Modeling of Protein Folding and Genetic Networks

Fredrik Sjunnesson

Department of Theoretical Physics
Lund University, Sweden

Thesis for the degree of Doctor of Philosophy

Thesis Advisor: Anders Irbäck

Faculty Opponent: Ugo Bastolla

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 Friday, the 3rd of October 2003, at 1.15 p.m.

DOKUMENTDATABLAD enl SIS 61 41 21

Organization
LUND UNIVERSITY
Department of Theoretical Physics
Sölvegatan 14A
SE-223 62 LUND

Document Name
DOCTORAL DISSERTATION

Date of issue
September 2003

CODEN:

Author(s)
Fredrik Sjunnesson

Sponsoring organization

Title and subtitle
Modeling of Protein Folding and Genetic Networks

Abstract
Models for potein folding are developed and applied to peptides and small proteins with both alpha-helix and beta-sheet structure. The energy functions, in which effective hydrophobicity forces and hydrogen bonds are taken to be the two central terms, are sequence-based and deliberately kept simple. The geometric representations of the protein chains are, by contrast, detailed and have torsion angles as the degrees of freedom. The thermodynamic properties of the models are studied using Monte Carlo methods and quantitative comparisons with experiments are carried out. To improve the sampling of compact states, a semi-local Monte Carlo update in the backbone torsion angles is developed. In addition, the thesis includes a study of a simple model for genetic networks, the Kauffman model.

Summary in Swedish
Modeller för proteinveckning utvecklas och tillämpas på peptider och små proteiner som har både alfahelix- och betablad-struktur. Energifunktionerna, i vilka effektiva hydrofobicitetskrafter och vätebindningar är de två centrala termerna, är helt sekvensbaserade och har medvetet hållits enkla. Den geometriska representationen av proteinkedjorna är däremot detaljerad och har torsionsvinklar som frihetsgrader. Modellernas termodynamiska egenskaper studeras med hjälp av Monte Carlo-metoder och kvantitativa jämförelser med experiment görs. För att förbättra samplingen av kompakta tillstånd utvecklas en semilokal Monte Carlo-uppdatering som arbetar i torsionsvinklar längs proteinkedjans ryggrad. Dessutom studeras en enkel model för genetiska nätverk, Kauffman-modellen.

Key words
Protein folding, all-atom model, two-state folding, Monte Carlo, local update, Kauffman model.

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language
English

ISSN and key title

ISBN
91-628-5783-5

Recipient's notes

Number of pages
114

Price

Security classification

Distribution by (name and address)
Fredrik Sjunnesson, Dept. of Theoretical Physics,
Sölveg. 14A, SE-223 62 Lund

To Maria

This thesis is based on the following publications:

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

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

III Anders Irbäck and Fredrik Sjunnesson,
Folding Thermodynamics of Three β-Sheet Peptides: A Model Study
LU TP 03-40.

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

V Sven Bilke and Fredrik Sjunnesson,
Stability of the Kauffman Model
Physical Review E **65**, 016129 (2002).

Contents

Introduction

Proteins are linear polymers, made from 20 small molecules called amino acids. The protein chains are normally between fifty and ten thousand amino acids long. Rather than being extended and flexible, however, naturally occurring proteins adopt distinct three-dimensional conformations that are critical to their function [1]. Despite that they all consist of the same, relatively simple building blocks, proteins are very complex biomolecules that affect virtually every property that characterizes a living organism.

Even though chemical differences between amino acids partly explains the functional diversity of proteins, the key to protein function is to be found in the three-dimensional structures formed by the amino acid chains [1]. Thus, a natural step when trying to understand the function of a protein is to find its three-dimensional structure. However, due to the small size of proteins and the fragility of their native conformations, this is a very delicate task. At present, the PDB [2] database contains about 22,000 experimentally determined structures. This is only a small fraction of the proteins with known amino acid sequence and since the number of known sequences grows much faster than the number of known structures, other means of determining the three-dimensional structure of proteins are highly desirable.

With this in mind, it is encouraging that the native structure of many proteins seem to be determined by their amino acid sequence alone and not depend on biological mechanisms in the cells. In fact, experiments [3] have shown that many proteins have the ability to fold spontaneously to their native state. Consequently, it should be possible to predict the native configuration of a protein if its amino acid sequence is known. This is the so-called *protein folding problem*, which has attracted large interest over the last decades. Today hundreds of research groups around the world work on this problem using different methods. Many of those groups have taken a statistical approach, which has proven powerful if structures of proteins with similar amino acid sequence are known. For predicting new folds, however, physical modeling is required. This remains

an open problem but progress is being made as seen by the improving result in the biannual CASP competitions [4].

Moreover, it is becoming increasingly clear that proteins cannot be characterized in terms of static structures only. In fact, many proteins are disordered and become structured only upon binding [5]. This illustrates that kinetic and thermodynamic properties play integral role in protein function, which is another reason why physical modeling is of great importance.

In this thesis we develop simplified protein models and apply them to peptides and small proteins. In contrast to statistical methods, which are only used to predict the native states of proteins, physical modeling makes it possible to study thermodynamic and kinetic properties of the proteins, too.

In the next few sections a short introduction to proteins will be given. The geometry of the protein chain will be discussed as well as the interactions that are the main driving forces for protein folding. Then follows a brief overview of models and methods used for modeling protein folding. Finally, the papers which this thesis is based upon will be introduced.

In Paper V a simple model for genetic networks, the Kauffman model, is studied. Even though proteins play an important role in genetic networks, as in virtually every biological system, this model is very different from our protein models and the introduction to this subject is deferred to the introduction of this paper.

The Protein Chain

Proteins consist of a linear backbone chain, onto which side chains are attached. The amino acids all have the same backbone atoms, see Fig. 1, while the side chains, indicated as R in Fig. 1, are of 20 different types.

The protein chain is held together by covalent bonds. This makes the structure relatively rigid, except for rotational degrees of freedom about the covalent bonds. Many covalent bonds, both along the backbone and in the side chains, have this torsional degree of freedom, called torsion angle. The backbone torsion angles are of particular importance, since these determine the overall structure of a protein. Each amino acid, except proline, has two backbone torsional degrees of freedom, the so-called *Ramachandran* [6] angles ϕ and ψ. In principle, rotation could occur even about the peptide bond, the bond connecting amino acids. This bond, however, has a partially double-bonded character, which restricts its flexibility.

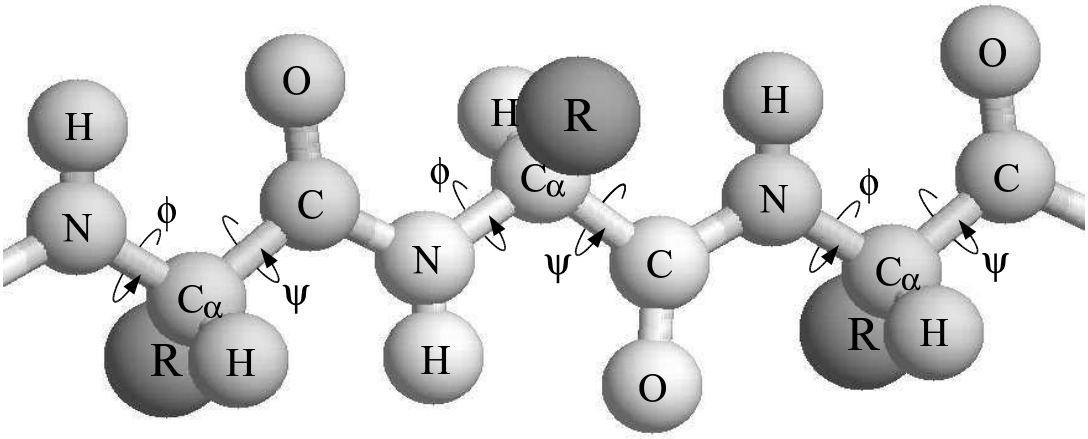

Figure 1: Schematic figure of a protein chain. R represents the side chains, which consist of one to 18 atoms each. The main degrees of freedom are the backbone torsion angles ϕ and ψ.

While there is considerable rotational freedom around the ϕ and ψ angles, not all torsion angles or combinations of angles are observed in proteins due to steric constraints. This is illustrated in Fig. 2, which is a so-called Ramachandran plot [6].

The asymmetry of the Ramachandran plot arises from the *right-handed chirality* of the protein chain. The side chains are always joined to the C_α atom at the position indicated in Fig. 1. There is no chemical reason why the C_α hydrogen atom and the side chain could not be interchanged but in nature only one of the two forms is synthesized. This has direct consequences for the local, or secondary, structure of proteins, which tend to show strong regularities, as exemplified in Fig. 2. The most common forms of secondary structure are the *α-helix* and the *β-sheet*. Because of the right-handed chirality, the α-helix only occurs in its right-handed form.

Since the local flexibility of the protein chain is highly affected by steric constraints, the amino acid glycine is of great importance. It is the smallest of the amino acids with a side chain consisting of one hydrogen atom only and because of its conformational flexibility it is often found in loop regions.

The amino acid with the most limited flexibility is proline. The reason is that the side chain forms a covalent bond with the nitrogen atom of the backbone. The resulting cyclic ring imposes rigid constraints on the flexibility of the backbone and effectively removes the ϕ torsion angle as a degree of freedom. Also,

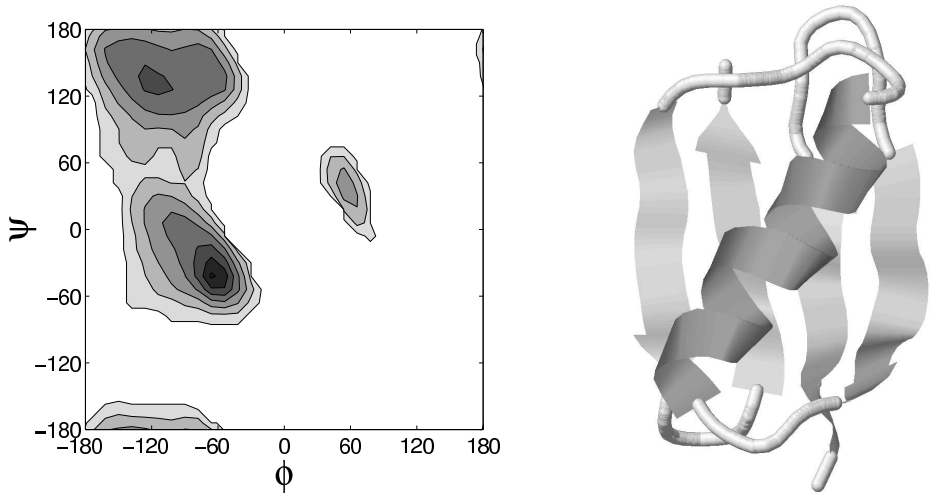

Figure 2: Left: A Ramachandran plot, which shows the distribution of torsion angles ϕ, ψ. The pronounced peaks at $(\phi \approx -60, \psi \approx -50)$ and $(\phi \approx -120, \psi \approx 130)$ corresponds to α-helix and β-sheet, respectively. The plot is based on structures in the PDB database. All glycine and proline amino acids are discarded due to their special properties, see text. Right: The PDB structure of Protein G (B1 domain), which contains both α-helix and β-sheet secondary structure. The two left-most β-strands in the figure form the β-hairpin that is studied in Papers II and III.

the backbone of the proline has no hydrogen atom bonded to the nitrogen atom, which affects what hydrogen bonds that can be formed. Prolines often start or break regions with secondary structure.

The side chains of the remaining 18 amino acids consist of 4 to 18 atoms and vary in chemical composition. Side chains that have charged or polar groups can interact with other charged or polar side chains or with the backbone, in which the so-called peptide units are polar. More important, however, is the favorable interactions that these side chains can make with the highly polar water molecules. Other side chains do not interact as favorably with water and are said to be *hydrophobic*. Their apparent tendency to avoid water give rise to the so-called *hydrophobic effect* and is thought to be the perhaps most important driving force behind the folding [7, 8] of proteins in aqueous environment.

Proteins that naturally occur in an aqueous environment, are generally referred to as *globular* since they, because of the hydrophobic effect, tend to adopt compact, roughly spherical shapes. In this thesis we restrict ourselves to globular proteins but there are also *membrane* and *fiber* proteins [1]. Because of the poor solubility of these proteins their characteristics differ in many ways from globular proteins.

The Folding Behavior

For small molecules, knowledge of their covalent structure is often enough to characterize their chemical properties. Proteins, however, are long and flexible enough to fold back onto themselves. The resulting three-dimensional structures as well their kinetic and thermodynamic properties are to a high degree determined by other, non-covalent, interactions.

Driving Forces

Except for disulfide bonds, that sometimes occur between pairs of cysteine amino acids, the folding and stability of proteins is governed by weak, non-covalent interaction, such as electrostatic interactions, van der Waals interactions and the very important hydrogen bonds.

A hydrogen bond occurs when two electro-negative atoms compete for the same hydrogen atom:

$$\text{— D — H} \cdots \text{A —}$$

The hydrogen is formally bonded covalently to one of the atoms, the donor (D), but it also interacts favorably with the other, the acceptor (A). The main component of the hydrogen bond is an electrostatic interaction between the D-H dipole, in which H has a partial positive charge, and a partial negative charge of A. The combination of electrostatic and covalent aspects of the hydrogen bond causes the most common and energetically most favorable hydrogen bonds to keep the three bonded atoms aligned [1].

Among the backbone atoms the nitrogen atom can act as a donor whereas the oxygen atom can act as an acceptor. Although some side chains also have groups that can take part in hydrogen bonds, the backbone-backbone hydrogen bonds are of special interest since both α-helices and β-sheets are stabilized through repeated patterns of such bonds.

The natural environment of globular proteins is soaked with water, so it is not only the hydrogen bonds within a protein chain that are relevant for the folding and stability. Protein-water and water-water hydrogen bonds also play important roles.

The oxygen atom of a water molecule has two valence electrons, each of which can form a hydrogen bond with hydrogens of two other water molecules. The

resulting attractive interactions between water molecules are relatively strong. Solutes that do not interact with the water as favorably as water-water interactions are said to be hydrophobic and tend to be expelled from from the water. For this reason, hydrophobic amino acids of proteins solved in water are in general buried in the core of the protein structures, while polar and charged amino acids tend to be located at the surface. This apparent attraction of hydrophobic amino acids is referred to as the hydrophobic effect.

The most obvious way to handle this in a model would be to explicitly include the water molecules and treat their interactions on a similar footing with the protein-protein interactions. However, for any reasonable amount of water surrounding a protein, the extra computational cost is dramatic and, in spite of its enormous biological importance, water is still poorly understood. For example, simulating the freezing of ordinary water remains a challenge [9].

For these reasons, most current models do not include water molecules explicitly. Instead, potentials are used that are meant to *implicitly* capture the effects of the water, of which the hydrophobic effect is thought to be most important. This reduces the computational cost but also introduces additional uncertainties about the form of the potential function.

The water has consequences also for protein-protein hydrogen bonds. Hydrogen bonds that are exposed to water can easily be replaced by hydrogen bonds with the water. This makes them more likely to break than hydrogen bonds buried in the core of a protein structure. Effectively, this makes the strength of the hydrogen bonds *context dependent*.

Thermodynamics and Kinetics

It is not only the native structure that determines the function of a protein, but thermodynamic and kinetic properties are also of great importance. A certain flexibility of the native structure often is a vital part of the biological function and an increasing number of severe and common diseases are being linked to protein misfolding and aggregation. Alzheimer's, Parkinson's, Huntington's and cystic fibrosis are a few examples.

Since the first arguments [10] that the number of possible configurations is too vast for proteins to fold by randomly searching the configuration space, researchers have been trying to characterize the transition from unfolded configurations to the native state.

