in polypeptide chains. Theoretical investigation with the aid of a replica approach. *Biophys Chem* 1989; 34; 187–199.

[14] Shakhnovich EI, Gutin AM. Engineering of stable and fast-folding sequences of model proteins. *Proc Natl Acad Sci USA* 1993; 90: 7195–7199.

[15] Shakhnovich EI. Proteins with selected sequences fold into unique native conformation. *Phys Rev Lett* 1994; 24: 3907–3910.

[16] Abkevich VI, Gutin AM, Shakhnovich EI. Free energy landscapes for protein folding kinetics - intermediates, traps and multiple pathways in theory and lattice model simulations. *J Chem Phys* 1994; 101: 6052–6062.

[17] Klimov DK, Thirumalai D. Factors governing the foldability of proteins. *Proteins Struct Funct Genet* 1996; 26: 411–441.

[18] Bastolla U, Frauenkron H, Gerstner E, Grassberger P, Nadler W. Testing a new Monte Carlo algorithm for protein folding. *Proteins Struct Funct Genet* 1998; 32: 52–66.

[19] Tiana G, Broglia RA, Roman HE, Vigezzi E, Shakhnovich EI. Folding and misfolding of designed protein-like chains with mutations. *J Chem Phys* 1998; 108: 757–761.

[20] Bastolla U, Roman HE, Vendruscolo M. Neutral evolution of model proteins: Diffusion in sequence space and overdispersion. *J Theor Biol* 1999; 200: 49–64.

[21] Park B, Levitt M. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. *J Mol Biol* 1996; 258: 367–392.

[22] Irbäck A, Sjunnesson F, Wallin S. Three-helix-bundle protein in a Ramachandran model. *Proc Natl Acad Sci USA* 2000; 97: 13614–13618.

[23] Irbäck A, Sjunnesson F, Wallin S. Hydrogen bonds, hydrophobicity forces and the character of the collapse transition. *J Biol Phys* 2001; 27: 169–179.

[24] Favrin G, Irbäck A, Wallin S. Folding of a small helical protein using hydrogen bonds and hydrophobicity forces. *Proteins Struct Funct Genet* 2002; 47: 99–105.

[25] Park BH, Levitt M. The complexity and accuracy of discrete state models of protein structure. *J Mol Biol* 1995; 249: 493–507.

[26] von Neumann J. Some matrix-inequalities and metrization of matric-space. *Tomsk Univ Rev* 1937; 1: 286–300.

[27] Kabsch W. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallog sect A* 1978; 34: 827–828.

[28] Maiorov VN, Crippen GM. Size-independent comparison of protein three-dimensional structures. *Proteins Struct Funct Genet* 1995; 22: 273–283.

[29] Levitt M. A simplified representation of protein conformations for rapid simulation of protein folding. *J Mol Biol* 1976; 104: 59–107.

[30] Cohen FE, Sternberg JE. On the prediction of protein structure: the significance of the root-mean-square deviation. *J Mol Biol* 1980; 138: 321–333.

[31] Maiorov VN, Crippen GM. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *J Mol Biol* 1994; 235: 625–634.

[32] Vendruscolo M, Subramanian B, Kanter I, Domany E, Lebowitz J. Statistical properties of contact maps. *Phys Rev E* 1999; 59: 977–984.

[33] Bastolla U, Frauenkron H, Grassberger P. Phase diagram of random heteropolymers: replica approach and application of a new Monte Carlo algorithm. *Jour Mol Liq* 2000; 84: 111–129.

[34] Chan HS, Kaya H, Shimizu S. Computational methods for protein folding: scaling a hierarchy of complexities. In: Jiang T, Xu Y, Zhang MQ, editors. Current topics in computational molecular biology. Cambridge, Massachusetts: MIT Press; 2002. p 403–447.

[35] Holm L, Sander C. Protein structure comparison by alignment of distance matrices. *J Mol Biol* 1993; 233: 123–138.

[36] Holm L, Sander C. Mapping the protein universe. *Science* 1996; 273: 595–602.

[37] Rooman MJ, Kocher JA, Wodak SJ. Prediction of protein backbone conformation based on seven structure assignments. *J Mol Biol* 1991; 221: 961–979.

[38] Bastolla U, Farwer J, Knapp EW, Vendruscolo M. How to guarantee optimal stability for most representative structures in the protein data bank. *Proteins Struct Funct Genet* 2001; 44: 79–96.

[39] Krantz BA, Srivastava AK, Nauli S, Baker D, Sauer RT, Sosnick TR. Understanding protein hydrogen bond formation with kinetic H/D amide isotope effects. *Nature Struct Biol* 2002; 9: 458–463.

Appendix

In this Appendix, we discuss the triangular inequality. We prove that it holds for the contact distance and find that violations of it are very rare for the power and Holm and Sander distances, although these ones do not strictly fulfill the inequality. The cRMSD and dRMSD measures are guaranteed to satisfy the triangular inequality since they are proportional to the ordinary Euclidean distance in $3N$- and $N(N-1)/2$-dimensional space, respectively.

Contact Distance

Let us denote by M_a, M_b and M_c the total number of contacts in structures a, b and c, respectively, and by M_{ab} the number of shared contacts between a and b, and so on. We assume that $M_a \geq M_b \geq M_c$ and consider here only the side ab of the triangle abc. For the other two sides, the triangular inequality can be proven analogously (and more easily). The inequality to be proven reads

$$1 - \frac{M_{ab}}{M_a} \leq 1 - \frac{M_{ac}}{M_a} + 1 - \frac{M_{bc}}{M_b} \,. \tag{3.23}$$

This can be readily transformed into

$$M_a \left(M_{bc|a} + M_{abc}\right) + M_b M_{ac|b} \leq M_a M_b + M_b M_{ab|c} \,, \tag{3.24}$$

where M_{abc} denotes the number of contacts present in all the three structures, and $M_{ab|c}$ denotes the number of contacts present in structures a and b but not in c. M_{abc} has been eliminated from both the left- and right-hand sides. Since $M_a \geq M_b$, the l.h.s. is not larger than $M_a(M_{ac|b} + M_{bc|a} + M_{abc})$, which in turn is not larger than $M_a M_c \leq M_a M_b$. This, finally, is not larger than the r.h.s., so the inequality must hold.

Power Distance

It is evident that the distance defined in Eq. (3.8) fulfills the triangular inequality. The normalized measure $D^{(1)}_{\mathrm{pow}}$ in Eq. (3.9) does so too, as can be seen by applying the inequality

$$\frac{|a-b|}{a+b} \leq \frac{|a-c|}{a+c} + \frac{|b-c|}{b+c} \tag{3.25}$$

to each term, where a, b and c are positive numbers. To prove this inequality, we assume that $a \geq b \geq c$ (the remaining two cases can be treated analogously).

In this case, the inequality (3.25) is equivalent to

$$(a+c)(b+c)(a-b) \leq 2(a+b)\left(ab-c^2\right) \ . \tag{3.26}$$

Here, since $b \geq c$, the r.h.s. is not smaller than $2(a+b)(ab-b^2)$. At the same time, the l.h.s. is not larger than $(a+b)(b+b)(a-b) = 2(a+b)(ab-b^2)$, since $a \geq b$ and $b \geq c$. Therefore the inequality is proven.

The triangular inequality does not hold strictly for the normalized version $D_{\rm pow}$ in Eq. (3.10) that we use in this paper. However, as mentioned earlier, no violation was detected for the (large) set of protein-like structures that we use.

The Holm and Sander Score

For the original version of the Holm and Sander distance, $D^*_{\rm HS}$ in Eq. (3.6), and for $N = 2$, the triangular inequality has the form

$$\frac{|a-b|}{a+b}\mathrm{e}^{-(a+b)^2} \leq \frac{|a-c|}{a+c}\mathrm{e}^{-(a+c)^2} + \frac{|b-c|}{b+c}\mathrm{e}^{-(b+c)^2} \ , \tag{3.27}$$

which is clearly violated if $c \gg a, b$. For the normalized version $D_{\rm HS}$ in Eq. (3.7) and $N = 2$, the triangular inequality is equivalent to the inequality [Eq. (3.25)], which, as we have seen, is satisfied. The triangular inequality for $D_{\rm HS}$ and $N = 3$ takes the form

$$\frac{\frac{|a_1-b_1|}{a_1+b_1}\mathrm{e}^{-(a_1+b_1)^2} + \frac{|a_2-b_2|}{a_2+b_2}\mathrm{e}^{-(a_2+b_2)^2} + \frac{|a_3-b_3|}{a_3+b_3}\mathrm{e}^{-(a_3+b_3)^2}}{\mathrm{e}^{-(a_1+b_1)^2} + \mathrm{e}^{-(a_2+b_2)^2} + \mathrm{e}^{-(a_3+b_3)^2}} \leq \frac{|a_1-c_1|}{a_1+c_1} + \frac{|b_1-c_1|}{b_1+c_1} \tag{3.28}$$

in the limit of one distant point (two distances $c_2, c_3 \to \infty$). This inequality can not be satisfied since the r.h.s becomes zero when $a_1 = b_1 = c_1$, and the l.h.s. depends on a_2, a_3, b_2 and b_3. For large N, however, $D_{\rm HS}$ consists of many terms and violations of the triangular inequality become very rare.

Thermodynamics of α- and β-Structure Formation in Proteins

Paper IV

Thermodynamics of α- and β-Structure Formation in Proteins

Anders Irbäck, Björn Samuelsson,
Fredrik Sjunnesson and Stefan Wallin

Complex Systems Division, Department of Theoretical Physics
Lund University, Sölvegatan 14A, SE-223 62 Lund, Sweden
http://www.thep.lu.se/complex/

Submitted to *Biophys. J.*

Abstract:
An atomic protein model with a minimalistic potential is developed and then tested on an α-helix and a β-hairpin, using exactly the same parameters for both peptides. We find that thermal unfolding curves for these sequences to a good approximation can be described by a simple two-state model, with parameters that are in reasonable *quantitative* agreement with experimental data. Despite the apparent two-state character of these curves, the energy distributions are found to lack a clear bimodal shape, which is discussed in some detail. We also perform a Monte Carlo-based kinetic study and find, in accord with experimental data, that the α-helix forms faster than the β-hairpin.

4.1 Introduction

Simulating protein folding at atomic resolution is a challenge, but no longer computationally impossible, as shown by recent studies [1, 2] of Gō-type [3] models with a bias towards the native structure. Extending these calculations to entirely sequence-based potentials remains, however, an open problem, due to well-known uncertainties about the form and relevance of different terms of the potential. In this situation, it is tempting to look into the properties of atomic models that are sequence-based and yet as simple and transparent as possible.

The development of models for protein folding is hampered by the fact that short amino acid sequences with protein-like properties are rare, which makes the calibration of potentials a non-trivial task. Breakthrough experiments in the past ten years have, however, found examples of such sequences. Of particular importance was the discovery of a peptide making β-structure on its own [4], the second β-hairpin from the protein G B1 domain, along with the finding that this 16-amino acid chain, like many small proteins, show two-state folding [5]. These experiments have stimulated many theoretical studies of the folding properties of this sequence, including simulations of atomic models with relatively detailed semi-empirical potentials [6–11]. Reproducing the melting behavior of the β-hairpin has, however, proven non-trivial, as was recently pointed out by Zhou *et al.* [11].

Here we develop and explore a simple sequence-based atomic model, which is found to provide a surprisingly good description of the thermodynamic behavior of this peptide. The same model, with unchanged parameters, is also applied to an α-helical peptide, the designed so-called F_S peptide with 21 amino acids [12,13]. We find that this sequence indeed makes an α-helix in the model, and our results for the stability of the helix agree reasonably well with experimental data [12–15]. Finally, we also study Monte Carlo-based kinetics for both these peptides. Here we investigate the relaxation of ensemble averages at the respective melting temperatures.

4.2 Model and Methods

4.2.1 The Model

Recently, we developed a simple sequence-based model with 5–6 atoms per amino acid for helical proteins [16–18]. Here we extend that model by incorporating all atoms. The interaction potential is deliberately kept simple. The chain representation is, by contrast, detailed; in fact, it is more detailed than in standard "all-atom" models as all hydrogens are explicitly included. The presence of the hydrogens has the advantage that local torsion potentials can be avoided. All bond lengths, bond angles and peptide torsion angles (180°) are held fixed, which means that each amino acid has the Ramachandran torsion angles ϕ, ψ and a number of side-chain torsion angles as its degrees of freedom (for Pro, ϕ is held fixed at $-65°$). The geometry parameters held constant are derived by statistical analysis of Protein Data Bank (PDB) [19] structures. A complete list of these parameters can be found as supplemental material.

The potential function

$$E = E_{\rm ev} + E_{\rm hb} + E_{\rm hp} \tag{4.1}$$

is composed of three terms, representing excluded-volume effects, hydrogen bonds and effective hydrophobicity forces (no explicit water), respectively. The remaining part of this section describes these different terms. Energy parameters are quoted in dimensionless units, in which the melting temperature $T_{\rm m}$, defined as the specific heat maximum, is given by $kT_{\rm m} = 0.4462 \pm 0.0014$ for the β-hairpin. In the next section, the energy scale of the model is set by fixing $T_{\rm m}$ for this peptide to the experimental midpoint temperature, $T_{\rm m} = 297\,{\rm K}$ [5].

The excluded-volume energy $E_{\rm ev}$ is given by

$$E_{\rm ev} = \epsilon_{\rm ev} \sum_{i<j} \left[\frac{\lambda_{ij}(\sigma_i + \sigma_j)}{r_{ij}} \right]^{12}, \tag{4.2}$$

where $\epsilon_{\rm ev} = 0.10$ and $\sigma_i = 1.77$, 1.71, 1.64, 1.42 and 1.00 Å for S, C, N, O and H atoms, respectively. Our choice of σ_i values is guided by the analysis of Tsai *et al.* [20]. The parameter λ_{ij} in Eq. 4.2 reduces the repulsion between non-local pairs; $\lambda_{ij} = 1$ for all pairs connected by three covalent bonds and for HH and OO pairs from adjacent peptide units, and $\lambda_{ij} = 0.75$ otherwise. The pairs for which $\lambda_{ij} = 1$ strongly influence the shapes of Ramachandran maps and rotamer potentials. The reason for using $\lambda_{ij} < 1$ for the large majority of all pairs is both computational efficiency and the restricted flexibility of chains with only torsional degrees of freedom. To speed up the calculations, the sum in Eq. 4.2 is evaluated using a pair dependent cutoff $r_{ij}^{\rm c} = 4.3\lambda_{ij}$ Å.

The hydrogen-bond energy $E_{\rm hb}$ has the form

$$E_{\rm hb} = \epsilon_{\rm hb}^{(1)} \sum_{\substack{j<i-2 \\ \text{or } j>i+1}} u(r_{ij}) v(\alpha_{ij}, \beta_{ij}) + \epsilon_{\rm hb}^{(2)} \sum u(r_{ij}) v(\alpha_{ij}, \beta_{ij}) \,, \tag{4.3}$$

where $\epsilon_{\rm hb}^{(1)} = 3.1$, $\epsilon_{\rm hb}^{(2)} = 2.0$ and the functions u and v are given by

$$u(r) = 5\left(\frac{\sigma_{\rm hb}}{r}\right)^{12} - 6\left(\frac{\sigma_{\rm hb}}{r}\right)^{10} \tag{4.4}$$

$$v(\alpha,\beta) = \begin{cases} (\cos\alpha\cos\beta)^{1/2} & \text{if } \alpha,\beta > 90^\circ \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

The first sum in Eq. 4.3 represents backbone-backbone hydrogen bonds. Term ij in this sum is an interaction between the NH and C′O groups of amino acids i and j, respectively. r_{ij} denotes the HO distance, and α_{ij} and β_{ij} are the NHO and HOC′ angles, respectively. The second sum in Eq. 4.3 is expressed in a schematic way. It represents interactions between oppositely charged side chains, and between charged side chains and the backbone. Both these types of interaction are, for convenience, taken to have the same form as backbone-backbone hydrogen bonds. The side chain atoms that can act as "donors" or "acceptors" in these interactions are the N atoms of Lys and Arg (donors) and the O atoms of Asp and Glu (acceptors). The second sum in Eq. 4.3 has a relatively weak influence on the thermodynamic behavior of the systems studied. The backbone-backbone hydrogen bonds are, by contrast, crucial and their strength, $\epsilon_{\rm hb}^{(1)}$, must be carefully chosen [17].

