Statistical physics of protein folding and aggregation

Giorgio Favrin

Department of Theoretical Physics
Lund University, Sweden

Thesis for the degree of Doctor of Philosophy

Thesis Advisor: Anders Irbäck

Faculty Opponent: Ulrich H. E. Hansmann
Department of Physics
Michigan Technological University

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 19th of May 2004, 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 2004 **CODEN:**
Author(s) Giorgio Favrin	**Sponsoring organization**

Title and subtitle

Statistical Physics of Protein Folding and Aggregation

Abstract

The mechanisms of protein folding and aggregation are investigated by computer simulations of all-atom and reduced models with sequence-based potentials. A quasi local Monte Carlo update is developed in order to efficiently sample proteins in the folded phase. A small helical protein, the B-domain of staphylococcal protein A, is studied using a reduced model. In the thermodynamically favoured topology, energy minimisation leads to a conformation whose root mean square deviation form the experimental structure is 1.8Å. We also study the thermodynamics and kinetics of small fast folding proteins without a clear free-energy barrier between the folded and unfolded states. Analytical calculations using a square well-potential unable us to predict the relaxation time within a factor of two. Finally using an all atom model, we study the aggregation properties of a 7-amino acid fragment of Alzheimer's amyloid beta peptide. We find that the system of three and six such fragments form aggregated structures with a high content of antiparallel beta-sheet structure, which is in line with experimental data

Key words

Protein folding, protein dynamics, two-state, three-helix bundle, Amyloid aggregation

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language
English

ISSN and key title

ISBN
91-628-6073-9

Recipient's notes

Number of pages
124

Price

Security classification

Distribution by (name and address)

Giorgio Favrin, Dept. of Theoretical Physics,
Sölveg. 14A, SE-223 62 Lund

A Rita

This thesis is based on the following publications:

I Giorgio Favrin, Anders Irbck and Fredrik Sjunnesson,
Monte Carlo Update for Chain Molecules: Biased Gaussian Steps in Torsional Space
Journal of Chemical Physics **114**, 8154-8158 (2001).

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 Giorgio Favrin, Anders Irbäck, Björn Samuelsson and Stefan Wallin,
Two-State Folding over a Weak Free-Energy Barrier,
Biophysical Journal. **85** 1457–1465 (2003).

IV Giorgio Favrin, Anders Irbäck and Stefan Wallin,
Sequence-Based Study of Two Related Proteins with Different Folding Behaviors,
Proteins: Structure,Function, and Genetics. **54**, 8–12. (2004).

V Giorgio Favrin, Anders Irbäck and Sandipan Mohanty,
Oligomerization of Amyloid Aβ_{16-22} Peptides Using Hydrogen Bonds and Hydrophobicity Forces ,
LU TP 04-18

Contents

Introduction

"The glory of him, who in motion sets all things,
seeps through the universe, and shines,
in one part more, and less elsewhere."
Dante Alighieri

This thesis is divided into two main parts, the introduction and the collection of papers I have been working on during my doctoral studies. The goal is to simulate the folding and aggregation mechanisms of proteins using physical potentials.

The diversity of functions displayed by proteins in living organisms suggests how important the folding process is. A few examples of functions carried out by proteins are: the expression of the genetic information by binding to specific sequences of nucleic acids, the transmission of information between cells or organs, and the antibody function in the immune system.

Proteins are polypeptide chains, which in their natural environment typically fold to a particular three-dimensional structure. This structure provides the unique environments and orientations of the functional groups that give proteins their many special properties. When a protein does not fold into the correct structure its function is compromised. Furthermore, "misfolding" of proteins have been linked to pathological conditions such as the Alzheimer's and the Parkinson's diseases.

In the papers included in this thesis we study the mechanisms by which proteins fold and aggregate. The introduction is structured as follows: First, I will introduce some basic concepts, then I will give an overview of various theoretical approaches to the folding problem, and finally I will briefly describe the

computational methods used in our studies.

Basics

Like the DNA, proteins are chains built from a relatively small collection of molecules that are repeated in different patterns (sequences). These molecules are called amino acids and only twenty types of them occur in nature [1]. The three-dimensional structure [2] that a certain protein assumes under biological conditions (*i.e. in vivo*), is encoded in its sequence and is commonly referred to as its *native* state.

The protein folding problem can then be seen as a complex decryption problem, a little like translating hieroglyphs. For proteins the encrypted text is the sequence and the *meaning* of it is the three-dimensional structure. A big improvement in the understanding of hieroglyphs came about with the discovery of the Rosetta Stone, which allowed linguists to compare the same text written in different languages, one of which was known. A similar methodology is used in the so called *homology* method [3]. Given the fact that the structure is resistant to mutation (roughly 30% of sequence similarity indicates structural identity), it is possible to check the similarity of a sequence with others whose folded structure is known, to predict its native configuration. Although this method leads to very good results in terms of structure prediction, when a homologous sequence can be found, it does not help to understand the *mechanism* by which the protein folds.

Understanding this mechanism is the goal of *ab initio* methods [4]. One starts from a model that, even if it does not match the full complexity of the real protein architecture, captures a core aspect of the physical protein folding problem. Within this model proteins are then simulated starting from the unfolded state until it reaches the native one. The simulation is repeated several times to gain the statistical confidence needed to reliably calculate observable physical quantities.

Aggregation and Misfolding

Murphy's Law of Thermodynamics:
"Things get worse under pressure"

Non-native conformations of proteins have recently attracted considerable attention. Bioinformatics studies of complete genomes suggest that a fairly large

fraction, perhaps up to one third, of the proteins encoded by the DNA of different organisms may be unstructured in the cell, or "natively unfolded" [5]. Unfolded states of proteins are usually studied experimentally through the addition of denaturants. The intra-chain non-covalent interactions that stabilize native structures of proteins become less favourable in the presence of denaturing agents, such as urea or guanidine. It has been suggested that natively unfolded proteins may have been selected through evolution because of their plasticity [6]. In order to function, these proteins should go through a disorder-to-order transition, which could be advantageous in certain situations as compared to more rigid ones. This property also gives them the ability to bind to different targets, as is often required in signalling cascades [7, 8], and a capability to overcome steric restrictions, enabling them to expose binding surfaces in complexes larger than those achieved by native proteins.

Another important issue concerning partially folded states of proteins is that they seem to be associated with "misfolding diseases", such as Alzheimer's, type II diabetes and the transmissible encephalopathies [9]. These diseases are characterised by highly organized amyloid aggregations. There is increasing evidence that amyloid aggregation is a generic feature of polypeptide chains, not just of the 20 or so proteins associated with pathological conditions. The ability of diverse peptides to form amyloid fibrils may be a consequence of the fact that the intermolecular bonds that stabilize the structure involve the peptide backbone. In addition, it suggests that the conformational properties of the multiple states that are accessible to proteins should be considered in order to provide an accurate description of their behaviour. When a mutation or a change in the folding conditions reduce the stability of the native state of a protein, the population of non-native states grows, thus increasing the probability of aggregation [10]. The study of unfolded states is thus interesting in order to better understand the folding mechanism and the properties of the free-energy landscape (*i.e.* variation of the free energy with protein configuration).

Physical Driving Forces

The non-covalent interactions of the protein chain with itself and with its environment (water) are the driving forces of protein folding. There are many variations on the same theme, so it is important to understand their physical nature. The interactions can essentially be divided into two groups: Short range repulsion and electrostatic interactions. That atoms repel each other at short distances is a consequence of the Pauli exclusion principle (two fermions cannot have the same quantum numbers). The electron clouds of both atoms occupy the lowest possible energy levels and if the clouds overlap then some

electrons must be elevated into excited states using the kinetic energy of the colliding atoms. Hence we observe a repulsive force at short distances [11]. All the other intramolecular interactions are thought to be essentially electrostatic in origin [1]; in other words attraction and repulsion of charges and dipoles. The simplest of these forces is the Coulomb [12] interaction between two charges:

$$F_c = \frac{1}{4\pi\epsilon}\frac{q_1 q_2}{r^2} \tag{1}$$

where the q's are the charges, r the distance between them and ϵ the dielectric constant of the medium. Atoms or molecules do not need to have a net charge to participate in electrostatic interactions. The particle nature of electrons makes an atom neutral only on *average*, effectively they can be described by rapidly fluctuating dipoles. These dipoles can attract the electrons of a nearby atom or molecule inducing another dipole. This interaction can be approximated with an expansion of the Coulomb potential in powers of the inverse distance between the charges, r^{-1}, using time-independent perturbation theory [13]. The first non-zero contribution turns out to be proportional to r^{-6} and attractive, and is commonly referred to as the van der Waals potential.

The non-covalent interactions are often represented by a potential as a function of distance r that includes both an attractive and a repulsive hard sphere potential. The most common of these is the so-called Lennard-Jones potential [14]:

$$E(r) = \epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - 2\left(\frac{\sigma}{r}\right)^{n}\right] \tag{2}$$

where σ is the minimum of $E(r)$ and $E(\sigma) = -\epsilon$. The r^{-n} dependence may model the electrostatic attraction between molecules described above. The r^{-12} represents the impenetrability of atoms, and has this form for computational efficiency. It is sometimes described by an exponential of $1/r$ instead, but there is little practical difference between these two choices [1]. These interactions are well understood in vacuum and in regular solids, but not in liquids. This lack of understanding is a consequence of the complexity of the liquid state, with its constantly changing interactions among many molecules in constant movement. The complexities of liquids are especially relevant for proteins, because their folded conformations usually occur only in a liquid-water environment.

In spite of its biological importance, water is not one of the best understood liquids. The forces between molecules dissolved in water are often more due to the properties of this solvent than to the intramolecular interactions themselves [1]. A more detailed description of these non-covalent interactions will include ionic bonds, hydrogen bonds and van der Waals bonds. Hydrogen bonds in particular stabilise the secondary structure of the protein. Furthermore the competition of hydrogen bonds between water molecules and the protein, with

hydrogen bonds among water molecules, is at the basis of what is commonly called the hydrophobic effect.Since both hydrogen bonds and hydrophobic interactions play central roles in our model, I will discuss them in some more detail.

Hydrogen Bonds

A hydrogen bond occurs when two electronegative atoms, such as nitrogen and oxygen, interact with the same hydrogen. The hydrogen is normally covalently attached to one atom, the donor, but interacts electrostatically with the other, the acceptor. This interaction is due to the dipole between the electronegative atoms and the proton. The covalent and electrostatic nature of the bond tends to keep the three atoms collinear. Hydrogen bonds are formed between different amino acids (residues), between amino acids and water and between water molecules. Hydrogen bonds between different amino acids are of special importance because they stabilise α-helices and β-sheets. The hydrogen bonds stabilising an α-helix are local, connecting the backbone CO group of amino acid i with backbone NH group of amino acid $i + 4$. The hydrogen bonds stabilising β-sheets may, by contrast, connect groups with a large separation along the sequence. Water has a strong propensity to form hydrogen bonds, and each water molecule can participate in two such bonds as an acceptor and two as a donor.

Hydrophobic effect

The *hydrophobic effect* is commonly considered to be the main physical driving force of the folding [15]. For water molecules it is rather unfavourable to form hydrogen bonds with non-polar (hydrophobic) molecules; they prefer to form such bonds between them. This property causes water in proximity of non-polar surfaces to adopt a particular configuration in order not to destroy the hydrogen bond network. Because of the ordering of water the system suffers an entropy loss. Trying to minimise this entropy loss ultimately causes the non-polar molecules to cluster together and thereby minimise the surface of contact with water. Another effect that leads to the minimisation of the hydrophobic surface is the fact that in this way water reduces the number of non-formed hydrogen bonds. Proteins might consist of both polar and non-polar amino acids. The cores of globular proteins are, due to the effective hydrophobic attraction, dominated by non-polar amino acids.

Theoretical Aspects

Many aspects of protein folding are still subjects of debate, in particular the amount of energetic *frustration*, the *two-state* character of the free-energy landscape and the *mechanism* of folding (nucleation *vs* diffusion collision).
In the following section I will describe these concepts and try to explain some of the open questions.

Frustration

> "Frustration arises when you buy a new boomerang
> and you realize you can't throw away the old one."
> thanks to *P. Johansson*

A fundamental property of the energy landscape of protein folding is frustration. In Giorgio Parisi's words, "frustration refers to an inability to satisfy simultaneously all the inclinations of all the microscopic entities" [16]. It arises typically in optimization problems involving competing quantities. A simple example is as follows. We have three players that we want to group into at most two teams, according to their mutual feelings; say love and hate. If they all hate each other, splitting them is obviously better even though two of them must be together, leaving one constraint unsatisfied. The system is frustrated. If they all love each other, we have the opposite situation where grouping them in the same team is better and does not lead to any unsatisfied constraint.

In general frustrated optimization problems have many equivalent solutions: solving our problem in different ways may lead to similar values of the quantity being optimised (*e.g.* energy). In the example with the three players, let us assume that the "energy" of the system will decrease by one unit for each satisfied constraint (pairing). In the *love* case, if all the players are equivalent, all constraints can be satisfied and only one minimum (all in the same team) is present. On the contrary, in the *hate* case, there are three equivalent configurations (i.e. the three ways to divide the players), that is three equivalent energy minima.

For proteins we can have two types of frustration, energetic and topological [17]. Energetic frustration can arise when two amino acids which are not close in the native structure form a stabilising contact (*e.g.* a hydrogen bond) as the chain folds. It can also arise because two amino acids happen to be close in the native structure although they do not interact favourably. Topological frustration is due to excluded volume and chain connectivity. It occurs when two configura-

tions similar in energy are not dynamically connected; that is, a transition from one to the other requires at least partial unfolding. While energetic frustration can be minimized by evolution (choosing particular sequences), the topological one is unavoidable.

Energy Landscape Theory

A major contribution to folding theories is represented by the *energy landscape picture* [17–19]. We call the variation of the energy with respect to protein configuration "energy landscape". Rather than viewing the folding of a protein as a sequence (pathway) of specific intermediates, the landscape theory views folding as a progression (diffusion) of ensembles. That is to say, this macroscopic pathway is an ensemble of microscopic ones. According to this theory, the energy landscape of a folding protein resembles a partially rough funnel (native contacts are stabilising). The local roughness of the funnel reflects transient trapping of the protein configurations in local energy minima. The overall funnel shape of the landscape, superimposed on this roughness, arises because the interactions present in the native structure of natural proteins conflict with each other much less than expected if there were no constraints of evolutionary design to achieve reliable and relatively fast folding: the frustration is minimised. This funnel shape is necessary to overcome the Levinthal paradox *i.e.* the statistical improbability of finding a unique stable folded conformation of proteins by random searches in a flat energy landscape.

The energy landscape picture is, nevertheless incomplete [4], as entropic effects are ignored. The growth in the number of thermally occupied states as we move away from the native structure favours disordered configurations. This represents a thermodynamic force opposing that derived form the energy landscape. The net result of these two opposing effects is represented by the *free*-energy landscape. The power of this theory lies in enabling us to project the high dimensionality of the folding problem down to one or few reaction coordinates. A good reaction coordinate describes how "close" a specific conformation is to the native state in terms of folding kinetics. There are various ways to quantify the similarity with the native state and some care is required in choosing one of them because some of this measures do not discriminate between topologies that are kinetically separated [20]. A typical example of two such topologies is given by the two mirror images of a three-helix bundle [21, 22]. By looking at the free-energy as a function of these reaction coordinates it is then possible to describe the kinetic behaviour of the system. If the motion in the free-energy landscape is assumed to be diffusive using Kramer's analysis [23] we can predict the folding time (see section on the "two state picture"). However when

projecting a high dimensional space into one or two reaction coordinates the dynamics along these coordinates tend to become non diffusive on short time scales. For many small proteins, the free-energy landscape is believed to have a simple shape with two minima, corresponding to one folded and one unfolded state. We define the folding temperature T_{f}, as the one where the unfolded population P_u equals the folded one P_{f}. The two minima are separated by a free-energy barrier, and the ensemble of states corresponding to this free-energy maximum is referred to as the *transition* state ensemble (TSE).

Folding mechanisms: Nucleation *vs* Diffusion Collision

"Is this good, or does it just *look* good?"
P. Dhonte

The folding mechanism still remains an open problem, with two major theories trying to explain it. One view, the diffusion collision [24] mechanism, wants secondary structure elements to be formed *before* the hydrophobic collapse occurs. In other words the folding kinetic is driven by secondary structure (α-helix or β strand) formation. Another view, the nucleation mechanism [4], argues that the tertiary structure is what drives the folding kinetic and the secondary one plays only a minor role. According to this view, the protein in the TSE forms a certain set of native interactions between specific residues, a *folding nucleus*. Hence a conformation in the TSE may look like a distorted native state with large unstructured loops. Some proteins, however, have a nucleus with partially formed secondary structure [25]. The nucleation does not require secondary structure elements to be formed before the transition state is reached but in some cases it could be formed simultaneously with the formation of the tertiary structure.

A number of theoretical and experimental studies [26–28] have found that the nucleation mechanism better describes the folding, ruling out the necessary kinetic role of secondary structure elements. Other studies [29] on the other hand, support the diffusion collision mechanism. It is possible, however, that there is not a single rule applying to all proteins and that both mechanisms explain the folding behaviour but for different proteins and different conditions. For a recent review of different folding mechanisms, see Ref. [30].

Two-state picture

Many small proteins behave in a two-state manner [31]. In the two-state picture, the free-energy landscape is formed out of two minima separated by a

barrier in between. Within this view, a generic observable is given by:

$$X(T) = \frac{X^{\mathrm{u}} + K(T)X^{\mathrm{f}}}{1 + K(T)} \tag{3}$$

where X^{u} and X^{f} are the values of the observable in the unfolded and in the folded state, respectively, and $K(T)$ is the ratio between the populations in the folded and the unfolded states. Another prediction of the two state picture is that the relaxation time of this observable, is a single exponential. Assuming diffusive motion along each reaction coordinate, it is possible to model the folding process with a Brownian motion in an external potential where the probability distribution of a generic reaction coordinate is described by the diffusion equation. With this assumption, the dependence of our observable X on time t, $X(t)$, can be written as a sum of exponentials:

$$X(t) = \langle X \rangle + \sum_{k=1}^{\infty} A_k e^{-t/\tau_k} \tag{4}$$

where $\tau_1 > \tau_2 > \tau_3 > \ldots$ and $\langle X \rangle$ denotes the equilibrium value of X. If the barrier is high then τ_1 is much bigger than the other τ_k, giving rise to an effectively single-exponential relaxation. The time constant for this slowest mode is then given by [23]

$$\tau_1 \propto e^{\beta \Delta F} \tag{5}$$

where ΔF is the height of the barrier and $\beta = 1/k_{\mathrm{B}}T$, k_{B} being Boltzmann's constant.