This used to be viewed as a relatively deterministic process and the folding was thought to be guided through a folding pathway via specific intermediate states,

in discrete steps. The presence of the intermediate states was thought to be a prerequisite for fast folding. In the "new view" of protein folding, ensembles play a more central role and the idea of folding pathways is replaced with the broader notions of energy landscapes and folding funnels [11].

This new view received important experimental support in 1991, when a small protein, chymotrypsin inhibitor 2 (CI2), was found to fold rapidly without significantly populating any well defined intermediate states [12]. In fact it was possible to characterize its folding behavior in terms of a simple two-state model, where only the unfolded state and the folded native state are populated. Since then, many small proteins, with different structures and stabilities, have been found to show this two-state behavior [13]. Therefore, the midpoint temperature and the energy difference between the two states, as obtained from a two-state fit, are often used to characterize the folding behavior of small, fast-folding proteins.

Models and Methods

Because of the size and complexity of proteins it is not possible to study protein folding from first principles. Different kinds of approximations must be made and models at varying levels of resolution exist. Although general concepts of protein folding have been studied using analytical methods [14–16], it is clear that more specific questions call for computer simulations.

Models

Even though there are many types of interaction that are relevant for protein folding, it is not unreasonable to believe that, at least for proteins with the most stable and well defined native structures, the different types of interactions mainly favor the same native structures. This so-called "principle of minimal frustration" [14], suggests that a few, well chosen, energy terms could be enough to capture the most important aspects of protein folding.

In this spirit, simple lattice-based models, such as the HP model [17], have been studied. Although these models can not be used for predicting specific structures, they have given important insight into generic physical principles of protein folding, such as the importance of the hydrophobic effect.

Another class of simplified models are the so-called Gō models [18]. Here interactions that do not favor the native structure are explicitly ignored, which

means that prior knowledge of the native structure is required. The fact that the force field depends on the native structure leads to some conceptual difficulties; amino acids of the same type do not have the same properties. However, models of this type may still be useful since some folding mechanisms appear to depend on the topology of the native state rather than details of the interatomic interactions [19]. One example is the *contact order* [20], i.e. the average sequence separation between amino acids that make contacts in the native structure, which has been found to be correlated with the folding time [20].

At the other end of the spectrum are the so-called "all-atom" models [21–23] with elaborate force fields, detailed chain geometries and sometimes even explicit water. Such models have been used to study peptides and other biomolecules as well as the unfolding of proteins [24]. However, so far it has not been possible to study the folding of proteins in a satisfactory way with this type of models.

The models we develop in this thesis have entirely sequence based potentials, which we try to keep as simple and transparent as possible. The most important interaction terms represent excluded volume, hydrogen bonds and effective hydrophobicity forces. The geometry, by contrast, is very detailed. In Papers II and III it is in fact more detailed than in standard "all-atom" models as even hydrogen atoms are explicitly included. The detailed geometry is in part a consequence of our ambition to keep the potential as simple as possible. We strongly believe that the local geometry is crucial to the overall properties of proteins. The degrees of freedom are the backbone torsion angles ϕ and ψ and torsion angles in the various side chains.

When determining the energy functions of these models, we looked at both local properties, such as the shape of the Ramachandran plot, and the overall thermodynamic behavior. There exist methods for parameter optimization, such as that of Ref. [25], that could have been used in this process. However, the functional form of the potential must also be chosen with care. Therefore, our determination of the energy functions was largely a trial and error process.

Simulation Methods

The most intuitive way to simulate protein folding is to model the time evolution by integration of Newton's equations of motion. This approach is commonly known as *molecular dynamics* and is the method that is most widely used, especially for all-atom models. It has the advantage that it gives a natural way to study kinetic properties, although an implicit treatment of water introduces some uncertainties. However, the fact that the state of a molecule

is monitored step by step is in some cases also a drawback. The generated configurations are highly correlated which means that it takes a very long time before the ensemble of generated configurations reaches the thermal equilibrium distribution, i.e. the Boltzmann distribution

$$p(\sigma) \propto \exp\left(\frac{-E(\sigma)}{kT}\right) \ , \tag{1}$$

where σ is the state of the system, $E(\sigma)$ is the energy of this state, T is the temperature and k is Boltzmann's constant. Therefore, if focus is on thermodynamics rather than kinetics, it is not obvious that molecular dynamics is a suitable choice.

In this thesis we use Monte Carlo methods rather than molecular dynamics. All the Monte Carlo methods used are of Metropolis [26] type, which is a general scheme to generate configuration of a given distribution. Each new configuration is obtained by a two-step procedure. First, a tentative configuration is generated by modifications of the previous one. Second, this configuration is subject to an accept/reject step. If it is rejected the old configuration is kept.

In the accept/reject step the tentative configurations is accepted with probability

$$P_{\text{acc}}(\sigma_i \to \sigma_j) = \min\left(1, \ \frac{W(\sigma_i|\sigma_j)\exp(-E(\sigma_j)/kT)}{W(\sigma_j|\sigma_i)\exp(-E(\sigma_i)/kT)}\right) \ , \tag{2}$$

where $W(\sigma_j|\sigma_i)$ is the distribution of tentative configurations σ_j given that the system is in configuration σ_i. It can easily be verified that this simple and very general scheme fulfills detailed balance, i.e. that in equilibrium there is no net flow between any states. Very often W is taken to be symmetric, that is $W(\sigma_j|\sigma_i) = W(\sigma_i|\sigma_j)$, but this is not necessary. It is important, however, that W is such that the algorithm is ergodic; starting from any configuration it should be possible to reach any other configuration. If both detailed balance and ergodicity are satisfied, the method is guaranteed to generate configurations that have the desired Boltzmann distribution.

Since each new configuration is based on the previous one there are correlations between configurations in this method, too. However, the updates do not have to follow any physical laws. So by finding suitable tentative updates it possible, at least in principle, to get decorrelation times that are drastically shorter than for molecular dynamics.

One method we use extensively in this thesis is the pivot update [27], where a random backbone torsion angle is assigned a random value between 0 and 2π. This simple method gives remarkably short decorrelation times of large-scale properties of extended chains [28]. For compact chains, it is also important to

be able to alter local properties without causing too drastic changes in their global structure. When the chains only have torsional degrees of freedom, this is a delicate task. For this purpose we develop in Paper IV a method, in which tentative updates are drawn from a conformation-dependent distribution that favors approximately local deformations of the chain. Since the tentative updates are conformation-dependent it is necessary to take the factors W in Eq. (2) properly into account, in order for the update to fulfill detailed balance.

At low temperatures, the system is likely to be found in compact low-energy configurations that are separated by high free-energy[1] barriers, which can make the time evolution of the system very slow. This is a general problem in thermodynamic simulations and is referred to as the *multiple minima* problem. For very simple systems it is sometimes possible to invent moves that take the system directly between such low-energy configurations. For proteins, however, other methods to address this problem must be used.

Two general methods that have been tried in protein simulations are the multicanonical [29] and dynamical-parameter [30,31] methods. In both methods one samples a generalized distribution rather than the desired Boltzmann distribution. In the multicanonical method the energy function is transformed in such a way that the energy distribution becomes flat. In the dynamical-parameter method, on the other hand, parameters of the model are introduced as degrees of freedom. Ordinary Metropolis simulations are then performed in this extended state space. We use a version of the dynamical-parameter method called *simulated tempering* [31, 32], in which the temperature is a dynamical parameter. The idea is to improve the sampling at low temperatures, where the free-energy landscape is rugged, by allowing the system to visit higher temperatures, where the free-energy landscape is much smoother. Since the updates in temperature are done using the Metropolis method detailed balance is retained at all times. In tests on protein models, simulated tempering, a variation of it called parallel tempering [33] and the multi-canonical method were found [34, 35] to be comparable in efficiency.

[1] The free energy, $F(X)$, for some coordinate X, is given by $F(X) = -kT \ln P(X)$, where $P(X)$ is the probability distribution.

The Papers

Paper I

In Paper I, we develop a protein model with a realistic backbone geometry. The model contains three types of amino acids. One type, representing glycine, has no side chain. The side chains of the other two amino acids have a simple single-atom representation and are polar and hydrophobic, respectively. The degrees of freedom are the Ramachandran torsion angles ϕ and ψ.

Since the backbone hydrogen bonds are vital for stabilizing secondary structures and hydrophobicity is widely held as the perhaps most important driving force in protein folding, hydrogen bonding and effective hydrophobicity forces are taken to be the two central terms of the energy function in this model, as well as those in Papers II and III. The energy function is sequence based and we have tried to keep it as simple as possible. This is facilitated by the use of a realistic backbone geometry, which simplifies the definition of hydrogen bonds and is crucial in order to obtain a good Ramachandran ϕ,ψ distribution.

The model is used to study a designed protein sequence, which has a three-helix bundle as its native state. Despite its simplicity, we find that the thermodynamics of the model is in qualitative agreement with what is expected for a small fast-folding helical protein; the folding transition is very abrupt and the full sequence forms more stable secondary structure than one- and two-helix segments.

In Ref. [36] a slightly extended version of this model, with five amino acid types rather than three, was applied to a real three-helix-bundle protein, the B domain of staphylococcal protein A.

Paper II

The model developed in Paper II contains all the atoms of all the 20 amino acids. The degrees of freedom are still the torsion angles, which in this extended model include both the Ramachandran angles of the backbone and a number of torsion angles in the various side chains.

Using this model, we study the thermodynamics of an α-helix and a β-hairpin, using exactly the same parameters for both peptides. We find that the melting curves are in reasonable *quantitative* agreement with experimental data. This contrasts sharply with prior all-atom studies of the β-hairpin which, as pointed

out in Ref. [37], all found a temperature dependence that was too weak.

The all-atom representation of course comes with a computational cost. However, we find that the detailed geometry is important in order to separate between two topologically distinct β-hairpin folds and it makes the formulation of the energy function simpler.

Paper III

In Paper III we refine the model of Paper II. Using this revised model, we investigate the folding thermodynamics of the β-hairpin from Paper II and the peptides LLM and Betanova, which both form three-stranded β-sheets in their native states.

The native populations obtained for these three sequences are in good agreements with experiments. We also find that the apparent native population depends on which of two observables that is used. This is interesting since different experimental techniques in fact have given different estimates for the β-hairpin.

Paper IV

Generating strictly local deformations of a chain molecule with only torsional degrees of freedom is a non-trivial task. The first (correct) Monte Carlo algorithm of this kind was the concerted-rotation method [38]. This method, however, is not easy to implement and tend to give large local deformations. As a result the acceptance rate may be low if the chain is folded and has bulky side chains.

In Paper IV, we develop a "small-step" Monte Carlo algorithm that generates semi-local moves. The method works with seven or eight adjacent torsion angles. These angles are updated using a biasing probability that favors approximately local deformations of the chain. Despite the presence of this conformation-dependent biasing probability, the model is almost as easy to implement as an unbiased move.

The possibility of combining our method with the concerted-rotation method to obtain a "small-step" and yet strictly local update has recently been described [39].

Paper V

Some proteins regulate the activity of genes, i.e. the rate at which the corresponding proteins are produced. This means that the activity of one gene may influence the activities of others. Such complex networks of interactions between genes are called *genetic networks*.

Already in 1969, a simple model of this process, the *Kauffman model* [40, 41], was introduced. In this model a gene is represented as a binary variable. The state of this variable at the next time step is determined by K other variables (genes) and a set of Boolean coupling functions. Depending upon the initial state, the system evolves to one of possibly several limit cycles, which in the biological picture are interpreted as different cell types.

It was found that for $K = 2$, these Boolean networks showed a remarkable stability; in most cases small perturbations of the state of the network did not change the trajectory to a different limit cycle. This is desirable in the biological interpretation, since stability of genetic networks against small fluctuations is a crucial property. Another striking observation was that the number of limit cycles seemed to grow as $\sqrt{N}$, where N is the system size. This was in analogy to multicellular organisms, where it is found empirically that the number of cell-types also grows approximately as the square-root of the genome-size.

In Paper V, we develop a method to decimate Kauffman networks by removing variables not relevant for the long-term dynamics of the system. We find that the reduced networks lack the well known stability observed in full Kauffman networks.

We also use the decimation method to facilitate full-enumeration studies of networks with $N \leq 32$. The results of this analysis show that the number of limit cycles grows faster with N than the often cited $\sqrt{N}$ behavior. In Ref. [42], it had been discussed that the asymptotic scaling of this observable could be faster than $\sqrt{N}$. Later it has been shown analytically that the asymptotic scaling in fact is faster than any polynomial in N [43].

Acknowledgments

First, I would like to thank my supervisor Anders for excellent and generous guidance. His expertise, patience and careful attention to details has been invaluable. I also would like to thank the other coauthors of the different papers, Björn, Giorgio, Stefan and Sven, for first-class teamwork and everybody else at the department for contributing to the great atmosphere. A special though of course goes to Stefan, with whom I have shared the office the past years, and to all our "part-time officemates", Peter, Giorgio, Pierre and others, who have never missed out on a lively discussion.

References

[1] Creighton, T.E. *Proteins: structures and molecular properties*, W.H. Freeman and Company, New York, 2nd edition.

[2] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E. (2000) "The protein data bank", *Nucleic Acids Res.* **28**, 235–242.

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

[4] For a special issue on the fourth critical assessment of techniques for protein folding (CASP4), see (2001), *Proteins: Struct. Funct. Genet.* **45(S5)**.

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

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

[7] Dill, K.A. (1990) "Dominant forces in protein folding", *Biochemistry* **29**, 7133–7155.

[8] Privalov, P.L. (1992) "Physical basis of the stability of the folded conformations of proteins", in *Protein Folding*, ed. Creighton, T.E. (Freeman, New York), 83–126.

[9] Matsumoto, M., Saito, S. & Ohmine, I. (2002) "Molecular dynamics simulation of the ice nucleation and growth process leading to water freezing", *Nature 416*, 409–413.

[10] Levinthal, C. (1968) "Are there pathways for protein folding?", *J. Chim. Phys.* **85**, 44–45.

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

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

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

[14] Bryngelson, J.D. & Wolynes, P.G. (1987) "Spin-glasses and the statistical-mechanics of protein folding", *Proc. Natl. Acad. Sci. USA* **84**, 7524–7528.

[15] Garel, T. & Orland, H. (1988) "Mean-field model for protein folding", *Europhys. Lett.* **6**, 307–310.

[16] Shakhnovich, E.I, & Gutin, A.M. (1989) "Frozen states of a disordered globular heteropolymer", *J. Phys. A-Math Gen* **22**, 1647–1659.

[17] Lau, K.F. & Dill, K.A. (1989) "A lattice statistical mechanics model of the conformational and sequence spaces of proteins", *Macromolecules* **22**, 3986–3997.

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

[19] Alm, E. & Baker, D. (1999) "Matching theory and experiment in protein folding", *Curr. Opin. Struct. Biol.* **9**, 189–196.

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

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

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

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

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

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

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

[27] Lal, M. (1969) "Monte Carlo computer simulation of chain molecules .I.", *Molec. Phys.* **17**, 57–64.

[28] Madras, N., Sokal, A.D. (1988) "The pivot algorithm - a highly efficient Monte-Carlo method for the self-avoiding walk", *J. Stat. Phys.* **50**, 109–186.

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

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

[31] Marianari, G., Parisi, G. (1992) "Simulated tempering: a new Monte Carlo scheme", *Europhys. Lett.* **19**, 451–458.

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

[33] Hukushima, K. & Nemoto, K. (1996) "Exchange Monte Carlo and application to spin glass simulations", *J. Phys. Soc. (Jap)* **65**, 1604–1608.