The functional form of the hydrogen-bond energy differs from that in our helix model [16–18] in that the exponent of the cosines is 1/2 instead of 2. The reason for this change is that the β-hairpin turned out to become too regular when using the exponent 2; the exponent 1/2 gives a more permissive angular dependence. The function $u(r)$ in Eq. 4.4 is calculated using a cutoff $r^{\rm c} = 4.5\,\text{Å}$ and $\sigma_{\rm hb} = 2.0\,\text{Å}$.

The last term of the potential, the hydrophobicity energy $E_{\rm hp}$, assigns to each amino acid pair an energy that depends on the amino acid types and the degree of contact between the side chains. It can be written as

$$E_{\rm hp} = \epsilon_{\rm hp} \sum M_{IJ} C_{IJ} \,, \tag{4.6}$$

where $\epsilon_{\rm hp} = 1.5$, and the sum runs over all possible amino acid pairs IJ except nearest neighbors along the chain. In the present study, the M_{IJ}'s (≤ 0) are given by the contact energies of Miyazawa and Jernigan [21] shifted to zero mean, provided that the amino acids I and J both are hydrophobic and that

	Ala	Val	Leu	Ile	Phe	Tyr	Trp	Met
Ala	0.00	0.44	1.31	0.98	1.21	0.00	0.22	0.34
Val		1.92	2.88	2.45	2.69	1.02	1.58	1.72
Leu			3.77	3.44	3.68	2.07	2.54	2.81
Ile				2.94	3.24	1.65	2.18	2.42
Phe					3.66	2.06	2.56	2.96
Tyr						0.57	1.06	1.31
Trp							1.46	1.95
Met								1.86

Table 4.1: The interaction matrix M_{IJ}, based on the shifted contact-energy matrix of Miyazawa and Jernigan [21]. The table shows absolute values ($M_{IJ} \leq 0$).

the shifted contact energy is negative; otherwise, $M_{IJ} = 0$. The statistical Miyazawa-Jernigan energies contain, of course, other contributions too, but receive a major contribution from hydrophobicity [22]. The matrix M_{IJ} is given in Table 4.1. Eight of the amino acids are classified as hydrophobic, namely Ala, Val, Leu, Ile, Phe, Tyr, Trp and Met. The geometry factor C_{IJ} in Eq. 4.6 is a measure of the degree of contact between amino acids I and J. To define C_{IJ}, we use a predetermined set of N_I atoms, denoted by A_I, for each amino acid I. For Phe, Tyr and Trp, the set A_I consists of the C atoms of the hexagonal ring. The other five hydrophobic amino acids each have an A_I containing all its non-hydrogen side-chain atoms. With these definitions, C_{IJ} can be written as

$$C_{IJ} = \frac{1}{N_I + N_J} \left[\sum_{i \in A_I} f(\min_{j \in A_J} r_{ij}^2) + \sum_{j \in A_J} f(\min_{i \in A_I} r_{ij}^2) \right], \tag{4.7}$$

where the function $f(x) = 1$ if $x < A$, $f(x) = 0$ if $x > B$, and $f(x) = (B - x)/(B - A)$ if $A < x < B$ [$A = (3.5\,\text{Å})^2$ and $B = (4.5\,\text{Å})^2$]. Roughly speaking, C_{IJ} is a measure of the fraction of atoms in A_I or A_J that are in contact with some atom from the opposite side chain.

4.2.2 Numerical Methods

To study the thermodynamic behavior of this model, we use the simulated-tempering method [23–25], in which the temperature is a dynamical variable. This method is chosen in order to speed up the calculations at low temperatures. Our simulations are started from random configurations, and eight different temperatures are studied, ranging from 273 K to 366 K.

Both the temperature update and all side-chain updates are standard Metropolis steps [26]. For the backbone degrees of freedom, we use three different elementary moves: first, the pivot move [27] in which a single torsion angle is turned; second, a semi-local method [28] that works with seven or eight adjacent torsion angles, which are turned in a coordinated way; and third, a symmetry-based update of three randomly chosen backbone torsion angles. To see how the third move works, consider the three bonds corresponding to the randomly chosen torsion angles. The idea is then to reflect the mid bond in the plane defined by the two others, keeping the directions of these two other bonds fixed. Both this update and the pivot move are non-local. They are included in our thermodynamic calculations in order to accelerate the evolution of the system at high temperatures.

Our kinetic simulations are also Monte Carlo-based, and only meant to mimic the time evolution of the system in a qualitative sense. They differ from our thermodynamic simulations in two ways: first, the temperature is held constant; and second, the two non-local backbone updates are not used, but only the semi-local method [28]. This restriction is needed in order to avoid large unphysical deformations of the chain. For the side-chain degrees of freedom, we use a Metropolis step in which the angle can change by any amount (same as in the thermodynamic runs). Thus, it is assumed that the torsion angle dynamics are much faster for the side chains than for the backbone.

In our thermodynamic analysis, statistical errors are obtained by analyzing data from ten independent runs, each containing 10^9 elementary steps and several folding/unfolding events. All errors quoted are 1σ errors. All fits of data discussed in the next section are carried out by using a Levenberg-Marquardt procedure [29].

4.3 Results and Discussion

Using the model described in the previous section, we first study the second β-hairpin from the protein G B1 domain (amino acids 41–56). Blanco *et al.* [4] analyzed this peptide in solution by NMR and found that the excised fragment adopts a structure similar to that in the full protein, although the NMR restraints were insufficient to determine a unique structure. In our calculations, in the absence of a complete structure for the isolated fragment, we monitor the root-mean-square deviation (rmsd) from the native β-hairpin of the full protein (PDB code 1GB1, first model), as determined by NMR [30]. The native β-hairpin contains a hydrophobic cluster consisting of Trp43, Tyr45, Phe52 and Val54. There is experimental evidence [31] that this cluster as well as

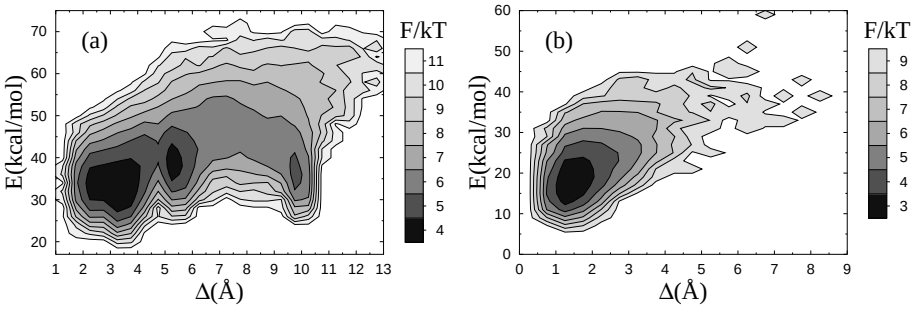


Figure 4.1: Free energy $F(\Delta, E) = -kT \ln P(\Delta, E)$ at $T = 273\,\mathrm{K}$ for (a) the β-hairpin and (b) the $\mathrm{F_s}$ peptide. E is energy and Δ denotes rmsd from the native β-hairpin and an ideal α-helix, respectively, calculated over all non-hydrogen atoms (a backbone rmsd would be unable to distinguish the two possible β-hairpin topologies).

sequence-specific hydrogen bonds in the turn are crucial for the stability of the isolated β-hairpin.

Fig. 4.1a shows the free energy $F(\Delta, E)$ as a function of rmsd from the native β-hairpin, Δ, and energy, E, at the temperature $T = 273\,\mathrm{K}$. For a β-hairpin there are two topologically distinct states with similar backbone folds but oppositely oriented side chains. The global minimum of $F(\Delta, E)$ is found at 2–4 Å in Δ and corresponds to a β-hairpin with the native topology and the native set of hydrogen bonds between the two strands. The main difference between structures within this minimum lies in the shape of the turn. The precise shape of the β-hairpin is, not unexpectedly, sensitive to details of the potential; in particular, we find that the second term in Eq. 4.3 does influence the shape of the turn, while having only a small effect on thermodynamic functions such as E_{hp}. Therefore, it is not unlikely that a more detailed potential would discriminate between different shapes of the turn, and thereby make the free-energy minimum more narrow.

Besides its global minimum, $F(\Delta, E)$ exhibits two local minima (see Fig. 4.1a), one corresponding to a β-hairpin with the non-native topology ($\Delta \approx 5\,\text{Å}$), and the other to an α-helix ($\Delta \approx 10\,\text{Å}$). A closer examination of structures from the two β-hairpin minima reveals that the C_β-C_β distances for Tyr45-Phe52 and Trp43-Val54 tend to be smaller in the non-native topology than in the native one. This is important because it makes it sterically difficult to achieve a proper contact between the aromatic side chains of Tyr45 and Phe52 in the non-native topology. As a result, this topology is hydrophobically disfavored.

This is the main reason why the model indeed favors the native topology over the non-native one.

We now turn to the melting behavior of the β-hairpin. By studying tryptophan fluorescence (Trp43), Muñoz *et al.* [5] found that the unfolding of this peptide with increasing temperature shows two-state character, with parameters $T_\mathrm{m} = 297\,\mathrm{K}$ and $\Delta E = 11.6\,\mathrm{kcal/mol}$, T_m and ΔE being the melting temperature and energy change, respectively. To study the character of the melting transition in our model, we monitor the hydrophobicity energy E_hp, a simple observable we expect to be strongly correlated with Trp43 fluorescence. Following Muñoz *et al.* [5], we fit our data for E_hp to a first-order two-state model. To reduce the number of parameters of the fit, T_m is held fixed, at the specific heat maximum (data not shown). The fit turns out not to be perfect, with a χ^2/dof of 4.5. The deviations from the fitted curve are nevertheless small, as can be seen from Fig. 4.2a; they can be detected only because the statistical errors are very small ($\sim 0.1\,\%$) at the highest temperatures. To further illustrate this point, we assign each data point an artifical uncertainty of $1\,\%$, an error size that is not uncommon for experimental data. With these errors, the same type of fit yields a χ^2/dof of 0.3, which confirms that the data indeed to a good approximation show two-state behavior. Our fitted value of ΔE is $9.3 \pm 0.3\,\mathrm{kcal/mol}$, which implies that the temperature dependence of the model is comparable to experimental data [5].

Several groups have simulated the same β-hairpin using atomic models with implicit [6, 7] or explicit [8–11] solvent. All these models have, in contrast to ours, given a very weak dependence on temperature, compared to experimental data [11]. Another important difference between at least some of these models [7, 9, 10] and ours, is that in our model there is no clear free-energy minimum corresponding to a hydrophobically collapsed state with few or no hydrogen bonds. A local free-energy minimum with helical content was found in one of these studies [10], but not in the others. Such a minimum exists in our model (see Fig. 4.1a), but the helix population is low.

In spite of its minimalistic potential, our model is able to make α-helices too. To show this, we consider the α-helical so-called $\mathrm{F_S}$ peptide, which has been extensively studied both experimentally [12–15] and theoretically [32]. This 21-amino acid peptide is given by $\mathrm{AAAAA(AAARA)_3A}$, where A is Ala and R is Arg. Using exactly the same model as before, with unchanged parameters, we find that the $\mathrm{F_S}$ sequence does make an α-helix. This can be seen from Fig. 4.1b, which shows the free energy $F(\Delta, E)$ at $T = 273\,\mathrm{K}$, Δ this time denoting rmsd from an ideal α-helix. $F(\Delta, E)$ has only one significant minimum, which indeed is helical. The melting behavior of this sequence is illustrated in Fig. 4.3a, which shows the temperature dependence of the hydrogen-bond energy. Data

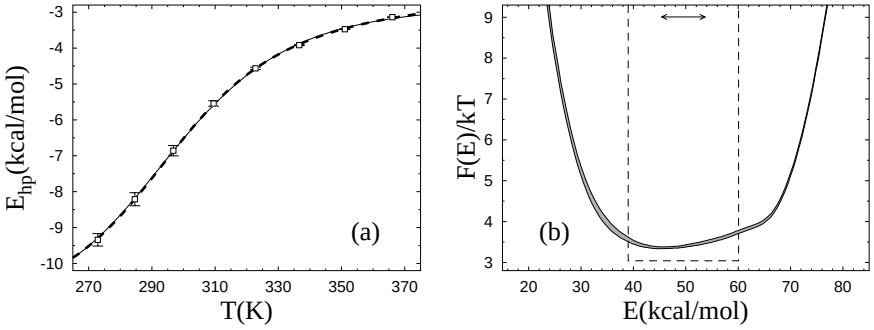


Figure 4.2: Unfolding of the β-hairpin sequence. (a) Temperature dependence of the hydrophobicity energy $E_{\rm hp}$ (see Eq. 4.6). The solid and dashed curves (essentially coinciding) are fits of the data to the two-state expression $E_{\rm hp} = (E^{\rm u}_{\rm hp} + K E^{\rm f}_{\rm hp})/(1+K)$ and the square-well model (see text), respectively. The effective equilibrium constant K is assumed to have the first-order form $K = \exp[(1/kT - 1/kT_{\rm m})\Delta E]$. Both fits have three free parameters, whereas $T_{\rm m} = 297\,{\rm K}$ is held fixed. (b) Free-energy profile $F(E) = -kT \ln P(E)$ at $T = T_{\rm m}$, obtained by reweighting [33] the data at a simulated T close to $T_{\rm m}$. The shaded band is centered around the expected value and shows statistical 1σ errors. The double-headed arrow indicates ΔE of the two-state fit. The dashed line shows $F(E)$ for the square-well fit.

are again quite well described by a first-order two-state model; the χ^2/dof for the fit is 20.5 and would be 1.7 if the errors were 1 %. Our fitted value of ΔE is $16.1 \pm 0.9\,{\rm kcal/mol}$ for $F_{\rm S}$, which may be compared to the result $\Delta E = 12 \pm 2\,{\rm kcal/mol}$ obtained by a two-state fit of infrared (IR) spectroscopy data [14]. As in the β-hairpin analysis, $T_{\rm m}$ is determined from the specific heat maximum (data not shown). For $F_{\rm S}$, we obtain $T_{\rm m} = 310\,{\rm K}$, which may be compared to the values $T_{\rm m} = 303$, $308\,{\rm K}$ and $T_{\rm m} = 334\,{\rm K}$ obtained by circular dichroism (CD) [13, 15] and IR spectroscopy [14], respectively. Let us stress that $T_{\rm m}$ for $F_{\rm S}$ is a prediction of the model; the energy scale of the model is set using $T_{\rm m}$ for the β-hairpin and then left unchanged in our study of $F_{\rm S}$.

The two-state fits shown in Figs. 4.2a and 4.3a are based on a first-order expression for the free energies of the two coexisting phases. The fits look good and can be improved by including higher order terms, which may give the impression that the behaviors of these systems can be fully understood in terms of a two-state model. However, the two-state picture is far from perfect. This can be seen from the free-energy profiles $F(E)$ shown in Figs. 4.2b and 4.3b,

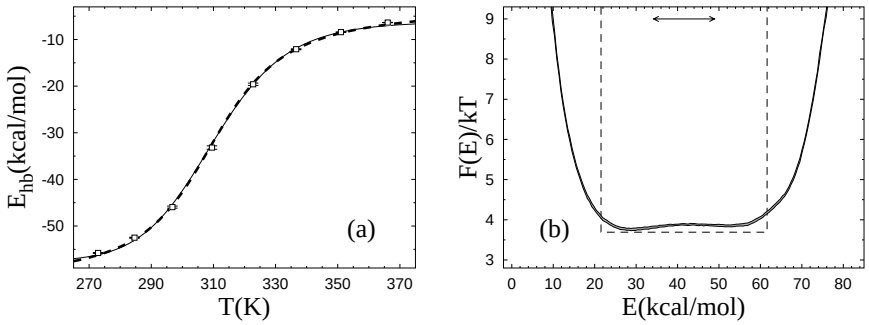


Figure 4.3: Unfolding of the F_S sequence. (a) Temperature dependence of the hydrogen-bond energy E_{hb} (see Eq. 4.3), with the same two types of fit as in Fig. 4.2a (same symbols). (b) Free-energy profile $F(E) = -kT \ln P(E)$ at $T = T_m$. Same symbols as in Fig. 4.2b.

which lack a clear bimodal shape. Clearly, this renders the parameters of a two-state model, such as ΔE, ambiguous. The analysis of these systems therefore shows that the results of a two-state fit must be interpreted with care. Given the actual shapes of $F(E)$, it is instructive to perform an alternative fit of the data in Figs. 4.2a and 4.3a, based on the assumptions that 1) $F(E)$ has the shape of a square well of width ΔE_{sw} at $T = T_m$, and that 2) the observable analyzed varies linearly with E.[1] These square-well fits are shown in Figs. 4.2a and 4.3a, and the corresponding free-energy profiles $F(E)$ (at $T = T_m$) are indicated in Figs. 4.2b and 4.3b. The square-well fits are somewhat better than the two-state fits. However, the fitted curves are strikingly similar, given the large difference between the underlying energy distributions. This shows that it is very hard to draw conclusions about the free-energy profile $F(E)$ from the temperature dependence of a single observable.