In Paper III we discuss the folding behaviour of proteins for which the barrier is very small. In the absence of a clear barrier there is no reason to expect a clear separation between τ_1 and the other τ_k. Recent experimental studies [32] on some very rapidly folding proteins (few μs) found indeed deviations from single-exponential relaxation; the results could, however, be fitted to a sum of two exponentials.

Potentials

> "It's simple physics. Calculate the velocity, v,
> in relation to the trajectory, t, in which g, gravity,
> of course remains a constant. It's not complicated."
> *G. Costanza*

The shape of the free-energy landscape observed in simulations is partly determined by the potential used to study the folding. As I discussed before, for

small proteins it is believed that the amount of energetic frustration is very low. An obvious starting point for an *effective* potential is thus the class of the Gō-like ones [33]. In the simplest form of Gō potentials, one ignores all interactions between parts that are not in contact in the native structure. In other words, the energetic frustration is completely removed and we are left only with the geometric one. Even if at a first glance this approach seems completely unphysical and oversimplified, it has two major advantages. First, its ability to fold essentially any protein. Second, properties of the near-native states might be well described, and in some cases [34] a good agreement with experiments has been obtained. It has recently been shown that non-native interactions may play a stabilising role along the normal folding pathways, in particular in the unfolded and transition states [35]. How much Gō potentials mimic true potentials is, however, still a subject of debate. Naturally, a clear disadvantage of this class of potentials is the lack of predictive power and their intrinsically unphysical nature.

Sequence-based potentials represent a completely different approach, in which no prior information on the native state is required. In this case the potential is formed out of many parts, each modelling a different physical driving force. However, the precise forms of these forces are not known. One consequence of this is that one has no control on the amount of energetic frustration. Indeed, trapping in local minima typically increases the complexity of these simulations compared to Gō calculations. It has to be noted, however, that the role and the amount of energetic frustration in *real* proteins are not completely understood.

Today we are still far from having the "perfect" general potential that folds every protein. It is of tremendous importance to have potentials able to fold several, structurally different proteins without any fine-tuning of the parameters. Only in this way, a clear understanding of the physical interactions can be achieved.

Computational Methods

"Before explaining how the computers
that I used for my simulations work,
I will give some insights on punch cards."
M. Ringnér

Computer simulations of protein folding rely on methods for sampling the space of all possible configurations. Monte Carlo and Molecular Dynamics are two such methods. The ideas behind them are quite different. In Monte Carlo,

one jumps stochastically between configurations. In Molecular Dynamics the motion in the configuration space is calculated by integration of the equations of motion. In the following section, I will give an overview of the basic concepts behind these methods.

Thermal Average

In an experiment we typically measure a quantity (observable) over a macroscopic sample. In other words, the measured quantity is an average over many microscopic entities. In a simulation we deal with a microscopic system, so we need to calculate this average explicitly.

In equilibrium statistical mechanics one computes averages of a quantity A from the Boltzmann distribution, *i.e.*

$$\langle A \rangle = \sum_l P_l^{\mathrm{eq}} A_l, \tag{6}$$

where l denotes a state, A_l is the value of A in that state, and P_l^{eq} is the equilibrium (Boltzmann) distribution, *i.e.*

$$P_l^{\mathrm{eq}} = \frac{e^{-\beta E_l}}{\sum_m e^{-\beta E_m}} \tag{7}$$

The number of states is exponentially large in the number of degrees of freedom, N, which makes it quite impractical to perform the sum in Eq. (6) except if N is really tiny. However, we are generally interested in large N. Rather than summing over *all* the states in Eq.(6), it is possible to sample a small fraction of these states. This leads to an *estimate* of the average, which will not be exact but will have statistical errors. We use an iterative procedure which, after some initial transient phase, generates states according to the Boltzmann distribution. Hence the estimate of the average is

$$\langle A \rangle_{\mathrm{est}} = \frac{1}{t_0} \sum_{t=1}^{t_0} A(t), \tag{8}$$

where t is simulation "time", $A(t)$ is the value of A at that particular time, and t_0 is the number of measurements. The difference between the estimate $\langle A \rangle_{\mathrm{est}}$ in Eq.(8) and the exact value $\langle A \rangle$ in Eq.(6) is proportional to $n^{-1/2}$ where n is the number of *statistically independent* measurements. The configurations generated by the algorithm will be *correlated* in general up to a certain *relaxation time* τ, so $n \sim t_0/\tau$ will be less than t_0.

Monte Carlo

The Monte Carlo method provides a way to compute thermal averages. The idea behind this approach is to sample the configuration space randomly. The method is called after the city in the Monaco principality, because of the game of roulette, a simple random number generator. A simple example of this method could be as follows. Let us suppose that we have a lake inside a square field (with side l) delimited by high walls. We want to measure the surface of the lake. A possible way to do that, would be to throw stones inside the field (randomly) and count how many of them end up on the lake (*i.e.* how many splashes we hear) and how many on the ground (no splash). Since we know the surface of the field (l^2) the ratio between the two numbers (stones in the lake/stones in the ground) will give us an estimate of the surface of the lake that becomes more and more precise the more stones we throw.

A Monte Carlo simulation begins with the system in some state, say l_0. We then generate stochastically a subsequent set of states (Markov chain), to which we give a *time* label $t = 0, 1, 2, 3 \ldots$ At $t = 0$ the system is definitely in state l_0 but at later times it can be in different states with non-zero probability $P_l(t)$. We desire that at long times $P_l(t)$ approaches the equilibrium distribution P_l^{eq}, *i.e.*

$$\lim_{t\to\infty} P_l(t) = P_l^{\text{eq}}. \tag{9}$$

In order to achieve this, the algorithm must be ergodic, *i.e.* starting from any state, after a sufficiently long time, there should be a non-zero probability for the system to be in *any other* state.

The initial distribution is made to converge to the equilibrium distribution after a certain *time* by a judicious choice of the "transition rates" $w_{l\to m}$, where $w_{l\to m}$ is the probability that, given that the system was in state l at time t, it will be in state m at time $t + 1$. The evolution of the probabilities $P_l(t)$ follows the *master equation*

$$P_l(t+1) - P_l(t) = \sum_m [P_m(t) w_{m\to l} - P_l(t) w_{l\to m}] \tag{10}$$

Clearly, a necessary condition for the method to work is that the Boltzmann distribution P^{eq}, is a *stationary* distribution, *i.e.* if $P_l(t) = P_l^{\text{eq}}$ for all l then $P_l(t+1) = P_l^{\text{eq}}$ for all l, which according to Eq.(10) requires that

$$\sum_m (P_l^{\text{eq}} w_{l\to m} - P_m^{\text{eq}} w_{m\to l}) = 0 \tag{11}$$

In practice, stationarity is usually accomplished by making *each* term in Eq.(11) vanish, *i.e.*

$$P_l^{\text{eq}} w_{l\to m} = P_m^{\text{eq}} w_{m\to l} \tag{12}$$

which is known as the *detailed balance* condition. Because the equilibrium distribution is given by Eq.(7), the detailed balance condition can be written

$$\frac{w_{l\to m}}{w_{m\to l}} = \frac{P_m^{\mathrm{eq}}}{P_l^{\mathrm{eq}}} = e^{-\beta(E_m - E_l)} = e^{-\beta\Delta E} \tag{13}$$

A common way of implementing a Monte Carlo move is to first draw a "trial" state m at time $t+1$ from the proposal matrix $U_{l\to m}$, given that the state at time t was l. This matrix $U_{l\to m}$ is usually chosen to be symmetric. The state m is then accepted as the state at $t+1$ with some probability $a_{l\to m}$. Otherwise the state at $t+1$ is the old state l. The transition matrix can be written as

$$w_{l\to m} = U_{l\to m} a_{l\to m}. \tag{14}$$

Assuming that $U_{l\to m}$ is symmetric, the detailed balance condition can be expressed as

$$\frac{a_{l\to m}}{a_{m\to l}} = e^{-\beta\Delta E} \tag{15}$$

The Metropolis algorithm [36] satisfies this equation by taking

$$a_{l\to m} = \min(e^{-\beta\Delta E}, 1). \tag{16}$$

In other words, one always accepts the move if the energy decreases but only accepts it with probability $\exp(-\beta\Delta E)$ if the energy increases.

Molecular Dynamics

The Molecular Dynamics method is based on integrating the equations of motion for all the particles of the system. This process is deterministic which makes Molecular Dynamics very different from Monte Carlo. The system is initialised with initial position and velocity for each particle. Then the forces acting on all of the particles are calculated. Finally, the system is evolved one step forward in time, using a suitable discretised form of the equations of motion, *e.g.*

$$r(t+\Delta t) \sim 2r(t) - r(t-\Delta t) + \frac{f(t)}{m}\Delta t^2 + O(\Delta t^4) \tag{17}$$

where r is a coordinate, f is the force, and Δt is the step size. This estimate of the new position contains an error that is of order Δt^4. To ensure numerical stability one should use a Δt smaller than the fastest time scale of the system. The time scales of protein folding are very broad, ranging from atom oscillations

at $O(10^{-15})$ seconds to folding times at $O(10^{-6})$ seconds and above. This large spread of time scales makes Molecular Dynamics simulation of protein folding typically very slow. The strength of this method relies on the possibility of performing "physical" updates. In thermodynamic simulations, where one is interested in the average behaviour of the system, it is much more efficient, however, to use generalised-ensemble techniques such as simulated tempering. In kinetic simulations, on the contrary one is interested in the evolution of the system. Molecular Dynamics, with its physical updates seems then the natural choice. However, whether the use of physical updates in a model without explicit solvent, models the real evolution of the system, still remains to be seen. Furthermore, it was recently noticed [37], that if kinetic events are separated by a large enough number of local Monte Carlo updates, and if they are observed in an ensemble of relaxation trajectories, they represent significant state population shifts and reflect properties of the free-energy landscape.

Improved Monte Carlo Methods

Even for proteins with minimal amount of energetic frustration, the free-energy landscape contains different local minima separated by barriers. From a computational point of view this means that the relaxation time of the system is very long; a longer simulation time is required in order to obtain statistically reliable results. A general way to overcome this problem is through the use of *umbrella sampling* [38] or *reweighting* [39] techniques, which make it possible to replace the desired Boltzmann distribution with another distribution that is easier to simulate. Suppose that there are two regions in the configuration space we want to sample, and the probability of visiting one of them is very small. One can then build a new distribution where the two regions are reweighted in order to make them equally probable. Two widely studied approaches to the problem of building this new probability distribution are the *multicanonical* [40, 41] and the *simulated tempering* [42, 43] methods. In the multicanonical method the probability distribution is defined in such a way that all energies are equally probable. This is achieved by dividing the energy axis into sub-intervals and adding a piece-wise linear term, $\alpha_E + \beta_E E$, to the true energy E. In this way the Boltzmann factor $P_B \propto \exp(-\hat{\beta}E)$ is replaced by the distribution

$$P_{\text{Mcan}} \propto \exp\{-\alpha_E - (\hat{\beta} + \beta_E)E\} \tag{18}$$

On each sub-interval this leads to a canonical weight factor with $\beta = \hat{\beta} + \beta_E$. The parameters β_E and α_E are chosen so as to make the energy distribution as flat as possible.

Since in our simulations we have used simulated tempering I will discuss this method in some more detail.

Simulated Tempering

The idea of simulated tempering is to add a second Markov chain to the usual one in order to make it easier for the system to move across free-energy barriers. More accurately, we want to equilibrate our statistical system with respect to the Boltzmann distribution Eq.(7) $P(\sigma)$ where σ represents a configuration of the system we want to study. We choose a new probability distribution, with an enlarged number of variables, $\tilde{P}(\sigma, \beta_\alpha)$, such that for fixed β_α, $\tilde{P}$ is proportional to the Boltzmann distribution at $\beta = \beta_\alpha$. In this way β becomes a dynamical variable with a predetermined set of allowed values β_α. The new equilibrium probability distribution is

$$P^{\mathrm{eq}}(\sigma, \beta_\alpha) \simeq e^{-\beta_\alpha E(\sigma) + g_\alpha} \tag{19}$$

The g_α are parameters which tune the relative probability of each β value; the probability of finding a given value of α is

$$P_\alpha = Z_\alpha e^{g_\alpha} \equiv e^{g_\alpha - \beta_\alpha F_\alpha} \tag{20}$$

where Z_α and F_α are respectively the partition function and the free energy at β_α. In order to have all β values equally probable (*i.e.* to visit all β with the same frequency) we need to set

$$g_\alpha = \beta_\alpha F_\alpha \tag{21}$$

It is possible to estimate g_α from a test run in which one monitors the probabilities of visiting the different temperatures. For this method to work the overlap of the probability distributions at adjacent temperatures has to be non-negligible.

A method closely related to simulated tempering is *parallel* tempering [44–47].

Paper I

As we have seen before the Monte Carlo method allows large unphysical moves which are extremely efficient to sample the configuration space at high temperatures when the chain is in an extended configuration. A modest change in a dihedral angle in the middle of the chain can lead to large movements of the atoms at the end. Near-native configurations of proteins are quite compact

objects. In this case large moves that lead to major rearrangement of part of the chain are inconvenient because they lead to steric clashes which reduce the acceptance rate. In a compact configuration a local update that leaves the rest of the chain fixed, is more efficient. In this paper we discuss a "quasi-local" Monte Carlo update. The method involves a change in the torsional angles of four consecutive amino acids, drawn from a Gaussian distribution that favours approximately local deformations of the chain. A bias parameter is introduced to control the degree of "locality" of the update. This method was successfully used in [48], combined with a chain closure algorithm [15], to produce quasi local pre-rotations.

Papers II and IV

In a recent paper [21], Irbäck, Sjunnesson and Wallin introduced an off-lattice model with a simplified geometry. Their potential does not rely on the Gō prescription. That model includes three types of amino acids, hydrophobic, polar, and glycine, the energy function is based on hydrogen bonds and hydrophobicity interactions. In paper (II) this model is extended through the introduction of alanine and proline. Alanine is taken to be intermediate in hydrophobicity between the hydrophobic and the polar amino acids, while proline has a special geometric representation in order to mimic its helix breaking properties. The only degrees of freedom are the backbone torsional angles. The potential is kept as simple as possible. It is composed of four terms, local, excluded volume, hydrogen bonds, and hydrophobicity. The extreme simplicity of the model is its strength because it focuses only on the essential ingredients needed to fold a three helix bundle. In paper (II) the 10-55-amino acid fragment of the B domain of staphylococcal protein A is studied with thermodynamic and kinetic simulations. Energy minimisation, restricted to the thermodynamically favoured topology, gives a configuration that has a root-mean-square deviation of 1.8Å from the experimentally determined structure. From the kinetic simulations we found that collapse is at least as fast as helix formation, ruling out the possibility of a diffusion collision pathway as suggested by Zhou and Karplus [50].

In paper (IV) we study the 9-54-amino acid fragment of the Z domain of staphylococcal protein A $Z_{\text{SPA-1}}$ and its wild-type. $Z_{\text{SPA-1}}$ was engineered from the wild type adding 13 mutations. Both fragments are compared with recent experimental data [51]. The model predicts that in $Z_{\text{SPA-1}}$ the helix content is lower and the melting behaviour is less cooperative as compared with the wild type. In the wild type of the Z domain, as well as in the B domain studied in paper (II), chain collapse and helix formation occur on similar time scales.

However, in the mutated sequence $Z_{\text{SPA-1}}$, chain collapse is faster than helix formation.

Paper III

In this paper we discussed what folding behaviour should be expected in case the energy barrier between the unfolded and native state is small or absent. The model and the protein studied are the same as in [21]. Thermodynamic simulations are used to study the free-energy behaviour as a function of the energy or as a function of the similarity to the native state. This free-energy landscape lacks a clear bimodal shape normally associated with two state systems; but in spite of that, the melting curves show a two-state character and the relaxation behaviour is close to a single exponential. Using a square-well approximation we made predictions of the relaxation time that are found to agree within a factor of two with the observed one. It is interesting to notice that the second and third terms in Eq.(4), solved for square-well potential, have time constants $\tau_2 = \tau_1/4$ and $\tau_3 = \tau_1/9$ at $T = T_{\text{f}}$, where τ_1 is given by Eq.(5). Due to statistical limitations, this could not be tested on our model protein. The double exponential fit of recent experimental data [32] on fast-folding proteins, that was previously discussed, gave $\tau_1/\tau_2 \approx 4$. However the lack of other studies showing this behaviour makes it impossible to draw any definitive conclusion.

Paper V

Using an all atom model [29–31] with a sequence based potential, the aggregation of the Aβ_{16-22} fragment of the Alzheimer's amyloid peptide is studied. We explore the behaviour of different systems with one, three, and six peptides, at constant peptide concentration. As indicated by experimental studies [18] we find a strong propensity for antiparallel beta sheet formations. In addition to the Aβ_{16-22} peptide, we study a few control sequences. Hydrophobicity is found to be a major driving force of the aggregation of this peptide.

A recent study [56] on the same peptide also found propensity for antiparallel beta-sheet formation. The authors suggested that this propensity is due to the Coulomb interactions between the opposite charges present at the extremities of the Aβ_{16-22}peptide. Since our potential Coulomb interactions are not included, our results indicate that other mechanisms, such as the geometry of backbone-backbone hydrogen bonds, might play an important role in determining the antiparallel orientation of Aβ_{16-22} peptides. We also simulate a system of three

peptides where the hydrophobic amino acids are distributed asymmetrically. In contrast to the Aβ_{16-22} peptide this system is able to form three parallel β-strands because of its hydrophobic moment. The hydrophobic moment was also found to be important by Gordon *et al.* in Ref. [19] who studied experimentally the Aβ_{16-22} peptide modified with the addition of octanoic acid, in order to increase its amphiphilicity. The octanoyl-Aβ_{16-22} is found to form fibrils with parallel β-strands.