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

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

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

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

[38] Dodd, L.R., Boone, T.D. & Theodorou, D.N. (1993) "A concerted rotation algorithm for atomistic Monte-Carlo simulation of polymer melts and glasses", *Molec. Phys.* **78**, 961–996.

[39] Ulmschneider, J.P. & Jorgensen, W.L. (2003) "Monte Carlo backbone sampling for polypeptides with variable bond angles and dihedral angles using concerted rotations and a Gaussian bias", *J. Chem. Phys.* **118**, 4261–4271.

[40] Kauffman, S.A. (1969) "Metabolic stability and epigenesis in randomly constructed genetic nets", *J. Theor. Biol.* **22**, 437–467.

[41] Kauffman, S.A. (1993) *The origins of order* (Oxford University Press).

[42] Bastolla, U. & Parisi, G. (1997) "A numerical study of the critical line of Kauffman networks", *J. Theor. Biol.* **187**, 117–133.

[43] Samuelsson, B. & Troein, C. (2003) "Superpolynomial growth in the number of attractors in Kauffman networks", *Phys. Rev. Lett.* **90**, 098701.

Three-Helix-Bundle Protein in a Ramachandran Model

Paper I

Three-Helix-Bundle Protein in a Ramachandran Model

Anders Irbäck, Fredrik Sjunnesson and Stefan Wallin

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

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

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

1.1 Introduction

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

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

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

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

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

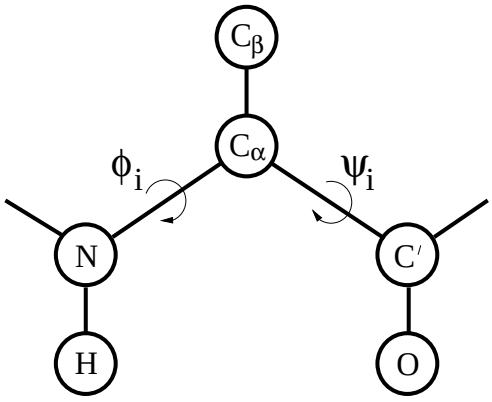

Figure 1.1: Schematic figure showing the representation of one amino acid.

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

1.2 The Model

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

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

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

Table 1.1: Geometry parameters.

Our energy function

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

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

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

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

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

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

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

where

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

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

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

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

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

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

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

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

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

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

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

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

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

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

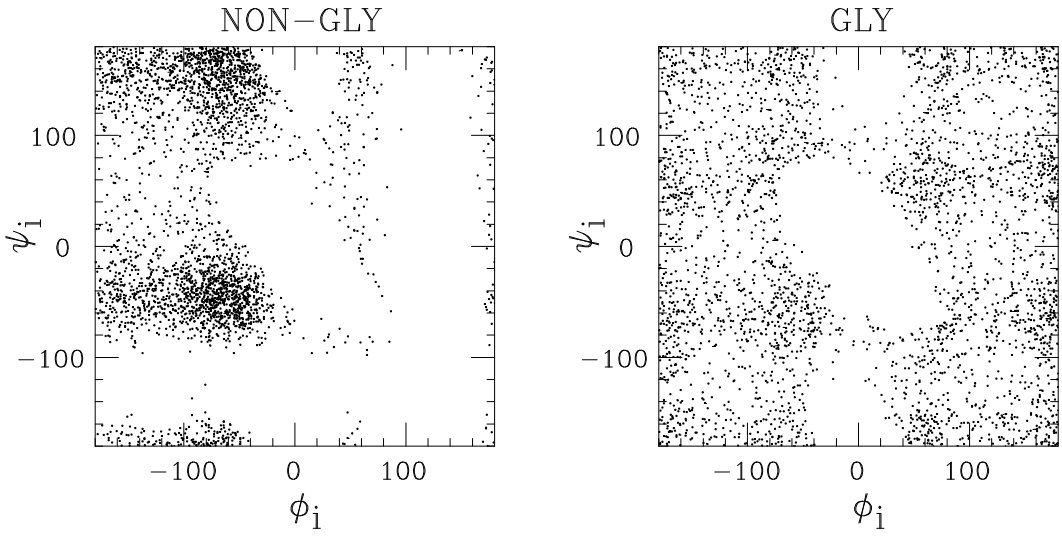

Figure 1.2: ϕ_i, ψ_i scatter plots for nonglycine and glycine, as obtained by simulations of the chains GXG for X=A/B and X=G, respectively, at $kT = 0.625$ (shown is ϕ_i, ψ_i for X).

In this model, we study the three sequences shown in Table 1.3, which contain 16, 35 and 54 amino acids, respectively. Following the strategy of Regan and DeGrado [11], the A and B amino acids are distributed along the sequence 1H in such a way that this segment can form a helix with all hydrophobic amino acids on the same side. The sequence 3H, consisting of three such stretches of As and Bs plus two GGG segments, is meant to form a three-helix bundle. This particular sequence was recently studied by Takada *et al.* [18], who used a more elaborate model with nonadditive forces.

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

Table 1.3: The sequences studied.

1.3 Results

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

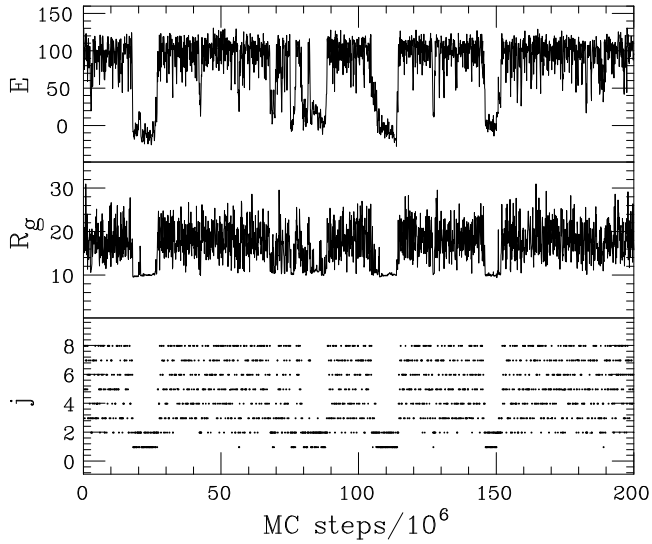

Figure 1.3: Monte Carlo evolution of the energy and radius of gyration in a typical simulation of the three-helix sequence. The bottom panel shows how the system jumps between the allowed temperatures T_j, which are given by $T_j = T_{\min}(T_{\max}/T_{\min})^{(j-1)/(J-1)}$ [34] with $kT_{\min} = 0.625$,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

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

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

In Fig. 1.4a we show the specific heat as a function of temperature for the one-, two- and three-helix sequences. A pronounced peak can be seen that gets stronger with increasing chain length. In fact, the increase in height is not inconsistent with a linear dependence on chain length, which is what one would have expected if it had been a conventional first-order phase transition with a latent heat.

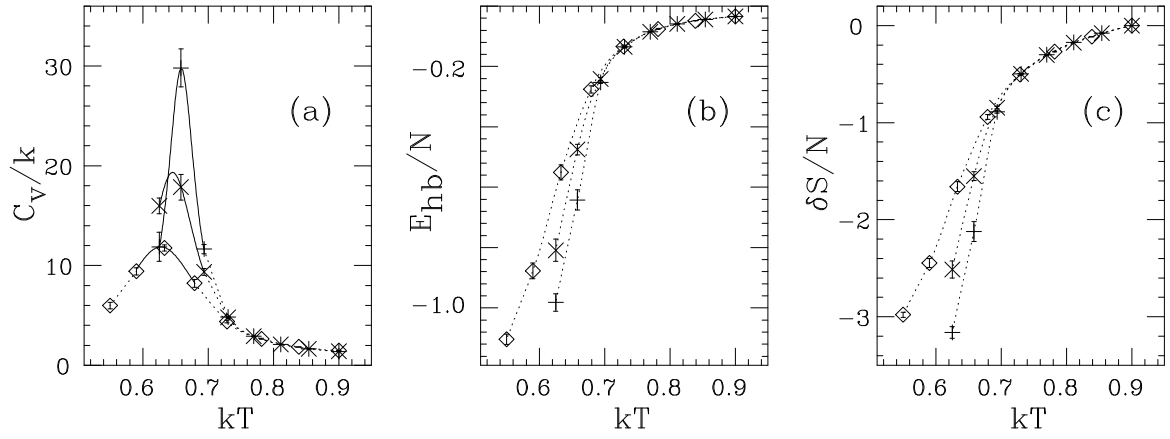

Figure 1.4: Thermodynamic functions against temperature for the sequences 1H ($\diamond$), 2H ($\times$) and 3H ($+$) in Table 1.3. (a) Specific heat $C_v = (\langle E^2\rangle - \langle E\rangle^2)/NkT^2$, N being the number of amino acids. (b) Hydrogen-bond energy per amino acid, E_{hb}/N. (c) Chain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

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

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

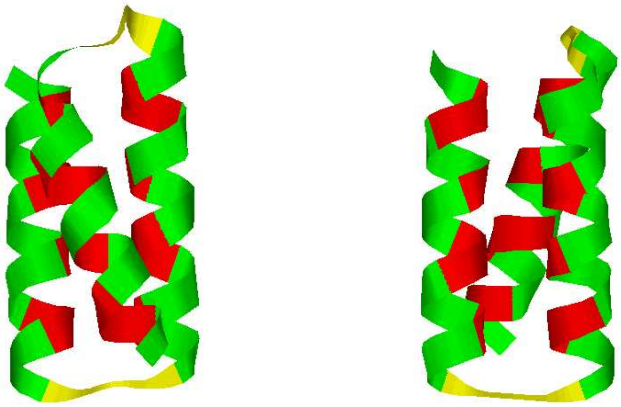

Figure 1.5: Representative low-temperature structures, FU and BU, respectively. Drawn with RasMol [36].

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

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

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

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

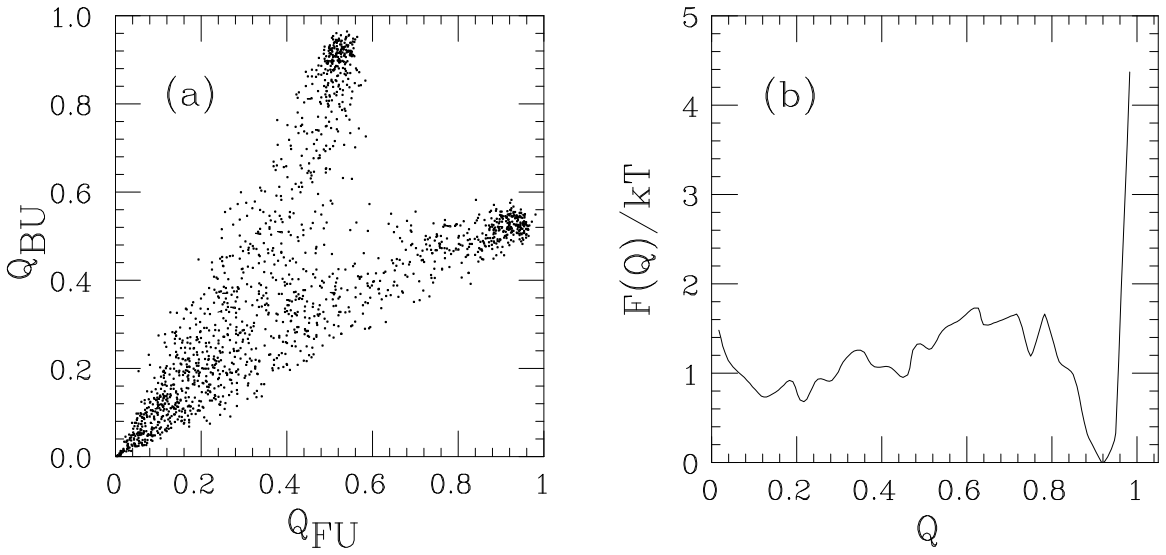

Figure 1.6: (a) $Q_{\mathrm{FU}}, Q_{\mathrm{BU}}$ (see Eq. 1.8) scatter plot at the specific heat maximum ($kT = 0.658$). (b) Free energy $F(Q)$ as a function of $Q = \max(Q_{\mathrm{FU}}, Q_{\mathrm{BU}})$ at the same temperature.

at the folding temperature. The free energy has a relatively sharp minimum at $Q \approx 0.9$, corresponding to $\delta \approx 3$Å. This is followed by a weak barrier around $Q = 0.7$, corresponding to $\delta \approx 6$Å. Finally, there is a broad minimum at small Q, where $Q = 0.2$ corresponds to $\delta \approx 13$Å.

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

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

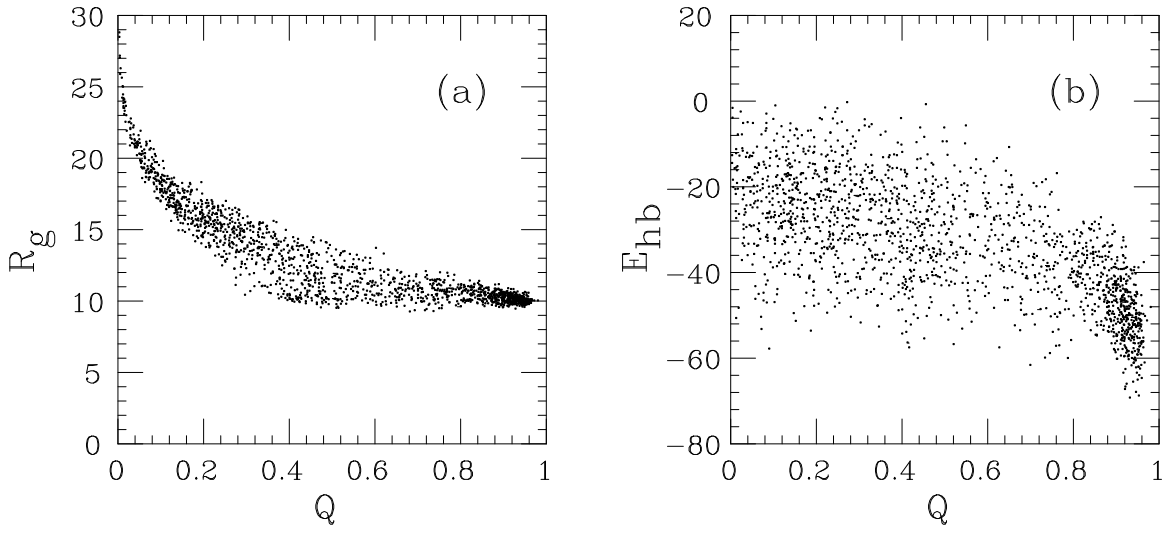

Figure 1.7: (a) $Q, R_{\rm g}$ and (b) $Q, E_{\rm hb}$ scatter plots at the folding temperature ($kT = 0.658$).

1.4 Summary and Outlook

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

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

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

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

work on homopolymers [26–29]. Note also that incorporation of full side chains makes the chains intrinsically stiffer, which might lead to a weaker $\epsilon_{\rm hb}/\epsilon_{\rm AA}$ dependence.