From Figs. 4.2b and 4.3b it can also be seen that the energy change ΔE obtained from the two-state fit is considerably smaller than the width of the energy distribution, which indicates that ΔE is smaller than the calorimetric energy change ΔE_{cal}. Scholtz *et al.* [34] determined ΔE_{cal} experimentally for an Ala-based helical peptide with 50 amino acids, and obtained a value of 1.3 kcal/mol per amino acid. This value corresponds to a ΔE_{cal} of 27.3

[1] With these two assumptions, one finds that the average value of an arbitrary observable O at temperature T is given by $O(T) = \int_0^1 (O^u(1-t) + O^f t)\lambda^t dt \Big/ \int_0^1 \lambda^t dt = O^u + (O^f - O^u)(\frac{\lambda}{\lambda-1} - \frac{1}{\ln\lambda})$, where $\lambda = \exp[(1/kT - 1/kT_m)\Delta E_{sw}]$ and O^u and O^f are the values of O at the respective edges of the square well.

kcal/mol for the F_S peptide. Comparing model results for $\Delta E_{\rm cal}$ with experimental data is not straightforward, due to uncertainties about what the relevant baseline subtractions are [35–37]. If we ignore baseline subtractions and simply define $\Delta E_{\rm cal}$ as the energy change between the highest and lowest temperatures studied, we obtain $\Delta E_{\rm cal} = 45.6 \pm 0.1\,\text{kcal/mol}$ for F_S, which is larger than the value of Scholtz *et al.* [34]. To get an idea of how much this result can be affected by a baseline subtraction, a fit of our specific heat data is performed, to a two-state expression supplemented with a baseline linear in T. The fit function is $C_{\rm v} = \Delta E_{\rm cal}(1+K)^{-2}\frac{dK}{dT} + c_0 + c_1(T - T_{\rm m})$, where c_0 and c_1 are baseline parameters and $K = \exp[(1/kT - 1/kT_{\rm m})\Delta E]$. With $\Delta E_{\rm cal}$, ΔE, c_0, c_1 and $T_{\rm m}$ as free parameters, this fit gives $\Delta E_{\rm cal} = 34.0 \pm 1.0\,\text{kcal/mol}$ ($\chi^2/\text{dof} = 5.2$), which is considerably closer to the value of Scholtz *et al.* [34]. It may be worth noting that the corresponding fit without baseline subtraction is much poorer ($\chi^2/\text{dof} \sim 300$). From these calculations, we conclude that the model may overestimate $\Delta E_{\rm cal}$, but it is not evident that the deviation is statistically significant, due to theoretical as well as experimental uncertainties.

The melting behavior of helical peptides is often analyzed using the Zimm-Bragg [38] or Lifson-Roig [39] models, which for large chain lengths are very different from the two-state model considered above. Our results for the F_S peptide are, nevertheless, quite well described by these models too. In fact, a fit of the helix content as a function of temperature to the Lifson-Roig model gives a χ^2/dof similar to that for the two-state fit above.[2] Our fitted Lifson-Roig parameters are $v = 0.016 \pm 0.009$ and $w(T = 273\,\text{K}) = 1.86 \pm 0.25$, corresponding to the Zimm-Bragg parameters $\sigma = 0.0003 \pm 0.0003$ and $s(T = 273\,\text{K}) = 1.83 \pm 0.25$ [40]. In this fit the temperature dependence of w is given by a first-order two-state expression, whereas v is held constant. The energy change ΔE_w has a fitted value of $1.33 \pm 0.17\,\text{kcal/mol}$. The statistical uncertainties on v and σ are large because the chain is small, which makes the dependence on these parameters weak. Thompson *et al.* [15] performed a Zimm-Bragg analysis of CD data for F_S, using the single-sequence approximation. Assuming a value of $\Delta E_s = 1.3\,\text{kcal/mol}$ for the energy change associated with helix propagation, they obtained a σ of 0.0012.

Our kinetic simulations of the two peptides are performed at their respective melting temperatures, $T_{\rm m}$. Starting from equilibrium conformations at $T = 366\,\text{K}$, we study the relaxation of ensemble averages under Monte Carlo dynamics (see Section 4.2.2). The ensemble consists of 1500 independent runs for each peptide. In Fig. 4.4, we show the "time" evolution of $\delta O(t) = O(t) - \langle O \rangle$,

[2] We define helix content in the following way. Each amino acid, except the two at the ends, is labeled h if $-90° < \phi < -30°$ and $-77° < \psi < -17°$, and c otherwise. j consecutive h's form a helical segment of length $j - 2$. The maximal number of amino acids in helical segments is then $N - 4$ for a chain with N amino acids.

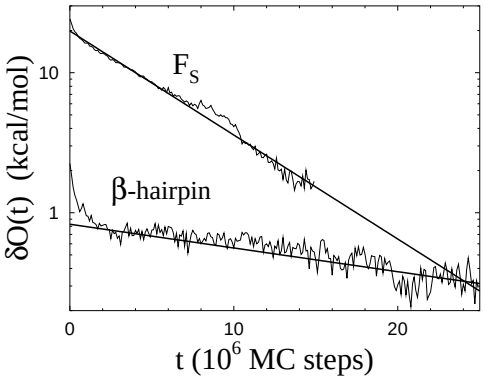


Figure 4.4: Monte Carlo relaxation of ensemble averages at $T = T_{\rm m}$ for the β-hairpin and the F_S peptide. The deviation $\delta O(t)$ from the equilibrium average (see text) is plotted against the number of elementary Monte Carlo steps, t. Straight lines are χ^2 fits of the data to a single exponential. Data for $t > 15{\cdot}10^6$ are omitted for F_S due to large statistical errors.

where $O(t)$ is an ensemble average after t Monte Carlo steps, $\langle O \rangle$ is the corresponding equilibrium average, and the observable O is $E_{\rm hp}$ for the β-hairpin and $E_{\rm hb}$ for F_S (same observables as in the thermodynamic calculations). Ignoring a brief initial period of rapid change, we find that the data, for both peptides, are fully consistent with single-exponential relaxation ($\chi^2/{\rm dof} \sim 1$), although the interval over which the signal $\delta O(t)$ can be followed is small in units of the relaxation time, especially for the β-hairpin. Nevertheless, assuming the single-exponential behavior to be correct, a statistically quite accurate determination of the relaxation times can be obtained. The fitted relaxation time is approximately a factor of 5 larger for the β-hairpin than for F_S. The corresponding factor is around 30 for experimental data [5, 14, 15]. A closer look at the β-hairpin data shows that the hydrophobic cluster and the hydrogen bonds, on average, form nearly simultaneously in our model. This is in agreement with the results of Zhou *et al.* [11], and in disagreement with the folding mechanism of Pande and Rokhsar [9] in which the collapse occurs before the hydrogen bonds form.

The two peptides studied in this paper make unusually clear-cut α- and β-structure, respectively. It is clear that refinements of the interaction potential will be required in order to obtain an equally good description of more general sequences. One interesting refinement would be to make the strength of the hydrogen bonds context-dependent, that is dependent on whether the hydrogen

bond is internal or exposed. This is probably needed in order for the model to capture, for example, the difference between the Ala-based F_S peptide and pure polyalanine. In fact, it has been argued [32, 41] that a major reason why F_S is a strong helix maker is that the Arg side chains shield the backbone from water and thereby make the hydrogen bonds stronger. The hydrogen bonds of a polyalanine helix lack this protection. In our model, the hydrogen bonds are context-independent, which could make polyalanine too helical. Although a direct comparison with experimental data is impossible due to its poor water solubility, simulations of polyalanine with 21 amino acids, A_{21}, seem to confirm this. For A_{21}, we obtain a helix content of about 80 % at $T = 273\,\mathrm{K}$, which is what we find for F_S too. Using a modified version of the Cornell *et al.* force field [42], García and Sanbonmatsu [32] obtained a helix content of 34 % at $T = 275\,\mathrm{K}$ for A_{21}; the unmodified force field was found [32] to give a value similar to ours at this temperature (but very different from ours at higher T). Our estimate that F_S is $\sim 80\,\%$ helical at $T = 273\,\mathrm{K}$ is consistent with experimental data [12, 15].

We also looked at two other helical peptides. The first of these is the Ala-based 16-amino acid peptide $(AEAAK)_3A$, where E is Glu and K is Lys. By CD, Marqusee and Baldwin [43] found this peptide to be $\sim 50\,\%$ helical at $T = 274\,\mathrm{K}$. In our model the corresponding value turns out to be $\sim 70\,\%$. Our last example is the 38–59-fragment of the B domain of staphylococcal protein A (PDB code 1BDD). This is a more general, not Ala-based sequence, containing three hydrophobic Leu. By CD, Bai *et al.* [44] obtained a helix content of $\sim 30\,\%$ at pH 5.2 and $T = 278\,\mathrm{K}$ for this fragment. In our model we obtain a helix content of $\sim 20\,\%$ at this temperature. So, the model predicts helix contents that are in approximate agreement with experimental data for F_S, $(AEAAK)_3A$ as well as the protein A fragment.

4.4 Summary and Outlook

We have developed and explored a protein model that combines an all-atom representation of the amino acid chain with a minimalistic sequence-based potential. The strength of the model is the simplicity of the potential, which at the same time, of course, means that there are many interesting features of real proteins that the model is unable to capture. One advantage of the model is that the calibration of parameters, which any model needs, becomes easier to carry out with fewer parameters to tune.

When calibrating the model, our goal was to ensure that, without resorting to parameter changes, our two sequences made a β-hairpin with the native topol-

ogy and an α-helix, respectively, which was not an easy task. Once this goal had been achieved, our thermodynamic and kinetic measurements were carried out without any further fine-tuning of the potential. Therefore, it is hard to believe that the generally quite good agreement between our thermodynamic results and experimental data is accidental. A more plausible explanation of the agreement is that the thermodynamics of these two sequences indeed are largely governed by backbone hydrogen bonding and hydrophobic collapse forces, as assumed by the model. The requirement that the two sequences make the desired structures is then sufficient to quite accurately determine the strengths of these two terms.

The main results of our calculations can be summarized as follows.

- Our thermodynamic simulations show first of all that the two sequences studied indeed make a β-hairpin with the native topology and an α-helix, respectively. The main reason why the model favors the native topology over the non-native one for the β-hairpin, is that the formation of the hydrophobic cluster is sterically difficult to accomplish in the non-native topology. The melting curves obtained for the two peptides are in reasonable agreement with experimental data, and can to a good approximation be described by a simple two-state model.

- A two-state description of the thermodynamic behavior is, nevertheless, found to be an oversimplification for both peptides, as can be seen from the energy distributions. Given that the systems are small and fluctuations therefore relatively large, this is maybe not surprising. What is striking is how difficult it is to detect these deviations from two-state behavior when studying the temperature dependence of a single observable.

- The results of our Monte Carlo-based kinetic runs at the respective melting temperatures are, for both peptides, consistent with single-exponential relaxation, and the relaxation time is found to be larger for the β-hairpin than for F_S.

Extending these calculations to larger chains will impose new conditions on the interaction potential, and thereby make it possible (and necessary) to refine it. Two interesting refinements would be to make the treatment of charged side chains and side-chain hydrogen bonds less crude and to introduce a mechanism for screening of hydrogen bonds [32,41,45,46]. Computationally, there is room for extending the calculations. In fact, simulating the thermodynamics of a chain with about 20 amino acids, with high statistics, does not take more than a few days on a standard desktop computer, in spite of the detailed geometry of

the model. This gives us hope to be able to look into the free-energy landscape and two-state character of small proteins in a not too distant future.

Acknowledgments

We thank Giorgio Favrin for stimulating discussions and help with computers. This work was in part supported by the Swedish Foundation for Strategic Research and the Swedish Research Council.

References

[1] Kussell, E., Shimada, J. & Shakhnovich, E.I. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 5343–5348.

[2] Clementi, C., García, A.E. & Onuchic, J.N. (2003) *J. Mol. Biol.* **326**, 933–954.

[3] Gō, N. & Abe, H. (1981) *Biopolymers* **20**, 991–1011.

[4] Blanco, F.J., Rivas, G. & Serrano, L. (1994) *Nat. Struct. Biol.* **1**, 584–590.

[5] Muñoz, V., Thompson, P.A., Hofrichter, J. & Eaton, W.A. (1997) *Nature* **390**, 196–199.

[6] Dinner, A.R., Lazaridis, T. & Karplus, M. (1999) *Proc. Natl. Acad. Sci. USA* **96**, 9068–9073.

[7] Zagrovic, B., Sorin, E.J. & Pande, V. (2001) *J. Mol. Biol.* **313**, 151–169.

[8] Roccatano, D., Amadei, A., Di Nola, A. & Berendsen, H.J.C. (1999) *Protein Sci.* **8**, 2130–2143.

[9] Pande, V.S. & Rokhsar, D.S. (1999) *Proc. Natl. Acad. Sci. USA* **96**, 9062–9067.

[10] García, A.E. & Sanbonmatsu, K.Y. (2001) *Proteins Struct. Funct. Genet.* **42**, 345–354.

[11] Zhou, R., Berne, B.J. & Germain, R. (2001) *Proc. Natl. Acad. Sci. USA* **98**, 14931–14936.

[12] Lockhart, D.J. & Kim, P.S. (1992) *Science* **257**, 947–951.

[13] Lockhart, D.J. & Kim, P.S. (1993) *Science* **260**, 198–202.

[14] Williams, S., Causgrove, T.P., Gilmanshin, R., Fang, K.S., Callender, R.H., Woodruff, W.H. & Dyer, R.B. (1996) *Biochemistry* **35**, 691–697.

[15] Thompson, P.A., Eaton, W.A. & Hofrichter, J. (1997) *Biochemistry* **36**, 9200–9210.

[16] Irbäck, A., Sjunnesson, F. & Wallin, S. (2000) *Proc. Natl. Acad. Sci. USA* **97**, 13614–13618.

[17] Irbäck, A., Sjunnesson, F. & Wallin, S. (2001) *J. Biol. Phys.* **27**, 169–179.

[18] Favrin, G., Irbäck, A. & Wallin, S. (2002) *Proteins Struct. Funct. Genet.* **47**, 99–105.

[19] Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112**, 535–542.

[20] Tsai, J., Taylor, R., Chothia, C. & Gerstein, M. (1999) *J. Mol. Biol.* **290**, 253–266.

[21] Miyazawa, S. & Jernigan, R.L. (1996) *J. Mol. Biol.* **256**, 623–644.

[22] Li, H., Tang, C. & Wingreen, N.S. (1997) *Phys. Rev. Lett.* **79**, 765–768.

[23] Lyubartsev, A.P., Martsinovski, A.A., Shevkunov, S.V. & Vorontsov-Velyaminov, P.N. (1992) *J. Chem. Phys.* **96**, 1776–1783.

[24] Marinari, E. & Parisi, G. (1992) *Europhys. Lett.* **19**, 451–458.

[25] Irbäck, A. & Potthast, F. (1995) *J. Chem. Phys.* **103**, 10298–10305.

[26] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. (1953) *J. Chem. Phys.* **21**, 1087–1092.

[27] Lal, M. (1969) *Molec. Phys.* **17**, 57–64.

[28] Favrin, G., Irbäck, A. & Sjunnesson, F. (2001) *J. Chem. Phys.* **114**, 8154–8158.