Acknowledgments

"Sorry, I wish I could help, but you know my policy . . . "
S. Wallin

Some persons around me were always able to support me in the bad days. Most importantly they shared with me the good ones. Thanks to all of them.
One of these friends is definitively Anders. Saying thanks to him for all the support, the enthusiasm, and guidance would be not enough, so I will not.

I have particularly enjoyed the early morning conversations with Carsten in front of a fresh cup of coffee. Just as in the case of Anders, it would be a mistake to thank Carsten for one thing in particular, because to do so would be to define, and to define is to limit.

I usually blame my office mates for the noise level in our office. Nevertheless, it has been a privilege to share the office with Sven, Markus, Thomas, Jari Peter and Björn.
Björn, among other qualities, has the capacity to work quietly; when I shared the office with him the noise level did not drop. Someone obviously deserves an apology, but as of yet I have not been able to figure out who.

Recently I have read three great manuals [58–60]; the authors, Stephen, Chafik, and Stefan have a special place in my heart.

Boh . . . to say something about Pierre let me use a metaphor
If Zidane makes a great goal people are happy, but not surprised. If Schillaci does the same, they cannot believe their own eyes. Analogously what is a fantastic behaviour for some people is normal for others.
. . . OK maybe I would be a bit surprised if Pierre played like Zidane, but that's not the point.

It has always been a pleasure to work with Björn, Fredrick, Sandipan, and Stefan. I was always impressed by their capacity to work hard with no apparent stress. My personal technique of screaming all over the place how tired I was, and how that *particular* problem was the toughest one I had ever encountered, wasn't as impressive, I believe.
Coffee is not coffee if you don't drink it *At Jari's* with Peter, Markus and the owner himself, Jaray. In this corner of "Bizarro World" life, not just coffee, has a special taste. (It is a bit like *Cheers*, but better)
An old friend of mine said. . . if you know someone that can whistle any tune magnificently, you just can't go and tell him that he whistles in a great way. . .
With them is the about the same.

I strongly believe Markus, Sandipan and Peter should thank *me* to let them proofread this great introduction.

There are three kinds of sisters, the good ones, the bad ones, and Valentina. I am lucky that I have got the last kind. Finally I have to express my gratitude to all my family who have always supported me during all these years, in particular to my father who could see my love for physics way before I did. It takes the love of a father to do so much for someone knowing that this person will never be able to pay it back even in a couple of lives.

References

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

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

[3] Bowie J.I. Lüthy R. & Eisemberg D. (1991)
"A method to identify protein sequences that fold into known three-dimensional structures",
Science **253**, 164.

[4] 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,
(IOS Press), pp. 37.

[5] V.N.Uversky (2002)
"Natively unfolded proteins: A point where biology waits for physics",
Protein Science, **11** 739.

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

[7] Meador E. Means A. & Quiocho F. (1993)
"Modulation of calmodulin plasticity in molecular recognition on the basis of x-ray structures",
Science **262** 1718.

[8] Crivici, A. Ikura, M. (1995)
"Molecular and structural basis of target recognition by calmodulin",
Annu. Rev. Biophys. Biomol. Struct. **24** 85.

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

[10] Dobson C.M. (1999)
"Protein Misfolding, Evolution and Disease",
Trends Biochem. Sci. **24**, 329.

[11] Povh B., Rith K., Scholz C., & Zetsche F. (1995)
"Paricles and Nuclei",
Springer-Verlag

[12] Jackson J.D. (1999)
"Classical Electrodynamics",
John Wiley & Sons, Inc

[13] Sakurai J.J (1985)
"Modern Quantum Mechanics",
Benjamin/Cummings Publishing

[14] Frenkel D. & Smit B. (1996)
"Understanding Molecular Simulation",
Academic Press

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

[16] Mezard M. Parisi G. & Virasoro M.A. (1987)
"Spin Glass Theory and Beyond",
World Scientific Publishing

[17] 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.

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

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

[20] Wallin S.,Farwer J., & Bastolla U. (2003)
"Testing similarity measures with continuous and discrete protein models"
Proteins Struct. Funct. Genet. **50**, 144.

[21] Irbäck A, Sjunnesson F.& Wallin S. (2000)
"Three helix-bundle in a Ramachandran model",
Proc. Natl. Acad. Sci. USA **97**, 13614.

[22] Favrin G., Irbäck A. & Wallin S. (2002)
"Folding of a small helical protein using hydrogen bonds and hydrophobicity forces",
Proteins Struct. Funct. Genet. **47**, 99.

[23] Kramers H.A. (1940)
"Brownian motion in a field of force and the diffusion model of chemical reactions",
Physica **7**, 284.

[24] Karplus M., Weaver D.L. (1994)
"Protein folding dynamics: The diffusion-collision model and experimental data" *Protein Science* **3**, 650.

[25] Prieto J. & Serrano L. (1997)
J. Mol. Biol. **274**, 276.

[26] Cregut D. Civera C., Macias M.J., Wallon G., & Serrano L. (1999)
"A tale of two secondary structure elements: when a beta-hairpin becomes an alpha-helix",
J. Mol. Biol. **292**, 389.

[27] Moran L.B., Schneider J.P., Kentsis A., Reddy G.A., & Sosnick T.R. (1999)
"Transition state heterogeneity in GCN4 coiled coil folding studied by using multisite mutations and crosslinking",
Proc. Natl. Acad. Sci. USA **96**, 10699.

[28] Kim D.E., Yi Q., Gladwin S.T., Goldberg J.M., & Baker D. (1998)
"The single helix in protein L is largely disrupted at the rate-limiting step in folding",
J. Mol. Biol. **284**, 807.

[29] Mayor U., Guydosh N.R., Johnson CM, Grossmann JG, Sato S, Jas GS, Freund SM, Alonso DO, Daggett V, Fersht AR. (2003) "The complete folding pathway of a protein from nanoseconds to microseconds". *Nature* **421**, 863.

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

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

[32] Yang W.Y., Gruebele M. (2003)
"Folding at the speed limit",
Nature **423**, 193.

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

[34] Clementi C., Nymeyer H., & Onuchic J.N. (2000)
" Topological and energetic factors: what determines the structural details of the transition state ensemble and *on-route* intermediates for protein folding? An investigation for small globular proteins",
J. Mol. Biol. **298**, 937.

[35] Paci E. Vendruscolo M. Karplus M. (2002)
"Native and non-native interactions along protein folding and unfolding pathways",
Proteins Struct. Funct. Genet. **47**,379.

[36] 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.

[37] Shimada J., Kussel E., Shakhnovich E.I. (2001)
"The folding thermodynamics and kinetics of crambin using an all-atom Monte Carlo simulation" *J. Mol. Biol.* **308**, 79.

[38] Torrie, G.M. & Valleu J.P. (1977)
"Non-physical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling",
J. Comp. Phys., **23**, 187.

[39] Ferrenberg A., Swendsen R. (1988) "New Monte Carlo technique for studying phase transitions"
Phys. Rev. Lett. **61**, 2635.

[40] Berg B.A. & Neuhaus T. (1991)
"Multicanonical ensemble: A new approach to simulate first-order phase transitions",
Phys. Lett. B **267**, 249.

[41] 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.

[42] 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.

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

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

[45] Hukushima K. & Nemoto K.(1995)
"Exchange Monte Carlo method and applications to spin glass simulations",
J. Phys. Soc. (Jap) **65**, 1604.

[46] Geyer C.J., & Thmopson E.A. (1995)
"Annealing Markov chain Monte Carlo with applications to ancestral inference "
J. Am. Stat. Ass. **90**, 909.

[47] Tesi M.C., Janse van Rensburg E.J. & Orlandini E. (1996)
"Monte Carlo study of the interacting self-avoiding walk model in three dimensions"
J. Stat. Phys. **82**, 155.

[48] Ulmschneider J.P., Jorgensen W.L. (2002)
"Monte Carlo backbone sampling for polypeptides with variable bond angles and dihedral angles using concerted rotations and a Gaussian bias",
J. Chem. Phys. **118**, 4261.

[49] Gō N., Sheraga H.A. (1970)
"Ring closure and local conformational deformations of chain molecules"
Macromolecules **3**, 178.

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

[51] 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.

[52] Irbäck A, Samuelsson B, Sjunnesson F, Wallin S.(2003)
Thermodynamics of α- and β-structure formation in proteins. *Biophys. J.* **85**, 1466.

[53] Irbäck A, Sjunnesson F. (2004)
Folding thermodynamics of three β-sheet peptides: A model study.
Proteins Struct. Funct. Genet. (in press).

[54] Irbäck A, Mohanty S.(2004)
manuscript in preparation.

[55] Balbach JJ, Ishii Y, Antzutkin ON, Leapman RD, Rizzo NW, Dyda F, Reed J, Tycko R. (2000) Amyloid fibril formation by Aβ_{16-22}, a seven-residue fragment of the Alzheimer's β-amyloid peptide, and structural characterization by solid state NMR. *Biochemistry* **39**, 13748.

[56] Klimov D.K., & Thirumalai D. (2003) *Structure* **11**, 295.

[57] Gordon DJ, Balbach JJ, Tycko R, Meredith SC. (2004)
Increasing the amphiphilicity of an amyloidogenic peptide changes the β-sheet structure in the fibrils from antiparallel to parallel.
Biophys. J. **86**,428.

[58] Burby S. (2004)
"Puking in a cab for dummies"
B.S. Press

[59] Driouichi C. (2004)
"How to find a post-doc in Bern"
B.S. Press

[60] Wallain S. (2004)
"Teach yourself crime witnessing in 21 days"
B.S. Press

Monte Carlo Update for Chain Molecules: Biased Gaussian Steps in Torsional Space

Paper I

Monte Carlo Update for Chain Molecules: Biased Gaussian Steps in Torsional Space

Giorgio Favrin, Anders Irbäck and Fredrik Sjunnesson

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

Journal of Chemical Physics **114**, 8154-8158 (2001)

Abstract:
We develop a new elementary move for simulations of polymer chains in torsion angle space. The method is flexible and easy to implement. Tentative updates are drawn from a (conformation-dependent) Gaussian distribution that favors approximately local deformations of the chain. The degree of bias is controlled by a parameter b. The method is tested on a reduced model protein with 54 amino acids and the Ramachandran torsion angles as its only degrees of freedom, for different b. Without excessive fine tuning, we find that the effective step size can be increased by a factor of three compared to the unbiased $b = 0$ case. The method may be useful for kinetic studies, too.

1.1 Introduction

Kinetic simulations of protein folding are notoriously difficult. Thermodynamic simulations may use unphysical moves and are therefore potentially easier, but existing methods need improvement. Three properties that a successful thermodynamic algorithm must possess are as follows. First and foremost, it must be able to alleviate the multiple-minima problem. Methods like the multicanonical algorithm [1, 2] and simulated tempering [33–35] try to do so by the use of generalized ensembles. Second, it must provide an efficient evolution of large-scale properties of unfolded chains. The simple pivot method [6] does remarkably well [7] in that respect. Third, it must be able to alter local properties of folded chains without causing too drastic changes in their global structure. This paper is concerned with the third problem, which is important if the backbone potentials are stiff and especially if the mobility is restricted to the biologically most relevant torsional degrees of freedom.

An update that rearranges a restricted section of the chain without affecting the remainder is local. For chains with flexible or semiflexible backbones, there exists a variety of local updates, ranging from simple single-site moves to more elaborate methods [8–12] where inner sections are removed and then regrown site by site in a configurational-bias manner [13, 14]. However, these methods break down if bond lengths and bond angles are completely rigid.

The problem of generating local deformations of chains with only torsional degrees of freedom was analyzed in a classic paper by Gō and Scheraga [15]. Based on this analysis, Dodd *et al.* [16] devised the first proper Monte Carlo algorithm of this type, the concerted-rotation method. This method works with seven adjacent torsion angles along the chain. One of these angles is turned by a random amount. Possible values of the remaining six angles are then determined by numerically solving a set of equations that guarantee that the move is local. The new conformation is finally drawn from the set of all possible solutions to this so-called rebridging problem. Variations and generalizations of this method have been discussed by several groups [17–19]. There are also methods [20–24] that combine elements of the configurational-bias and concerted-rotation approaches. One of these methods [23] uses an analytical rebridging scheme, inspired by the solution for a similar problem in robotic control [25].

The concerted-rotation approach is a powerful method that can generate large local deformations by finding the discrete solutions to the rebridging problem. However, the method is not easy to implement and large local deformations

may be difficult to accomplish if, for example, the chain is folded and has bulky side groups. Hence, there are situations where this method is not the obvious choice.

In this paper, we discuss a different and less sophisticated type of Monte Carlo move in torsion angle space. This algorithm is by nature a "small-step" algorithm so large local deformations cannot take place. Drastic global changes would still occur if the steps were random. To avoid that, a biasing probability is introduced. The method becomes approximately local if the bias is made strong. Compared to a strictly local update, this method has the disadvantage that a much smaller part of the energy function is left unchanged, so the CPU time per update is larger. However, this problem is not too severe for moderate chain lengths. Moreover, both our method and strictly local ones are typically combined with some truly nonlocal update like pivot, and such an update is not faster than ours.

The algorithm proceeds as follows. We consider n torsion angles ϕ_i, where $n = 8$ in our calculations. To update these angles, we introduce a conformation-dependent $n \times n$ matrix $\mathbf{G}$ such that $\delta\bar{\phi}^T \mathbf{G} \delta\bar{\phi} \approx 0$ for changes $\delta\bar{\phi} = (\delta\phi_1, \ldots, \delta\phi_n)$ that correspond to local deformations. The steps $\delta\bar{\phi}$ are then drawn from the Gaussian distribution

$$P(\delta\bar{\phi}) \propto \exp\left[-\frac{a}{2}\delta\bar{\phi}^T(\mathbf{1} + b\mathbf{G})\delta\bar{\phi}\right] , \tag{1.1}$$

where $\mathbf{1}$ denotes the $n \times n$ unit matrix and a and b are tunable parameters. The parameter a controls the acceptance rate, whereas b sets the degree of bias. The new conformation is finally subject to an accept/reject step. Important to the implementation of the algorithm is that the matrix $\mathbf{G}$ is non-negative and symmetric. Hence, it is possible to take the "square root" of $\mathbf{1} + b\mathbf{G}$, which facilitates the calculations.

This method, which is quite general, is tested on a reduced model protein [26] with 54 amino acids and the Ramachandran torsion angles as its only degrees of freedom. This chain forms a three-helix bundle in its native state and exhibits an abrupt collapse transition that coincides with its folding transition. The performance of the method is studied both above and below the folding temperature, for different values of the parameters a and b. For a suitable choice of b, we find that the effective step size can be increased by a factor of three in the folded phase, compared to the unbiased $b = 0$ case. The optimal value of b corresponds to a relatively strong bias, that is an approximately local update.

1.2 The Model

In our calculations, we consider a reduced protein model [26] where each amino acid is represented by five or six atoms. The three backbone atoms N, C_α and C' are all included, whereas the side chain is represented by a single atom, C_β. The C_β atom can be hydrophobic, polar or absent, which means that there are three different types of amino acids in the model. For a schematic illustration of the chain representation, see Fig. 1.1.

All bond lengths, bond angles and peptide torsion angles (180°) are held fixed, which leaves us with two degrees of freedom per amino acid, the Ramachandran torsion angles (see Fig. 1.1).

The energy function

$$E = E_{\rm loc} + E_{\rm sa} + E_{\rm hb} + E_{\rm AA} \tag{1.2}$$

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

$$E_{\rm loc} = \frac{\epsilon_{\rm loc}}{2} \sum_i (1 + \cos 3\phi_i)\,. \tag{1.3}$$

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

$$E_{\rm sa} = \epsilon_{\rm sa} {\sum_{i<j}}' \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}, \tag{1.4}$$

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

$$E_{\rm hb} = \epsilon_{\rm hb} \sum_{ij} u(r_{ij}) v(\alpha_{ij}, \beta_{ij})\,, \tag{1.5}$$

where i and j represent H and O atoms (see Fig. 1.1), respectively, and

$$u(r_{ij}) = 5\left(\frac{\sigma_{\rm hb}}{r_{ij}}\right)^{12} - 6\left(\frac{\sigma_{\rm hb}}{r_{ij}}\right)^{10} \tag{1.6}$$

$$v(\alpha_{ij}, \beta_{ij}) = \begin{cases} \cos^2\alpha_{ij} \cos^2\beta_{ij} & \alpha_{ij}, \beta_{ij} > 90^\circ \\ 0 & \text{otherwise} \end{cases} \tag{1.7}$$

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

$$E_{\rm AA} = \epsilon_{\rm AA} \sum_{i<j} \left[\left(\frac{\sigma_{\rm AA}}{r_{ij}}\right)^{12} - 2\left(\frac{\sigma_{\rm AA}}{r_{ij}}\right)^{6}\right], \tag{1.8}$$

where both i and j represent hydrophobic C_β. In the following, kT is given in dimensionless units, in which $\epsilon_{\rm hb} = 2.8$ and $\epsilon_{\rm AA} = 2.2$. Further details of the model, including numerical values of all the parameters, can be found in Ref. [26].

In this model, we study a designed three-helix-bundle protein with 54 amino acids. In Ref. [26], it was demonstrated that this sequence indeed forms a stable three-helix bundle, except for a twofold topological degeneracy, and that it has a first-order-like folding transition that coincides with the collapse transition. It should be noted that these properties are found without resorting to the widely used but drastic Gō approximation [27], where interactions that do not favor the desired structure are ignored.

1.3 The Algorithm

We now turn to the algorithm, which we describe assuming the particular chain geometry defined in Sec. 2. That this scheme can be easily generalized to other types of chains will be evident.