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

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

Acknowledgements

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Thermodynamics of α- and β-Structure Formation in Proteins

Paper II

Thermodynamics of α- and β-Structure Formation in Proteins

Anders Irbäck, Björn Samuelsson,
Fredrik Sjunnesson and Stefan Wallin

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

In press: *Biophysical Journal*

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

2.1 Introduction

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

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

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

2.2 Model and Methods

2.2.1 The Model

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

The potential function

$$E = E_{\rm ev} + E_{\rm hb} + E_{\rm hp} \tag{2.1}$$

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

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

$$E_{\rm ev} = \epsilon_{\rm ev} \sum_{i<j} \left[\frac{\lambda_{ij}(\sigma_i + \sigma_j)}{r_{ij}} \right]^{12}, \tag{2.2}$$

where $\epsilon_{\rm ev} = 0.10$ and $\sigma_i = 1.77$, 1.71, 1.64, 1.42 and 1.00 Å for S, C, N, O and H atoms, respectively. Our choice of σ_i values is guided by the analysis of Tsai *et al.* [20]. The parameter λ_{ij} in Eq. 2.2 reduces the repulsion between non-local pairs; $\lambda_{ij} = 1$ for all pairs connected by three covalent bonds and for HH and OO pairs from adjacent peptide units, and $\lambda_{ij} = 0.75$ otherwise. The pairs for which $\lambda_{ij} = 1$ strongly influence the shapes of Ramachandran maps and rotamer potentials. The reason for using $\lambda_{ij} < 1$ for the large majority of all pairs is both computational efficiency and the restricted flexibility of chains with only torsional degrees of freedom. To speed up the calculations, the sum in Eq. 2.2 is evaluated using a pair dependent cutoff $r_{ij}^{\rm c} = 4.3\lambda_{ij}$ Å.

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

$$E_{\rm hb} = \epsilon_{\rm hb}^{(1)} \sum_{\substack{j<i-2 \\ \text{or } j>i+1}} u(r_{ij}) v(\alpha_{ij}, \beta_{ij}) + \epsilon_{\rm hb}^{(2)} \sum u(r_{ij}) v(\alpha_{ij}, \beta_{ij}) \,, \tag{2.3}$$

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

$$u(r) = 5\left(\frac{\sigma_{\rm hb}}{r}\right)^{12} - 6\left(\frac{\sigma_{\rm hb}}{r}\right)^{10} \tag{2.4}$$

$$v(\alpha,\beta) = \begin{cases} (\cos\alpha\cos\beta)^{1/2} & \text{if } \alpha,\beta > 90^\circ \\ 0 & \text{otherwise} \end{cases} \tag{2.5}$$

The first sum in Eq. 2.3 represents backbone-backbone hydrogen bonds. Term ij in this sum is an interaction between the NH and C′O groups of amino acids i and j, respectively. r_{ij} denotes the HO distance, and α_{ij} and β_{ij} are the NHO and HOC′ angles, respectively. The second sum in Eq. 2.3 is expressed in a schematic way. It represents interactions between oppositely charged side chains, and between charged side chains and the backbone. Both these types of interaction are, for convenience, taken to have the same form as backbone-backbone hydrogen bonds. The side chain atoms that can act as "donors" or "acceptors" in these interactions are the N atoms of Lys and Arg (donors) and the O atoms of Asp and Glu (acceptors). The second sum in Eq. 2.3 has a relatively weak influence on the thermodynamic behavior of the systems studied. The backbone-backbone hydrogen bonds are, by contrast, crucial and their strength, $\epsilon_{\rm hb}^{(1)}$, must be carefully chosen [17].

The functional form of the hydrogen-bond energy differs from that in our helix model [16–18] in that the exponent of the cosines is 1/2 instead of 2. The reason for this change is that the β-hairpin turned out to become too regular when using the exponent 2; the exponent 1/2 gives a more permissive angular dependence. The function $u(r)$ in Eq. 2.4 is calculated using a cutoff $r^{\rm c} = 4.5\,\text{Å}$ and $\sigma_{\rm hb} = 2.0\,\text{Å}$.

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

$$E_{\rm hp} = \epsilon_{\rm hp} \sum M_{IJ} C_{IJ} \,, \tag{2.6}$$

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

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

Table 2.1: The interaction matrix M_{IJ}, based on the shifted contact-energy matrix of Miyazawa and Jernigan [21]. The table shows absolute values ($M_{IJ} \leq 0$).

the shifted contact energy is negative; otherwise, $M_{IJ} = 0$. The statistical Miyazawa-Jernigan energies contain, of course, other contributions too, but receive a major contribution from hydrophobicity [22]. The matrix M_{IJ} is given in Table 2.1. Eight of the amino acids are classified as hydrophobic, namely Ala, Val, Leu, Ile, Phe, Tyr, Trp and Met. The geometry factor C_{IJ} in Eq. 2.6 is a measure of the degree of contact between amino acids I and J. To define C_{IJ}, we use a predetermined set of N_I atoms, denoted by A_I, for each amino acid I. For Phe, Tyr and Trp, the set A_I consists of the C atoms of the hexagonal ring. The other five hydrophobic amino acids each have an A_I containing all its non-hydrogen side-chain atoms. With these definitions, C_{IJ} can be written as

$$C_{IJ} = \frac{1}{N_I + N_J} \left[\sum_{i \in A_I} f(\min_{j \in A_J} r_{ij}^2) + \sum_{j \in A_J} f(\min_{i \in A_I} r_{ij}^2) \right], \tag{2.7}$$

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

2.2.2 Numerical Methods

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

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

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

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

2.3 Results and Discussion

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

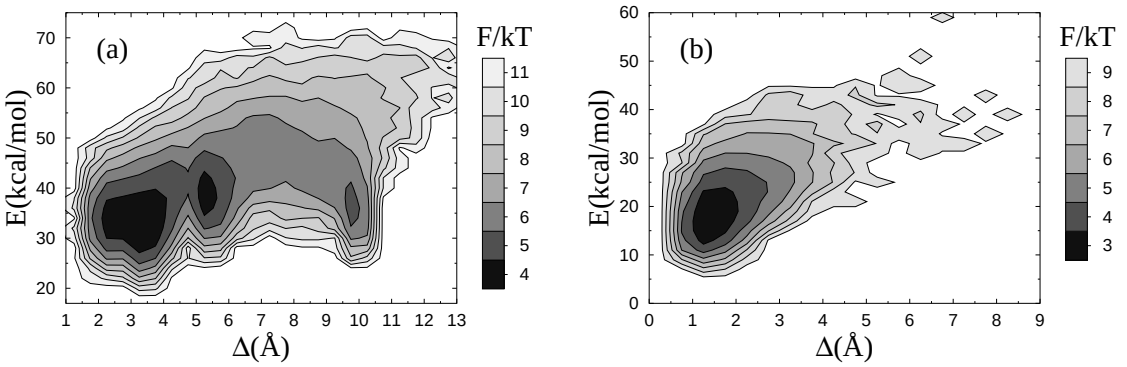

Figure 2.1: Free energy $F(\Delta, E) = -kT \ln P(\Delta, E)$ at $T = 273\,\mathrm{K}$ for (a) the β-hairpin and (b) the $\mathrm{F_S}$ peptide. E is energy and Δ denotes rmsd from the native β-hairpin and an ideal α-helix, respectively, calculated over all non-hydrogen atoms (a backbone rmsd would be unable to distinguish the two possible β-hairpin topologies).

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

Fig. 2.1a shows the free energy $F(\Delta, E)$ as a function of rmsd from the native β-hairpin, Δ, and energy, E, at the temperature $T = 273\,\mathrm{K}$. For a β-hairpin there are two topologically distinct states with similar backbone folds but oppositely oriented side chains. The global minimum of $F(\Delta, E)$ is found at 2–4 Å in Δ and corresponds to a β-hairpin with the native topology and the native set of hydrogen bonds between the two strands. The main difference between structures within this minimum lies in the shape of the turn. The precise shape of the β-hairpin is, not unexpectedly, sensitive to details of the potential; in particular, we find that the second term in Eq. 2.3 does influence the shape of the turn, while having only a small effect on thermodynamic functions such as E_{hp}. Therefore, it is not unlikely that a more detailed potential would discriminate between different shapes of the turn, and thereby make the free-energy minimum more narrow.

Besides its global minimum, $F(\Delta, E)$ exhibits two local minima (see Fig. 2.1a), one corresponding to a β-hairpin with the non-native topology ($\Delta \approx 5\,\mathrm{Å}$), and the other to an α-helix ($\Delta \approx 10\,\mathrm{Å}$). A closer examination of structures from the two β-hairpin minima reveals that the C_β-C_β distances for Tyr45-Phe52 and Trp43-Val54 tend to be smaller in the non-native topology than in the native one. This is important because it makes it sterically difficult to achieve a proper contact between the aromatic side chains of Tyr45 and Phe52 in the non-native topology. As a result, this topology is hydrophobically disfavored.

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

We now turn to the melting behavior of the β-hairpin. By studying tryptophan fluorescence (Trp43), Muñoz *et al.* [5] found that the unfolding of this peptide with increasing temperature shows two-state character, with parameters $T_\mathrm{m} = 297\,\mathrm{K}$ and $\Delta E = 11.6\,\mathrm{kcal/mol}$, T_m and ΔE being the melting temperature and energy change, respectively. To study the character of the melting transition in our model, we monitor the hydrophobicity energy E_hp, a simple observable we expect to be strongly correlated with Trp43 fluorescence. Following Muñoz *et al.* [5], we fit our data for E_hp to a first-order two-state model. To reduce the number of parameters of the fit, T_m is held fixed, at the specific heat maximum (data not shown). The fit turns out not to be perfect, with a χ^2/dof of 4.5. The deviations from the fitted curve are nevertheless small, as can be seen from Fig. 2.2a; they can be detected only because the statistical errors are very small ($\sim 0.1\,\%$) at the highest temperatures. To further illustrate this point, we assign each data point an artifical uncertainty of $1\,\%$, an error size that is not uncommon for experimental data. With these errors, the same type of fit yields a χ^2/dof of 0.3, which confirms that the data indeed to a good approximation show two-state behavior. Our fitted value of ΔE is $9.3 \pm 0.3\,\mathrm{kcal/mol}$, which implies that the temperature dependence of the model is comparable to experimental data [5].

Several groups have simulated the same β-hairpin using atomic models with implicit [6, 7] or explicit [8–11] solvent. All these models have, in contrast to ours, given a very weak dependence on temperature, compared to experimental data [11]. Another important difference between at least some of these models [7, 9, 10] and ours, is that in our model there is no clear free-energy minimum corresponding to a hydrophobically collapsed state with few or no hydrogen bonds. A local free-energy minimum with helical content was found in one of these studies [10], but not in the others. Such a minimum exists in our model (see Fig. 2.1a), but the helix population is low.

In spite of its minimalistic potential, our model is able to make α-helices too. To show this, we consider the α-helical so-called $\mathrm{F_S}$ peptide, which has been extensively studied both experimentally [12–15] and theoretically [32]. This 21-amino acid peptide is given by $\mathrm{AAAAA(AAARA)_3A}$, where A is Ala and R is Arg. Using exactly the same model as before, with unchanged parameters, we find that the $\mathrm{F_S}$ sequence does make an α-helix. This can be seen from Fig. 2.1b, which shows the free energy $F(\Delta, E)$ at $T = 273\,\mathrm{K}$, Δ this time denoting rmsd from an ideal α-helix. $F(\Delta, E)$ has only one significant minimum, which indeed is helical. The melting behavior of this sequence is illustrated in Fig. 2.3a, which shows the temperature dependence of the hydrogen-bond energy. Data

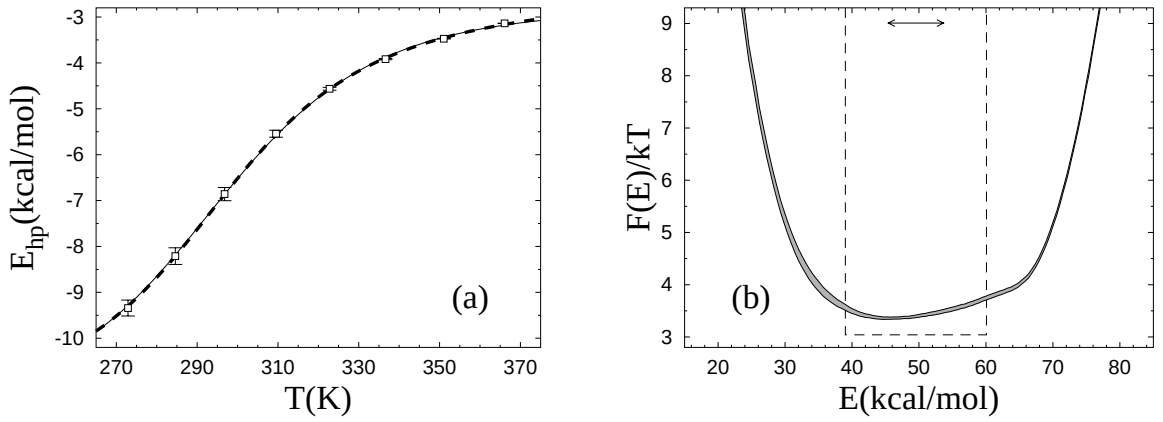

Figure 2.2: Unfolding of the β-hairpin sequence. (a) Temperature dependence of the hydrophobicity energy $E_{\rm hp}$ (see Eq. 2.6). The solid and dashed curves (essentially coinciding) are fits of the data to the two-state expression $E_{\rm hp} = (E^{\rm u}_{\rm hp} + K E^{\rm f}_{\rm hp})/(1+K)$ and the square-well model (see text), respectively. The effective equilibrium constant K is assumed to have the first-order form $K = \exp[(1/kT - 1/kT_{\rm m})\Delta E]$. Both fits have three free parameters, whereas $T_{\rm m} = 297\,{\rm K}$ is held fixed. (b) Free-energy profile $F(E) = -kT \ln P(E)$ at $T = T_{\rm m}$, obtained by reweighting [33] the data at a simulated T close to $T_{\rm m}$. The shaded band is centered around the expected value and shows statistical 1σ errors. The double-headed arrow indicates ΔE of the two-state fit. The dashed line shows $F(E)$ for the square-well fit.

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

The two-state fits shown in Figs. 2.2a and 2.3a are based on a first-order expression for the free energies of the two coexisting phases. The fits look good and can be improved by including higher order terms, which may give the impression that the behaviors of these systems can be fully understood in terms of a two-state model. However, the two-state picture is far from perfect. This can be seen from the free-energy profiles $F(E)$ shown in Figs. 2.2b and 2.3b,

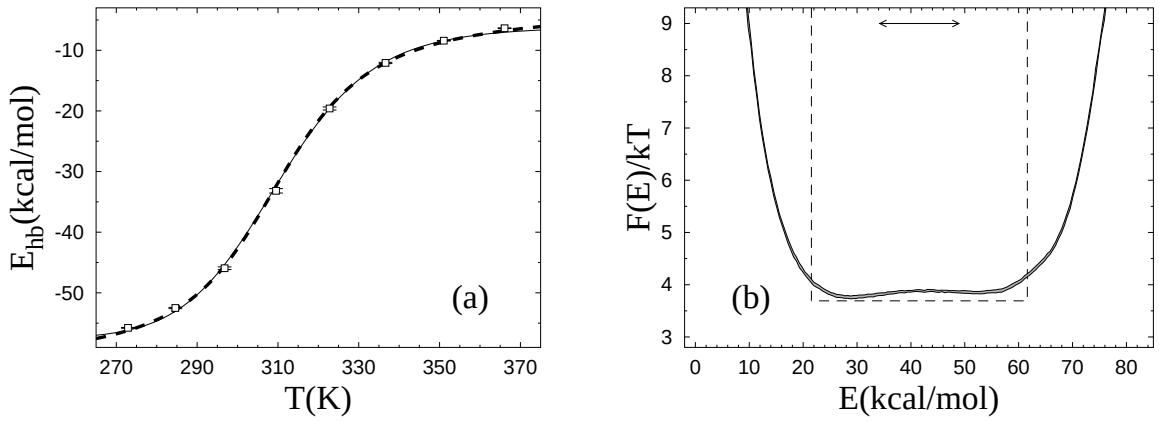

Figure 2.3: Unfolding of the F_S sequence. (a) Temperature dependence of the hydrogen-bond energy E_{hb} (see Eq. 2.3), with the same two types of fit as in Fig. 2.2a (same symbols). (b) Free-energy profile $F(E) = -kT \ln P(E)$ at $T = T_m$. Same symbols as in Fig. 2.2b.

which lack a clear bimodal shape. Clearly, this renders the parameters of a two-state model, such as ΔE, ambiguous. The analysis of these systems therefore shows that the results of a two-state fit must be interpreted with care. Given the actual shapes of $F(E)$, it is instructive to perform an alternative fit of the data in Figs. 2.2a and 2.3a, based on the assumptions that 1) $F(E)$ has the shape of a square well of width ΔE_{sw} at $T = T_m$, and that 2) the observable analyzed varies linearly with E.[1] These square-well fits are shown in Figs. 2.2a and 2.3a, and the corresponding free-energy profiles $F(E)$ (at $T = T_m$) are indicated in Figs. 2.2b and 2.3b. The square-well fits are somewhat better than the two-state fits. However, the fitted curves are strikingly similar, given the large difference between the underlying energy distributions. This shows that it is very hard to draw conclusions about the free-energy profile $F(E)$ from the temperature dependence of a single observable.