[29] Press, W.H., Flannery, B.P., Teukolsky, S.A. & Vetterling, W.T. (1992) *Numerical Recipes in C: The Art of Scientific Computing*, (Cambridge University Press, Cambridge).

[30] Gronenborn, A.M., Filpula, D.R., Essig, N.Z., Achari, A., Whitlow, M., Wingfield, P.T. & Clore, G.M. (1991) *Science* **253**, 657–661.

[31] Kobayashi, N., Honda, S., Yoshii, H. & Munekata, E. (2000) *Biochemistry* **39**, 6564–6571.

[32] García, A.E. & Sanbonmatsu, K.Y. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 2782–2787.

[33] Ferrenberg, A.M. & Swendsen R.H. (1988) *Phys. Rev. Lett.* **61**, 2635–2638, and erratum (1989) **63**, 1658, and references given in the erratum.

[34] Scholtz, J.M., Marqusee, S., Baldwin, R.L., York, E.J., Stewart, J.M., Santaro, M. & Bolen, D.W. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 2854–2858.

[35] Zhou, Y., Hall, C.K. & Karplus, M. (1999) *Protein Sci.* **8**, 1064–1074.

[36] Chan, H.S. (2000) *Proteins Struct. Funct. Genet.* **40**, 543–571.

[37] Kaya, H. & Chan, H.S. (2000) *Proteins Struct. Funct. Genet.* **40**, 637–661.

[38] Zimm, B.H. & Bragg, J.K. (1959) *J. Chem. Phys.* **31**, 526–535.

[39] Lifson, S. & Roig, A. (1960) *J. Chem. Phys.* **34**, 1963–1974.

[40] Qian, H. & Schellman, J.A. (1992) *J. Phys. Chem.* **96**, 3987–3994.

[41] Vila, J.A., Ripoll, D.R. & Scheraga, H.A. (2000) *Proc. Natl. Acad. Sci. USA* **97**, 13075–13079.

[42] Cornell, W.D, Cieplak, P, Bayly, C.I., Gould, I.R, Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W. & Kollman, P.A. (1995) *J. Am. Chem. Soc.* **117**, 5179–5197.

[43] Marqusee, S. & Baldwin, R.L. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 8898–8902.

[44] Bai, Y., Karimi, A., Dyson, H.J. & Wright, P.E. (1997) *Protein Sci.* **6**, 1449–1457.

[45] Takada, S., Luthey-Schulten, Z. & Wolynes, P.G. (1999) *J. Chem. Phys.* **110**, 11616–11629.

[46] Guo, C., Cheung, M.S., Levine, H. & Kessler, D.A. (2002) *J. Chem. Phys.* **116**, 4353–4365.

Supplemental Material

Bond lengths (in Å)

------ BACKBONE -------		
N,CA	1.46	
CA,C	1.52	
C,N	1.33	
CA,CB	1.53	
C,O	1.23	
N,H	0.98	
CA,HA	1.08	
CA,2HA	1.08	only GLY
--------- VAL ---------		
CB,CGX	1.52	X=1,2
--------- LEU ---------		
CB,CG	1.53	
CG,CDX	1.52	X=1,2
--------- ILE ---------		
CB,CG1	1.53	
CB,CG2	1.53	
CG1,CD1	1.52	
--------- SER ---------		
CB,OG	1.42	
--------- THR ---------		
CB,OG1	1.43	
CB,CG2	1.52	
--------- CYS ---------		
CB,SD	1.81	
--------- MET ---------		
CB,CG	1.52	
CG,SD	1.81	
SD,CE	1.79	
--------- PRO ---------		
CB,CG	1.51	
CG,CD	1.51	
--------- ASP ---------		
CB,CG	1.52	
CG,ODX	1.25	X=1,2
--------- ASN ---------		
CB,CG	1.52	
CG,OD1	1.23	
CG,ND2	1.33	
--------- GLU ---------		
CB,CG	1.52	
CG,CD	1.52	
CD,OEX	1.25	X=1,2
--------- GLN ---------		
CB,CG	1.52	
CG,CD	1.52	
CD,OE1	1.23	
CG,NE2	1.33	
--------- LYS ---------		
CB,CG	1.52	
CG,CD	1.52	
CD,CE	1.52	
CE,NZ	1.49	
--------- ARG ---------		
CB,CG	1.52	
CG,CD	1.52	
CD,NE	1.46	
NE,CZ	1.33	
CZ,NHX	1.33	X=1,2
--------- HIS ---------		
CB,CG	1.52	
CG,ND1	1.35	*)
--------- PHE ---------		
CB,CG	1.52	
CG,CD1	1.39	*)
--------- TYR ---------		
CB,CG	1.51	
CG,CD1	1.39	*)
CZ,OH	1.38	
--------- TRP ---------		
CB,CG	1.50	
CG,CD1	1.39	*)

All bonds between an H and a side-chain atom have length 1.00 Å.
*) Rings are regular pentagons/hexagons.

Bond angles (in degrees)

```
------ BACKBONE -------
N,CA,C            111.0
CA,C,N            116.6
C,N,CA            121.7
N,CA,CB           110.0
CA,C,O            121.7
C,N,H             119.2
N,CA,HA           109.0
N,CA,2HA          109.0  only GLY
--------- ALA ---------
CA,CB,XHB         109.5  X=1,2,3
--------- VAL ---------
CA,CB,CGX         110.7  X=1,2
CA,CB,HB          109.1
CB,CGY,XHGY       109.5  X=1,2,3; Y=1,2
--------- LEU ---------
CA,CB,CG          117.1
CB,CG,CDX         110.1  X=1,2
CA,CB,XHB         107.0  X=1,2
CB,CG,HG          109.3
CG,CDY,XHDY       109.5  X=1,2,3; Y=1,2
--------- ILE ---------
CA,CB,CG1         110.4
CA,CB,CG2         110.4
CB,CG1,CD1        113.6
CA,CB,HB          109.2
CB,CG1,XHG1       108.1  X=1,2
CB,CG2,XHG2       109.5  X=1,2,3
CG1,CD1,XHD1      109.5  X=1,2,3
--------- SER ---------
CA,CB,OG          110.6
CA,CB,XHB         109.1  X=1,2
CB,OG,HG          108.0
--------- THR ---------
CA,CB,OG1         108.6
CA,CB,CG2         111.5
CA,CB,HB          109.3
CB,OG1,HG1        108.0
CB,CG2,XHG2       109.5  X=1,2,3
--------- CYS ---------
CA,CB,SG          113.4
CA,CB,XHB         108.1  X=1,2
CB,SG,HG          108.0
--------- MET ---------
CA,CB,CG          113.5
CB,CG,SD          111.9
CG,SD,CE          100.5
CA,CB,XHB         108.1  X=1,2
CB,CG,XHG         108.7  X=1,2
SD,CE,XHE         109.5  X=1,2,3
--------- PRO ---------
CA,CB,CG          103.3
CB,CG,CD          110.8
CA,CB,XHB         111.6  X=1,2
CB,CG,XHG         109.0  X=1,2
CG,CD,XHD         110.7  X=1,2
--------- ASP ---------
CA,CB,CG          113.2
CB,CG,ODX         118.6  X=1,2
CA,CB,XHB         108.2  X=1,2
--------- ASN ---------
CA,CB,CG          112.6
CB,CG,OD1         120.9
CB,CG,ND2         117.0
CA,CB,XHB         108.4  X=1,2
CG,ND2,XHD2       120.0  X=1,2
--------- GLU ---------
CA,CB,CG          114.1
CB,CG,CD          113.2
CG,CD,OEX         118.5  X=1,2
CA,CB,XHB         108.0  X=1,2
CB,CG,XHG         108.2  X=1,2
--------- GLN ---------
CA,CB,CG          113.7
CB,CG,CD          112.6
CG,CD,OE1         121.0
CG,CD,NE2         116.9
CA,CB,XHB         108.1  X=1,2
CB,CG,XHG         108.4  X=1,2
CD,NE2,XHE2       120.0  X=1,2
```

Bond angles cont.

```
--------- LYS ---------
CA,CB,CG          113.8
CB,CG,CD          111.6
CG,CD,CE          111.6
CD,CE,NZ          111.6
CA,CB,XHB         108.1  X=1,2
CB,CG,XHG         108.8  X=1,2
CG,CD,XHD         108.8  X=1,2
CD,CE,XHE         108.8  X=1,2
CE,NZ,XHZ         109.5  X=1,2,3
--------- ARG ---------
CA,CB,CG          113.7
CB,CG,CD          111.5
CG,CD,NE          111.5
CD,NE,CZ          124.4
NE,CZ,NHX         120.0  X=1,2
CA,CB,XHB         108.1  X=1,2
CB,CG,XHG         108.8  X=1,2
CG,CD,XHD         108.8  X=1,2
CZ,NE,HE          120.0
CZ,NHY,XHHY       120.0  X=1,2; Y=1,2
--------- HIS ---------
CA,CB,CG          113.2
CB,CG,ND1         126.0
CA,CB,XHB         108.2  X=1,2
--------- PHE ---------
CA,CB,CG          113.7
CB,CG,CD1         120.0
CA,CB,XHB         108.1  X=1,2
--------- TYR ---------
CA,CB,CG          113.6
CB,CG,CD1         120.0
CE1,CZ,OH         120.0
CA,CB,XHB         108.1  X=1,2
CZ,OH,HH          108.0
--------- TRP ---------
CA,CB,CG          113.8
CB,CG,CD1         126.0
CA,CB,XHB         108.1  X=1,2
```

The rings of HIS, PHE, TYR and TRP are regular pentagons/hexagons with hydrogens pointing in the radial direction.

Torsion angles (in degrees)

```
------ BACKBONE -------
CA,C,N,CA               180.0
C,N,CA,C - C,N,CA,CB    120.9
C,N,CA,C - C,N,CA,HA   -118.7
C,N,CA,C - C,N,CA,2HA   118.7  only GLY
```

For side-chain branch points, we assume exact 2-fold or 3-fold torsional symmetry. The rings of PRO, HIS, PHE, TYR and TRP as well as the atom group NE, CZ, NH1 and NH2 of ARG are planar.

Number of side-chain DOFs (χ_i)

```
GLY               0
ALA               1
VAL               3
LEU               4
ILE               4
SER               2
THR               3
CYS               2
MET               4
PRO               0
ASP               2
ASN               3
GLU               3
GLN               4
LYS               5
ARG               4
HIS               2
PHE               2
TYR               3
TRP               2
```

Two-State Folding over a Weak Free-Energy Barrier

Paper V

Two-State Folding over a Weak Free-Energy Barrier

Giorgio Favrin, Anders Irbäck,
Björn Samuelsson and Stefan Wallin

Complex Systems Division, Department of Theoretical Physics
Lund University, Sölvegatan 14A, SE-223 62 Lund, Sweden
http://www.thep.lu.se/complex/

Submitted to *Biophys. J.*

Abstract:
We present a Monte Carlo study of a model protein with 54 amino acids that folds directly to its native three-helix-bundle state without forming any well-defined intermediate state. The free-energy barrier separating the native and unfolded states of this protein is found to be weak, even at the folding temperature. Nevertheless, we find that melting curves to a good approximation can be described in terms of a simple two-state system, and that the relaxation behavior is close to single exponential. The motion along individual reaction coordinates is roughly diffusive on timescales beyond the reconfiguration time for an individual helix. A simple estimate based on diffusion in a square-well potential predicts the relaxation time within a factor of two.

5.1 Introduction

In a landmark paper in 1991, Jackson and Fersht [1] demonstrated that chymotrypsin inhibitor 2 folds without significantly populating any meta-stable intermediate state. Since then, it has become clear that this protein is far from unique; the same behavior has been observed for many small single-domain proteins [2]. It is tempting to interpret the apparent two-state behavior of these proteins in terms of a simple free-energy landscape with two minima separated by a single barrier, where the minima represent the native and unfolded states, respectively. If the barrier is high, this picture provides an explanation of why the folding kinetics are single exponential, and why the folding thermodynamics show two-state character.

However, it is well-known that the free-energy barrier, ΔF, is not high for all these proteins. In fact, assuming the folding time τ_{f} to be given by $\tau_{\mathrm{f}} = \tau_0 \exp(\Delta F/kT)$ with $\tau_0 \sim 1\,\mu s$ [3], it is easy to find examples of proteins with ΔF values of a few kT [2] (k is Boltzmann's constant and T the temperature). It should also be mentioned that Garcia-Mira *et al.* [4] recently found a protein that appears to fold without crossing any free-energy barrier.

Suppose the native and unfolded states coexist at the folding temperature and that there is no well-defined intermediate state, but that a clear free-energy barrier is missing. What type of folding behavior should one then expect? In particular, would such a protein, due to the lack of a clear free-energy barrier, show easily detectable deviations from two-state thermodynamics and single-exponential kinetics? Here we investigate this question based on Monte Carlo simulations of a designed three-helix-bundle protein [5–7].

Our study consists of three parts. First, we investigate whether or not melting curves for this model protein show two-state character. Second, we ask whether the relaxation behavior is single exponential or not, based on ensemble kinetics at the folding temperature. Third, inspired by energy-landscape theory (for a recent review, see Refs. [8, 9]), we try to interpret the folding dynamics of this system in terms of simple diffusive motion in a low-dimensional free-energy landscape.

5.2 Model and Methods

5.2.1 The Model

Simulating atomic models for protein folding remains a challenge, although progress is currently being made in this area [10–16]. Here, for computational efficiency, we consider a reduced model with 5–6 atoms per amino acid [5], in which the side chains are replaced by large C_β atoms. Using this model, we study a designed three-helix-bundle protein with 54 amino acids.

The model has the Ramachandran torsion angles ϕ_i, ψ_i as its degrees of freedom, and is sequence-based with three amino acid types: hydrophobic (H), polar (P) and glycine (G). The sequence studied consists of three identical H/P segments with 16 amino acids each (PPHPPHHPPHPPHHPP), separated by two short GGG segments [17, 18]. The H/P segment is such that it can make an α-helix with all the hydrophobic amino acids on the same side.

The interaction potential

$$E = E_{\rm loc} + E_{\rm ev} + E_{\rm hb} + E_{\rm hp} \tag{5.1}$$

is composed of four terms. The local potential $E_{\rm loc}$ has a standard form with threefold symmetry,

$$E_{\rm loc} = \frac{\epsilon_\phi}{2}\sum_i (1+\cos 3\phi_i) + \frac{\epsilon_\psi}{2}\sum_i (1+\cos 3\psi_i)\,. \tag{5.2}$$

The excluded-volume term $E_{\rm ev}$ is given by a hard-sphere potential of the form

$$E_{\rm ev} = \epsilon_{\rm ev}\sum_{i<j}{}' \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}, \tag{5.3}$$

where the sum runs over all possible atom pairs except those consisting of two hydrophobic C_β. The parameter σ_{ij} is given by $\sigma_{ij} = \sigma_i + \sigma_j + \Delta\sigma_{ij}$, where $\Delta\sigma_{ij} = 0.625\,$Å for $C_\beta C'$, $C_\beta N$ and $C_\beta O$ pairs that are connected by a sequence of three covalent bonds, and $\Delta\sigma_{ij} = 0\,$Å otherwise. The introduction of the parameter $\Delta\sigma_{ij}$ can be thought of as a change of the local potential.

The hydrogen-bond term $E_{\rm hb}$ has the form

$$E_{\rm hb} = \epsilon_{\rm hb}\sum_{ij} u(r_{ij})v(\alpha_{ij},\beta_{ij})\,, \tag{5.4}$$

where the functions $u(r)$ and $v(\alpha,\beta)$ are given by

$$u(r) \;=\; 5\left(\frac{\sigma_{\rm hb}}{r}\right)^{12} - 6\left(\frac{\sigma_{\rm hb}}{r}\right)^{10} \tag{5.5}$$

$$v(\alpha,\beta) \quad = \quad \begin{cases} \cos^2\alpha\cos^2\beta & \alpha,\beta > 90^\circ \\ 0 & \text{otherwise} \end{cases} \tag{5.6}$$

The sum in Eq. 5.4 runs over all possible HO pairs, and r_{ij} denotes the HO distance, α_{ij} the NHO angle, and β_{ij} the HOC′ angle. The last term of the potential, the hydrophobicity term $E_{\rm hp}$, is given by

$$E_{\rm hp} = \epsilon_{\rm hp} \sum_{i<j} \left[\left(\frac{\sigma_{\rm hp}}{r_{ij}} \right)^{12} - 2 \left(\frac{\sigma_{\rm hp}}{r_{ij}} \right)^{6} \right], \tag{5.7}$$

where the sum runs over all pairs of hydrophobic C_β.