Consider a segment of four adjacent amino acids k, $k+1$, $k+2$ and $k+3$ along the chain, and let the corresponding eight Ramachandran angles (see Fig. 1.1) form a vector $\bar{\phi} = (\phi_1, \ldots, \phi_n)$, where $n = 8$. A change $\delta\bar{\phi}$ of $\bar{\phi}$ will, by construction, leave all amino acids $k' < k$, as well as the N, H and C_α atoms of amino acid k, unaffected. For all amino acids $k' > k+3$ to remain unaffected too, it is sufficient to require that the three atoms C_α, C' and O of amino acid $k+3$ (see Fig. 1.1) do not move. If this condition is fulfilled, the deformation of the chain is local.

Denote the position vectors of the C_α, C' and O atoms of amino acid $k+3$ by $\mathbf{r}_I$, $I = 1, 2, 3$. A bias toward local deformations can be obtained by favoring changes $\delta\bar{\phi}$ that correspond to small values of the quantity

$$\Delta^2 = \sum_{I=1}^{3} (\delta \mathbf{r}_I)^2 \,, \tag{1.9}$$

which for small $\delta\phi_i$ can be written as

$$\Delta^2 \approx \delta\bar{\phi}^T \mathbf{G} \delta\bar{\phi} = \sum_{i,j=1}^{n} \delta\phi_i G_{ij} \delta\phi_j \,, \tag{1.10}$$

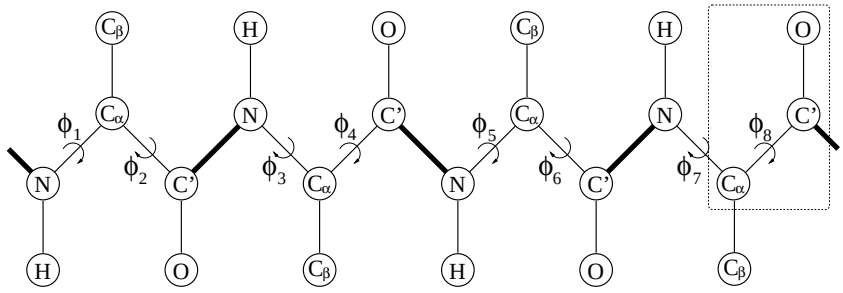

Figure 1.1: Update of the chain defined in Sec. 1.2. Eight torsion angles ϕ_i are turned. Turns such that the three atoms in the box are left unaffected are favored. Thick lines represent peptide bonds and the peptide torsion angles are fixed.

where

$$G_{ij} = \sum_{I=1}^{3} \frac{\partial \mathbf{r}_I}{\partial \phi_i} \cdot \frac{\partial \mathbf{r}_I}{\partial \phi_j} \,. \tag{1.11}$$

Note that the three vectors $\mathbf{r}_I$ can be described in terms of six independent parameters, since bond lengths and angles are fixed. This implies that the $n \times n$ matrix $\mathbf{G}$, which by construction is non-negative and symmetric, has eigenvectors with eigenvalue zero for $n = 8 > 6$. A bias toward small Δ^2 means that these soft modes are favored.

We can now define the update, which consists of the following two steps.

1. Draw a tentative new $\bar{\phi}$, $\bar{\phi}'$, from the Gaussian distribution

$$W(\bar{\phi} \to \bar{\phi}') = \frac{(\det \mathbf{A})^{1/2}}{\pi^3} \exp\left[-(\bar{\phi}' - \bar{\phi})^T \mathbf{A} (\bar{\phi}' - \bar{\phi})\right] \,, \tag{1.12}$$

where the matrix

$$\mathbf{A} = \frac{a}{2}(\mathbf{1} + b\mathbf{G}) \tag{1.13}$$

is a linear combination of the $n \times n$ unit matrix $\mathbf{1}$ and the matrix $\mathbf{G}$ defined by Eq. 1.11. The shape of this distribution depends on the parameters $a > 0$ and $b \geq 0$. The parameter b sets the degree of bias toward small Δ^2. The bias is strong for large b and disappears in the limit $b \to 0$. The parameter a is a direction-independent scale factor that is needed to control the acceptance rate. Larger a means higher acceptance rate, for fixed b. If $b = 0$, then the components $\delta\phi_i$ are independent Gaussian random numbers with zero mean and variance a^{-1}.

Note that $W(\bar{\phi} \to \bar{\phi}') \neq W(\bar{\phi}' \to \bar{\phi})$ since the matrix $\mathbf{G}$ is conformation dependent.

2. Accept/reject $\bar{\phi}'$ with probability

$$P_{\text{acc}} = \min\left(1, \frac{W(\bar{\phi}' \to \bar{\phi})}{W(\bar{\phi} \to \bar{\phi}')} \exp[-(E' - E)/kT]\right) \tag{1.14}$$

for acceptance. The factor $W(\bar{\phi}' \to \bar{\phi})/W(\bar{\phi} \to \bar{\phi}')$ is needed for detailed balance to be fulfilled, since W is asymmetric.

It should be stressed that this scheme is quite flexible. For example, it can be immediately applied to chains with nonplanar peptide torsion angles. The use of the concerted-rotation method for simulations of such chains has recently been discussed [28].

A convenient and efficient implementation of the algorithm can be obtained if one takes the "square root" of the matrix $\mathbf{A}$, which can be done because $\mathbf{A}$ is symmetric and positive definite. More precisely, it is possible to find a lower triangular matrix $\mathbf{L}$ (with nonzero elements only on the diagonal and below) such that

$$\mathbf{A} = \mathbf{L}\mathbf{L}^T . \tag{1.15}$$

An efficient routine for this so-called Cholesky decomposition can be found in [29].

1.3.1 Implementing step 1

Given the Cholesky decomposition of the matrix $\mathbf{A}$, the first step of the algorithm can be implemented as follows.

- Draw a $\bar{\psi} = (\psi_1, \ldots, \psi_n)$ from the distribution $P(\bar{\psi}) \propto \exp(-\bar{\psi}^T\bar{\psi})$. The components ψ_i are independent Gaussian random numbers and can be generated, for example, by using the Box-Muller method

$$\psi_i = (-\ln R_1)^{1/2} \cos 2\pi R_2 , \tag{1.16}$$

where R_1 and R_2 are uniformly distributed random numbers between 0 and 1.

- Given $\bar{\psi}$, solve the triangular system of equations

$$\mathbf{L}^T \delta\bar{\phi} = \bar{\psi} \tag{1.17}$$

for $\delta\bar{\phi}$. It can be readily verified that the $\delta\bar{\phi} = \bar{\phi}' - \bar{\phi}$ obtained this way has the desired distribution Eq. 1.12.

1.3.2 Implementing step 2

The Cholesky decomposition is also useful when calculating the acceptance probability in the second step of the algorithm. The factor $W(\bar{\phi} \to \bar{\phi}')$ can be easily computed by using that

$$(\det \mathbf{A})^{1/2} = \prod_{i=1}^{n} L_{ii} \tag{1.18}$$

and that $\exp[-(\bar{\phi}' - \bar{\phi})^T \mathbf{A}(\bar{\phi}' - \bar{\phi})] = \exp(-\bar{\psi}^T \bar{\psi})$. The reverse probability $W(\bar{\phi}' \to \bar{\phi})$ depends on $\mathbf{A}(\bar{\phi}')$ and can be obtained in a similar way, if one makes a Cholesky decomposition of that matrix, too.

1.3.3 Pivot update

Previous simulations [26] of the model protein defined in Sec. 1.2 were carried out by using simulated tempering with pivot moves as the elementary conformation update. With this algorithm, the system was successfully studied down to temperatures just below the folding transition. However, the performance of the pivot update, where a single angle ϕ_i is turned, deteriorates in the folded phase. What we hope is that the exploration of this phase can be made more efficient by alternating the pivot moves with moves of the type described previously.

1.4 Results

The character of the proposed update depends strongly on the bias parameter b. The suggested steps have a random direction if $b = 0$. The distribution $W(\bar{\phi} \to \bar{\phi}')$ in Eq. 1.12 is, by contrast, highly asymmetric in the limit $b \to \infty$, with nonzero width only in directions corresponding to eigenvalue zero of the matrix $\mathbf{G}$. In particular, this implies that the reverse probability $W(\bar{\phi}' \to \bar{\phi})$ in the acceptance criterion Eq. 1.14 tends to be small for large b.

For the acceptance rate to be reasonable, it is necessary to use a very small step size if b is small or large. The question is whether the step size can be increased by a better choice of b. To find that out, we performed a set of simulations of the three-helix-bundle protein defined in Sec. 1.2 for different a and b. Two different temperatures were studied, $kT = 0.6$ and 0.7, one on either side of the folding temperature $kT_f \approx 0.66$ [26].

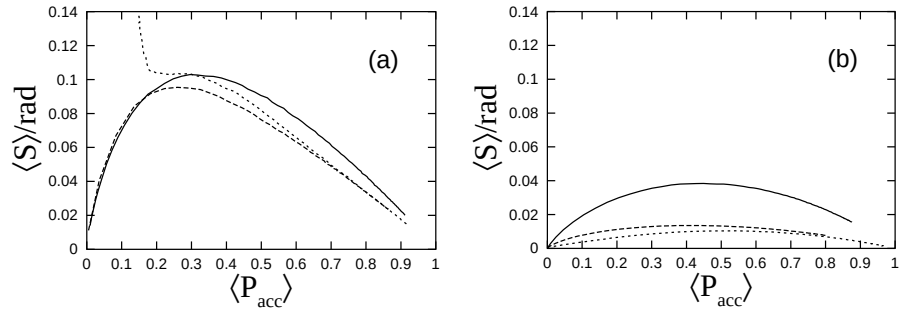

Figure 1.2: Average step size, $\langle S \rangle$, against average acceptance rate, $\langle P_{\rm acc} \rangle$, for different updates at (a) $kT = 0.7$ and (b) $kT = 0.6$. Shown are results for the $b = b_{\rm max}$ (full lines), $b = 0$ (dashed lines) and pivot (dotted lines) updates.

In these runs, we monitored the step size S, where

$$S = |\delta\bar{\phi}| = \left[\sum_{i=1}^{n} (\delta\phi_i)^2 \right]^{1/2} \tag{1.19}$$

for accepted moves and $S = 0$ for rejected ones. Measurements were taken only when the $n = 8$ angles all were in the segment that makes the middle helix of the three. We focus on this segment because it is the most demanding part to update.

The average step size, $\langle S \rangle$, depends strongly on b. A rough optimization of b was carried out by maximizing $\langle S \rangle$ as a function of a for different fixed $b = 10^k$ (k integer). The best values found were $b_{\rm max} = 10$ (rad/Å)^2 and $b_{\rm max} = 0.1$ (rad/Å)^2 at $kT = 0.6$ and $kT = 0.7$, respectively. Note that the preferred degree of bias is higher in the folded phase.

In Fig. 1.2, we show $\langle S \rangle$ against the average acceptance rate, $\langle P_{\rm acc} \rangle$, for $b = 0$ and $b = b_{\rm max}$ at the two temperatures; $\langle P_{\rm acc} \rangle$ is an increasing function of a for fixed b and T. Also shown are the corresponding results for the pivot update, where only one angle ϕ_i is turned ($S = |\delta\phi_i|$ if the change is accepted). At the higher temperature, we find that the $b = b_{\rm max}$ and $b = 0$ updates show similar behaviors. The pivot update is somewhat better and has its maximum $\langle S \rangle$ at low $\langle P_{\rm acc} \rangle$, where the proposed change $\delta\phi_i$ is drawn from the uniform distribution between 0 and 2π. This is consistent with the finding [7] that the pivot update is a very efficient method for self-avoiding walks, in spite of a low acceptance rate. The situation is different at the lower temperature, which is

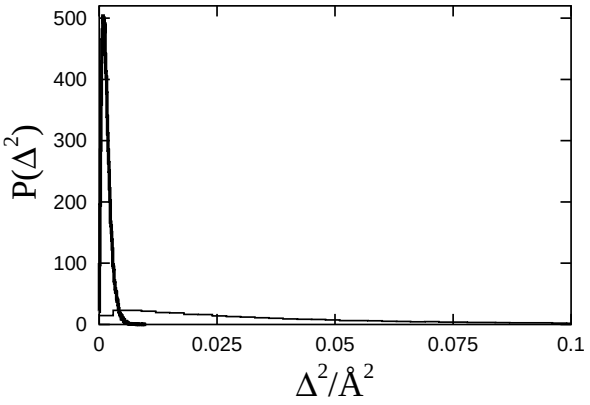

Figure 1.3: Distributions of Δ^2 (see Eq. 1.9) for the $b = b_{\max}$ (thick line) and $b = 0$ (thin line) updates at $kT = 0.6$. The values used for the parameter a correspond to maximum $\langle S \rangle$.

much harder to simulate. Here, the $b = b_{\max}$ update is the best. The maximum $\langle S \rangle$ is approximately three times higher for this method than for the other two. This shows that the biasing probability Eq. 1.12 is indeed useful in the folded phase.

The $b = 0$ update can be compared with the moves used by Shimada *et al.* [30] in a recent all-atom study of kinetics and thermodynamics for the protein crambin with 46 amino acids. These authors updated sets of two, four or six backbone torsion angles, using independent Gaussian steps with a standard deviation of 2°. Our $b = 0$ update has maximum $\langle S \rangle$ at $a \approx 6400$ (rad)$^{-2}$ for $kT = 0.6$, which corresponds to a standard deviation of 0.7°. This value is in line with that used by Shimada *et al.*, since we turn eight angles.

How local is the method for $b = b_{\max}$? To get an idea of that, we calculated the distribution of Δ^2 (see Eq. 1.9) for accepted moves, for $b = b_{\max}$ and $b = 0$ at $kT = 0.6$. As was previously the case, we restricted ourselves to angles in the middle helix. The two distributions are shown in Fig. 1.3 and we see that the one corresponding to $b = b_{\max}$ is sharply peaked near $\Delta^2 = 0$. This shows that the $b = b_{\max}$ update is much more local than the unbiased $b = 0$ update, although the average step size, $\langle S \rangle$, is considerably larger for $b = b_{\max}$.

So far, we have discussed static (one-step) properties of the updates. We also

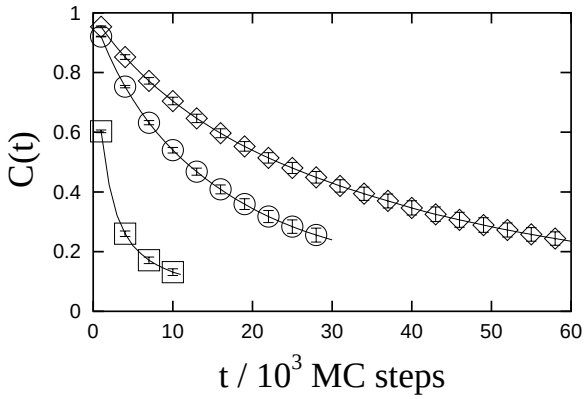

Figure 1.4: The autocorrelation function $C(t)$ (see the text) at $kT = 0.6$ for the $b = b_{\max}$ ($\Box$), $b = 0$ ($\circ$) and pivot ($\diamond$) updates. Step-size parameters correspond to maximum $\langle S \rangle$.

estimated the dynamic autocorrelation function

$$C_i(t) = \frac{\langle \cos\phi_i(t) \cos\phi_i(0) \rangle - \langle \cos\phi_i(0) \rangle^2}{\langle \cos^2\phi_i(0) \rangle - \langle \cos\phi_i(0) \rangle^2} \tag{1.20}$$

for different ϕ_i. This measurement is statistically very difficult at low temperatures. However, the sixteen most central angles ϕ_i in the sequence, all belonging to the middle helix, were found to be effectively frozen at $kT = 0.6$, and the time scale for the small fluctuations of these angles about their mean values was possible to estimate. In Fig. 1.4, we show the average $C_i(t)$ for these sixteen angles, denoted by $C(t)$, against Monte Carlo time t, for the $b = 0$, $b = b_{\max}$ and pivot updates. One time unit corresponds to one elementary move, accepted or rejected, at a random position along the chain. We see that $C(t)$ decays most rapidly for the $b = b_{\max}$ update. So, the larger step size of this update does make the exploration of these degrees of freedom more efficient.

Let us finally comment on our choice to work with $n = 8$ angles. This number can be easily altered and some calculations were done with $n = 6$ and $n = 7$, too. For $n = 6$, the performance was worse, which is not unexpected because there are no soft modes available; there are not more variables than constraints. The results obtained for $n = 7$ were, by contrast, comparable to or slightly better than the $n = 8$ results.

1.5 Discussion

Straightforward Monte Carlo updates of torsional degrees of freedom tend to cause large changes in the global structure of the chains unless the step size is made very small, which is a problem in simulations of dense polymer systems. The strictly local concerted-rotation approach provides a solution to this problem but is rather complicated to implement. In this paper, we have discussed a method that may be less powerful but is much easier to implement, which suppresses rather than eliminates nonlocal deformations.

The method is flexible and not much harder to implement than simple unbiased updates. However, compared to such updates, it has two distinct advantages: the step size can be increased and the update becomes more local, as shown by our simulations of the three-helix-bundle protein in its folded phase.

Making the update more local is important in order to be able to increase the step size and thereby improve the efficiency. At the same time, it makes the dynamics more realistic; the proposed method is, in contrast to the other methods mentioned, tailored to avoid drastic deformations both locally and globally. Therefore, although this paper was focused on thermodynamic simulations, it should be noted that this method may be useful for kinetic studies, too.

Acknowledgments

This work was supported in part by the Swedish Foundation for Strategic Research. G.F. acknowledges support from Universitá degli studi di Cagliari and the EU European Social Fund.

References

[1] B.A. Berg and T. Neuhaus, *Phys. Rev. Lett.* **68**, 9 (1992).

[2] U.H.E. Hansmann and Y. Okamoto, *J. Comput. Chem.* **14**, 1333 (1993).

[3] A.P. Lyubartsev, A.A. Martsinovski, S.V. Shevkunov and P.V. Vorontsov-Velyaminov, *J. Chem. Phys.* **96**, 1776 (1992).

[4] E. Marinari and G. Parisi, *Europhys. Lett.* **19**, 451 (1992).

[5] A. Irbäck and F. Potthast, *J. Chem. Phys.* **103**, 10298 (1995).

[6] M. Lal, *Molec. Phys.* **17**, 57 (1969).

[7] N. Madras and A.D. Sokal, *J. Stat. Phys.* **50**, 109 (1988).