From Figs. 2.2b and 2.3b it can also be seen that the energy change ΔE obtained from the two-state fit is considerably smaller than the width of the energy distribution, which indicates that ΔE is smaller than the calorimetric energy change ΔE_{cal}. Scholtz *et al.* [34] determined ΔE_{cal} experimentally for an Ala-based helical peptide with 50 amino acids, and obtained a value of 1.3 kcal/mol per amino acid. This value corresponds to a ΔE_{cal} of 27.3 kcal/mol for the

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

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

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

Our kinetic simulations of the two peptides are performed at their respective melting temperatures, $T_{\rm m}$. Starting from equilibrium conformations at $T = 366\,{\rm K}$, we study the relaxation of ensemble averages under Monte Carlo dynamics (see Section 2.2). The ensemble consists of 1500 independent runs for each peptide. In Fig. 2.4, we show the "time" evolution of $\delta O(t) = O(t) - \langle O \rangle$,

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

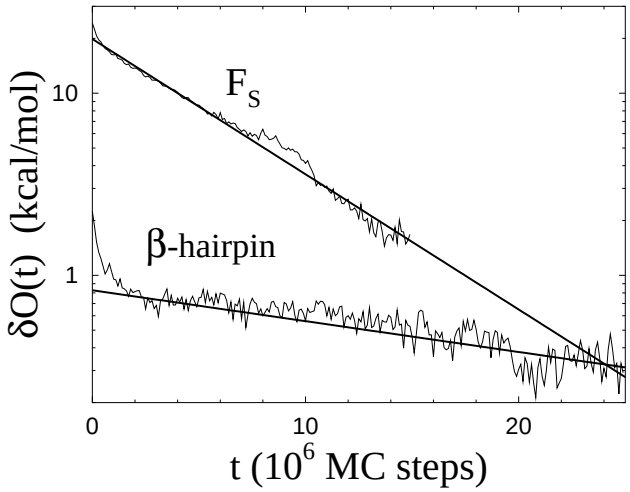

Figure 2.4: Monte Carlo relaxation of ensemble averages at $T = T_{\mathrm{m}}$ for the β-hairpin and the $\mathrm{F_S}$ peptide. The deviation $\delta O(t)$ from the equilibrium average (see text) is plotted against the number of elementary Monte Carlo steps, t. Straight lines are χ^2 fits of the data to a single exponential. Data for $t > 15{\cdot}10^6$ are omitted for $\mathrm{F_S}$ due to large statistical errors.

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

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

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

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

2.4 Summary and Outlook

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

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

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

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

- Our thermodynamic simulations show first of all that the two sequences studied indeed make a β-hairpin with the native topology and an α-helix, respectively. The main reason why the model favors the native topology over the non-native one for the β-hairpin, is that the formation of the hydrophobic cluster is sterically difficult to accomplish in the non-native topology. The melting curves obtained for the two peptides are in reasonable agreement with experimental data, and can to a good approximation be described by a simple two-state model.
- A two-state description of the thermodynamic behavior is, nevertheless, found to be an oversimplification for both peptides, as can be seen from the energy distributions. Given that the systems are small and fluctuations therefore relatively large, this is maybe not surprising. What is striking is how difficult it is to detect these deviations from two-state behavior when studying the temperature dependence of a single observable.
- The results of our Monte Carlo-based kinetic runs at the respective melting temperatures are, for both peptides, consistent with single-exponential relaxation, and the relaxation time is found to be larger for the β-hairpin than for F_S.

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

Acknowledgments

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Supplemental Material

Bond lengths (in Å)

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

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

Bond angles (in degrees)

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

Bond angles cont.

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

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

Torsion angles (in degrees)

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

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

Number of side-chain DOFs (χ_i)

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

Folding Thermodynamics of Three β-Sheet Peptides: A Model Study

Paper III

Folding Thermodynamics of Three β-Sheet Peptides: A Model Study

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/

Abstract:
We study the folding thermodynamics of a β-hairpin and two three-stranded β-sheet peptides using a simplified sequence-based all-atom model, in which folding is driven mainly by backbone hydrogen bonding and effective hydrophobic attraction. The native populations obtained for these three sequences are in good agreement with experimental data. We also show that the apparent native population depends on which observable is studied; the hydrophobicity energy and the number of native hydrogen bonds give different results. The magnitude of this dependence matches well with the results obtained in two different experiments on the β-hairpin.

3.1 Introduction

Peptide folding is currently attracting considerable attention. Recent advances in this area include the *de novo* design of two monomeric three-stranded antiparallel β-sheet peptides, Betanova [1, 2] and Beta3s [3]. Peptides that have the ability to fold on their own and are well characterized experimentally are valuable not least as a testbed for theoretical models and methods for protein folding. β-sheet peptides are particularly interesting in this respect, as β-sheet formation is more challenging to model than α-helix formation. Therefore, it is no surprise that both Betanova [4,5] and Beta3s [6] have become the subject of computational studies. Simulations of peptide sequences that are somewhat similar to these and occur in natural proteins, so-called WW domains, have been reported, too [7]. For a recent review of computational studies of peptide folding, see Ref. [8].

Here we present a study of the C-terminal β-hairpin from the protein G B1 domain and a triple mutant of Betanova called LLM [2]. The original Betanova, which is less stable than the peptide LLM [2], is considered too. These different sequences are studied using an all-atom model with a simplified interaction potential. An earlier version of this model was tested on the same β-hairpin and an α-helix, the designed so-called F_S [9, 10], with encouraging results [11]. The model was able to fold these sequences and the folded populations showed a temperature dependence comparable with experimental data. It should be pointed out that the interaction potential of this model, like that in Ref. [12] but unlike many other simplified potentials for protein folding, is sequence-based.

The paper is organized as follows. In Sec. 2 we describe the model and the computational methods used. The results obtained for the three β-sheet peptides are discussed in Sec. 3. A brief summary can be found in Sec. 4.

3.2 Model and Methods

The model we study is a revised version of that developed in Ref. [11]. It contains all atoms of the polypeptide chain, including hydrogens, but no explicit water molecules. All bond lengths, bond angles and peptide torsion angles (180°) are held fixed, so each amino acid has the Ramachandran torsion angles ϕ, ψ and a number of side-chain torsion angles as its degrees of freedom (for Pro, ϕ is held fixed at $-65°$). Numerical values of all the geometry parameters can be found in Ref. [11].

The potential function

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

is composed of four terms. The remaining part of this section describes these different terms, with emphasis on what is new compared with Ref. [11]. Energy parameters are quoted in dimensionless units. To set the energy scale of the model, we use the midpoint temperature for the β-hairpin as determined by Muñoz *et al.* [13], $T_{\rm m} = 297\,{\rm K}$, which corresponds to $kT \approx 0.440$ in the model (see Sec. 3.1).

The first term in Eq. 3.1, $E_{\rm ev}$, represents excluded-volume effects and has the form

$$E_{\rm ev} = \kappa_{\rm ev} \sum_{i<j} \left[\frac{\lambda_{ij}(\sigma_i + \sigma_j)}{r_{ij}} \right]^{12} , \tag{3.2}$$

where $\kappa_{\rm ev} = 0.10$ 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 role of the parameter λ_{ij} is to reduce the repulsion between non-local pairs; $\lambda_{ij} = 1$ for all pairs connected by three covalent bonds and $\lambda_{ij} = 0.75$ otherwise. The reason for using $\lambda_{ij} < 1$ for non-local pairs is both computational efficiency and the restricted flexibility of chains with only torsional degrees of freedom. To speed up the calculations, the sum in Eq. 3.2 is evaluated using a cutoff of $r_{ij}^{\rm c} = 4.3\lambda_{ij}$ Å.

The second interaction term, $E_{\rm loc}$, is new compared with the earlier model. By introducing this term and modifying σ_i for C and N, we slightly adjusted the shape of the Ramachandran ϕ, ψ distribution. $E_{\rm loc}$ is a local electrostatic energy given by

$$E_{\rm loc} = \kappa_{\rm loc} \sum_I \rho_I \left(\sum \frac{q_i q_j}{r_{ij}^{(I)}/\text{Å}} \right) , \tag{3.3}$$

where the outer sum runs over all non-Pro amino acids along the chain, and the inner sum represents the interaction between the partial charges of the backbone NH and C′O groups within one amino acid (the sum has four terms: NC′, NO, HC′ and HO). The partial charges are $q_i = \pm 0.20$ for H and N and $q_i = \pm 0.42$ for C′ and O [14]. We put $\kappa_{\rm loc} = 125$, which corresponds to a dielectric constant of $\epsilon_r \approx 2.0$ if $\rho_I = 1$. The factor ρ_I reduces the interaction strength for the two end amino acids and Gly, which can be viewed as a crude form of context dependence; $\rho_I = 0.25$ for end amino acids, $\rho_I = 0.5$ for Gly, and $\rho_I = 1$ otherwise. A similar factor is used for hydrogen bonds (see below).

The third term in Eq. 3.1, $E_{\rm hp}$, is an effective attraction between hydrophobic side chains that are not nearest or next-nearest neighbors along the chain. It

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

Table 3.1: The interaction matrix M_{IJ} (see Eq. 3.4). All amino acid pairs not occurring in the table have $M_{IJ} = 0$.

has the pairwise additive form

$$E_{\rm hp} = -\sum M_{IJ} C_{IJ} \,, \tag{3.4}$$

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 contact gets. The contact measure C_{IJ} is a number between 0 and 1, defined as in Ref. [11]. The interaction matrix M_{IJ} is given in Table 3.1 and differs from that used in Ref. [11], which was based on the Miyazawa-Jernigan contact energies [15]. With an all-atom representation, this cannot be expected to be a good choice for more general sequences, since the Miyazawa-Jernigan contact energies were derived using a different, reduced chain representation [15]. The new matrix M_{IJ} (see Table 3.1) has a simpler structure than the previous one in that the hydrophobic amino acids are grouped into three classes. The M_{IJ} values are taken to be large for the aromatic class (Phe, Trp, Tyr), which in part is an attempt to compensate for the fact that it is relatively difficult for these large side chains with few degrees of freedom to make proper contacts.

The last term of the potential, the hydrogen-bond energy $E_{\rm hb}$, is given by

$$E_{\rm hb} = \epsilon_{\rm hb}^{(1)} \sum_{\rm bb-bb} \rho_{ij} u(r_{ij}) v(\alpha_{ij}, \beta_{ij}) + \epsilon_{\rm hb}^{(2)} \sum_{\rm sc-bb} \rho_{ij} u(r_{ij}) v(\alpha_{ij}, \beta_{ij}) \,, \tag{3.5}$$

where the two terms represent backbone-backbone interactions and interactions between the backbone and charged side chains, respectively. We do not include any side chain-side chain interactions, as was done in Ref. [11]. Apart from that, the only difference compared with the earlier model is the factor ρ_{ij}, which like ρ_I in Eq. 3.3 can be seen as a simple form of context dependence. We put $\rho_{ij} = 0.25$ if any of the two amino acids involved is an end amino acid, $\rho_{ij} = 0.5$ if any of them is a Gly, and $\rho_{ij} = 1$ otherwise. The constants $\epsilon_{\rm hb}^{(1)} = 3.1$ and $\epsilon_{\rm hb}^{(2)} = 2.0$ as well as the functions u and v are exactly the same as in Ref. [11].

To study the thermodynamic behavior of this model, we use simulated tempering [16, 17], in which the temperature is a dynamical variable. Details on our

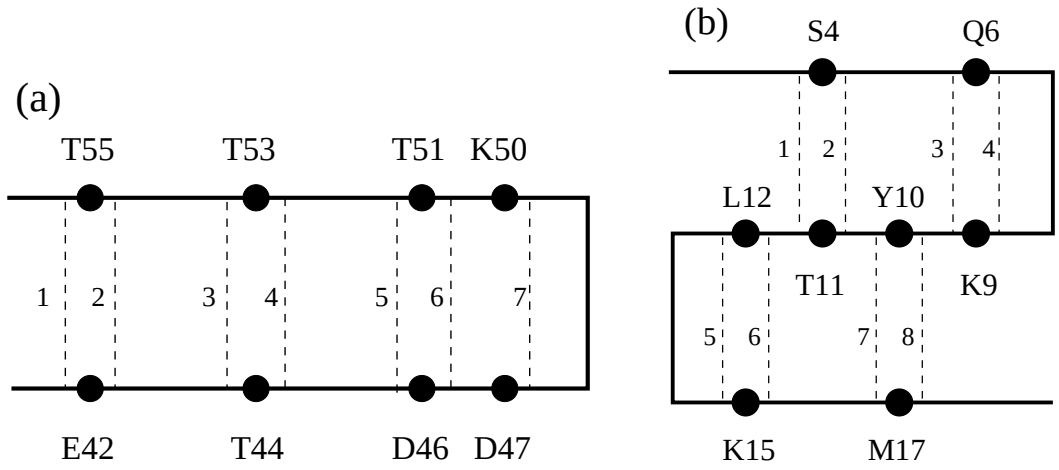

Figure 3.1: Schematic illustration of the backbone-backbone hydrogen bonds taken as native for (a) the C-terminal β-hairpin from the protein G B1 domain [13, 23], and (b) the mutant LLM of Betanova [2]. Diagram (b) is used for the original Betanova, too (with L12 and M17 replaced by N12 and T17, respectively).

implementation of this method can be found in Ref. [18]. For a review of simulated tempering and other generalized-ensemble techniques for protein folding, see Ref. [19]. Eight different temperatures are studied, ranging from 284 K to 371 K. For the backbone degrees of freedom, we use three different elementary moves: first, the pivot move [20] in which a single torsion angle is turned; second, a semi-local method [21] that works with seven or eight adjacent torsion angles, which are turned in a coordinated way; and third, a symmetry-based update of three randomly chosen backbone torsion angles [11]. For the side-chain degrees of freedom, we use simple Metropolis updates of individual angles. All our simulations are started from random conformations and contain several folding/unfolding events. All statistical errors quoted are 1σ errors based on the results from eight independent runs for each peptide. The fits of data discussed in the next section are carried out by using a Levenberg-Marquardt procedure [22].