To speed up the calculations, a cutoff radius r_c is used, which is taken to be 4.5 Å for $E_{\rm ev}$ and $E_{\rm hb}$, and 8 Å for $E_{\rm hp}$. Numerical values of all energy and geometry parameters can be found elsewhere [5].

The thermodynamic behavior of this three-helix-bundle protein has been studied before [5, 6]. These studies demonstrated that this model protein has the following properties:

- It does form a stable three-helix bundle, except for a twofold topological degeneracy. These two topologically distinct states both contain three right-handed helices. They differ in how the helices are arranged. If we let the first two helices form a U, then the third helix is in front of the U in one case (FU), and behind the U in the other case (BU). The reason that the model is unable to discriminate between these two states is that their contact maps are effectively very similar [19].

- It makes more stable helices than the corresponding one- and two-helix sequences, which is in accord with the experimental fact that tertiary interactions generally are needed for secondary structure to become stable.

- It undergoes a first-order-like folding transition directly from an expanded state to the three-helix-bundle state, without any detectable intermediate state. At the folding temperature $T_{\rm f}$, there is a pronounced peak in the specific heat.

Here we analyze the folding dynamics of this protein in more detail, through an extended study of both thermodynamics and kinetics.

As a measure of structural similarity with the native state, we monitor a parameter Q that we call nativeness. To calculate Q, we use representative conformations for the FU and BU topologies, respectively, obtained by energy

minimization. For a given conformation, we compute the root-mean-square deviations δ_{FU} and δ_{BU} from these two representative conformations (calculated over all backbone atoms). The nativeness Q is then obtained as

$$Q = \max\left[\exp\left(-\delta_{\mathrm{FU}}^2/(10\text{Å})^2\right), \exp\left(-\delta_{\mathrm{BU}}^2/(10\text{Å})^2\right)\right], \tag{5.8}$$

which makes Q a dimensionless number between 0 and 1.

Energies are quoted in units of kT_{f}, with the folding temperature T_{f} defined as the specific heat maximum. In the dimensionless energy unit used in our previous study [5], this temperature is given by $kT_{\mathrm{f}} = 0.6585 \pm 0.0006$.

5.2.2 Monte Carlo Methods

To simulate the thermodynamic behavior of this model, we use simulated tempering [20–22], in which the temperature is a dynamic variable. This method is chosen in order to speed up the calculations at low temperatures. Our simulations are started from random configurations. The temperatures studied range from $0.95\,T_{\mathrm{f}}$ to $1.37\,T_{\mathrm{f}}$.

The temperature update is a standard Metropolis step. In conformation space we use two different elementary moves: first, the pivot move in which a single torsion angle is turned; and second, a semi-local method [23] that works with seven or eight adjacent torsion angles, which are turned in a coordinated manner. The non-local pivot move is included in our calculations in order to accelerate the evolution of the system at high temperatures.

Our kinetic simulations are also Monte Carlo-based, and only meant to mimic the time evolution of the system in a qualitative sense. They differ from our thermodynamic simulations in two ways: first, the temperature is held constant; and second, the non-local pivot update is not used, but only the semi-local method [23]. This restriction is needed in order to avoid large unphysical deformations of the chain.

Statistical errors on thermodynamic results are obtained by jackknife analysis [24] of results from ten or more independent runs, each containing several folding/unfolding events. All errors quoted are 1σ errors. The fits of data discussed below are carried out by using a Levenberg-Marquardt procedure [25].

5.2.3 Analysis

Melting curves for proteins are often described in terms of a two-state picture. In the two-state approximation, the average of a quantity X at temperature T is given by

$$X(T) = \frac{X_{\mathrm{u}} + X_{\mathrm{n}} K(T)}{1 + K(T)}\,, \tag{5.9}$$

where $K(T) = P_{\mathrm{n}}(T)/P_{\mathrm{u}}(T)$, $P_{\mathrm{n}}(T)$ and $P_{\mathrm{u}}(T)$ being the populations of the native and unfolded states, respectively. Likewise, X_{n} and X_{u} denote the respective values of X in the native and unfolded states. The effective equilibrium constant $K(T)$ is to leading order given by $K(T) = \exp[(1/kT - 1/kT_{\mathrm{m}})\Delta E]$, where T_{m} is the midpoint temperature and ΔE the energy difference between the two states. With this $K(T)$, a fit to Eq. 5.9 has four parameters: ΔE, T_{m} and the two baselines X_{u} and X_{n}.

A simple but powerful method for quantitative analysis of the folding dynamics is obtained by assuming the motion along different reaction coordinates to be diffusive [26, 27]. The folding process is then modeled as one-dimensional Brownian motion in an external potential given by the free energy $F(r) = -kT \ln P_{\mathrm{eq}}(r)$, where $P_{\mathrm{eq}}(r)$ denotes the equilibrium distribution of r. Thus, it is assumed that the probability distribution of r at time t, $P(r,t)$, obeys Smoluchowski's diffusion equation

$$\frac{\partial P(r,t)}{\partial t} = \frac{\partial}{\partial r}\left[D(r)\left(\frac{\partial P(r,t)}{\partial r} + \frac{P(r,t)}{kT}\frac{\partial F(r)}{\partial r}\right)\right]\,, \tag{5.10}$$

where $D(r)$ is the diffusion coefficient.

This picture is not expected to hold on short timescales, due to the projection onto a single coordinate r, but may still be useful provided that the diffusive behavior sets in on a timescale that is small compared to the relaxation time. By estimating $D(r)$ and $F(r)$, it is then possible to predict the relaxation time from Eq. 5.10. Such an analysis has been successfully carried through for a lattice protein [27].

The relaxation behavior predicted by Eq. 5.10 is well understood when $F(r)$ has the shape of a double well with a clear barrier. In this situation, the relaxation is single exponential with a rate constant given by Kramers' well-known result [28]. However, this result cannot be applied to our model, in which the free-energy barrier is small or absent, depending on which reaction coordinate is used. Therefore, we perform a detailed study of Eq. 5.10 for some relevant choices of $D(r)$ and $F(r)$, using analytical as well as numerical methods.

	$\Delta E/kT_\mathrm{f}$	$T_\mathrm{m}/T_\mathrm{f}$
E	40.1 ± 3.3	1.0050 ± 0.0020
E_hb	41.0 ± 2.6	1.0024 ± 0.0017
E_hp	45.4 ± 3.3	1.0056 ± 0.0017
R_g	45.7 ± 3.8	1.0099 ± 0.0018
Q	53.6 ± 2.1	0.9989 ± 0.0008

Table 5.1: Parameters ΔE and T_m obtained by fitting results from our thermodynamic simulations to the two-state expression in Eq. 5.9. This is done individually for each of the quantities in the first column; the energy E, the hydrogen-bond energy E_hb, the hydrophobicity energy E_hp, the radius of gyration R_g (calculated over all backbone atoms), and the nativeness Q (see Eq. 5.8). The fits are performed using seven data points in the temperature interval $0.95\,T_\mathrm{f} \leq T \leq 1.11\,T_\mathrm{f}$.

5.3 Results

5.3.1 Thermodynamics

In our thermodynamic analysis, we study the five different quantities listed in Table 5.1. The first question we ask is to what extent the temperature dependence of these quantities can be described in terms of a first-order two-state system (see Eq. 5.9).

Fits of our data to this equation show that the simple two-state picture is not perfect ($\chi^2/\mathrm{dof} \sim 10$), but this can be detected only because the statistical errors are very small at high temperatures ($< 0.1\%$). In fact, if we assign artifical statistical errors of 1% to our data points, an error size that is not uncommon for experimental data, then the fits become perfect with a χ^2/dof close to unity. Fig. 5.1 shows the temperature dependence of the hydrogen-bond energy E_hb and the radius of gyration R_g, along with our two-state fits.

Table 5.1 gives a summary of our two-state fits. In particular, we see that the fitted values of both the energy change ΔE and the midpoint temperature T_m are similar for the different quantities. It is also worth noting that the T_m values fall close to the folding temperature T_f, defined as the maximum of the specific heat. The difference between the highest and lowest values of T_m is less than 1%. There is a somewhat larger spread in ΔE, but this parameter has a larger statistical error.

So, the melting curves show two-state character, and the fitted parameters ΔE

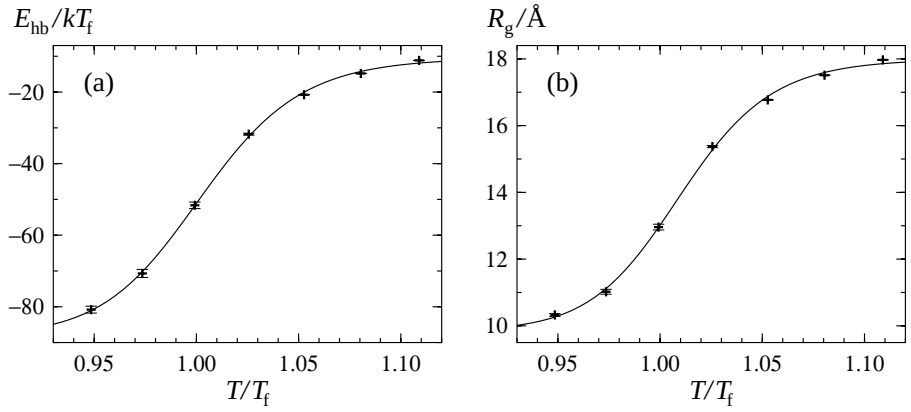


Figure 5.1: Temperature dependence of (a) the hydrogen-bond energy $E_{\rm hb}$ and (b) the radius of gyration $R_{\rm g}$. The lines are fits to Eq. 5.9.

and $T_{\rm m}$ are similar for different quantities. From this it may be tempting to conclude that the thermodynamic behavior of this protein can be fully understood in terms of a two-state system. The two-state picture is, nevertheless, an oversimplification, as can be seen from the shapes of the free-energy profiles $F(E)$ and $F(Q)$. Fig. 5.2 shows these profiles at $T = T_{\rm f}$. First of all, these profiles show that the native and unfolded states coexist at $T = T_{\rm f}$, so the folding transition is first-order-like. However, there is no clear free-energy barrier separating the two states; $F(Q)$ exhibits a very weak barrier, $< 1\,kT$, whereas $F(E)$ shows no barrier at all. In fact, $F(E)$ has the shape of a square well rather than a double well.

5.3.2 Kinetics

Our kinetic study is performed at $T = T_{\rm f}$. Using Monte Carlo dynamics (see Model and Methods), we study the relaxation of ensemble averages of various quantities. For this purpose, we performed a set of 3000 folding simulations, starting from equilibrium conformations at temperature $T_0 \approx 1.06\,T_{\rm f}$. At this temperature, the chain is extended and has a relatively low secondary-structure content (see Fig. 5.1).

In the absence of a clear free-energy barrier (see Fig. 5.2), it is not obvious whether or not the relaxation should be single exponential. To get an idea of what to expect for a system like this, we consider the relaxation of the energy

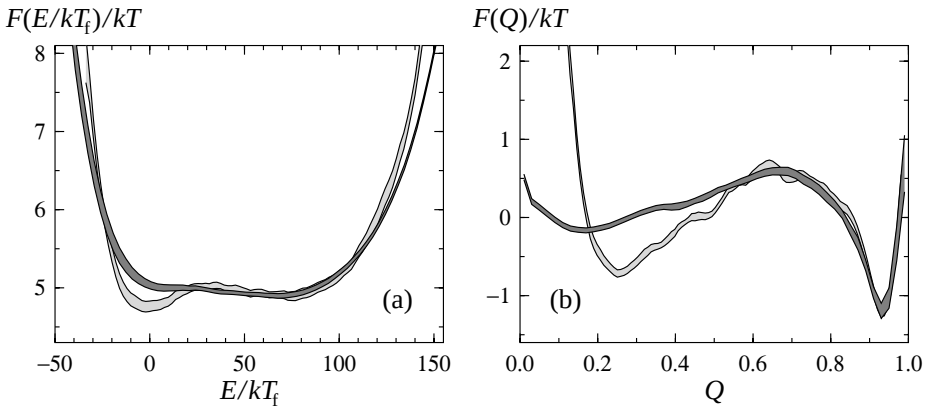


Figure 5.2: Free-energy profiles at $T = T_\mathrm{f}$ for (a) the energy E and (b) the nativeness Q (dark bands). The light-grey bands show free energies F_b for block averages (see Eq. 5.12), using a block size of $\tau_\mathrm{b} = 10^6$ MC steps. Each band is centered around the expected value and shows statistical 1σ errors.

E in a potential $F(E)$ that has the form of a perfect square well at $T = T_\mathrm{f}$. For this idealized $F(E)$, it is possible to solve Eq. 5.10 analytically for relaxation at an arbitrary temperature T. This solution is given in Appendix A, for the initial condition that $P(E, t = 0)$ is the equilibrium distribution at temperature T_0. Using this result, the deviation from single-exponential behavior can be mapped out as a function of T_0 and T, as is illustrated in Fig. 5.3. The size of the deviation depends on both T_0 and T, but is found to be small for a wide range of T_0, T values. This clearly demonstrates that the existence of a free-energy barrier is not a prerequisite to observe single-exponential relaxation.

Let us now turn to the results of our simulations. Fig. 5.4 shows the relaxation of the average energy E and the average nativeness Q in Monte Carlo (MC) time. In both cases, the large-time data can be fitted to a single exponential, which gives relaxation times of $\tau \approx 1.7 \cdot 10^7$ and $\tau \approx 1.8 \cdot 10^7$ for E and Q, respectively, in units of elementary MC steps. The corresponding fits for the radius of gyration and the hydrogen-bond energy (data not shown) give relaxation times of $\tau \approx 2.1 \cdot 10^7$ and $\tau \approx 1.8 \cdot 10^7$, respectively. The fit for the radius of gyration has a larger uncertainty than the others, because the data points have larger errors in this case.

The differences between our four fitted τ values are small and most probably due to limited statistics for the large-time behavior. Averaging over the four

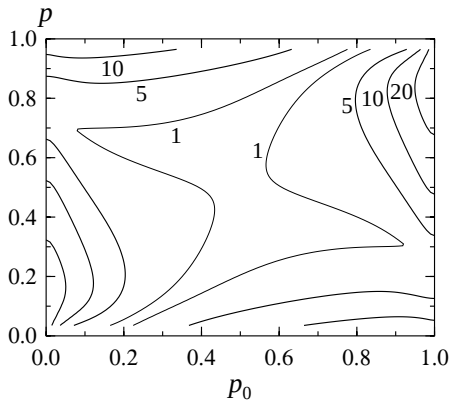


Figure 5.3: Level diagram showing the deviation (in %) from a single exponential for diffusion in energy in a square well, based on the exact solution in Appendix A. The system relaxes at temperature T, starting from the equilibrium distribution at temperature T_0. p is defined as $p = (\langle E\rangle - E_\mathrm{n})/\Delta E_\mathrm{sw}$, where $\langle E\rangle$ is the average energy at temperature T, and E_n and ΔE_sw denote the lower edge and the width, respectively, of the square well. p can be viewed as a measure of the unfolded population at temperature T, and is 0.5 if $T = T_\mathrm{f}$. p_0 is the the corresponding quantity at temperature T_0. As a measure of the deviation from a single exponential, we take $\delta_\mathrm{max}/\delta E(t_0)$, where δ_max is the maximum deviation from a fitted exponential and $\delta E(t_0) = E(t_0) - \langle E\rangle$, $E(t_0)$ being the mean at the smallest time included in the fit, t_0. Data at times shorter than 1% of the relaxation time were excluded from the fit.

different variables, we obtain a relaxation time of $\tau \approx 1.8\cdot 10^7$ MC steps for this protein. The fact that the relaxation times for the hydrogen-bond energy and the radius of gyration are approximately the same shows that helix formation and chain collapse proceed in parallel for this protein. This finding is in nice agreement with recent experimental results for small helical proteins [29].