[8] F.A. Escobedo and J.J. de Pablo, *J. Chem. Phys.* **102**, 2636 (1995).

[9] D. Frenkel and B. Smit, *Understanding Molecular Simulations*, (Academic, New York, 1996).

[10] M. Vendruscolo, *J. Chem. Phys.* **106**, 2970 (1997).

[11] C.D. Wick and J.I. Siepmann, *Macromolecules* **33**, 7207 (2000).

[12] Z. Chen and F.A. Escobedo, *J. Chem. Phys.* **113**, 11382 (2000).

[13] D. Frenkel, G.C.A.M. Mooij and B. Smit, *J. Phys.: Condens. Matter* **4**, 3053 (1992).

[14] J.J. de Pablo, M. Laso and U.W. Suter, *J. Chem. Phys.* **96**, 6157 (1992).

[15] N. Gō and H.A. Scheraga, *Macromolecules* **3**, 178 (1970).

[16] L.R. Dodd, T.D. Boone and D.N. Theodorou, *Molec. Phys.* **78**, 961 (1993).

[17] D. Hoffmann and E.-W. Knapp, *Eur. Biophys. J.* **24**, 387 (1996).

[18] P.V.K. Pant and D.N. Theodorou, *Macromolecules* **28**, 7224 (1995).

[19] V.G. Mavrantzas, T.D. Boone, E. Zervopoulou and D.N. Theodorou, *Macromolecules* **32**, 5072 (1999).

[20] E. Leonitidis, J.J. de Pablo, M. Laso and U.W. Suter, *Adv. Polym. Sci.* **116**, 283 (1994).

[21] M.W. Deem and J.S. Bader, *Molec. Phys.* **87**, 1245 (1996).

[22] M.G. Wu and M.W. Deem, *Molec. Phys.* **97**, 559 (1999).

[23] M.G. Wu and M.W. Deem, *J. Chem. Phys.* **111**, 6625 (1999).

[24] A. Uhlherr, *Macromolecules* **33**, 1351 (2000).

[25] D. Manocha and J.F. Canny, *IEEE Trans. Rob. Autom.* **10**, 648 (1994).

[26] A. Irbäck, F. Sjunnesson and S. Wallin, *Proc. Natl. Acad. Sci. USA* **97**, 13614 (2000).

[27] N. Gō and H. Taketomi, *Proc. Natl. Acad. Sci. USA* **75**, 559 (1978).

[28] A.R. Dinner, *J. Comput. Chem.* **21**, 1132 (2000).

[29] W.H. Press, S.A. Teukolsky, W.T. Vetterling and B.P. Flannery, *Numerical Recipes in C* (Cambridge University Press, Cambridge, 1992).

[30] J. Shimada, E.L. Kussell and E.I. Shakhnovich, e-print cond-mat/0011369.

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

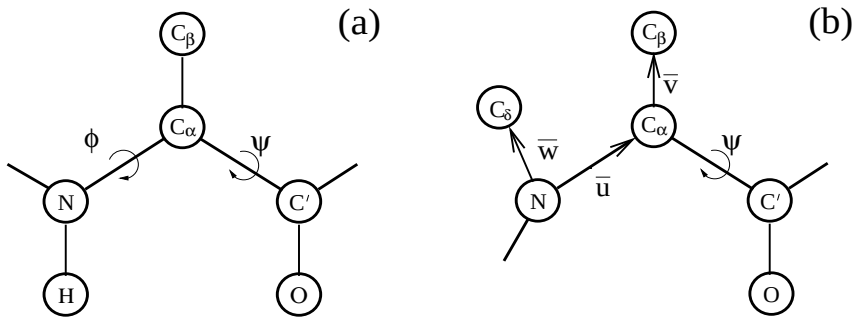

Figure 2.1: (a) Schematic figure showing the common geometric representation for all amino acids except glycine and proline. (b) The representation of proline. The C_δ atom is assumed to lie in the plane of the N, C_α and C_β atoms. The N-C_δ bond vector $\bar{w}$ is given by $\bar{w} = -0.596\bar{u} + 0.910\bar{v}$, 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

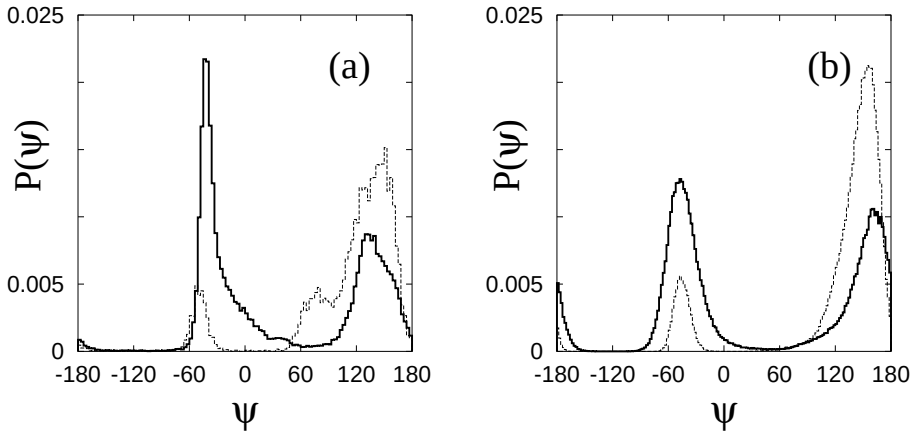

Figure 2.2: (a) Distributions of the Ramachandran angle ψ, based on PDB data. The full (dashed) line represents non-glycine, non-proline amino acids that are followed by a non-proline (proline) in the sequence. (b) The corresponding histograms for the model, as obtained by simulations of Gly-X-X (full 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).

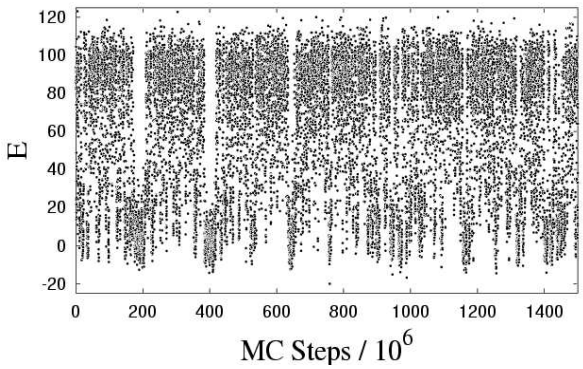

Figure 2.3: Monte Carlo evolution of the energy in a simulated-tempering run.

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.

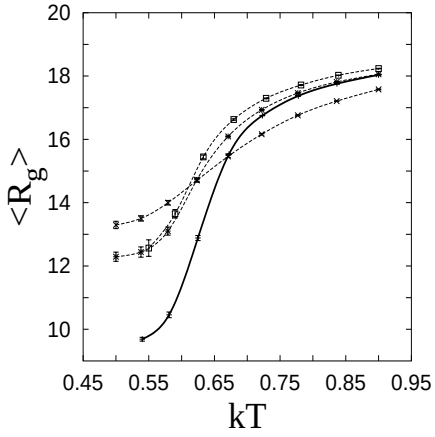

Figure 2.4: The radius of gyration (in Å) against temperature. Full and dashed lines represent the protein A sequence and the three random sequences (see the text), respectively.

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-

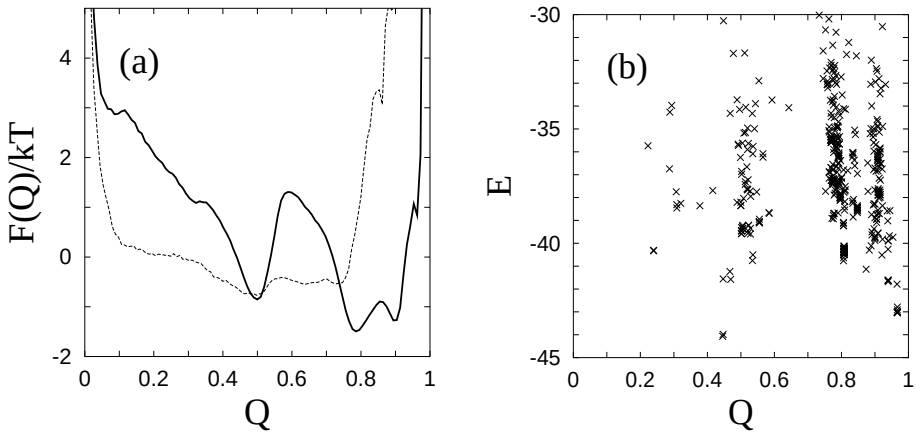

Figure 2.5: (a) Free-energy profile $F(Q) = -kT \ln P(Q)$ at $kT = 0.54$ (full line), where $P(Q)$ is the probability distribution of Q. Also shown (dashed line) is the result for one of the random sequences at $kT = 0.50$. (b) Q, E scatter plot for quenched conformations with low energy.

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).

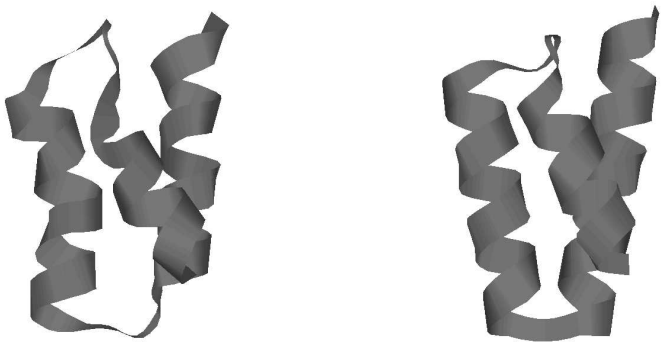

Figure 2.6: Schematic illustrations of the native structure (left) and our minimum-energy structure for the native topology (right). Drawn with RasMol [33].

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

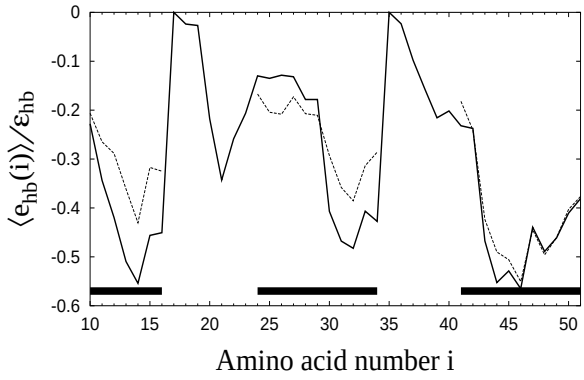

Figure 2.7: Hydrogen-bond profile showing the normalized average energy of α-helical hydrogen bonds, $\langle e_{\rm hb}(i)\rangle/\epsilon_{\rm hb}$, against amino acid number i, at $kT = 0.58$. The full line represents the protein A sequence, whereas the dashed lines represent the corresponding three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

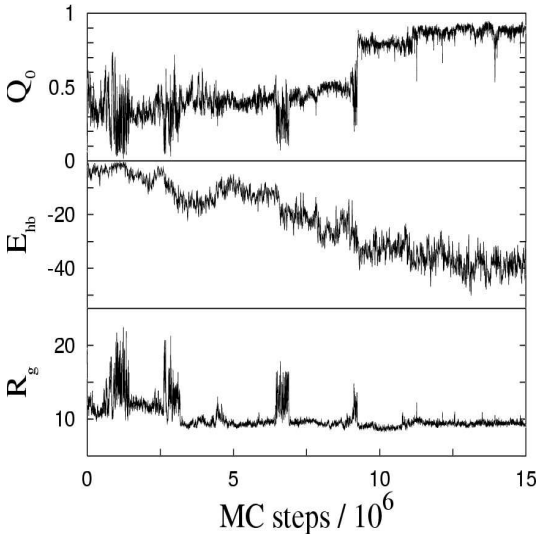

Figure 2.8: Monte Carlo evolution of the similarity parameter Q_0 (top), the hydrogen-bond energy E_{hb} (middle) and the radius of gyration R_{g} (bottom) in a kinetic simulation at $kT = 0.54$.

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.

Two-State Folding over a Weak Free-Energy Barrier

Paper III

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/

Biophysical Journal. **85**, 1457-1465 (2003)

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.

3.1 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.

3.2 Model and Methods

3.2.1 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

$$E = E_{\rm loc} + E_{\rm ev} + E_{\rm hb} + E_{\rm hp} \tag{3.1}$$

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

$$E_{\rm loc} = \frac{\epsilon_\phi}{2}\sum_i (1+\cos 3\phi_i) + \frac{\epsilon_\psi}{2}\sum_i (1+\cos 3\psi_i)\,. \tag{3.2}$$

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

$$E_{\rm ev} = \epsilon_{\rm ev}{\sum_{i<j}}' \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}, \tag{3.3}$$

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

$$E_{\rm hb} = \epsilon_{\rm hb}\sum_{ij} u(r_{ij})v(\alpha_{ij},\beta_{ij})\,, \tag{3.4}$$

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

$$u(r) \;=\; 5\left(\frac{\sigma_{\rm hb}}{r}\right)^{12} - 6\left(\frac{\sigma_{\rm hb}}{r}\right)^{10} \tag{3.5}$$

$$v(\alpha,\beta) \quad = \quad \begin{cases} \cos^2\alpha\cos^2\beta & \alpha,\beta > 90^\circ \\ 0 & \text{otherwise} \end{cases} \tag{3.6}$$

The sum in Eq. 3.4 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

$$E_{\rm hp} = \epsilon_{\rm hp} \sum_{i<j} \left[\left(\frac{\sigma_{\rm hp}}{r_{ij}} \right)^{12} - 2 \left(\frac{\sigma_{\rm hp}}{r_{ij}} \right)^{6} \right], \tag{3.7}$$

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

$$Q = \max\left[\exp\left(-\delta_{\mathrm{FU}}^2/(10\text{Å})^2\right), \exp\left(-\delta_{\mathrm{BU}}^2/(10\text{Å})^2\right)\right] , \tag{3.8}$$

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 $kT_{\mathrm{f}} = 0.6585 \pm 0.0006$.

3.2.2 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].

3.2.3 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

$$X(T) = \frac{X_{\mathrm{u}} + X_{\mathrm{n}} K(T)}{1 + K(T)}\,, \tag{3.9}$$

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. 3.9 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

$$\frac{\partial P(r,t)}{\partial t} = \frac{\partial}{\partial r}\left[D(r)\left(\frac{\partial P(r,t)}{\partial r} + \frac{P(r,t)}{kT}\frac{\partial F(r)}{\partial r}\right)\right]\,, \tag{3.10}$$

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. 3.10. Such an analysis has been successfully carried through for a lattice protein [27].

The relaxation behavior predicted by Eq. 3.10 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. 3.10 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 3.1: Parameters ΔE and T_m obtained by fitting results from our thermodynamic simulations to the two-state expression in Eq. 3.9. 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. 3.8). The fits are performed using seven data points in the temperature interval $0.95\,T_\mathrm{f} \leq T \leq 1.11\,T_\mathrm{f}$.

3.3 Results

3.3.1 Thermodynamics

In our thermodynamic analysis, we study the five different quantities listed in Table 3.1. 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. 3.9).

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. 3.1 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 3.1 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

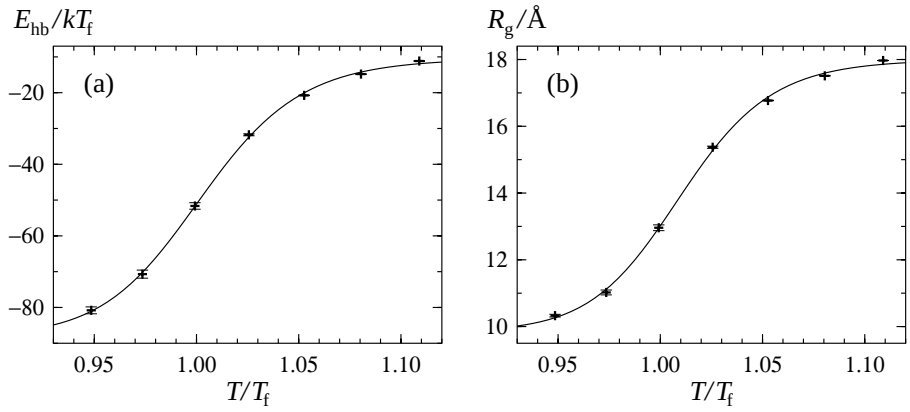

Figure 3.1: Temperature dependence of (a) the hydrogen-bond energy $E_{\rm hb}$ and (b) the radius of gyration $R_{\rm g}$. The lines are fits to Eq. 3.9.

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. 3.2 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.

3.3.2 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. 3.1).

In the absence of a clear free-energy barrier (see Fig. 3.2), 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

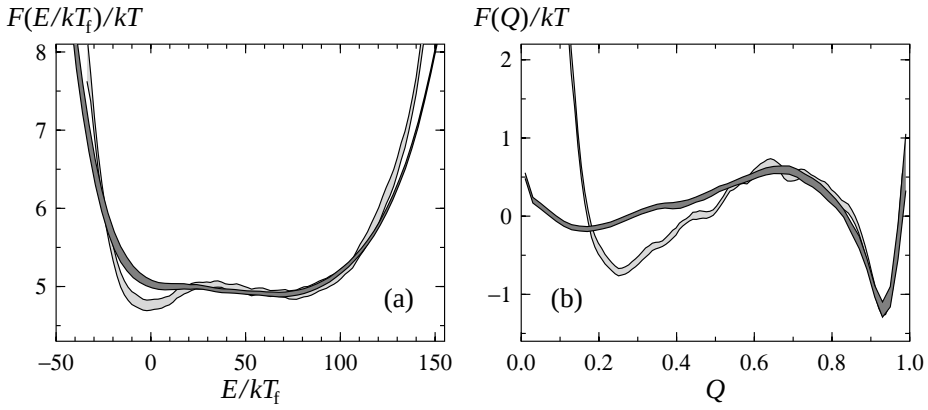

Figure 3.2: Free-energy profiles at $T = T_\mathrm{f}$ for (a) the energy E and (b) the nativeness Q (dark bands). The light-grey bands show free energies F_b for block averages (see Eq. 3.12), using a block size of $\tau_\mathrm{b} = 10^6$ MC steps. Each band is centered around the expected value and 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. 3.10 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. 3.3. 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. 3.4 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

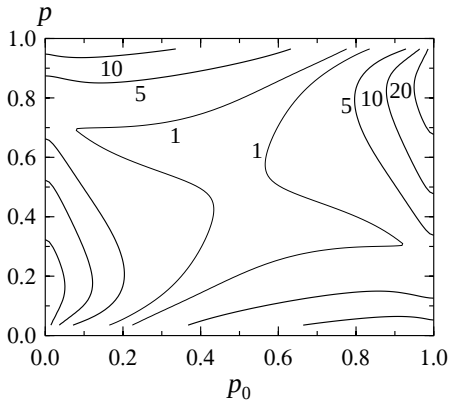

Figure 3.3: Level diagram showing the deviation (in %) from a single exponential for diffusion in energy in a square well, based on the exact solution in Appendix A. The system relaxes at temperature T, starting from the equilibrium distribution at temperature T_0. p is defined as $p = (\langle E\rangle - E_\mathrm{n})/\Delta E_\mathrm{sw}$, 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. 3.4). 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.