For a given protein structure, there generally exist alternative structures with similar secondary-structure content but different overall topologies. This holds true even for a small β-hairpin, for which a flip of the side chains gives rise to a topologically distinct structure. To make models discriminate between different topologies is a delicate task. To assess whether or not a model is able to do that, it is necessary to make a suitable choice of observables. In our calculations, we monitor two variables that can be used for this purpose: first, the root-mean-square deviation (rmsd) from the folded structure, Δ, calculated over all non-H atoms (a backbone rmsd is much less informative); and second,

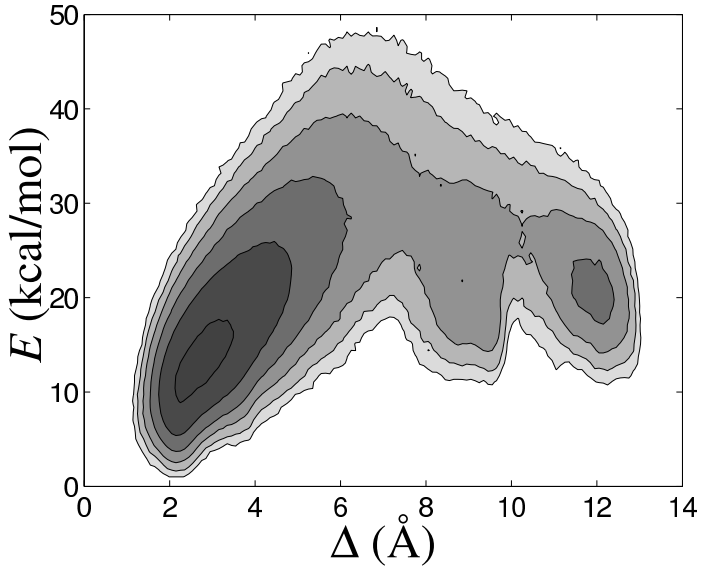

Figure 3.2: Free energy $F(\Delta, E)$ for $\mathrm{F_S}$ at $T = 284\,\mathrm{K}$, where Δ denotes heavy-atom rmsd from an ideal α-helix and E is energy. The contours are spaced at intervals of $1\,kT$ and dark tone corresponds to low free energy. Contours more than $6\,kT$ above the minimum free energy are not shown.

the number of native backbone-backbone hydrogen bonds, $N_{\mathrm{hb}}^{\mathrm{nat}}$. Figure 3.1 illustrates which hydrogen bonds we take to be present in the native states of the peptides studied. In our calculations, a hydrogen bond is considered formed if the energy is less than $-\epsilon_{\mathrm{hb}}^{(1)}/3$ (see Eq. 3.5).

Clearly, this simplified potential is not expected to be able to fold arbitrary sequences. For example, the pairwise additive ansatz for the hydrophobicity potential is most likely insufficient for long chains. On the other hand, the results obtained using the first version of the potential, for $\mathrm{F_S}$ and the β-hairpin, were quite good, and by confronting the model with new sequences it should be possible refine the potential.

Here, using the revised model described above, we investigate the folding thermodynamics of the β-hairpin from our earlier study and the three-stranded β-sheet peptides LLM and Betanova. Before turning to these results, it should be pointed out that the $\mathrm{F_S}$ sequence still makes an α-helix in the revised model, as can be seen from the free energy $F(\Delta, E)$ in Fig. 3.2. $F(\Delta, E)$ has a pronounced, dominating minimum at $\Delta \approx 2$–$3.5\,\mathrm{Å}$, which corresponds to α-helix. In addition, there are weakly populated minima corresponding to β-sheet structures at $\Delta \approx 9\,\mathrm{Å}$ and $\Delta \approx 12\,\mathrm{Å}$.

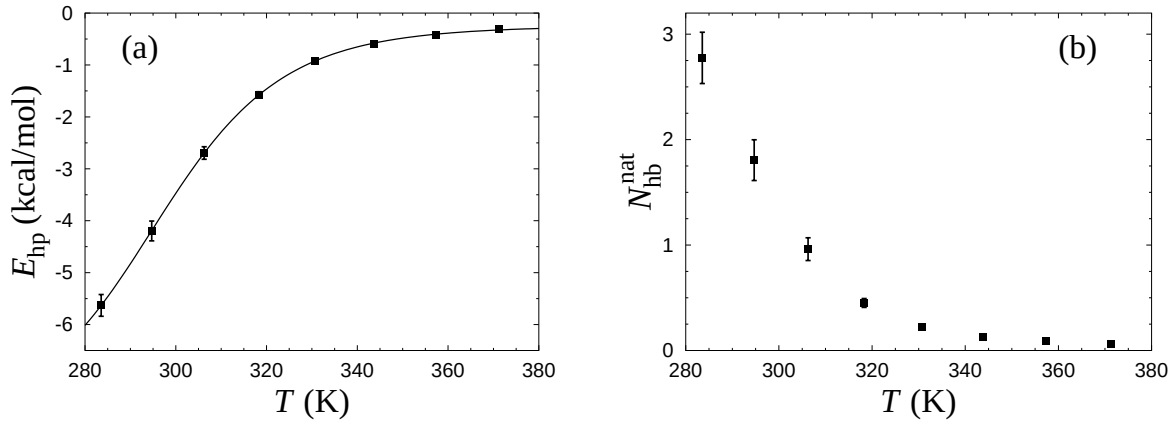

Figure 3.3: The temperature dependence of (a) the hydrophobicity energy $E_{\rm hp}$ and (b) the number of native hydrogen bonds, $N_{\rm hb}^{\rm nat}$, for the β-hairpin. The line in (a) is a first-order two-state fit.

3.3 Results and Discussion

3.3.1 β-Hairpin

We now turn to our simulations of the 16-amino acid C-terminal β-hairpin from the protein G B1 domain. In Ref. [11] this peptide was studied using the first version of our model. An important quantity in that study was the hydrophobicity energy $E_{\rm hp}$. This variable should be strongly correlated with Trp fluorescence, which Muñoz *et al.* [13] used to characterize the melting behavior of this peptide. The temperature dependence of $E_{\rm hp}$ was indeed in reasonable agreement with the data of Muñoz *et al.* Several other groups have performed atomic simulations of the same β-hairpin, with [24–27] or without [12, 28, 29] explicit water. In contrast to ours, most models seem to require further calibration in order not to show a temperature dependence much weaker than that of experimental data.

Figure 3.3a shows the temperature dependence of $E_{\rm hp}$ in the revised model. The line is a fit of the data to a simple (first-order) two-state expression. The parameters of the fit are the midpoint temperature $T_{\rm m}$, the energy difference ΔE, and two baselines. We use the parameter $T_{\rm m}$ to set the energy scale of the model; this parameter is taken as $T_{\rm m} = 297\,{\rm K}$ as determined by Muñoz *et al.* [13]. For the energy difference, we then obtain $\Delta E = 13.1\,{\rm kcal/mol}$. These values of the two-state parameters $T_{\rm m}$ and ΔE correspond to a native population of 74% at $T = 284\,{\rm K}$, which agrees well with the result of Muñoz *et al.*, 72% at $T = 284\,{\rm K}$ [13]. The NMR analysis of Blanco *et al.* [23] gave, by contrast, a lower native population, 42% at $T = 278\,{\rm K}$. A possible explanation of this dis-

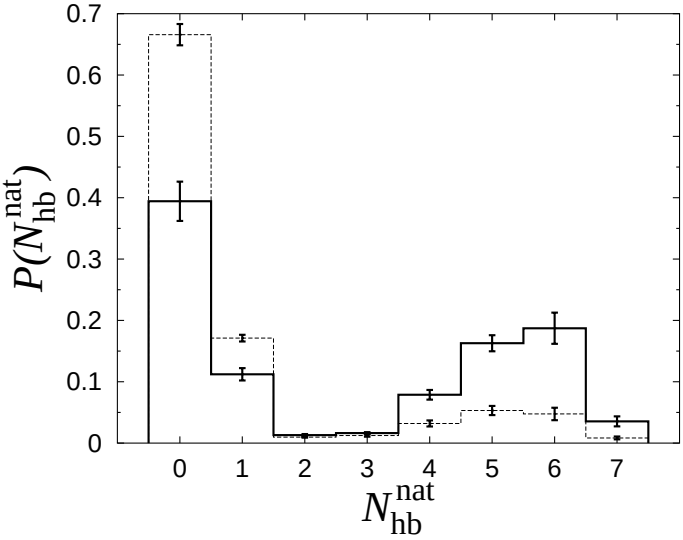

Figure 3.4: Histogram of the number of native hydrogen bonds, $N_{\rm hb}^{\rm nat}$, at $T = 284\,\mathrm{K}$ (full line) and $T = 306\,\mathrm{K}$ (dashed line) for the β-hairpin.

crepancy would be that this peptide does not show a clear two-state behavior; the apparent native population may then very well depend on which quantity is studied. At first glance, this explanation may seem unlikely, given that the temperature dependence of the Trp fluorescence data to a good approximation shows two-state character [13]. Let us therefore stress that this sequence does not behave as a simple two-state system in our model, despite that the two-state fit in Fig. 3.3a looks quite good. This can be seen, for example, from the energy distribution, which lacks a clear bimodal shape. This was shown in Ref. [11], and holds in the revised model as well. For a detailed discussion of a fast-folding model proteins without a clear free-energy barrier, which makes a three-helix bundle, see Ref. [30].

In Fig. 3.3b we show the temperature dependence of the number of native hydrogen bonds, $N_{\rm hb}^{\rm nat}$, which we expect to be more strongly correlated than $E_{\rm hp}$ with the NMR measurements of Blanco *et al.* For $N_{\rm hb}^{\rm nat}$, a two-state fit is not meaningful; for that, further data at lower temperatures would be needed. However, it is possible to determine an apparent native population from the probability distribution of $N_{\rm hb}^{\rm nat}$. Figure 3.4 shows this distribution at $T = 284\,\mathrm{K}$ and $T = 306\,\mathrm{K}$. At $T = 284\,\mathrm{K}$, it shows a clear bimodal character. If we define conformations that lack at most two of the seven native hydrogen bonds (see Fig. 3.1a) as native, that is $N_{\rm hb}^{\rm nat} \geq 5$, we obtain a native population of 39% at $T = 284\,\mathrm{K}$. This shows first of all that we do find different native populations depending on which observable we study. Moreover, the result obtained using $N_{\rm hb}^{\rm nat}$ is in fact close to the NMR-based estimate of Blanco *et al.*, whereas that obtained using $E_{\rm hp}$ is close to the Trp fluorescence-based estimate

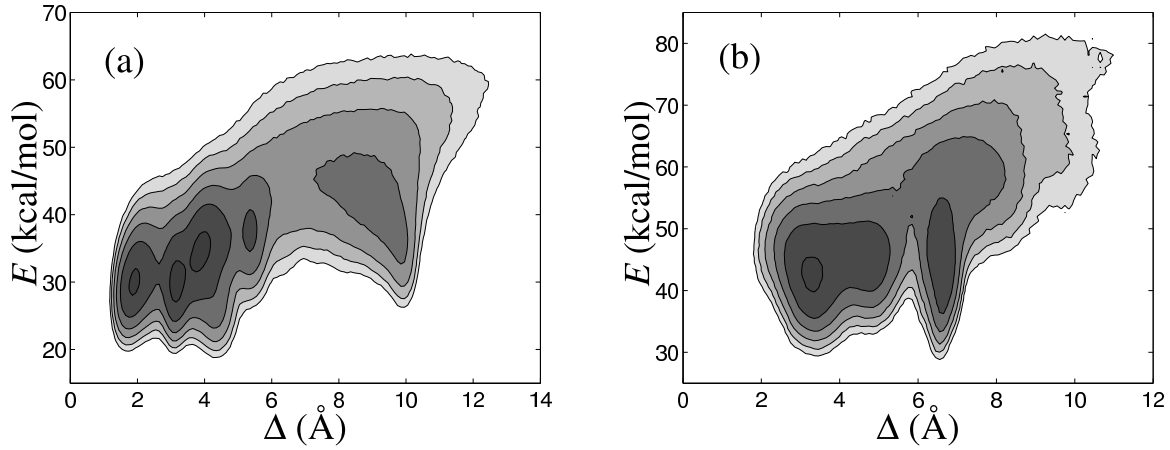

Figure 3.5: Free energy $F(\Delta, E)$ at $T = 284\,\mathrm{K}$ for (a) the β-hairpin and (b) the peptide LLM. E is energy and Δ is a heavy-atom rmsd, calculated using all the 16 amino acids for the β-hairpin and amino acids 3–18 for LLM. The first two and last two amino acids of LLM do not take part in the β-sheet structure. The contour levels are as in Fig. 3.2.

of Muñoz *et al.* At the second temperature studied in Fig. 3.4, $T = 306\,\mathrm{K}$, the fraction of conformations with $N_{\mathrm{hb}}^{\mathrm{nat}} \geq 5$ is down to 11%.

The two-state parameter ΔE extracted from our E_{hp} data is somewhat smaller here, $\Delta E = 13.1\,\mathrm{kcal/mol}$, than it was in our earlier study, $\Delta E = 16.1\,\mathrm{kcal/mol}$ [11]. The reason for this is not so much that the model has changed, but that the fits were done in different ways. In our previous study, T_{m} was held fixed at the specific heat maximum. Here, following the analysis of Muñoz *et al.* more closely, we take T_{m} to be a parameter of the fit. The fitted value of T_{m} turns out to lie slightly below (1–2%) the specific heat maximum. Our new analysis improves the agreement with the result of Muñoz *et al.*, which was $\Delta E = 11.6\,\mathrm{kcal/mol}$ [13].

Although the precise shape of the structures with lowest energy is sensitive to the details of the model, it is also interesting to make an rmsd-based comparison with experimental data. For this purpose, we use the NMR structure for the full protein G B1 domain [31] (PDB code 1GB1, first model), as the NMR restraints for the isolated β-hairpin were insufficient to determine a unique structure. Figure 3.5a shows the free energy $F(\Delta, E)$ at $T = 284\,\mathrm{K}$. Three distinct, highly populated minima can be seen. The two minima with lowest E are found at $\Delta \approx 2.0$Å and $\Delta \approx 3.1$Å, respectively. Both these correspond to β-hairpin structures with a high $N_{\mathrm{hb}}^{\mathrm{nat}}$. That $N_{\mathrm{hb}}^{\mathrm{nat}}$ is high implies, in particular, that the topology of the β-hairpin is the native one. The main difference between these two minima lies in the shape of turn. The third minimum, at $\Delta \approx 4.0$Å, is

somewhat higher in E than the first two. This minimum is also dominated by β-hairpin structures with the native topology and many hydrogen bonds, but the two strands tend to be out of register with each other, so $N_{\rm hb}^{\rm nat}$ is low. Largely, it is the existence of this third minimum that makes the apparent native population depend on which of the observables $E_{\rm hp}$ and $N_{\rm hb}^{\rm nat}$ we use. Finally, there are also two weakly populated free-energy minima corresponding to β-sheet structures with the non-native topology ($\Delta \approx 5.3\,\text{Å}$) and α-helix ($\Delta \approx 8$–$10\,\text{Å}$), respectively.

3.3.2 Three-Stranded β-Sheets

The *de novo* design of the 20-amino acid three-stranded antiparallel β-sheet peptide Betanova was reported in 1998 [1]. Recently, mutants of this peptide with higher stability were created by López de la Paz *et al.* [2]. Among the most stable mutants found was the triple mutant LLM (Val5Leu, Asn12Leu, Thr17Met). The peptide LLM and the original Betanova were estimated [2] to have native populations of 36% and 9%, respectively, at $T = 283\,\text{K}$, based on NMR data. Melting curves have, as far as we know, not been reported for these peptides.