For Q, it is necessary to go to very short times in order to see any significant deviation from a single exponential (see Fig. 5.4). For E, we find that the single-exponential behavior sets in at roughly $\tau/3$, which means that the deviation from this behavior is larger than in the analytical calculation above. On the other hand, for comparisons with experimental data, we expect the behavior of Q to be more relevant than that of E. The simulations confirm that the relaxation can be approximately single exponential even if there is no clear free-energy barrier.

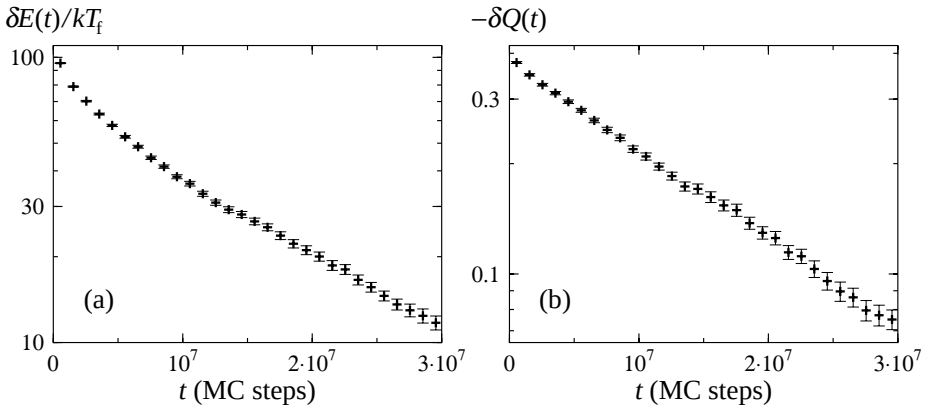


Figure 5.4: Relaxation behavior of the three-helix-bundle protein at the folding temperature T_f, starting from the equilibrium ensemble at $T_0 \approx 1.06T_\mathrm{f}$. (a) $\delta E(t) = E(t) - \langle E\rangle$ against simulation time t, where $E(t)$ is the average E after t MC steps (3000 runs) and $\langle E\rangle$ denotes the equilibrium average (at T_f). (b) Same plot for the nativeness Q.

To translate the relaxation time for this protein into physical units, we compare with the reconfiguration time for the corresponding one-helix segment. To that end, we performed a kinetic simulation of this 16-amino acid segment at the same temperature, $T = T_\mathrm{f}$. This temperature is above the midpoint temperature for the one-helix segment, which is $0.95\,T_\mathrm{f}$ [5]. So, the isolated one-helix segment is unstable at $T = T_\mathrm{f}$, but makes frequent visits to helical states with low hydrogen-bond energy, E_hb. To obtain the reconfiguration time, we fitted the large-time behavior of the autocorrelation function for E_hb,

$$C_\mathrm{hb}(t) = \langle E_\mathrm{hb}(t)E_\mathrm{hb}(0)\rangle - \langle E_\mathrm{hb}(0)\rangle^2\,, \tag{5.11}$$

to an exponential. The exponential autocorrelation time, which can be viewed as a reconfiguration time, turned out to be $\tau_\mathrm{h} \approx 1.0 \cdot 10^6$ MC steps. This is roughly a factor 20 shorter than the relaxation time τ for the full three-helix bundle. Assuming the reconfiguration time for an individual helix to be $\sim 0.2\,\mu$s [30, 31], we obtain relaxation and folding times of $\sim 4\,\mu$s and $\sim 8\,\mu$s, respectively, for the three-helix bundle. This is fast but not inconceivable for a small helical protein [2]. In fact, the B domain of staphylococcal protein A is a three-helix-bundle protein that has been found to fold in $< 10\,\mu$s, at 37°C [32].

5.3.3 Relaxation-Time Predictions

We now turn to the question of whether the observed relaxation time can be predicted based on the diffusion equation, Eq. 5.10. For that purpose, we need to know not only the free energy $F(r)$, but also the diffusion coefficient $D(r)$. Socci *et al.* [27] successfully performed this analysis for a lattice protein that exhibited a relatively clear free-energy barrier. Their estimate of $D(r)$ involved an autocorrelation time for the unfolded state. The absence of a clear barrier separating the native and unfolded states makes it necessary to take a different approach in our case.

The one-dimensional diffusion picture is not expected to hold on short timescales, but only after coarse-graining in time. A computationally convenient way to implement this coarse-graining in time is to study block averages $b(t)$ defined by

$$b(t) = \frac{1}{\tau_\mathrm{b}} \sum_{t \le s < t+\tau_\mathrm{b}} r(s) \qquad\qquad t = 0, \tau_\mathrm{b}, 2\tau_\mathrm{b}, \ldots \tag{5.12}$$

where τ_b is the block size and r is the reaction coordinate considered. The diffusion coefficient can then be estimated using $D_\mathrm{b}(r) = \langle(\delta b)^2\rangle / 2\tau_\mathrm{b}$, where the numerator is the mean-square difference between two consecutive block averages, given that the first of them has the value r.

In our calculations, we use a block size of $\tau_\mathrm{b} = 10^6$ MC step, corresponding to the reconfiguration time τ_h for an individual helix. We do not expect the dynamics to be diffusive on timescales shorter than this, due to steric traps that can occur in the formation of a helix. In order for the dynamics to be diffusive, the timescale should be such that the system can escape from these traps.

Using this block size, we first make rough estimates of the relaxation times for E and Q based on the result in Appendix A for a square-well potential and a constant diffusion coefficient. These estimates are given by $\tau_{\mathrm{pred},0} = \Delta r_\mathrm{sw}^2 / D_\mathrm{b} \pi^2$, where Δr_sw is the width of the potential and D_b is the average diffusion coefficient. [1] Our estimates of Δr_sw and D_b can be found in Table 5.2, along with the resulting predictions $\tau_{\mathrm{pred},0}$. We find that these simple predictions agree with the observed relaxation times τ within a factor of two.

We also did the same calculation for smaller block sizes, $\tau_\mathrm{b} = 10^0, 10^1, \ldots, 10^5$ MC steps. This gave $\tau_{\mathrm{pred},0}$ values smaller or much smaller than the observed τ, signaling non-diffusive dynamics. This confirms that for $b(t)$ to show diffu-

[1] Eq. 5.15 in Appendix A can be applied to other observables than E. The predicted relaxation time $\tau_{\mathrm{pred},0}$ is given by τ_1.

	$\Delta r_{\rm sw}$	$D_{\rm b}$	$\tau_{\rm pred,0}$	$\tau_{\rm pred}$	τ
E:	$140kT_{\rm f}$	$(9.3 \pm 0.2) \cdot 10^{-5}(kT_{\rm f})^2$	$2.1 \cdot 10^7$	$1.9 \cdot 10^7$	$1.7 \cdot 10^7$
Q:	1.0	$(1.00 \pm 0.02) \cdot 10^{-8}$	$1.0 \cdot 10^7$	$0.8 \cdot 10^7$	$1.8 \cdot 10^7$

Table 5.2: The predictions $\tau_{\rm pred,0}$ and $\tau_{\rm pred}$ (see text) along with the observed relaxation time τ, as obtained from the data in Fig. 5.4, for the energy E and the nativeness Q. $\Delta r_{\rm sw}$ is the width of the square-well potential and $D_{\rm b}$ is the average diffusion coefficient.

sive dynamics, $\tau_{\rm b}$ should not be smaller than the reconfiguration time for an individual helix.

Having seen the quite good results obtained by this simple calculation, we now turn to a more detailed analysis, illustrated in Fig. 5.5a. The block size is the same as before, $\tau_{\rm b} = 10^6$ MC steps, but the space dependence of the diffusion coefficient $D_{\rm b}(r)$ is now taken into account, and the potential, $F_{\rm b}(r)$, reflects the actual distribution of block averages. This potential is similar but not identical to that for the unblocked variables, as can be seen from Fig. 5.2. Fig. 5.5b shows the diffusion coefficient $D_{\rm b}(E)$, which is largest at intermediate values between the native and unfolded states. The behavior of $D_{\rm b}(Q)$ (not shown) is the same in this respect. Hence, there is no sign of a kinetic bottleneck to folding for this protein.

Given $D_{\rm b}(r)$ and $F_{\rm b}(r)$, we solve Eq. 5.10 for $P(r,t)$ by using the finite-difference scheme in Appendix B. The initial distribution $P(r, t = 0)$ is taken to be the same as in the kinetic simulations. We find that the mean of $P(r,t)$ shows single-exponential relaxation to a good approximation. An exponential fit of these data gives us a new prediction, $\tau_{\rm pred}$, for the relaxation time.

From Table 5.2 it can be seen that the predictions obtained through this more elaborate calculation, $\tau_{\rm pred}$, are not better than the previous ones, $\tau_{\rm pred,0}$. This shows that the underlying diffusion picture is not perfect, although the relaxation time can be predicted within a factor of two.

It might be possible to obtain better predictions by simply increasing the block size. However, for the calculation to be useful, the block size must remain small compared to the relaxation time. A more interesting possibility is to refine the simple diffusion picture used here, in which, in particular, non-Markovian effects are ignored. Such effects may indeed affect folding times [9, 33].

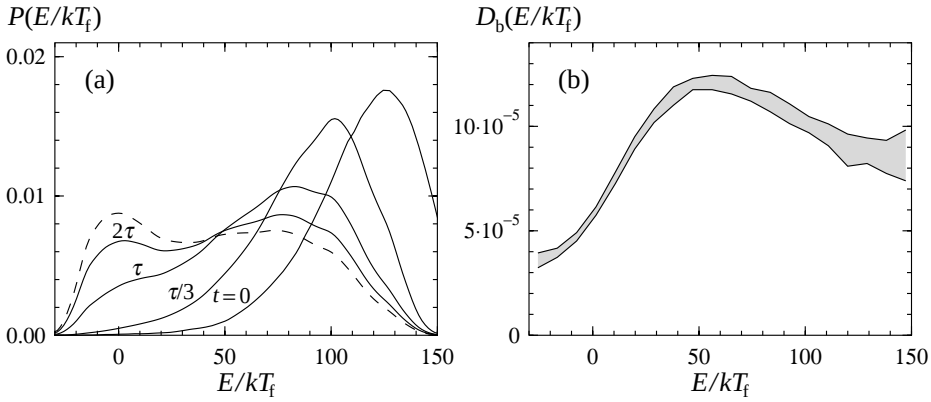


Figure 5.5: (a) Numerical solution of Eq. 5.10 with the energy as reaction coordinate. The distribution $P(E,t)$ is shown for $t = 0$, $\tau/3$, τ and 2τ (full lines), where τ is the relaxation time. The dashed line is the equilibrium distribution. The diffusion coefficient $D_{\rm b}(E)$ and the potential $F_{\rm b}(E)$ (light-gray band in Fig. 5.2a) were both determined from numerical simulations, using a block size of $\tau_{\rm b} = 10^6$ MC steps (see Eq. 5.12). (b) The space dependence of the diffusion coefficient $D_{\rm b}(E)$. The band is centered around the expected value and shows the statistical 1σ error.

5.4 Summary and Discussion

We have analyzed the thermodynamics and kinetics of a designed three-helix-bundle protein, based on Monte Carlo calculations. We found that this model protein shows two-state behavior, in the sense that melting curves to a good approximation can be described by a simple two-state system and that the relaxation behavior is close to single exponential. A simple two-state picture is, nevertheless, an oversimplification, as the free-energy barrier separating the native and unfolded states is weak ($\lesssim 1kT$). The weakness of the barrier implies that a fitted two-state parameter such as ΔE has no clear physical meaning, despite that the two-state fits look good.

Reduced [18, 34–37] and all-atom [10, 11, 14, 38–40] models for small helical proteins have been studied by many other groups. However, we are not aware of any other model that exhibits a first-order-like folding transition without resorting to the so-called Gō prescription [41]; our model is sequence-based.

Using an extended version of this model that includes all atoms, we recently

found similar results for two peptides, an α-helix and a β-hairpin [16]. Here the calculated melting curves could be directly compared with experimental data, and a reasonable quantitative agreement was found.

The smallness of the free-energy barrier prompted us to perform an analytical study of diffusion in a square-well potential. Here we studied the relaxation behavior at temperature T, starting from the equilibrium distribution at temperature T_0, for arbitrary T and T_0. We found that this system shows a relaxation behavior that is close to single exponential for a wide range of T_0, T values, despite the absence of a free-energy barrier. We also made relaxation-time predictions based on this square-well approximation. Here we took the diffusion coefficient to be constant. It was determined assuming the dynamics to be diffusive on timescales beyond the reconfiguration time for an individual helix. The predictions obtained this way were found to agree within a factor of two with observed relaxation times, as obtained from the kinetic simulations. So, this calculation, based on the two simplifying assumptions that the potential is a square well and that the diffusion coefficient is constant, gave quite good results. A more detailed calculation, in which these two additional assumptions were removed, did not give better results. This shows that the underlying diffusion picture leaves room for improvement.

Our kinetic study focused on the behavior at the folding temperature T_f, where the native and unfolded states, although not separated by a clear barrier, are very different. This makes the folding mechanism transparent. We found that this model protein folds without the formation of any obligatory intermediate state and that helix formation and chain collapse occur in parallel, which is in accord with experimental data by Krantz *et al.* [29]. The difference between the native and unfolded states is much smaller at the lowest temperature we studied, $0.95T_\mathrm{f}$, because the unfolded state is much more native-like here. Mayor *et al.* [42] recently reported experimental results on a three-helix-bundle protein, the engrailed homeodomain [43], including a characterization of its unfolded state. In particular, the unfolded state was found to have a high helix content. This study was performed at a temperature below $0.95T_\mathrm{f}$. In our model, there is a significant decrease in helix content of the unfolded state as the temperature increases from $0.95T_\mathrm{f}$ to T_f. It would be very interesting to see what the unfolded state of this protein looks like near T_f.

It is instructive to compare our results with those of Zhou and Karplus [35], who discussed two folding scenarios for helical proteins, based on a Gō-type C_α model. In their first scenario, folding is fast, without any obligatory intermediate, and helix formation occurs before chain collapse. In the second scenario, folding is slow with an obligatory intermediate on the folding pathway, and helix formation and chain collapse occur simultaneously. The behavior of our

model does not match any of these two scenarios, in spite of a recent statement to the contrary [40]. Our model shows fast folding despite that helix formation and chain collapse cannot be separated.

Acknowledgments

This work was in part supported by the Swedish Foundation for Strategic Research and the Swedish Research Council.

Appendix A: Diffusion in a square well

Here we discuss Eq. 5.10 in the situation when the reaction coordinate r is the energy E, and the potential $F(E)$ is a square well of width $\Delta E_{\rm sw}$ at $T = T_{\rm f}$. This means that the equilibrium distribution is given by $P_{\rm eq}(E) \propto \exp(-\delta\beta E)$ if E is in the square well and $P_{\rm eq}(E) = 0$ otherwise, where $\delta\beta = 1/kT - 1/kT_{\rm f}$. Eq. 5.10 then becomes

$$\frac{\partial P(E,t)}{\partial t} = \frac{\partial}{\partial E}\left[D\left(\frac{\partial P(E,t)}{\partial E} + \delta\beta P(E,t)\right)\right] . \tag{5.13}$$

For simplicity, the diffusion coefficient is assumed to be constant, $D(E) = D$. The initial distribution $P(E, t = 0)$ is taken to be the equilibrium distribution at some temperature T_0, and we put $\delta\beta_0 = 1/kT_0 - 1/kT_{\rm f}$.