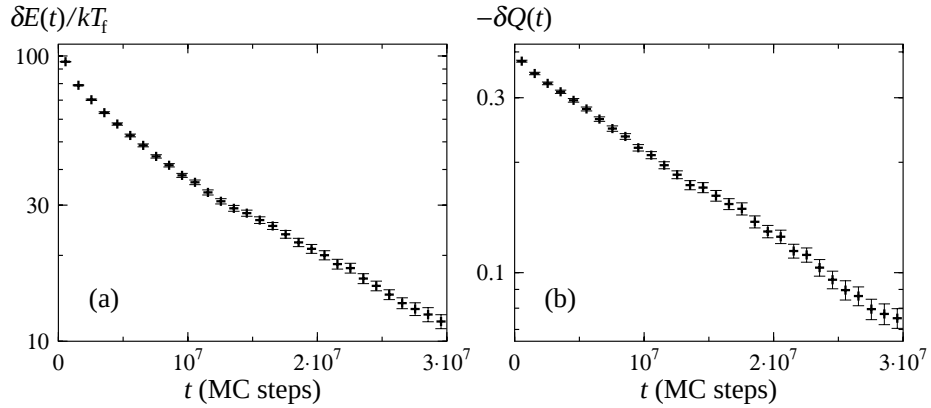

Figure 3.4: Relaxation behavior of the three-helix-bundle protein at the folding temperature T_f, starting from the equilibrium ensemble at $T_0 \approx 1.06 T_\mathrm{f}$. (a) $\delta E(t) = E(t) - \langle E \rangle$ against simulation time t, where $E(t)$ is the average E after t MC steps (3000 runs) 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,

$$C_\mathrm{hb}(t) = \langle E_\mathrm{hb}(t) E_\mathrm{hb}(0) \rangle - \langle E_\mathrm{hb}(0) \rangle^2 \,, \tag{3.11}$$

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].

3.3.3 Relaxation-Time Predictions

We now turn to the question of whether the observed relaxation time can be predicted based on the diffusion equation, Eq. 3.10. 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

$$b(t) = \frac{1}{\tau_\mathrm{b}} \sum_{t \le s < t+\tau_\mathrm{b}} r(s) \qquad\qquad t = 0, \tau_\mathrm{b}, 2\tau_\mathrm{b}, \ldots \tag{3.12}$$

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 3.2, 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. 3.15 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 3.2: 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. 3.4, 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. 3.5a. 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. 3.2. Fig. 3.5b 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. 3.10 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 3.2 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].

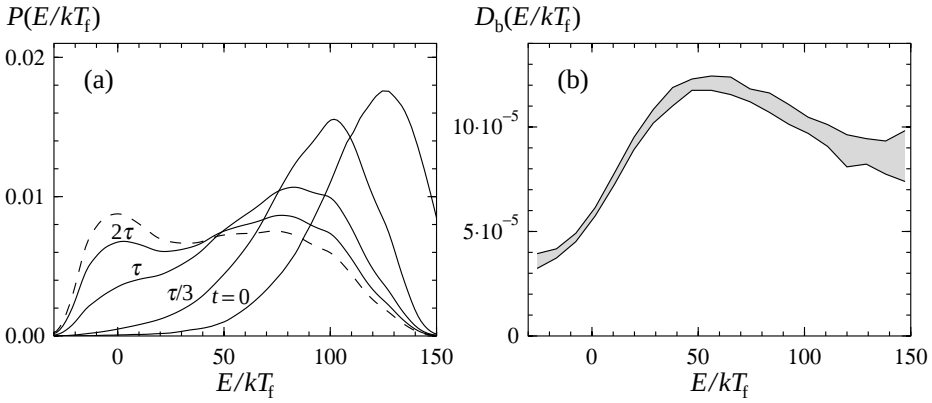

Figure 3.5: (a) Numerical solution of Eq. 3.10 with the energy as reaction coordinate. The distribution $P(E,t)$ is shown for $t = 0$, $\tau/3$, τ and 2τ (full lines), where τ is the relaxation time. The dashed line is the equilibrium distribution. The diffusion coefficient $D_{\rm b}(E)$ and the potential $F_{\rm b}(E)$ (light-gray band in Fig. 3.2a) were both determined from numerical simulations, using a block size of $\tau_{\rm b} = 10^6$ MC steps (see Eq. 3.12). (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.

3.4 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. 3.10 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. 3.10 then becomes

$$\frac{\partial P(E,t)}{\partial t} = \frac{\partial}{\partial E}\left[D\left(\frac{\partial P(E,t)}{\partial E} + \delta\beta P(E,t)\right)\right] . \tag{3.13}$$

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. 3.13 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

$$E(t) = \langle E\rangle + \sum_{k=1}^{\infty} A_k \mathrm{e}^{-t/\tau_k} , \tag{3.14}$$

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

$$1/\tau_k = \frac{D}{\Delta E_{\rm sw}^2}\left(\pi^2 k^2 + \tfrac{1}{4}\delta\beta^2 \Delta E_{\rm sw}^2\right) \tag{3.15}$$

and the expansion coefficients by

$$A_k = B_k \Delta E_{\rm sw} \frac{\pi^2 k^2 \left(\delta\beta - \delta\beta_0\right)\Delta E_{\rm sw}}{\left(\pi^2 k^2 + (\delta\beta_0 - \tfrac{1}{2}\delta\beta)^2 \Delta E_{\rm sw}^2\right)\left(\pi^2 k^2 + \tfrac{1}{4}\delta\beta^2\Delta E_{\rm sw}^2\right)^2} , \tag{3.16}$$

where

$$B_k = \frac{4\delta\beta_0 \Delta E_{\rm sw}}{\sinh \tfrac{1}{2}\delta\beta_0 \Delta E_{\rm sw}} \times \begin{cases} \cosh\left(\tfrac{1}{2}(\delta\beta_0 - \tfrac{1}{2}\delta\beta)\Delta E_{\rm sw}\right)\cosh \tfrac{1}{4}\delta\beta\Delta E_{\rm sw} & \text{if } k \text{ odd} \\ \sinh\left(\tfrac{1}{2}(\delta\beta_0 - \tfrac{1}{2}\delta\beta)\Delta E_{\rm sw}\right)\sinh \tfrac{1}{4}\delta\beta\Delta E_{\rm sw} & \text{if } k \text{ even} \end{cases} \tag{3.17}$$

Finally, the equilibrium average is

$$\langle E\rangle = \frac{E_{\rm n} + E_{\rm u}}{2} + \frac{1}{\delta\beta} - \frac{\Delta E_{\rm sw}}{2}\coth \tfrac{1}{2}\delta\beta\Delta E_{\rm sw} , \tag{3.18}$$

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

$$A_k \approx B_k \Delta E_{\rm sw} \frac{\pi^2 k^2 (\delta\beta - \delta\beta_0)\Delta E_{\rm sw}}{\left(\pi^2 k^2 + \frac{1}{4}\delta\beta^2 \Delta E_{\rm sw}^2\right)^3} \tag{3.19}$$

with

$$B_k \approx \frac{4\delta\beta_0 \Delta E_{\rm sw}}{\sinh \frac{1}{2}\delta\beta_0 \Delta E_{\rm sw}} \times \begin{cases} \cosh^2 \frac{1}{4}\delta\beta \Delta E_{\rm sw} & \text{if } k \text{ odd} \\ \sinh^2 \frac{1}{4}\delta\beta \Delta E_{\rm sw} & \text{if } k \text{ even} \end{cases} \tag{3.20}$$

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. 3.10 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. 3.10 with suitable finite differences, we obtain

$$\frac{d\mathbf{p}}{dt} = \mathbf{A}\mathbf{p}(t)\,, \tag{3.21}$$

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

$$\begin{aligned} (\mathbf{A}\mathbf{p}(t))_j &= \frac{1}{\Delta r^2}\left[D_{j+1/2}(p_{j+1}(t) - p_j(t)) - D_{j-1/2}(p_j(t) - p_{j-1}(t))\right] \\ &+ \frac{1}{4kT\Delta r^2}\left[D_{j+1}p_{j+1}(t)(F_{j+2} - F_j) - D_{j-1}p_{j-1}(t)(F_j - F_{j-2})\right] \end{aligned} \tag{3.22}$$

Let now $\mathbf{p}^n = \mathbf{p}(t_n)$, where $t_n = n\Delta t$. By applying the trapezoidal rule for integration to Eq. 3.21, we obtain

$$\mathbf{p}^{n+1} - \mathbf{p}^n = \frac{\Delta t}{2}\left(\mathbf{A}\mathbf{p}^n + \mathbf{A}\mathbf{p}^{n+1}\right)\,. \tag{3.23}$$

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 IV

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/

Proteins: Structure, Function, and Genetics **54**, 8-12 (2004)

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.

4.1 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}.

4.2 Materials and Methods

4.2.1 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

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

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 4.1 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 4.1: Amino acids 9 to 35 for Z_{SPA-1} and the wild-type Z domain.

4.2.2 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.

4.3 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

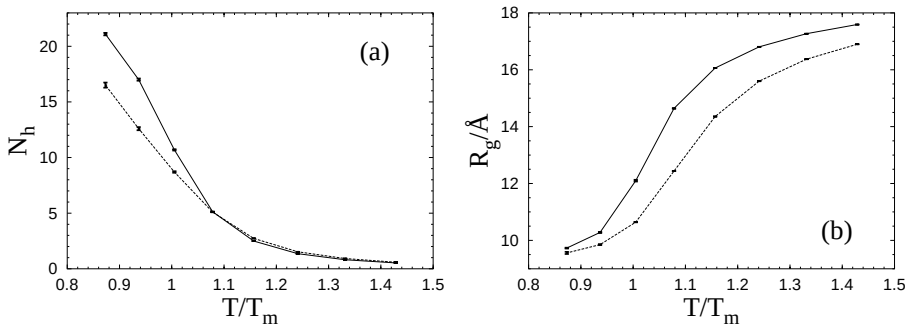

Figure 4.1: Helix formation and chain collapse for the Z_{SPA-1} sequence (dashed line) and the wild-type sequence (full line). (a) The number of helical amino acids, N_h, against temperature. (b) The radius of gyration (calculated over all backbone atoms), R_g, against temperature. T_m denotes the melting 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 4.1a 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 4.1b 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 4.1a shows that chain collapse occurs before helix formation for Z_{SPA-1}. The results in Figures 4.1a and 4.1b 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.

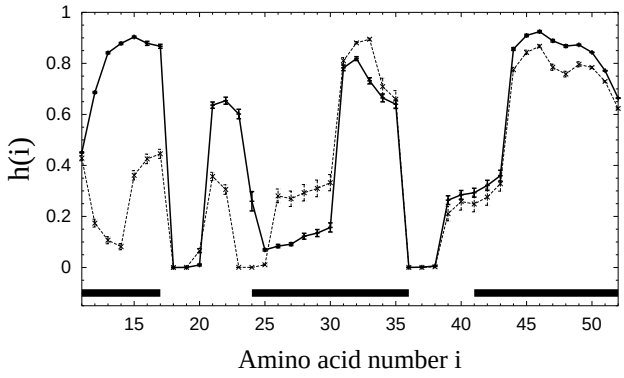

Figure 4.2: Helix content along the chain, $h(i)$, for the Z_{SPA-1} sequence (dashed line) and the wild-type sequence (full line) at $T = 0.87\,T_m$, where T_m is the melting temperature for the wild-type sequence. $h(i)$ denotes the probability that amino acid i is helical (for the definition of helical, 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 4.2 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

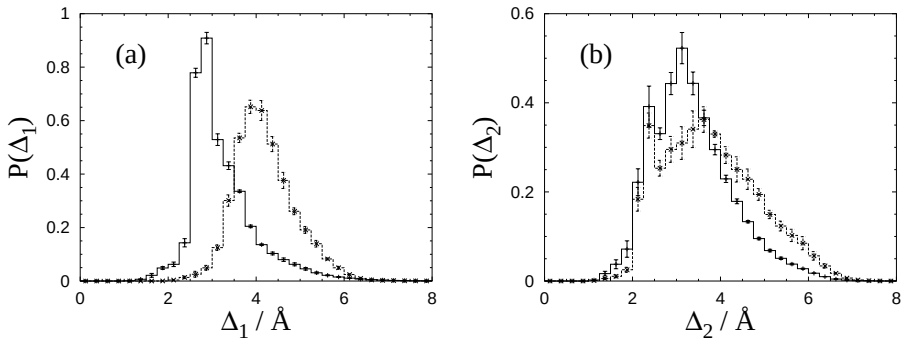

Figure 4.3: RMSD distributions for the Z_{SPA-1} sequence (dashed line) and the wild-type sequence (full line). (a) The distribution of Δ_1 (amino acids 9–31). (b) The distribution of Δ_2 (amino acids 32-54). Both Δ_1 and Δ_2 are backbone RMSDs. The temperature is the same as in Figure 4.2.

I contains seven mutations (see Table 4.1), 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 4.3 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 4.2, 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.

4.4 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.

Oligomerization of Amyloid Aβ_{16-22} Peptides Using Hydrogen Bonds and Hydrophobicity Forces

Paper V

Oligomerization of Amyloid Aβ_{16-22} Peptides Using Hydrogen Bonds and Hydrophobicity Forces

Giorgio Favrin, Anders Irbäck and Sandipan Mohanty Complex

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

Abstract:
The 16–22-amino acid fragment of the β-amyloid peptide associated with the Alz-heimer's disease, Aβ, is capable of forming amyloid fibrils. Here we analyze the aggregation mechanism of Aβ_{16-22} peptides by unbiased thermodynamic simulations at the atomic level of systems of one, three or six Aβ_{16-22} peptides. We find that the Aβ_{16-22} peptide is unstructured in isolation, whereas the three- and six-chain systems form aggregated structures with a high content of antiparallel β-sheet structure, which is consistent with experimental data. The aggregated structures can have many different shapes, but certain particularly stable shapes can be identified.

5.1 Introduction

The fibrillar aggregates that characterize amyloid diseases, such as the Alzheimer's disease, are formed by specific peptides or proteins. However, it is known that several non-disease-related proteins are capable of forming similar amyloid structures [1, 2], and that the aggregation of such proteins can be cytotoxic [3]. This suggests, first, that polypeptide chains have a general tendency to form amyloid structures and, second, that natural proteins should have evolved mechanisms to avoid this tendency. Such mechanisms have indeed been proposed [4–6]. The propensity of a given polypeptide chain to form amyloid fibrils depends, nevertheless, on its amino acid sequence [7–11]. Recently, a three-amino acid motif promoting amyloidogenesis was extracted by mutagenesis analysis [12].

While the structure of amyloid fibrils is not known in atomic detail, there is ample evidence from X-ray fiber diffraction studies that the core of the typical amyloid fibril is composed of β-sheets whose strands run perpendicular to the fibril axis [13]. More detailed information is available, for example, for fibrils made from different fragments of the Alzheimer's Aβ peptide. In particular, there is evidence from solid-state NMR studies for a parallel organization of the β-strands in $A\beta_{10-35}$ [14] and $A\beta_{1-40}$ [15] fibrils, and for an antiparallel organization in $A\beta_{34-42}$ [16], $A\beta_{11-25}$ [17] and $A\beta_{16-22}$ fibrils [18, 19].

Small peptides like $A\beta_{16-22}$ are well suited as model systems for probing the mechanisms of aggregation and fibril formation, and are being studied not only *in vitro* but also *in silico*. Computational studies of the oligomerization properties of different peptide systems have been performed using both simplified [20–23] and atomic [24–28] models. These studies have provided useful insights into the behavior of these systems. To properly sample multichain systems is, nevertheless, a computational challenge. This goal needs to be achieved in order to exploit the full potential of numerical simulations.

Here we investigate the formation and properties of $A\beta_{16-22}$ oligomers by unbiased Monte Carlo simulations of systems with up to six chains, using a sequence-based atomic model. The same model has previously been used to study the folding of individual peptides [29–31]. It was shown that this model is able to fold five different peptides, two α-helix and three β-sheet peptides, for one and the same choice of parameters. The calculated melting behaviors were, moreover, in good agreement with experimental data for all these five peptides.

5.2 Model and Methods

The amino acid sequence of the peptide we study, Aβ_{16-22}, is Lys-Leu-Val-Phe-Phe-Ala-Glu. Experimentally, the Aβ_{16-22} peptide was studied [18, 19] with acetyl and amide groups at the N and C termini, respectively. In our calculations, we ignore these end groups, which are relatively small.

We study systems of one, three or six Aβ_{16-22} peptides. The multichain systems are contained in periodic boxes. The box sizes are $(35\text{Å})^3$ and $(44\text{Å})^3$ for three and six chains, respectively, corresponding to a constant peptide concentration. For computational efficiency, the peptide concentration is taken to be high.

Except at the chain ends, our model contains all atoms of the chains [29–31], including hydrogens. All bond lengths, bond angles and peptide torsion angles (180°) are held fixed, so that each amino acid only has the Ramachandran torsion angles ϕ, ψ and a number of side-chain torsion angles as its degrees of freedom. Numerical values of the geometrical parameters held constant can be found elsewhere [29].