Our simulations of LLM show first of all that this sequence does make a three-stranded antiparallel β-sheet in this model. This can be seen from Fig. 3.5b, which shows the free energy $F(\Delta, E)$ at $T = 284\,\text{K}$. The free energy has a broad minimum at $\Delta \approx 3$–$5\,\text{Å}$, corresponding to β-sheet structures with the native topology and a high $N_{\rm hb}^{\rm nat}$. The shape of the β-sheet varies within the minimum. At $\Delta \approx 3.4\,\text{Å}$, where the free energy is lowest, the β-sheet has a bent shape, which enables the chain to make strong hydrophobic contacts. At $\Delta \approx 4.5\,\text{Å}$, the β-sheet tends to be much flatter, which is hydrophobically disfavored but makes it possible for the chain to form more perfect hydrogen bonds. There is also a free-energy minimum at $\Delta \approx 6.5\,\text{Å}$, which corresponds to three-stranded antiparallel β-sheet structures with the non-native topology. However, the native topology is the thermodynamically favored one. Note that the native and non-native topologies exhibit non-overlapping sets of backbone-backbone hydrogen bonds, so $N_{\rm hb}^{\rm nat}$ is low at the $\Delta \approx 6.5\,\text{Å}$ minimum.

The main reason why the model favors the native topology over the non-native one lies in the side-chain orientations for the hydrophobic pairs Trp3-Leu12 and Leu5-Tyr10. The C_α-C_β vectors of these pairs point inwards in the non-native topology, which makes it difficult to achieve proper contacts between the side chains. This is much easier to accomplish in the native topology, where the C_α-C_β vectors point outwards. Interestingly, the situation is similar for the

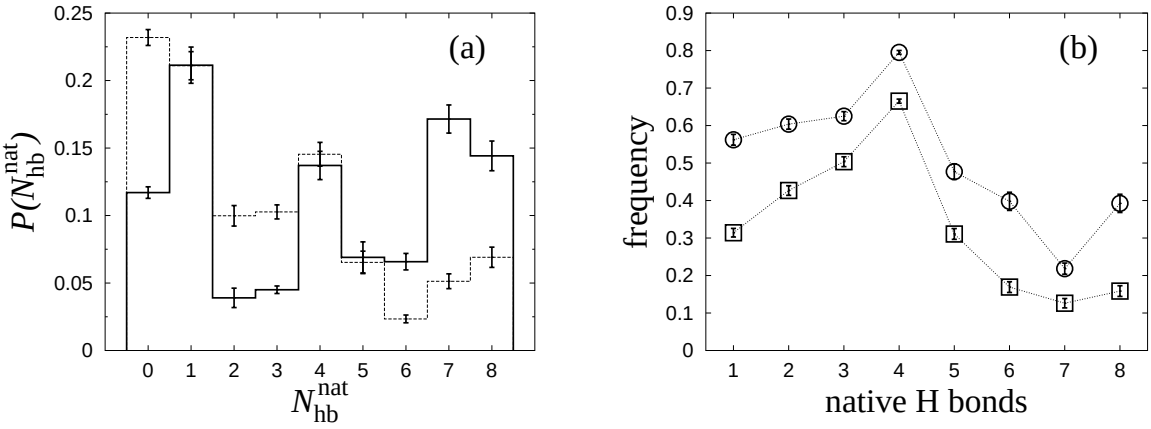

Figure 3.6: (a) Histogram of the number of native hydrogen bonds, $N_{\rm hb}^{\rm nat}$, at $T = 284\,\mathrm{K}$ for LLM (full line) and Betanova (dashed line). (b) The frequency of occurrence for the eight native hydrogen bonds (labeled according to Fig. 3.1b) for LLM (○) and Betanova (□) at $T = 284\,\mathrm{K}$.

β-hairpin above [11]. The β-hairpin also has two pairs of hydrophobic side chains that are 'bow-legged' in the native topology and 'knock-kneed' in the non-native one.

Next we estimate the native population for LLM. As we want to compare with the NMR-based results of López de la Paz *et al.* [2], we consider $N_{\rm hb}^{\rm nat}$ rather than $E_{\rm hp}$. Figure 3.6a shows the $N_{\rm hb}^{\rm nat}$ distribution at $T = 284\,\mathrm{K}$. In addition to the native and non-native peaks at high and low $N_{\rm hb}^{\rm nat}$, respectively, this distribution exhibits a third peak at $N_{\rm hb}^{\rm nat} = 4$. The typical conformation at this peak contains only the first of the two native β-turns (see Fig. 3.1b). The second β-turn is less stable, as will be discussed below. As before, we take conformations as native if at most two native hydrogen bonds are missing, that is if $N_{\rm hb}^{\rm nat} \geq 6$. We then find a native population of 38% at $T = 284\,\mathrm{K}$. We also performed simulations of the original Betanova, and Fig. 3.6a shows the $N_{\rm hb}^{\rm nat}$ distribution for this sequence, too. From the figure it is evident that Betanova is less stable than LLM. The probability that $N_{\rm hb}^{\rm nat} \geq 6$ is 14% for Betanova at $T = 284\,\mathrm{K}$. This means that the native populations obtained using this criterion are similar to the NMR-based results of López de la Paz *et al.* [2] for LLM as well as Betanova. Let us stress that the energy scale of the model is set using melting data for the β-hairpin and is then held fixed in our study of the other sequences.

Figure 3.6b shows the frequencies of occurrence for the different native hydrogen bonds (see Fig. 3.1b) for LLM and Betanova. For Betanova, there is a clear difference between the hydrogen bonds involved in the first β-turn (1–4) and those involved in the second β-turn (5–8). The latter four occur infrequently,

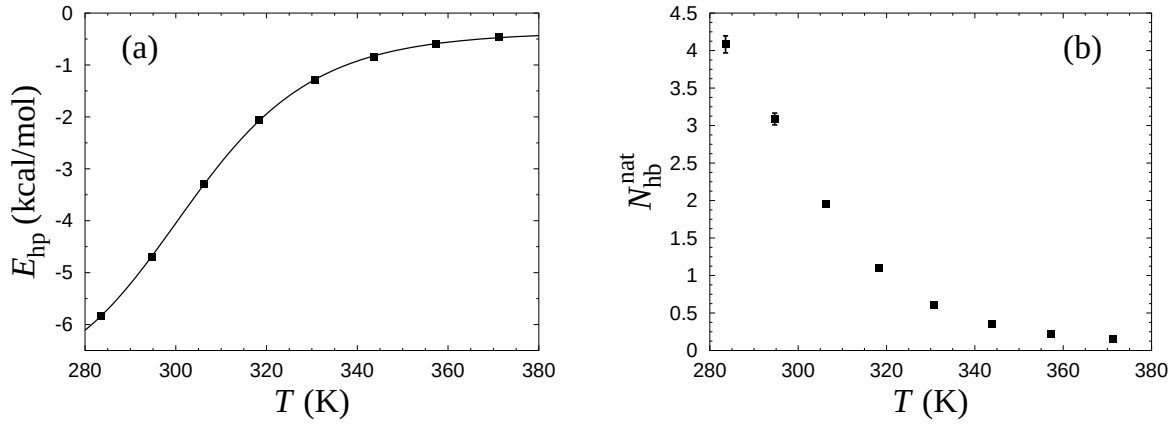

Figure 3.7: The temperature dependence of (a) the hydrophobicity energy $E_{\rm hp}$ and (b) the number of native hydrogen bonds, $N_{\rm hb}^{\rm nat}$, for the LLM peptide. The line in (a) is a first-order two-state fit.

showing that the second β-turn is quite unstable, which is in line with the conclusions of López de la Paz *et al.* [2]. For LLM, the difference in stability between the two β-turns is less pronounced. However, hydrogen bond 7, which connects Met17 to Tyr10 (see Fig. 3.1b), is quite unstable. The reason for this is that the side chain of Met17 can make better contacts with other hydrophobic side chains if the strand is slightly bent. This bend makes it difficult for hydrogen bond 7 to form.

Finally, in Fig. 3.7 we show the temperature dependence of $E_{\rm hp}$ and $N_{\rm hb}^{\rm nat}$ for LLM. As in the β-hairpin case, we find that a simple two-state fit provides a good description of the data for $E_{\rm hp}$. The fitted values of the parameters $T_{\rm m}$ and ΔE are $T_{\rm m} = 303\,{\rm K}$ and $\Delta E = 13.0\,{\rm kcal/mol}$, which means that the native population obtained from this fit is significantly higher than that obtained from the $N_{\rm hb}^{\rm nat}$ distribution (see Fig. 3.6a). So, the model predicts that the apparent native population depends on which observable is used for this sequence, too. We are not aware of any existing experimental data that support, or refute, this conclusion for LLM.

3.4 Summary

Using a novel all-atom model with a simplified sequence-based potential, we have investigated the equilibrium behavior as a function of temperature for three β-sheet peptides. For each of these peptides, several independent Monte Carlo simulations were performed, starting from different random conformations. After comparing the results from the different runs, we feel confident

	Model, 284 K		Experiment	
	$N_{\mathrm{hb}}^{\mathrm{nat}}$	E_{hp}	NMR	Trp fluorescence
β-hairpin	39%	74%	42%, 278 K [23]	72%, 284 K [13]
LLM	38%		36%, 283 K [2]	
Betanova	14%		9%, 283 K [2]	

Table 3.2: Summary of apparent native populations obtained from simulations and experimental data, respectively (see the text). The model results have statistical errors of 1–4%.

that the Monte Carlo methods employed were capable of mapping out all relevant free-energy minima.

We determined native populations for these sequences in two ways, from the distribution of the number of native hydrogen bonds ($N_{\mathrm{hb}}^{\mathrm{nat}}$) and from the temperature dependence of the hydrophobicity energy (E_{hp}). These estimates were compared with experimental results based on NMR and Trp fluorescence, respectively. This comparison is summarized in Table 3.2. NMR-based native populations have been reported for all the three sequences, and our $N_{\mathrm{hb}}^{\mathrm{nat}}$-based estimates are in good agreement with these results. For the β-hairpin, there is also an experimental result based on Trp fluorescence, and this result is close to what we find using data for E_{hp}. That we find different native populations depending on whether we use $N_{\mathrm{hb}}^{\mathrm{nat}}$ or E_{hp} reflects the fact that the melting transition is not a clear two-state transition for these sequences. It is worth noting that the temperature dependence of a quantity such as E_{hp} is quite well described by a simple two-state expression, despite that the two-state picture is an oversimplification.

The results obtained for these three sequences, including the dependence on temperature, are encouraging, especially since they were achieved while keeping the interaction potential relatively simple. In order to extend these calculations to more general sequences, it is clear that refinement of the potential will be required. To what extent this goal can be accomplished remains to be seen.

Acknowledgments

We thank Luis Serrano and Manuela López de la Paz for providing NMR data for LLM and Betanova. This work was in part supported by the Swedish Foundation for Strategic Research and the Swedish Research Council.

References

[1] Kortemme, T., Ramírez-Alvarado, M. & Serrano, L. (1998) *Science* **281**, 253–256.

[2] López de la Paz, M., Lacroix, E., Ramírez-Alvarado, M. & Serrano, L. (2001) *J. Mol. Biol.* **312**, 229–246.

[3] de Alba, E., Santaro, J., Rico, M. & Jiménez, M.A. (1999) *Protein Sci.* **8**, 854–865.

[4] Bursulaya, B.D. & Brooks, C.L. III (1999) *J. Am. Chem. Soc.* **121**, 9947–9951.

[5] Colombo, C., Roccatano, D. & Mark, A.E. (2002) *Proteins: Struct. Funct. Genet.* **46**, 380–392.

[6] Cavalli, A., Haberthür, U., Paci, E. & Caflisch, A. (2003) *Protein Sci.* **12**, 1801–1803.

[7] Karanicolas, J. & Brooks, C.L. III (2003) *Proc. Natl. Acad. Sci. USA* **100**, 3954–3959.

[8] Granakaran, S., Nymeyer, H., Portman, J., Sanbonmatsu, K.Y. & García, A.E. (2003) *Curr. Opin. Struct. Biol.* **13**, 168–174.

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

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

[11] Irbäck, A., Samuelsson, B., Sjunnesson, F. & Wallin, S. (2003) in press: *Biophys. J.*

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

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

[14] Branden, C. & Tooze J (1991) *Introduction to Protein Structure* (Garland Publishing, New York).

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

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

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

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

[19] Hansmann, U.H.E. & Okamoto, Y. (1999) *Curr. Opin. Struct. Biol.* **9**, 177–183.

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

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

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

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

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

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

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

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

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

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

[30] Favrin, G., Irbäck, A., Samuelsson, B. & Wallin, S. (2003) in press: *Biophys. J.*

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

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

Paper IV

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.

4.1 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 [3–5] 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

$$P(\delta\bar{\phi}) \propto \exp\left[-\frac{a}{2}\delta\bar{\phi}^T(\mathbf{1} + b\mathbf{G})\delta\bar{\phi}\right] , \tag{4.1}$$

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.

4.2 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. 4.1.

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. 4.1).

The energy function

$$E = E_{\rm loc} + E_{\rm sa} + E_{\rm hb} + E_{\rm AA} \tag{4.2}$$

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

$$E_{\rm loc} = \frac{\epsilon_{\rm loc}}{2} \sum_i (1 + \cos 3\phi_i)\,. \tag{4.3}$$

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

$$E_{\rm sa} = \epsilon_{\rm sa} {\sum_{i<j}}' \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}, \tag{4.4}$$

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

$$E_{\rm hb} = \epsilon_{\rm hb} \sum_{ij} u(r_{ij}) v(\alpha_{ij}, \beta_{ij})\,, \tag{4.5}$$

where i and j represent H and O atoms (see Fig. 4.1), respectively, and

$$u(r_{ij}) = 5\left(\frac{\sigma_{\rm hb}}{r_{ij}}\right)^{12} - 6\left(\frac{\sigma_{\rm hb}}{r_{ij}}\right)^{10} \tag{4.6}$$

$$v(\alpha_{ij}, \beta_{ij}) = \begin{cases} \cos^2\alpha_{ij} \cos^2\beta_{ij} & \alpha_{ij}, \beta_{ij} > 90^\circ \\ 0 & \text{otherwise} \end{cases} \tag{4.7}$$

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

$$E_{\rm AA} = \epsilon_{\rm AA} \sum_{i<j} \left[\left(\frac{\sigma_{\rm AA}}{r_{ij}}\right)^{12} - 2\left(\frac{\sigma_{\rm AA}}{r_{ij}}\right)^{6}\right], \tag{4.8}$$

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.