By separation of variables, it is possible to solve Eq. 5.13 with this initial condition analytically for arbitrary values of the initial and final temperatures T_0 and T, respectively. In particular, this solution gives us the relaxation behavior of the average energy. The average energy at time t, $E(t)$, can be expressed in the form

$$E(t) = \langle E\rangle + \sum_{k=1}^{\infty} A_k \mathrm{e}^{-t/\tau_k} , \tag{5.14}$$

where $\langle E\rangle$ denotes the equilibrium average at temperature T. A straightforward calculation shows that the decay constants in this equation are given by

$$1/\tau_k = \frac{D}{\Delta E_{\rm sw}^2}\left(\pi^2 k^2 + \tfrac{1}{4}\delta\beta^2 \Delta E_{\rm sw}^2\right) \tag{5.15}$$

and the expansion coefficients by

$$A_k = B_k \Delta E_{\rm sw} \frac{\pi^2 k^2 \left(\delta\beta - \delta\beta_0\right)\Delta E_{\rm sw}}{\left(\pi^2 k^2 + (\delta\beta_0 - \tfrac{1}{2}\delta\beta)^2 \Delta E_{\rm sw}^2\right)\left(\pi^2 k^2 + \tfrac{1}{4}\delta\beta^2\Delta E_{\rm sw}^2\right)^2} , \tag{5.16}$$

where

$$B_k = \frac{4\delta\beta_0 \Delta E_{\rm sw}}{\sinh \tfrac{1}{2}\delta\beta_0 \Delta E_{\rm sw}} \times \begin{cases} \cosh\left(\tfrac{1}{2}(\delta\beta_0 - \tfrac{1}{2}\delta\beta)\Delta E_{\rm sw}\right)\cosh \tfrac{1}{4}\delta\beta\Delta E_{\rm sw} & \text{if } k \text{ odd} \\ \sinh\left(\tfrac{1}{2}(\delta\beta_0 - \tfrac{1}{2}\delta\beta)\Delta E_{\rm sw}\right)\sinh \tfrac{1}{4}\delta\beta\Delta E_{\rm sw} & \text{if } k \text{ even} \end{cases} \tag{5.17}$$

Finally, the equilibrium average is

$$\langle E\rangle = \frac{E_{\rm n} + E_{\rm u}}{2} + \frac{1}{\delta\beta} - \frac{\Delta E_{\rm sw}}{2}\coth \tfrac{1}{2}\delta\beta\Delta E_{\rm sw} , \tag{5.18}$$

where $E_{\rm n}$ and $E_{\rm u}$ are the lower and upper edges of the square well, respectively.

It is instructive to consider the behavior of this solution when $|\delta\beta - \delta\beta_0| \ll 1/\Delta E_{\rm sw}$. The expression for the expansion coefficients can then be simplified to

$$A_k \approx B_k \Delta E_{\rm sw} \frac{\pi^2 k^2 (\delta\beta - \delta\beta_0)\Delta E_{\rm sw}}{\left(\pi^2 k^2 + \frac{1}{4}\delta\beta^2 \Delta E_{\rm sw}^2\right)^3} \tag{5.19}$$

with

$$B_k \approx \frac{4\delta\beta_0 \Delta E_{\rm sw}}{\sinh \frac{1}{2}\delta\beta_0 \Delta E_{\rm sw}} \times \begin{cases} \cosh^2 \frac{1}{4}\delta\beta \Delta E_{\rm sw} & \text{if } k \text{ odd} \\ \sinh^2 \frac{1}{4}\delta\beta \Delta E_{\rm sw} & \text{if } k \text{ even} \end{cases} \tag{5.20}$$

Note that A_k scales as k^2 if $k \ll \frac{1}{2\pi}|\delta\beta|\Delta E_{\rm sw}$, and as $1/k^4$ if $k \gg \frac{1}{2\pi}|\delta\beta|\Delta E_{\rm sw}$. Note also that the last factor in B_k suppresses A_k for even k if T is close to $T_{\rm f}$. From these two facts it follows that $|A_1|$ is much larger than the other $|A_k|$ if T is near $T_{\rm f}$. This makes the deviation from a single exponential small.

Appendix B: Numerical solution of the diffusion equation

To solve Eq. 5.10 numerically for arbitrary $D(r)$ and $F(r)$, we choose a finite-difference scheme of Crank-Nicolson type with good stability properties. To obtain this scheme we first discretize r. Put $r_j = j\Delta r$, $D_j = D(r_j)$ and $F_j = F(r_j)$, and let $\mathbf{p}(t)$ be the vector with components $p_j(t) = P(r_j, t)$. Approximating the RHS of Eq. 5.10 with suitable finite differences, we obtain

$$\frac{d\mathbf{p}}{dt} = \mathbf{A}\mathbf{p}(t)\,, \tag{5.21}$$

where $\mathbf{A}$ is a tridiagonal matrix given by

$$\begin{aligned} (\mathbf{A}\mathbf{p}(t))_j &= \frac{1}{\Delta r^2}\left[D_{j+1/2}(p_{j+1}(t) - p_j(t)) - D_{j-1/2}(p_j(t) - p_{j-1}(t))\right] \\ &+ \frac{1}{4kT\Delta r^2}\left[D_{j+1}p_{j+1}(t)(F_{j+2} - F_j) - D_{j-1}p_{j-1}(t)(F_j - F_{j-2})\right] \end{aligned} \tag{5.22}$$

Let now $\mathbf{p}^n = \mathbf{p}(t_n)$, where $t_n = n\Delta t$. By applying the trapezoidal rule for integration to Eq. 5.21, we obtain

$$\mathbf{p}^{n+1} - \mathbf{p}^n = \frac{\Delta t}{2}\left(\mathbf{A}\mathbf{p}^n + \mathbf{A}\mathbf{p}^{n+1}\right)\,. \tag{5.23}$$

This equation can be used to calculate how $P(r, t)$ evolves with time. It can be readily solved for $\mathbf{p}^{n+1}$ because the matrix $\mathbf{A}$ is tridiagonal.

References

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

[2] Jackson S.E. 1998. How do small single-domain proteins fold? *Fold. Des.* 3:R81-R91.

[3] Hagen S.J., J. Hofrichter, A. Szabo, and W.A. Eaton. 1996. Diffusion-limited contact formation in unfolded cytochrome C: Estimating the maximum rate of protein folding. *Proc. Natl. Acad. Sci. USA* 93:11615-11617.

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

[5] Irbäck A., F. Sjunnesson, and S. Wallin. 2000. Three-helix-bundle protein in a Ramachandran model. *Proc. Natl. Acad. Sci. USA* 97:13614-13618.

[6] Irbäck A., F. Sjunnesson, and S. Wallin. 2001. Hydrogen bonds, hydrophobicity forces and the character of the collapse transition. *J. Biol. Phys.* 27:169-179.

[7] Favrin G., A. Irbäck, and S. Wallin. 2002. Folding of a small helical protein using hydrogen bonds and hydrophobicity forces. *Proteins Struct. Funct. Genet.* 47:99-105.

[8] Plotkin S.S., and J.N. Onuchic. 2002. Understanding protein folding with energy landscape theory. Part I: Basic concepts. *Q. Rev. Biophys.* 35:111-167.

[9] Plotkin S.S., and J.N. Onuchic. 2002. Understanding protein folding with energy landscape theory. Part II: Quantitative aspects. *Q. Rev. Biophys.* 35:205-286.

[10] Kussell E., J. Shimada, and E.I. Shakhnovich. 2002. A structure-based method for derivation of all-atom potentials for protein folding. *Proc. Natl. Acad. Sci. USA* 99:5343-5348.

[11] Shen M.Y., and K.F. Freed. 2002. All-atom fast protein folding simulations: The villin headpiece. *Proteins Struct. Funct. Genet.* 49:439-445.

[12] Zhou R., and B.J. Berne. 2002. Can a continuum solvent model reproduce the free energy landscape of a β-hairpin folding in water? *Proc. Natl. Acad. Sci. USA* 99:12777-12782.

[13] Shea J.-E., J.N. Onuchic, and C.L. Brooks III. 2002. Probing the folding free energy landscape of the src-SH3 protein domain. *Proc. Natl. Acad. Sci. USA* 99:16064-16068.

[14] Zagrovic B., C.D. Snow, M.R. Shirts, and V.S. Pande. 2002. Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. *J. Mol. Biol.* 323:927-937.

[15] Clementi C., A.E. García, and J.N. Onuchic. 2003. Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: all-atom representation study of Protein L. *J. Mol. Biol.* 326:933-954.

[16] Irbäck A., B. Samuelsson, F. Sjunnesson, and S. Wallin. 2003. Thermodynamics of α- and β-structure formation in proteins. Preprint submitted to *Biophys. J.*

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

[18] Takada S., Z. Luthey-Schulten, and P.G. Wolynes. 1999. Folding dynamics with nonadditive forces: A simulation study of a designed helical protein and a random heteropolymer", *J. Chem. Phys.* 110:11616-11628.

[19] Wallin S., J. Farwer, and U. Bastolla. 2003. Testing similarity measures with continuous and discrete protein models. *Proteins Struct. Funct. Genet.* 50:144-157.

[20] Lyubartsev A.P., A.A. Martsinovski, S.V. Shevkunov, and P.N. Vorontsov-Velyaminov. 1992. New approach to Monte Carlo calculation of the free energy: Method of expanded ensembles. *J. Chem. Phys.* 96:1776-1783.

[21] Marinari E., and G. Parisi. 1992. Simulated tempering: A new Monte Carlo scheme. *Europhys. Lett.* 19:451-458.

[22] Irbäck A., and F. Potthast. 1995. Studies of an off-lattice model for protein folding: Sequence dependence and improved sampling at finite temperature. *J. Chem. Phys.* 103:10298-10305.

[23] Favrin G, A. Irbäck, and F. Sjunnesson. 2001. Monte Carlo update for chain molecules: Biased Gaussian steps in torsional space, *J. Chem. Phys.* 114:8154-8158.

[24] Miller R.G. 1974. The jackknife - a review. *Biometrika* 61:1-15.

[25] Press W.H., B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. 1992. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge.

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

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

[28] Kramers H.A. 1940. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica* 7:284-304.

[29] Krantz B.A., A.K. Srivastava, S. Nauli, D. Baker, R.T. Sauer, and T.R. Sosnick. 2002. Understanding protein hydrogen bond formation with kinetic H/D amide isotope effects. *Nat. Struct. Biol.* 9:458-463.

[30] Williams S., T.P. Causgrove, R. Gilmanshin, K.S. Fang, R.H. Callender, W.H. Woodruff, and R.B. Dyer. 1996. Fast events in protein folding: Helix melting and formation in a small peptide. *Biochemistry* 35:691-697.

[31] Thompson P.A., W.A. Eaton, and J. Hofrichter. 1997. Laser temperature jump study of the helix⇌coil kinetics of an alanine peptide interpreted with 'kinetic zipper' model. *Biochemistry* 36:9200-9210.

[32] Myers J.K., and T.G. Oas. 2001. Preorganized secondary structure as an important determinant of fast folding. *Nat. Struct. Biol.* 8:552-558.

[33] Plotkin S.S., and P.G. Wolynes. 1998. Non-Markovian configurational diffusion and reaction coordinates for protein folding. *Phys. Rev. Lett.* 80:5015-5018.

[34] Kolinski A., W. Galazka, and J. Skolnick. 1998. Monte Carlo studies of the thermodynamics and kinetics of reduced protein models: Application to small helical, β, and α/β proteins. *J. Chem. Phys.* 108:2608-2617.

[35] Zhou Y., and M. Karplus. 1999. Interpreting the folding kinetics of helical proteins. *Nature* 401:400-403.

[36] Shea J.-E., J.N. Onuchic, and C.L. Brooks III. 1999. Exploring the origins of topological frustration: Design of a minimally frustrated model of fragment B of protein A. *Proc. Natl. Acad. Sci. USA* 96:12512-12517.

[37] Berriz G.F., and E.I. Shakhnovich. 2001. Characterization of the folding kinetics of a three-helix bundle protein via a minimalist Langevin model. *J. Mol. Biol.* 310:673-685.

[38] Guo Z., C.L. Brooks III, and E.M. Boczko. 1997. Exploring the folding free energy surface of a three-helix bundle protein. *Proc. Natl. Acad. Sci. USA* 94:10161-10166.

[39] Duan Y., and P.A. Kollman. 1998. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* 282:740-744.

[40] Linhananta A., and Y. Zhou. 2003. The role of sidechain packing and native contact interactions in folding: Discrete molecular dynamics folding simulations of an all-atom Gō model of fragment B of Staphylococcal protein A. *J. Chem. Phys.* 117:8983-8995.

[41] Gō N., and H. Abe. 1981. Noninteracting local-structure model of folding and unfolding transition in globular proteins. *Biopolymers* 20:991-1011.

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

[43] Clarke N.D., C.R. Kissinger, J. Desjarlais, G.L. Gilliland, and C.O. Pabo. 1994. Structural studies of the engrailed homeodomain. *Protein Sci.* 3:1779-1787.

Sequence-Based Study of Two Related Proteins with Different Folding Behaviors

Paper VI

Sequence-Based Study of Two Related Proteins with Different Folding Behaviors

Giorgio Favrin, Anders Irbäck and Stefan Wallin

Complex Systems Division, Department of Theoretical Physics
Lund University, Sölvegatan 14A, SE-223 62 Lund, Sweden
http://www.thep.lu.se/complex/

Submitted to *Proteins Struct. Funct. Genet.*

Abstract:
Z_{SPA-1} is an engineered protein that binds to its parent, the three-helix-bundle Z domain of staphylococcal protein A. Uncomplexed Z_{SPA-1} shows a reduced helix content and a melting behavior that is less cooperative, compared with the wild-type Z domain. Here we show that the difference in folding behavior between these two sequences can be partly understood in terms of a minimalistic model, in which folding is driven by backbone hydrogen bonding and effective hydrophobic attraction.

6.1 Introduction

It is becoming increasingly clear that unstructured proteins play an important biological role [1, 2]. In many cases, such proteins adopt a specific structure upon binding to their biological targets. Recently, it was demonstrated that the *in vitro* evolved Z_{SPA-1} protein [3] exhibits coupled folding and binding [4, 5].

Z_{SPA-1} is derived from the Z domain of staphylococcal protein A, a 58-amino acid, well characterized [6] three-helix-bundle protein. Z_{SPA-1} was engineered [3] by randomizing 13 amino acid positions and selecting for binding to the Z domain itself. Subsequently, the structure of the Z:Z_{SPA-1} complex was determined both in solution [4] and by crystallography [5]. In the complex, both Z_{SPA-1} and the Z domain adopt structures similar to the solution structure of the Z domain. However, in solution, Z_{SPA-1} does not behave as the Z domain; Wahlberg *et al.* [4] found that uncomplexed Z_{SPA-1} lacks a well-defined structure, and that its melting behavior is less cooperative than that of the wild-type sequence.

The Z domain is a close analog of the B domain of protein A, a chain that is known to show two-state folding without any meta-stable intermediate state [7, 8]. The folding behavior of the B domain has also been studied theoretically by many different groups, including ourselves, using both all-atom [9–12] and reduced [13–17] models. In many cases, it was possible to fold this chain, but to achieve that most models rely on the so-called Gō prescription [18]. Our model [17] folds this chain in a cooperative, approximately two-state manner without resorting to this prescription. Our model is thus entirely sequence-based. This makes it possible for us to study both Z_{SPA-1} and the wild-type Z domain and compare their behaviors, using one and the same model.

The purpose of this note is twofold. First, we check whether our model can explain the difference in melting behavior between Z_{SPA-1} and the wild-type sequence. Second, using this model, we study the structural properties of Z_{SPA-1}.

6.2 Materials and Methods

6.2.1 Model

The model we study [17] is an extension of a model with three amino acids [19–21] to a five-letter alphabet. The five amino acid types are hydrophobic (Hyd), polar (Pol), Ala, Pro and Gly. Hyd, Pol and Ala share the same geometric representation but differ in hydrophobicity. Pro and Gly have their own geometric representations.

The Hyd, Pol and Ala representation contains six atoms. The three backbone atoms N, C_α and C' and the H and O atoms of the peptide unit are all included. The H and O atoms are used to define hydrogen bonds. The sixth atom is a large C_β that represents the side chain. The representation of Gly is the same except that C_β is missing. The representation of Pro differs from that of Hyd, Pol and Ala in that the H atom is replaced by a side-chain atom, C_δ, and that the Ramachandran angle ψ is held fixed at $-65°$.

The degrees of freedom of our model are the Ramachandran torsion angles ϕ and ψ, with the exception that ψ is held fixed for Pro. All bond lengths, bond angles and peptide torsion angles (180°) are held fixed.

The interaction potential

$$E = E_{\rm loc} + E_{\rm ev} + E_{\rm hb} + E_{\rm hp} \tag{6.1}$$

is composed of four terms. The first term is a local ϕ, ψ potential. The other three terms represent excluded volume, backbone hydrogen bonds and effective hydrophobic attraction, respectively (no explicit water). For simplicity, the hydrophobicity potential is taken to be pairwise additive. Only Hyd-Hyd and Hyd-Ala C_β pairs experience this type of interaction. In particular, this means that Ala is intermediate in hydrophobicity between Hyd and Pol. The amino acids in the Hyd class are Val, Leu, Ile, Phe, Trp and Met, whereas those in the Pol class are Arg, Asn, Asp, Cys, Gln, Glu, His, Lys, Ser, Thr and Tyr. A complete description of the model, including numerical values of all the parameters, can be found in our earlier study [17].