The interaction potential

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

is composed of four terms, which we describe next. Energy parameters are given on a scale such that a temperature of $T = 300\,\text{K}$ corresponds to $kT \approx 0.44$ (k is Boltzmann's constant) [31].

The first term in Eq. 5.1, E_{ev}, represents excluded-volume effects and has the form

$$E_{\text{ev}} = \kappa_{\text{ev}} \sum_{i<j} \left[\frac{\lambda_{ij}(\sigma_i + \sigma_j)}{r_{ij}} \right]^{12}, \tag{5.2}$$

where $\kappa_{\text{ev}} = 0.10$, and $\sigma_i = 1.77$, 1.75, 1.55, 1.42 and 1.00 Å for S, C, N, O and H atoms, respectively. The parameter λ_{ij} has the value 0.75 for all pairs except those connected by three covalent bonds, for which $\lambda_{ij} = 1$. When the two atoms belong to different chains, we always use $\lambda_{ij} = 0.75$. To speed up the calculations, Eq. 5.2 is evaluated using a cutoff of $r_{ij}^{\text{c}} = 4.3\lambda_{ij}$ Å, and pairs with fixed separation are omitted.

The second energy term, E_{loc}, is a local intrachain potential. It has the form

$$E_{\text{loc}} = \kappa_{\text{loc}} \sum_{I} \left(\sum \frac{q_i q_j}{r_{ij}^{(I)}/\text{Å}} \right), \tag{5.3}$$

where the inner sum represents the interactions between the partial charges of the backbone NH and C′O groups in one amino acid, I. The inner sum has four terms (NC′, NO, HC′ and HO) which depend only on ϕ and ψ for amino acid I. The partial charges are taken as $q_i = \pm 0.20$ for H and N and $q_i = \pm 0.42$ for C′ and O [32], and we put $\kappa_{\rm loc} = 100$, corresponding to a dielectric constant of $\epsilon_r \approx 2.5$.

The third term of the energy function is the hydrogen-bond energy $E_{\rm hb}$, which has the form

$$E_{\rm hb} = \epsilon_{\rm hb}^{(1)} \sum_{\rm bb-bb} u(r_{ij}) v(\alpha_{ij}, \beta_{ij}) + \epsilon_{\rm hb}^{(2)} \sum_{\rm sc-bb} u(r_{ij}) v(\alpha_{ij}, \beta_{ij}) \,, \tag{5.4}$$

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

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

$$v(\alpha, \beta) = \begin{cases} (\cos\alpha \cos\beta)^{1/2} & \text{if } \alpha, \beta > 90^\circ \\ 0 & \text{otherwise} \end{cases} \tag{5.6}$$

We consider only hydrogen bonds between NH and CO groups, and r_{ij} denotes the HO distance, α_{ij} the NHO angle and β_{ij} the HOC angle. The parameters $\epsilon_{\rm hb}^{(1)}$, $\epsilon_{\rm hb}^{(2)}$ and $\sigma_{\rm hb}$ are taken as 3.1, 2.0 and 2.0 Å, respectively. The function $u(r)$ is calculated using a cutoff of $r^{\rm c} = 4.5$ Å. The first sum in Eq. 5.4 contains backbone-backbone interactions, while the second sum contains interactions between charged side chains (Asp, Glu, Lys and Arg) and the backbones. For intrachain hydrogen bonds we make two restrictions. First, we disallow backbone NH (C′O) groups to make hydrogen bonds with the two nearest backbone C′O (NH) groups on each side of them. Second, we forbid hydrogen bonds between the side chain of one residue with the two nearest donors or acceptors on either side of its C_α. For interchain hydrogen bonds, we make no such restrictions.

The fourth energy term, $E_{\rm hp}$, represents an effective hydrophobic attraction between nonpolar side chains (no explicit water molecules). It has the pairwise additive form

$$E_{\rm hp} = -\sum_{I<J} M_{IJ} C_{IJ} \,, \tag{5.7}$$

where C_{IJ} is a measure of the degree of contact between side chains I and J, and M_{IJ} sets the energy that a pair in full contact gets. The matrix M_{IJ} is defined in Table 5.1. To calculate C_{IJ} we use a predetermined set of atoms, A_I, for each side chain I. We define C_{IJ} as

$$C_{IJ} = \frac{1}{N_I + N_J} \left[\sum_{i \in A_I} f(\min_{j \in A_J} r_{ij}^2) + \sum_{j \in A_J} f(\min_{i \in A_I} r_{ij}^2) \right] , \tag{5.8}$$

		I	II	III
I	Ala	0.0	0.1	0.1
II	Ile, Leu, Met, Pro, Val		0.9	2.8
III	Phe, Trp, Tyr			3.2

Table 5.1: The matrix M_{IJ} that sets the strength of the effective hydrophobic attraction (see Eq. 5.7).

where the function $f(x)$ is given by $f(x) = 1$ 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 the fraction of atoms in A_I or A_J that are in contact with some atom from the other side chain. The definition of A_I for the different hydrophobic side chains has been given elsewhere [29,31]. For pairs that are nearest or next-nearest neighbors along the same chain, we use a reduced strength for the hydrophobic attraction; M_{IJ} is reduced by a factor of 2 for next-nearest neighbors, and taken to be 0 for nearest neighbors.

In our previous studies of the folding of different peptides [30, 31], we introduced a simple form of context dependence in the local potential E_{loc} and the hydrogen-bond energy E_{hb}; these interactions were taken to be weaker at the chain ends, which tend to be more exposed to water. In the present calculations, we do not make such a reduction of these interactions. The motivation for keeping the strength unchanged at the ends of the chains is that the experiments on Aβ_{16-22} [18, 19], as mentioned above, were carried out with acetyl and amide capping groups at the ends.

To study the thermodynamic behavior of this model, we use simulated tempering [33–35] in which the temperature is a dynamical variable. For a review of simulated tempering and other generalized-ensemble techniques for protein folding, see Ref. [36]. We study the one- and three-chain systems at eight different temperatures, ranging from 279 K to 374 K, and the six-chain system at seven temperatures, ranging from 291 K to 374 K.

Our simulations are carried out using two different elementary moves for the backbone degrees of freedom: first, the highly non-local pivot move in which a single backbone torsion angle is turned; and second, a semi-local method [37] that works with up to eight adjacent torsion angles, which are turned in a coordinated manner. Side-chain angles are updated one by one. In addition to these updates, we also use rigid-body translations and rotations of whole chains. Every update involves a Metropolis accept/reject step, thus ensuring detailed balance. All our simulations are started from random configurations. All statistical errors quoted are 1σ errors obtained from the variation between

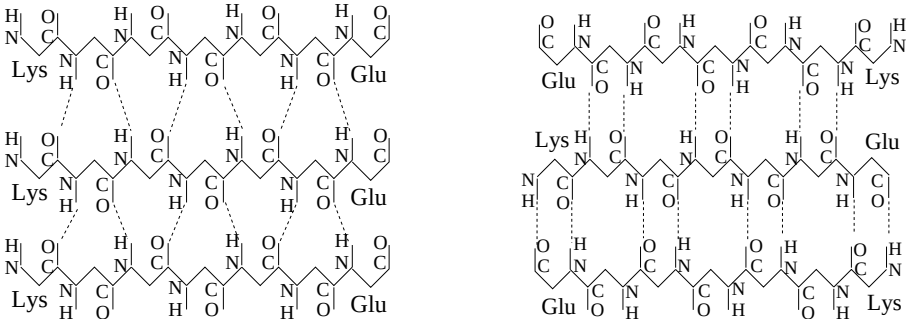

Figure 5.1: Schematic illustrations of the hydrogen bond patterns for in-register, parallel β-strands (left) and in-register, antiparallel β-strands (right).

independent runs.

In our calculations, we monitor several quantities. For a chain with N amino acids, we define the α-helix and β-strand contents as the fractions of the $N-2$ inner amino acids with their (ϕ, ψ) pair in the α-helix and β-strand regions of the Ramachandran space. We assume that α-helix corresponds to $-150° < \phi < -90°$, $90° < \psi < 150°$ and that β-strand corresponds to $-80° < \phi < -48°$, $-59° < \psi < -27°$ [26,38]. The average α-helix and β-strand contents, over all the chains of the system, are denoted by H and S, respectively.

To characterize the aggregated structures, we determine all pairs of chains with an interchain hydrogen bond energy less than $-1.5\epsilon_{\rm hb}^{(1)}$ (see Eq. 5.4). For each such interacting pair of chains, we calculate the scalar product of their normalized end-to-end unit vectors. If this scalar product is greater than 0.7 (less than -0.7), we say that the two chains are parallel (antiparallel). For a given multichain configuration, we denote the numbers of parallel and antiparallel pairs of chains by n_+ and n_-, respectively. Fig. 5.1 illustrates the hydrogen-bond patterns in parallel and antiparallel β-sheets, respectively.

The propensity of the peptides to aggregate depends strongly on temperature. To examine the character of the transition/crossover from high- to low-temperature behavior, we determine the specific heat

$$C_{\rm v}(T) = \frac{1}{N_{\rm c}N}\frac{d\langle E\rangle}{dT} = \frac{1}{N_{\rm c}NkT^2}\left(\langle E^2\rangle - \langle E\rangle^2\right) , \qquad (5.9)$$

where $N_{\rm c}$ is the number of chains, N is the number of amino acids per chain, and $\langle O\rangle$ denotes a Boltzmann average of variable O. A peak in $C_{\rm v}$ whose height grows linearly with $N_{\rm c}$ would signal that the system develops a first-order phase transition in the limit $N_{\rm c} \to \infty$.

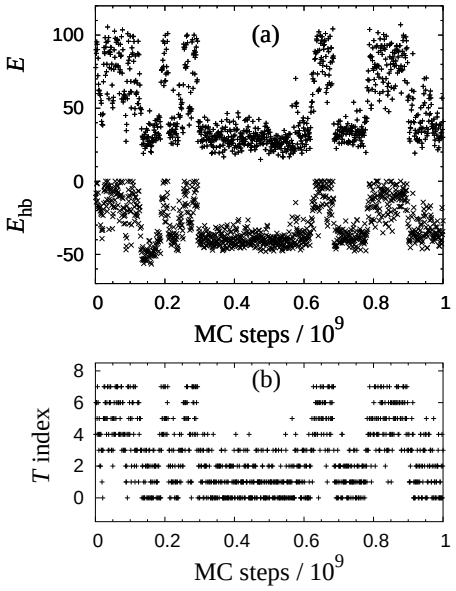

Figure 5.2: Monte Carlo evolution in typical run for $N_c = 3$ Aβ_{16-22} peptides. (a) Energy E (+) and hydrogen-bond energy E_{hb} (×), both in kcal/mol. (b) The temperature index k. There are eight allowed temperatures T_k, satisfying $T_0 = 279\,\mathrm{K} < T_1 < \ldots < T_7 = 374\,\mathrm{K}$. Measurements were taken every 10^6 Monte Carlo steps.

5.3 Results and Discussion

Using the model described in the previous section, we study the thermodynamics of systems of N_c Aβ_{16-22} peptides for $N_c = 1$, 3 and 6. Fig. 5.2 illustrates the evolution of the $N_c = 3$ system in a typical simulated-tempering run. In the course of the run, aggregated low-energy structures form and dissolve several times.

5.3.1 Secondary Structure

Fig. 5.3 shows the α-helix and β-strand contents H and S, as defined in the previous section, against temperature for different N_c. For $N_c = 1$, we see that both H and S are low at all temperatures, which in particular implies

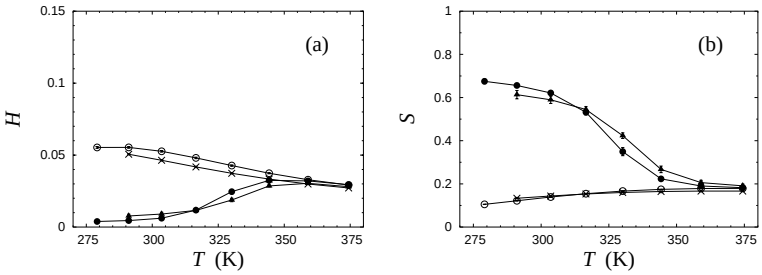

Figure 5.3: (a) The α-helix content H against temperature T for $A\beta_{16-22}$ for $N_c = 1$ (○), $N_c = 3$ (●) and $N_c = 6$ (▲). Also shown are data for $N_c = 6$ S_r peptides (×). Lines joining data points are only a guide for the eye. (b) Same for the β-strand content S. Note that the scales in (a) and (b) are different.

that the $A\beta_{16-22}$ monomer is mainly a random coil even at low temperatures. For $N_c > 1$, the β-strand content shows a qualitatively different behavior; S increases sharply with decreasing temperature, to values of $S = 0.6$ and higher. This shows that the three- and six-chain systems form ordered β-sheet structure as the temperature decreases. The α-helix content, on the other hand, remains small for $N_c = 3$ and 6.

Our results for $N_c = 1$ and $N_c = 3$ can be compared with results from molecular dynamics simulations with explicit water by Klimov and Thirumalai [26]. Using somewhat different definitions of H and S and a temperature of $T = 300\,\mathrm{K}$, these authors found that $H = 0.11$ and $S = 0.33$ for $N_c = 1$, and $H = 0.26$ and $S = 0.30$ for $N_c = 3$. Our $N_c = 1$ results (see Fig. 5.3) are in reasonable agreement with theirs, given that we use stricter definitions of α-helix and β-strand. However, our $N_c = 3$ results disagree with theirs. They obtained a smaller β-strand content and a larger α-helix content compared to their own $N_c = 1$ results; whereas we observe a much larger β-strand content for $N_c = 3$ compared to $N_c = 1$.

5.3.2 β-Strand Organization

To find out whether the low-energy configurations, rich in β-strands, show a preference for either parallel or antiparallel β-sheets, we study the joint probability distribution $P(n_+, n_-)$, where n_+ and n_- count the numbers of interacting chain pairs that are parallel and antiparallel, respectively (see Sec. 5.2). Table 5.2 shows this distribution for the $N_c = 3$ system at $T = 279\,\mathrm{K}$. At this temperature, we find that the typical configuration is a three-stranded β-sheet

n_+	n_-		
	0	1	2
0	< 0.01	≈ 0.01	0.57 ± 0.13
1	< 0.01	0.41 ± 0.13	
2	< 0.01		

Table 5.2: The probability distribution $P(n_+, n_-)$ (see Sec. 5.2) for $N_c = 3$ Aβ_{16-22} peptides at $T = 279\,\mathrm{K}$. Values not shown correspond to (n_+, n_-) combinations never observed.

n_+	n_-					
	0	1	2	3	4	5
0	< 0.01	< 0.01	< 0.01	0.07 ± 0.04	0.32 ± 0.12	0.10 ± 0.06
1	< 0.01	< 0.01	0.04 ± 0.03	0.12 ± 0.06	0.12 ± 0.07	
2	< 0.01	< 0.01	0.19 ± 0.11	0.04 ± 0.02		

Table 5.3: Same as Table 5.2 for $N_c = 6$ Aβ_{16-22} peptides at $T = 291\,\mathrm{K}$. Values not shown are less than 10^{-3}.

with $n_+ + n_- = 2$. No configuration with $n_+ + n_- > 2$ was observed. The most probable combination is $(n_+, n_-) = (0, 2)$, corresponding to a maximum number of antiparallel pairs. Mixed configurations with $(n_+, n_-) = (1, 1)$ also occur with a high frequency. The probability of having two parallel pairs, $(n_+, n_-) = (2, 0)$, is, by contrast, tiny.

Table 5.3 shows the $P(n_+, n_-)$ distribution for $N_c = 6$ Aβ_{16-22} peptides at $T = 291\,\mathrm{K}$. Although the statistical uncertainties are somewhat large, it is clear that n_- tends to be larger than n_+ for this system size, too. In fact, configurations with $n_+ > 2$ are extremely rare, whereas the probability of having $n_- \geq 4$ is roughly 50%.

These results for $N_c = 3$ and $N_c = 6$ strongly suggest that Aβ_{16-22} peptides are very unlikely to form extended parallel β-sheets in our model. By contrast, nothing seems to prevent the propagation of antiparallel β-sheets. This finding is in agreement with the solid-state NMR results [18, 19] mentioned in the introduction, which suggest that Aβ_{16-22} fibrils contain antiparallel β-sheets.

These experimental studies [18, 19] furthermore suggest that the antiparallel β-strands are in register in Aβ_{16-22} fibrils (see Fig. 5.4b). A study of the hydrogen-bond patterns (see Fig. 5.1) in our calculations shows that a significant fraction of the antiparallel β-strand pairs are in register, for $N_c = 3$ as well as $N_c = 6$. However, an even larger fraction of these pairs exhibit the one-

Figure 5.4: Schematic representations of three different registries for an antiparallel pair of $A\beta_{16-22}$ peptides. In our model, (a) occurs more frequently than (b), which in turn occurs more frequently than (c).

step shifted registry shown in Fig. 5.4a. The opposite one-step shift, shown in Fig. 5.4c, is, by contrast, very rare. In our model, we believe that the relative frequencies of occurrence for the three registries shown in Fig. 5.4 can be largely understood in terms of steric effects; the large Phe side chains are difficult to accommodate in (c), easier to accommodate in (b), and unproblematic in (a). The experimental results on full fibrils indicate, however, that other mechanisms that are ignored in our model might play an important role for the registry. It is interesting to note that recent experiments on $A\beta_{11-25}$ fibrils [17] found evidence for the registry in Fig. 5.4a at pH 7.4, and for that in Fig. 5.4c at pH 2.4.

5.3.3 The Role of Hydrophobicity

It is widely held that hydrophobic attraction is a major driving force behind the formation of amyloid fibrils. To get an idea of the importance of hydrophobic attraction in the aggregation of $A\beta_{16-22}$ peptides, we consider a randomly selected example of a sequence without strongly hydrophobic side chains, given by Asp-Ala-His-Glu-Arg-Lys-Glu. For this peptide, to be called S_r, we performed simulations similar to those for $A\beta_{16-22}$, using $N_c = 6$. From Fig. 5.3 it can be seen that both the α-helix and β-strand contents are small for the system of six S_r peptides. In fact, our results for six S_r peptides are very similar to those for an $A\beta_{16-22}$ peptide in isolation. Hence, we find that the system of six S_r peptides does not form aggregated structures. The inability of this system to form aggregated structures strongly suggests that hydrophobic attraction indeed plays a key role in the aggregation of $A\beta_{16-22}$ peptides. This finding is in agreement with the conclusions of Klimov and Thirumalai [26].