Following previous calculations for the B domain of protein A [9–17], we consider the 9–54-amino acid fragments of $Z_{\rm SPA-1}$ and the wild-type Z domain (corresponding to the 10–55-amino acid fragment of the B domain), rather than the full sequences. $Z_{\rm SPA-1}$ differs from the wild-type sequence at 13 positions, all of which are found in the section 9–35. Table 6.1 shows this part of the sequences.

Z_{SPA-1}	QQN	AFY	EIL	HLP	NLN	EEQ	RNA	FIQ	SLK
wild-type	LSV	AGR	EIV	TLP	NLN	DPQ	KKA	FIF	SLW

Table 6.1: Amino acids 9 to 35 for Z_{SPA-1} and the wild-type Z domain.

6.2.2 Numerical Methods

To simulate the thermodynamic behavior of this model, we use simulated tempering [22–24], in which the temperature is a dynamic variable. This method is chosen in order to speed up the calculations at low temperature. The temperature update is a standard Metropolis step. In conformation space we use two different elementary moves: first, the pivot move in which a single torsion angle is turned; and second, a semi-local method [25] that works with seven or eight adjacent torsion angles, which are turned in a coordinated manner. The non-local pivot move is included in our calculations in order to accelerate the evolution of the system at high temperature, whereas the semi-local method improves the performance at low temperature.

Our simulations are started from random configurations. All statistical errors quoted are 1σ errors obtained by analyzing data from eight independent runs.

The temperatures studied range from $0.87\,T_{\mathrm{m}}$ to $1.43\,T_{\mathrm{m}}$, T_{m} being the melting temperature for the wild-type Z domain. The experimental value of this temperature is $T_{\mathrm{m}} = 75°\mathrm{C}$ [4]. Hence, the lowest and highest temperatures studied correspond to 31°C and 225°C, respectively. In the dimensionless energy unit used in our earlier study [17], T_{m} is given by $kT_{\mathrm{m}} = 0.630 \pm 0.001$ (k is Boltzmann's constant). In the model we define T_{m} as the maximum of the specific heat.

6.3 Results and Discussion

Using the model described in the previous section, we study the 9–54-amino acid fragments of Z_{SPA-1} and the wild-type Z domain. Both calculations are carried out using exactly the same parameters as in our earlier study of the B domain [17].

The most striking conclusions reached by Wahlberg *et al.* [4] in their study of the solution behavior of Z_{SPA-1} concern the helix content and the absence of a well-defined structure. By CD, they found the helix content to be smaller for Z_{SPA-1} than for the wild-type sequence, the mean residue ellipticity for

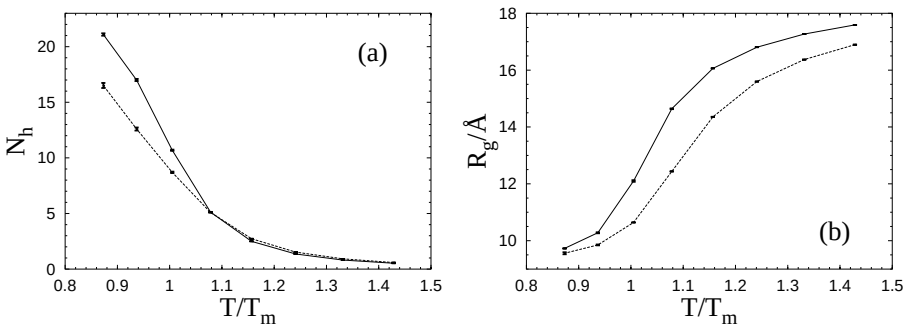


Figure 6.1: Helix formation and chain collapse for the Z_{SPA-1} sequence (dashed line) and the wild-type sequence (full line). (a) The number of helical amino acids, N_h, against temperature. (b) The radius of gyration (calculated over all backbone atoms), R_g, against temperature. T_m denotes the melting temperature for the wild-type sequence. The NMR structure for the wild-type Z domain has $N_h = 29$ and $R_g = 9.0$ Å.

Z_{SPA-1} being 60% of that for the wild-type sequence. Furthermore, the helix formation was found to set in at a lower temperature and to be less cooperative for Z_{SPA-1} than for the wild-type Z domain. Figure 6.1a shows our results for the helix content as a function of temperature for the two sequences.[1] In agreement with the experimental results, we find that Z_{SPA-1} has a lower helix content, and that the helix formation is shifted toward lower temperature for this sequence. Figure 6.1b shows the temperature dependence of our data for the radius of gyration. We find that Z_{SPA-1} is more compact than the wild-type sequence. A comparison with Figure 6.1a shows that chain collapse occurs before helix formation for Z_{SPA-1}. The results in Figures 6.1a and 6.1b demonstrate in particular that the melting behavior is less cooperative for Z_{SPA-1} than for the wild-type sequence. This conclusion is supported by our data for the specific heat (not shown). The peak in the specific heat turns out to be more pronounced for the wild-type sequence than for Z_{SPA-1}.

That the model predicts Z_{SPA-1} to be more compact than the wild-type sequence is not surprising, given that the number of hydrophobic amino acids is larger for Z_{SPA-1} (14) than for the wild-type sequence (11). In addition,

[1] We define helix content in the following way. Each amino acid, except the two at the ends, is labeled h if $-90° < \phi < -30°$ and $-77° < \psi < -17°$, and c otherwise. The two amino acids at the ends are labeled c. An amino acid is said to be helical if both the amino acid itself and its nearest neighbors are labeled h. The total number of helical amino acids is denoted by N_h. The maximum value of N_h is $N - 4$ for a chain with N amino acids.

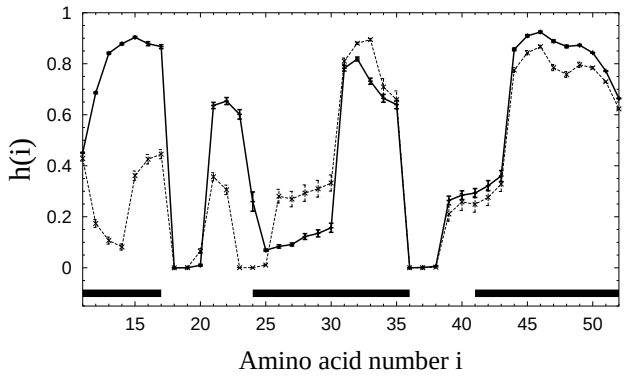


Figure 6.2: Helix content along the chain, $h(i)$, for the Z_{SPA-1} sequence (dashed line) and the wild-type sequence (full line) at $T = 0.87\,T_m$, where T_m is the melting temperature for the wild-type sequence. $h(i)$ denotes the probability that amino acid i is helical (for the definition of helical, see footnote). Thick horizontal lines indicate helical parts of the NMR structure [6] for the wild-type Z domain.

Z_{SPA-1} has one more Pro than the wild-type sequence, which does change the local properties of the chain and could affect the overall size, too. It should be pointed out that the effect of a Pro on the overall size may be poorly described by the model, because the prolyl peptide bond is held fixed in the model (trans).

The reduced total helix content of Z_{SPA-1} shows that this sequence does not make a perfect three-helix bundle, but does not tell in what way the structure differs from a three-helix bundle. It could be that one of the three helices is missing and that the other two are still there, but it could also be that the disorder is more uniform along the chain, so that all three helices are present but partially disordered. The NMR analysis of Z_{SPA-1} [4] does not exclude any of these two possibilities. Figure 6.2 shows how the helix content varies along the chains in our model. The helix profile for the wild-type Z domain can be compared with experimental data [6]. The comparison shows that helix II is somewhat distorted in the model, whereas our data for helices I and III match the experimental data well. These two helices, I and III, respond very differently to the mutations leading to Z_{SPA-1}; helix III remains stable whereas helix I becomes unstable. Although helix III is free from mutations, this helix could, of course, have become unstable, too. Our results suggest that this is not the case; the stability of helix III is very similar for the two sequences. Helix

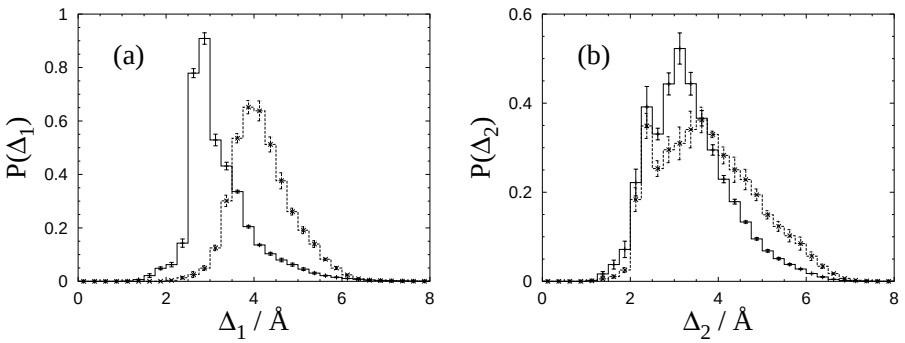


Figure 6.3: RMSD distributions for the Z_{SPA-1} sequence (dashed line) and the wild-type sequence (full line). (a) The distribution of Δ_1 (amino acids 9–31). (b) The distribution of Δ_2 (amino acids 32-54). Both Δ_1 and Δ_2 are backbone RMSDs. The temperature is the same as in Figure 6.2.

I contains seven mutations (see Table 6.1), which change the hydrophobicity pattern and make it less helical. Furthermore, this part of the chain is more flexible in Z_{SPA-1} due to the mutation Phe13Gly. Our model predicts that helix I, as a result these of two changes, is unstable in Z_{SPA-1}.

To further investigate the structural effects of the mutations, we monitor root-mean-square deviations (RMSDs) from the NMR structure [6] for the wild-type Z domain (PDB code 2SPZ, model 1). For a given conformation, we compute two RMSD values, Δ_1 and Δ_2, for the first and second halves of the chain, respectively. The two parts of the chain are separately superimposed on the NMR structure. Figure 6.3 shows the probability distributions of Δ_1 and Δ_2 for Z_{SPA-1} and the wild-type sequence. In line with the results in Figure 6.2, we find that the two Δ_2 distributions are similar, although the distribution for Z_{SPA-1} is slightly wider. By contrast, the two Δ_1 distributions differ markedly; the mean is significantly higher for Z_{SPA-1} than for the wild-type sequence. This clearly shows that in our model the main difference between the two sequences lies in the behavior of the first half of the chain.

6.4 Conclusion

Using one and the same model, with unchanged parameters, we have compared the thermodynamic behaviors of an engineered sequence and its parent. In spite of its minimalistic potential, the model is able to capture important effects of the mutations; in line with experimental data, we find that the mutated sequence, Z_{SPA-1}, shows a reduced helix content and a melting behavior that is less cooperative than for the wild-type sequence. The model predicts that chain collapse occurs before helix formation sets in for Z_{SPA-1}. It also predicts that the main structural difference between Z_{SPA-1} and the wild-type sequence lies in the behavior of helix I, which is less stable in Z_{SPA-1}. To decide whether or not these predictions are correct requires further experimental data.

Acknowledgments

We thank Torleif Härd for a helpful discussion. This work was in part supported by the Swedish Research Council.

References

[1] Wright PE, Dyson HJ. Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *J. Mol. Biol.* 1999; 293: 321–331.

[2] Dyson HJ, Wright PE. Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.* 2002; 12: 54–60.

[3] Eklund M, Axelsson L, Uhlén M, Nygren P-Å. Anti-idiotypic protein domains selected from protein A-based affibody libraries. *Proteins Struct. Funct. Genet.* 2002; 48: 454–462.

[4] Wahlberg E, Lendel C, Helgstrand M, Allard P, Dincbas-Renqvist V, Hedqvist A, Berglund H, Nygren P-Å, Härd T. An affibody in complex with a target protein: Structure and coupled folding. *Proc. Natl. Acad. Sci. USA* 2003; 100: 3185–3190.

[5] Högbom M, Eklund M, Nygren P-Å, Nordlund P. Structural basis for recognition by an *in vitro* evolved affibody. *Proc. Natl. Acad. Sci. USA* 2003; 100: 3191–3196.

[6] Tashiro M, Tejero R, Zimmerman DE, Celda B, Nilsson B, Montelione GT. High-resolution solution NMR structure of the Z domain of staphylococcal protein A. *J. Mol. Biol.* 1997; 272: 573–590.

[7] Bai Y, Karimi A, Dyson HJ, Wright PE. Absence of a stable intermediate on the folding pathway of protein A. *Protein Sci.* 1997; 6: 1449–1457.

[8] Myers JK, Oas TG. Preorganized secondary structure as an important determinant of fast protein folding. *Nat. Struct. Biol.* 2001; 8: 552–558.

[9] Boczko EM, Brooks CL III. First-principles calculation of the folding free energy of a three-helix bundle protein. *Science* 1995; 269: 393–396.

[10] Guo Z, Brooks CL III, Boczko EM. Exploring the folding free energy surface of a three-helix bundle protein. *Proc. Natl. Acad. Sci. USA* 1997; 94: 10161–10166.

[11] Kussell EL, Shimada J, Shakhnovich EI. A structure-based method for derivation of all-atom potentials for protein folding. *Proc. Natl. Acad. Sci. USA* 2002; 99: 5343–5348.

[12] Linhananta A, Zhou Y. The role of sidechain packing and native contact interactions in folding: Discrete molecular dynamics folding simulations of an all-atom Gō model of fragment B of staphylococcal protein A. *J. Chem. Phys.* 2002; 117: 8983–8995.

[13] Kolinski A, Galazka W, Skolnick J. Monte Carlo studies of the thermodynamics and kinetics of reduced protein models: Application to small helical, β, and α/β proteins. *J. Chem. Phys.* 1998; 108: 2608–2617.

[14] Zhou Y, Karplus M. Interpreting the folding kinetics of helical proteins. *Nature* 1999; 401: 400–403.

[15] Shea J-E, Onuchic JN, Brooks CL III. Exploring the origins of topological frustration: Design of a minimally frustrated model of fragment B of protein A. *Proc. Natl. Acad. Sci. USA* 1999; 96: 12512–12517.

[16] Berriz GF, Shakhnovich EI. Characterization of the folding kinetics of a three-helix bundle protein via a minimalist Langevin model. *J. Mol. Biol.* 2001; 310: 673–685.

[17] Favrin G, Irbäck A, Wallin S. Folding of a small helical protein using hydrogen bonds and hydrophobicity forces. *Proteins Struct. Funct. Genet.* 2002; 47: 99–105.

[18] Gō N, Abe H. Noninteracting local-structure model of folding and unfolding transition in globular proteins. *Biopolymers* 1981; 20: 991–1011.

[19] Irbäck A, Sjunnesson F, Wallin S. Three-helix-bundle protein in a Ramachandran model. *Proc. Natl. Acad. Sci. USA* 2000; 97: 13614–13618.

[20] Irbäck A, Sjunnesson F, Wallin S. Hydrogen bonds, hydrophobicity forces and the character of the collapse transition. *J. Biol. Phys.* 2001; 27: 169–179.

[21] Favrin G, Irbäck A, Samulesson B, Wallin S. Two-state folding over a weak free-energy barrier. Lund Preprint LU TP 03-07.

[22] Lyubartsev AP, Martsinovski AA, Shevkunov SV, Vorontsov-Velyaminov PV. New approach to Monte Carlo calculation of the free energy: Method of expanded ensembles. *J. Chem. Phys.* 1992; 96: 1776–1783.

[23] Marinari E, Parisi G. Simulated tempering: A new Monte Carlo scheme. *Europhys. Lett.* 1992; 19: 451–458.

[24] Irbäck A, Potthast F. Studies of an off-lattice model for protein folding: Sequence dependence and improved sampling at finite temperature. *J. Chem. Phys.* 1995; 103: 10298–10305.

[25] Favrin G, Irbäck A, Sjunnesson F. Monte Carlo update for chain molecules: Biased Gaussian steps in torsional space. *J. Chem. Phys.* 2001; 114: 8154–8158.