An open and very important question is what makes the organization of the β-strands parallel in some fibrils and antiparallel in others. Klimov and Thirumalai [26] found that Coulomb interactions between charged side chains are the main determinant of the organization of the β-strands in $A\beta_{16-22}$ oligomers. This conclusion may seem reasonable because the two end side chains of the

Aβ_{16-22} peptide carry opposite charges, which indeed should make an antiparallel organization electrostatically favorable. However, our model completely ignores charges and still strongly favors the antiparallel organization. Other mechanisms than Coulomb interactions between side-chain charges might therefore play a significant role. A recent experimental study [19] demonstrates the importance of the amphiphilic structure. This study showed that the β-sheet structure of Aβ_{16-22} fibrils can be changed from antiparallel to parallel by adding an end group that increases the amphiphilicity of the peptide. In the absence of a clear amphiphilicity, the antiparallel organization may be favored because of sequence-independent factors such as the geometry of backbone-backbone hydrogen bonds. The fact that our model, with its relatively simple potential, strongly favors the antiparallel organization supports this view.

We also performed simulations of a system of three peptides with the amino acid sequence Lys-Phe-Phe-Ala-Ala-Ala-Glu. This peptide has a much more asymmetric distribution of hydrophobicity than Aβ_{16-22}. In contrast to the corresponding Aβ_{16-22} system, this system is able to make three-stranded parallel β-sheets, corresponding to $(n_+, n_-) = (2, 0)$. The probability of having $(n_+, n_-) = (0, 2)$ is, however, non-negligible, because an out-of-register arrangement makes it possible for antiparallel strands to form some hydrophobic Phe-Phe contacts. By making the peptide larger, it should be possible to reduce such contacts, and thereby strongly favor the parallel orientation.

5.3.4 The Onset of Aggregation

At high temperatures, Aβ_{16-22} peptides do not self-assemble into aggregated structures (see Fig. 5.3). The temperature at which the aggregation sets in depends strongly on peptide concentration. Exploring that dependence is beyond the scope of the present study. Nevertheless, it is interesting to examine the character of the transition from high- to low-temperature behavior. Fig. 5.5 shows the temperature dependence of the specific heat $C_{\mathrm{v}}(T)$ for $N_{\mathrm{c}} = 1$, 3 and 6. For $N_{\mathrm{c}} = 1$, $C_{\mathrm{v}}(T)$ shows a very weak dependence on temperature. For $N_{\mathrm{c}} = 3$ and $N_{\mathrm{c}} = 6$, on the other hand, we see that $C_{\mathrm{v}}(T)$ exhibits a pronounced peak, and the height of the peak grows with N_{c}. Near the peak, the energy distribution is very broad (data not shown), so aggregated low-energy and unstructured high-energy states coexist at these temperatures. Intermediate energies are significantly populated as well.

To get an idea of the character of the states with intermediate energies, we divided the energy axis into bins and calculated the average α-helix and β-strand contents for each bin. Fig. 5.6 shows the resulting α-helix and β-strand profiles $H(E)$ and $S(E)$, respectively. We see that the β-strand content increases

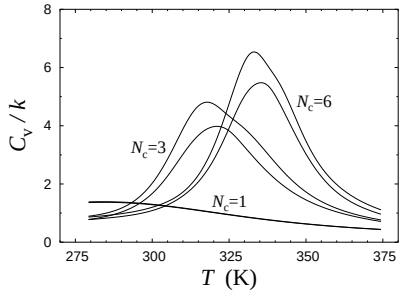

Figure 5.5: Specific heat C_v (see Eq. 5.9) against temperature T for $N_c = 1$, 3 and 6 $A\beta_{16-22}$ peptides, as obtained using histogram reweighting techniques [39]. The bands are centered around the expected values and show statistical 1σ errors.

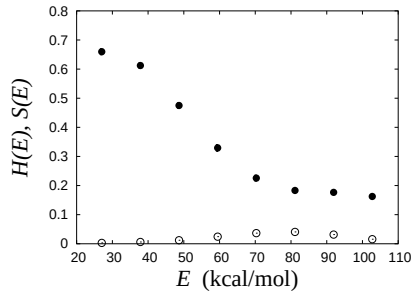

Figure 5.6: α-helix profile $H(E)$ ($\circ$) and β-strand profile $S(E)$ ($\bullet$), as defined in the text, for $N_c = 3$ $A\beta_{16-22}$ peptides at $T = 330\,\mathrm{K}$.

steadily with decreasing energy, whereas the α-helix content is low at all energies. In their study of $N_c = 3$ $A\beta_{16-22}$ peptides, Klimov and Thirumalai [26] found evidence for an obligatory α-helical intermediate. Our data for $H(E)$ show a maximum at $E \sim 80\,\mathrm{kcal/mol}$ (see Fig. 5.6), but the maximum value of $H(E)$ is very small. Hence, we see no sign of an obligatory α-helical intermediate in our model. Most of the amino acids in a typical configuration at intermediate energies are either random coil or β-strand.

5.3.5 Examples of Low-Energy Structures

Although the systems studied are small, it is interesting examine the shape of the aggregated structures, especially since they represent potential seeds for

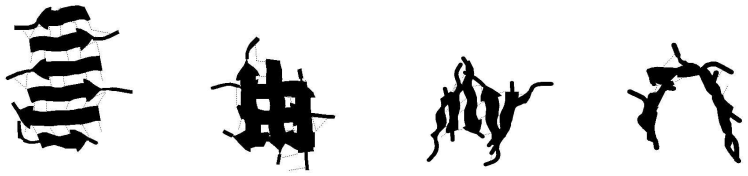

Figure 5.7: Examples low-energy structures from our simulations. From left to right: (i) An array of antiparallel β-strands. (ii) Two arrays of three antiparallel β-strands "sandwiching" several of their hydrophobic side-chains between them. (iii) An array with a definite curvature which indicates possible closed "barrel" structures for larger systems. (iv) A different angle of the same structure as (iii) to help in visualization.

the fibril formation. Also, it is known that relatively small assemblies formed early in the aggregation of full-length Aβ [40–42] as well as non-disease-related proteins [3] can be toxic.

We find that the three- and six-chain Aβ_{16-22} systems do not exhibit a single dominating free-energy minimum but rather a number of more or less degenerate local minima. Fig. 5.7 shows a few snapshots of such minima taken from our six-chain simulations. The β-strand content is, as noted earlier, high, and the structures shown in Fig. 5.7 illustrate this property. Both parallel and antiparallel arrangements of these β-strands occur in our model, with a definite preference for the antiparallel arrangements.

In the simplest class of typical structures observed in our simulations, the β-strands are stacked together in relatively flat sheets. For $N_c = 6$, six-stranded β-sheets occur with a significant frequency, as indicated by the $P(n_+, n_-)$ distribution in Table 5.3. Further, for the six-chain system, we observe the emergence of new non-trivial structures with no analogs in the three-chain simulations. Examples of such structures are given in Fig. 5.7. It is possible to enhance the stability of the system by stacking two linear arrays of β-strands together, bringing hydrophobic side chains from different arrays in close contact. Such "sandwiches" occur with a non-negligible frequency in our simulations. There are also indications that a closed cyclic array of chains, or a "barrel", might emerge as a possible low-energy state for larger systems.

In none of our simulations did we find any indication of a free-energy minimum in which the β-strands are joined end-to-end to form the so-called β-helix [43]. In our model, stability is enhanced by increasing the number of hydrogen bonds or by increasing hydrophobic contacts. For system sizes as small as those we

examined, the β-helix is inferior to many competing structures in both these respects, and hence its absence is expected.

5.4 Conclusion

Using a sequence-based atomic model which was originally developed for the study of the folding behavior of single peptides [29–31], we studied the thermodynamics of small systems of $A\beta_{16-22}$ peptides, without changing any parameter of the model. We found that $A\beta_{16-22}$ peptides form oligomers with a high content of antiparallel β-sheet structure, which is the same type of structure that has been experimentally found in $A\beta_{16-22}$ amyloid fibrils [18, 19].

Instead of an absolute free-energy minimum, we observed several nearly degenerate minima corresponding to different supra-molecular structures, all consisting of arrangements of β-strands. Apart from single β-sheets, laminated multi-sheet structures were found near free-energy minima for the six-chain system.

In addition to the $A\beta_{16-22}$ peptide, we also studied a few control sequences. Our results strongly support the view that hydrophobic attraction is a major driving force behind the oligomerization of $A\beta_{16-22}$ peptides; hydrogen bonding alone is not sufficient in order to obtain stable aggregated structures. That $A\beta_{16-22}$ β-sheets tend to be antiparallel is not surprising, given that antiparallel β-sheets are widely held to be intrinsically more stable than parallel ones. On the other hand, it has been suggested [26] that the main mechanism responsible for the antiparallel orientation in $A\beta_{16-22}$ oligomers is Coulomb interactions between oppositely charged side chains. In our model, charges are completely ignored, and yet the antiparallel orientation is strongly favored. The β-sheet registry that occurs most frequently in our simulations is different from that which has been found in full fibrils of $A\beta_{16-22}$ [18, 19]. The registry favored by the model is sterically advantageous. The mechanism responsible for the registry found in full fibrils remains to be uncovered.

Acknowledgments

This work was in part supported by the Swedish Research Council and the Knut and Alice Wallenberg Foundation through the Swegene consortium.

References

[1] Rochet JC, Lansbury PT Jr. Amyloid fibrillogenesis: themes and variations. *Curr. Opin. Struct. Biol.* 2000; 10: 60–68.

[2] Dobson CM. Protein folding and misfolding. *Nature* 2003; 426: 884–890.

[3] Bucciantini M, Giannoni E, Chiti F, Baroni F, Formigli L, Zurdo J, Taddei N, Ramponi G, Dobson CM, Stefani M. Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature* 2002; 416: 507–511.

[4] Otzen DE, Kristensen O, Oliveberg M. Designed protein tetramer zipped together with a hydrophobic Alzheimer homology: A structural clue to amyloid assembly. *Proc. Natl. Acad. Sci. USA* 2000; 97: 9907–9912.

[5] Broome BM, Hecht MH. Nature disfavors sequences of alternating polar and non-polar amino acids: Implications for amyloidogenesis. *J. Mol. Biol.* 2000; 296: 961–968.

[6] Richardson JS, Richardson DC. Natural β-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc. Natl. Acad. Sci. USA* 2002; 99: 2754–2759.

[7] West MW, Wang W, Patterson J, Mancias JD, Beasley JR, Hecht MH. *De novo* amyloid proteins from designed combinatorial libraries. *Proc. Natl. Acad. Sci. USA* 1999; 96: 11211–11216.

[8] Villegas V, Zurdo J, Filimonov VV, Avilés FX, Dobson CM, Serrano L. Protein engineering as a strategy to avoid formation of amyloid fibrils. *Protein Sci.* 2000; 9: 1700-1708.

[9] Hammarström P, Jiang X, Hurshman AR, Powers ET, Kelly JW. Sequence-dependent denaturation energetics: A major determinant in amyloid disease diversity. *Proc. Natl. Acad. Sci. USA* 2002; 99: 16427–16432.

[10] López de la Paz M, Goldie K, Zurdo J, Lacroix E, Dobson CM, Hoenger A, Serrano L. *De novo* designed peptide-based amyloid fibrils. *Proc. Natl. Acad. Sci. USA* 2002; 99: 16052–16057.

[11] Chiti F, Stefani M, Taddei N, Ramponi G, Dobson CM. Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature* 2003; 424: 805–808.

[12] López de la Paz M, Serrano L. Sequence determinants of amyloid formation. *Proc. Natl. Acad. Sci. USA* 2004; 101: 87–92.

[13] Sunde M, Blake C. The structure of amyloid fibrils by electron microscopy and X-ray diffraction. *Adv. Protein Chem.* 1997; 50: 123–159.

[14] Burkoth TS, Benzinger T, Urban V, Morgan DM, Gregory DM, Thiyagarajan P, Botto RE, Meredith SC, Lynn DG. Structure of the β-amyloid$_{(10-35)}$ fibril. *J. Am. Chem. Soc.* 2000; 122: 7883–7889.

[15] Petkova AT, Ishii Y, Balbach JJ, Antzutkin ON, Leapman RD, Delaglio F, Tycko R. A structural model for Alzheimer's β-amyloid fibrils based on experimental constraints from solid state NMR. *Proc. Natl. Acad. Sci. USA* 2002; 99: 16742–16747.

[16] Lansbury PT, Costa PR, Griffiths JM, Simon EJ, Auger M, Halverson KJ, Kocisko DA, Hendsch ZS, Ashburn TT, Spencer RG, Tidor B, Griffin RG. Structural model for the β-amyloid fibril based on interstrand alignment of an antiparallel-sheet comprising a C-terminal peptide. *Nat. Struct. Biol.* 1995; 2: 990–998.

[17] Petkova AT, Buntkowsky G, Dyda F, Leapman RD, Yau W-M, Tycko R. Solid state NMR reveals a pH-dependent antiparallel β-sheet registry in fibrils formed by a β-amyloid peptide. *J. Mol. Biol.* 2004; 335: 247–260.

[18] Balbach JJ, Ishii Y, Antzutkin ON, Leapman RD, Rizzo NW, Dyda F, Reed J, Tycko R. Amyloid fibril formation by $A\beta_{16-22}$, a seven-residue fragment of the Alzheimer's β-amyloid peptide, and structural characterization by solid state NMR. *Biochemistry* 2000; 39: 13748–13759.

[19] Gordon DJ, Balbach JJ, Tycko R, Meredith SC. Increasing the amphiphilicity of an amyloidogenic peptide changes the β-sheet structure in the fibrils from antiparallel to parallel. *Biophys. J.* 2004; 86: 428–434.

[20] Bratko D, Blanch HW. Competition between protein folding and aggregation: A three-dimensional lattice-model simulation. *J. Chem. Phys.* 2001; 114: 561–569.

[21] Harrison PM, Chan HS, Prusiner SB, Cohen FE. Conformational propagation with prion-like characteristics in a simple model for protein folding. *Protein Sci.* 2001; 10: 819–835.

[22] Dima RI, Thirumalai D. Exploring protein aggregation and self-propagation using lattice models: phase diagram and kinetics. *Protein Sci.* 2002; 11: 1036–1049.

[23] Jang H, Hall CK, Zhou Y. Assembly and kinetic folding pathways of a tetrameric β-sheet complex: Molecular dynamics simulations on simplified off-lattice protein models. *Biophys. J.* 2004; 86: 31–49.

[24] Ma B, Nussinov R. Stabilities and conformations of Alzheimer's β-amyloid peptide oligomers ($A\beta_{16-22}$, $A\beta_{16-35}$, and $A\beta_{10-35}$): Sequence effects. *Proc. Natl. Acad. Sci. USA* 2002; 99: 14126–14131.

[25] Ma B, Nussinov R. Molecular dynamics simulations of alanine rich β-sheet oligomers: Insight into amyloid formation. *Protein Sci.* 2002; 11: 2335–2350.

[26] Klimov DK, Thirumalai D. Dissecting the assembly of $A\beta_{16-22}$ amyloid peptides into antiparallel β sheets. *Structure* 2003; 11: 295–307.

[27] Gsponer J, Haberthür U, Caflisch A. The role of side-chain interactions in the early steps of aggregation: Molecular dynamics simulations of an amyloid-forming peptide from the yeast prion Sup35. *Proc. Natl. Acad. Sci. USA* 2003; 100: 5154–5159.

[28] Paci E, Gsponer J, Salvatella X, Vendruscolo M. Molecular dynamics studies of the process of amyloid aggregation of peptide fragments of transthyrin. Preprint.

[29] Irbäck A, Samuelsson B, Sjunnesson F, Wallin S. Thermodynamics of α- and β-structure formation in proteins. *Biophys. J.* 2003; 85: 1466–1473.

[30] Irbäck A, Sjunnesson F. Folding thermodynamics of three β-sheet peptides: A model study. *Proteins Struct. Funct. Genet.* 2004 (in press).

[31] Irbäck A, Mohanty S. manuscript in preparation.

[32] Branden C, Tooze J. *Introduction to Protein Structure*. New York: Garland Publishing; 1991.

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

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

[35] 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.

[36] Hansmann UHE, Okamoto Y. *Curr. Opin. Struct. Biol.* 1999; 9: 177–183.

[37] 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.

[38] Srinivasan R, Rose GD. LINUS: a hierarchic procedure to predict the fold of a protein. *Proteins Struct. Funct. Genet.* 1995; 22: 81–99.

[39] Ferrenberg AM, Swendsen, RH. New Monte Carlo technique for studying phase transitions. *Phys. Rev. Lett.* 1988; 61: 2635–2638.

[40] Lambert MP, Barlow AK, Chromy BA, Edwards C, Freed R, Liosatos M, Morgan TE, Rozovsky I, Trommer B, Viola KL, Wals P, Zhang C, Finch CE, Krafft GA, Klein WL. Diffusive, nonfibrillar ligands derived from $A\beta_{1-42}$ are potent nervous system neurotoxins. *Proc. Natl. Acad. Sci. USA* 1998; 95: 6448–6453.

[41] Walsh DM, Hartley DM, Kusumoto Y, Fezoui Y, Condron MM, Lomakin A, Benedek GB, Selkoe DJ, Teplow DB. Amyloid β-protein fibrillogenesis — Structure and biological activity of protofibrillar intermediates. *J. Biol. Chem.* 1999; 274: 25945–25952.

[42] Walsh DM, Klyubin I, Fadeeva JV, Cullen WK, Anwyl R, Wolfe MS, Rowan MJ, Selkoe DJ. Naturally secreted oligomers of amyloid β protein potently inhibit hippocampal long-term potentiation *in vivo*. *Nature* 2002; 416: 535–539.

[43] Wetzel R. Ideas of order for amyloid fibril structure. *Structure* 2002; 10: 1031–1036.