4.3 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. 4.1) 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. 4.1) 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

$$\Delta^2 = \sum_{I=1}^{3} (\delta \mathbf{r}_I)^2 \,, \tag{4.9}$$

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

$$\Delta^2 \approx \delta\bar{\phi}^T \mathbf{G} \delta\bar{\phi} = \sum_{i,j=1}^{n} \delta\phi_i G_{ij} \delta\phi_j \,, \tag{4.10}$$

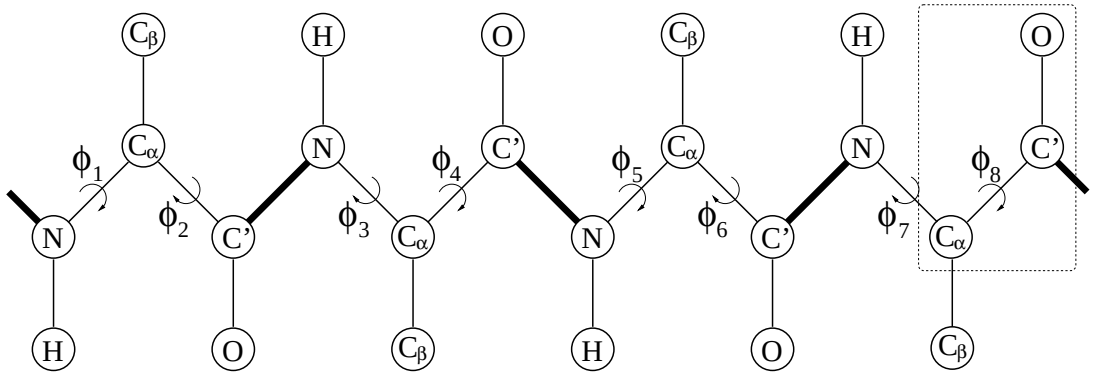

Figure 4.1: Update of the chain defined in Sec. 4.2. Eight torsion angles ϕ_i are turned. Turns such that the three atoms in the box are left unaffected are favored. Thick lines represent peptide bonds and the peptide torsion angles are fixed.

where

$$G_{ij} = \sum_{I=1}^{3} \frac{\partial \mathbf{r}_I}{\partial \phi_i} \cdot \frac{\partial \mathbf{r}_I}{\partial \phi_j} \,. \tag{4.11}$$

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

$$W(\bar{\phi} \to \bar{\phi}') = \frac{(\det \mathbf{A})^{1/2}}{\pi^3} \exp\left[-(\bar{\phi}' - \bar{\phi})^T \mathbf{A} (\bar{\phi}' - \bar{\phi})\right] \,, \tag{4.12}$$

   where the matrix

$$\mathbf{A} = \frac{a}{2}(\mathbf{1} + b\mathbf{G}) \tag{4.13}$$

   is a linear combination of the $n \times n$ unit matrix $\mathbf{1}$ and the matrix $\mathbf{G}$ defined by Eq. 4.11. 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

$$P_{\text{acc}} = \min\left(1, \frac{W(\bar{\phi}' \to \bar{\phi})}{W(\bar{\phi} \to \bar{\phi}')} \exp[-(E' - E)/kT]\right) \tag{4.14}$$

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

$$\mathbf{A} = \mathbf{L}\mathbf{L}^T . \tag{4.15}$$

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

4.3.1 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

$$\psi_i = (-\ln R_1)^{1/2} \cos 2\pi R_2 , \tag{4.16}$$

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

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

$$\mathbf{L}^T \delta\bar{\phi} = \bar{\psi} \tag{4.17}$$

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. 4.12.

4.3.2 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

$$(\det \mathbf{A})^{1/2} = \prod_{i=1}^{n} L_{ii} \tag{4.18}$$

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.

4.3.3 Pivot update

Previous simulations [26] of the model protein defined in Sec. 4.2 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.

4.4 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. 4.12 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. 4.14 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. 4.2 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].

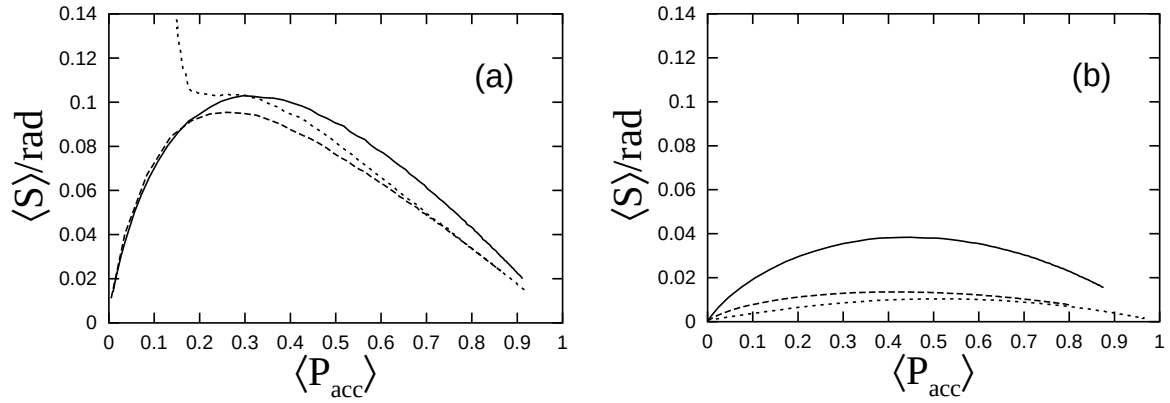

Figure 4.2: Average step size, $\langle S \rangle$, against average acceptance rate, $\langle P_{\rm acc} \rangle$, for different updates at (a) $kT = 0.7$ and (b) $kT = 0.6$. Shown are results for the $b = b_{\rm max}$ (full lines), $b = 0$ (dashed lines) and pivot (dotted lines) updates.

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

$$S = |\delta\bar{\phi}| = \left[\sum_{i=1}^{n} (\delta\phi_i)^2 \right]^{1/2} \tag{4.19}$$

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. 4.2, 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

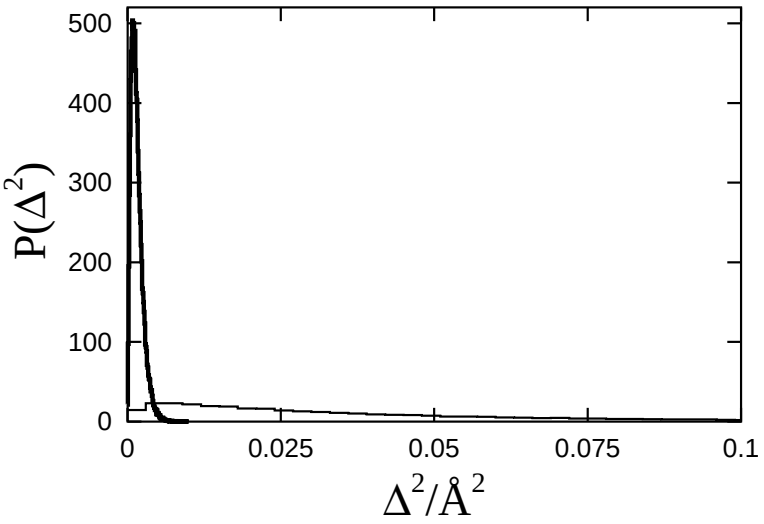

Figure 4.3: Distributions of Δ^2 (see Eq. 4.9) for the $b = b_{\max}$ (thick line) and $b = 0$ (thin line) updates at $kT = 0.6$. The values used for the parameter a correspond to maximum $\langle S \rangle$.

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. 4.12 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. 4.9) 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. 4.3 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

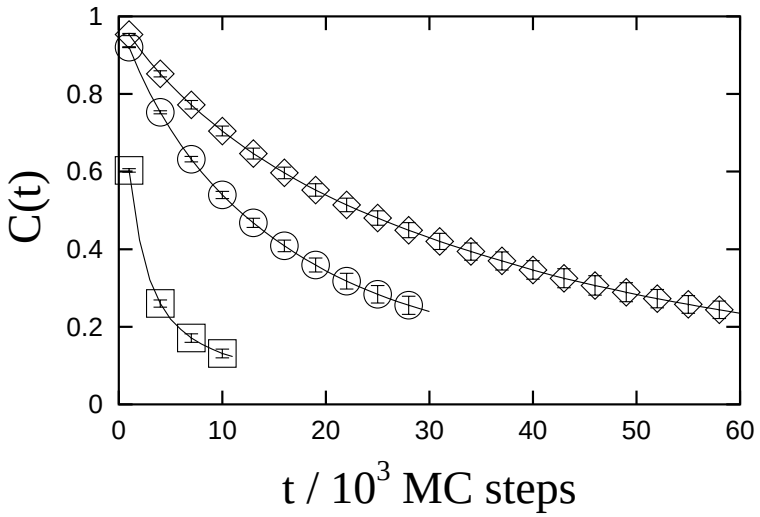

Figure 4.4: The autocorrelation function $C(t)$ (see the text) at $kT = 0.6$ for the $b = b_{\max}$ ($\Box$), $b = 0$ ($\circ$) and pivot ($\diamond$) updates. Step-size parameters correspond to maximum $\langle S \rangle$.

estimated the dynamic autocorrelation function

$$C_i(t) = \frac{\langle \cos\phi_i(t) \cos\phi_i(0) \rangle - \langle \cos\phi_i(0) \rangle^2}{\langle \cos^2\phi_i(0) \rangle - \langle \cos\phi_i(0) \rangle^2} \tag{4.20}$$

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. 4.4, 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.

4.5 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.

Stability of the Kauffman Model

Paper V

Stability of the Kauffman Model

Sven Bilke 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/

Physical Review E **65**, 016129 (2002)

Abstract:
Random Boolean networks, the Kauffman model, are revisited by means of a novel decimation algorithm, which removes variables that cannot be relevant to the asymptotic dynamics of the system. The major part of the removed variables have the same fixed state in all limit cycles. These variables are denoted the stable core of the network and their number grows approximately linearly with N, the number of variables in the original network. The sensitivity of the attractors to perturbations is investigated. We find that reduced networks lack the well known insensitivity observed in full Kauffman networks. We conclude that, somewhat counter-intuitive, this remarkable property of full Kauffman networks is generated by the dynamics of their stable core. The decimation method is also used to simulate large critical Kauffman networks. For networks up to $N = 32$ we perform *full enumeration* studies. Strong evidence is provided for that the number of limit cycles grows linearly with N. This result is in sharp contrast to the often cited $\sqrt{N}$ behavior.

5.1 Introduction

Boolean networks were introduced by Kauffman [1,2] as simplified models of the complex interaction in the regulatory networks of living cells. The binary variable σ_i encodes the activity of the effective "gene" i; expressed or not expressed. Depending upon the initial state, the system evolves to one of possibly several limit cycles. In the biological picture, the different limit cycles are interpreted as different cell types. One of Kauffman's motivations for investigating these networks was the idea that the structure of genetic networks present in nature is not only determined by selection. Rather, a good fraction of the network functionality is inherent in the ensemble of regulatory networks as such. In fact, he found an ensemble of critical Boolean networks "on the edge of chaos" that captures some features observed in nature. These Boolean networks show a remarkable stability; in most cases small perturbations of the state of the network do not change the trajectory to a different limit cycle. This is desirable in the biological interpretation since stability of genetic regulatory networks against small fluctuations is a crucial property. Another striking observation is that the number of limit cycles for the critical Boolean networks grows as a square-root of the system size [1, 2]. This is an analogy to multicellular organisms, where it is found empirically that the number of cell-types also grows approximately as the square-root of the genome-size.

The model also exhibits analogies [3] with infinite range spin glasses [4]. In the framework of an annealed approximation [5], some of the previous numerical observations concerning a phase transition between a *frozen* and a *chaotic* phase in the model could be understood. The average limit Hamming distance d_h, the number of bit-wise differences between two random configurations, was used as an order parameter. In the frozen phase one has $d_h = 0$ for infinite systems whereas in the chaotic phase one has $d_h \neq 0$. The parameter driving the transition is the probability that two different inputs to a Boolean variable σ_i give raise to different values. In [6] the annealed approximation was extended to provide distributions for the number and the length of the limit cycles. Also, good agreement between the results from the annealed approximation and the numerical calculations was demonstrated in the chaotic phase. An alternative order parameter s, the fraction of variables that are stable, i.e. evolve to the same fixed state independently of the initial state, was introduced in [7]. These stable variables are said to constitute the stable core of the network. In the infinite size limit one has $s = 1$ in the frozen phase, whereas $s \neq 1$ in the chaotic phase. In [8] the concept of relevant variables was introduced. A variable σ_i is *not* relevant, if it is stable and/or no variable's state depends on σ_i. The relevant variables are of interest since they contain all information about the

asymptotic dynamics of the network, i.e. the number of limit cycles and their cycle lengths.

In this work we focus on the stability of the Kauffman model and how this property is related to the stable core of the network. The probability that inversion of a single variable will make the system end up in a different limit cycle is known to be small and approaches zero for large networks. However, we find that if the network is reduced to its relevant variables, this probability is drastically raised and increases slightly with the system size.

To facilitate this study we introduce a decimation method that removes variables that *cannot* be relevant by inspection of transition functions and network connectivity. The resulting reduced network contains *all* relevant variables and possibly some irrelevant ones. Since all relevant variables are included it will have exactly the same asymptotic dynamics as the original network even though the total number of variables is drastically reduced. We find that resulting number of variables is close to the true number of relevant variables. This indicates that properties of the stable core can mostly be understood by the comparatively trivial interactions detected in the decimation procedure.

The decimation procedure can also be used to reduce the bias in the estimate of some observables like for example the number n_c of limit cycles. Different from earlier works we do not observe a $\sqrt{N}$ scaling, but rather a linear growth of n_c with the system size.

5.2 The Kauffman Model

A random Boolean network is essentially a cellular automaton with N binary state variables σ_i. These evolve synchronously according to the transition functions $f_i(\{\sigma\})$, which are chosen randomly at time $t = 0$ and are then kept fixed. In the Kauffman model f_i are constrained to depend on at most K different randomly chosen input variables:

$$\sigma_i(t) = f_i[\sigma_{v_i^1}(t-1), \ldots, \sigma_{v_i^K}(t-1)], \tag{5.1}$$

for every variable σ_i. The integers $\{v_i^1, \ldots, v_i^K\}$ define the input connections to variable σ_i.

The transition function f_i maps each possible combination of input signals to Boolean output values. These output values are independently set to $true$ or $false$ with probabilities p and $1 - p$ respectively. This makes some functions independent of some or all of its K input variables. Furthermore, depending

on K and p, a finite fraction of the state variables σ_i are not used by any of the transition functions.

The random Boolean network is a deterministic system. Given the state variables at some time, the future trajectory of the σ_i is known. The volume of the state space is finite, therefore all trajectories must posses a limit cycle. Besides the stability of the system the number of limit cycles, the length distribution of the cycles and transient trajectories are well established observables for this model. In numerical simulations it is in general not possible to probe the models whole state space, except for very small systems. The volume of the $\{\sigma\}$ state space grows exponentially and the number of graphs $\{f_i\}$ even super-exponentially. The commonly used strategy for exploring this model therefore contains two approximations.

1. A small fraction of all possible networks is used as a representative ensemble.
2. For each network only a subset of the state space is probed.

Point 2 introduces a systematic bias to the number of limit cycles since not all of them will be found. In the results section we will re-analyze the number of cycles after decimation of irrelevant nodes. This allows *full enumeration* of state space for up to $N = 32$. In this way we get an improved estimate for the scaling of the respective observables with the system size.

5.3 The Decimation Procedure

It is well known that some variables in a Kauffman network evolve to the same steady state independently of the initial configuration. These *stable* variables are clearly irrelevant for the asymptotic behavior of the network. The same holds for those variables that do not regulate any other variable, i.e. no transition functions is dependent on them. In fact, as pointed out in [8], for a variable to be relevant it has to be unstable and regulate some unstable variables that in turn regulate others and so on.

By definition, in the frozen phase the fraction of stable variables goes to unity as N goes to infinity. Therefore, a large fraction of the variables are likely to be stable even for finite N. Since the irrelevant variables includes all stable variables, a considerable part of a network does not affect the asymptotic dynamics at all. The process of identifying the irrelevant variables can be divided into two separate steps. Firstly, the stable variables are identified. Secondly, the variables that do not regulate any unstable variables are identified.

Identifying stable variables is in principal easy, but computationally demanding. In [8] this was done by performing simulations of the dynamics of the system and monitoring which variables were in the same state in all probed limit cycles. However, finding all limit cycles essentially means that all 2^N possible states have to be probed, which is possible only for very small networks. Since a variable that is stable within the probed limit cycles may change state within some of the unprobed limit cycles, searching a fraction of state space will in some cases overestimate number of stable variables.

Here we introduce an alternative method, which by pure inspection of the connectivity and the transition functions of a network identifies variables that *must* be stable. The basis for our approach is that transition functions dependent on no input variables give a constant output, i.e. the corresponding variable is stable.

As stated above, some transition functions are independent of all their input variables, i.e. they are constants. This means that the corresponding variables will be stable (after the initial time step) and a transition function that is dependent on such variable will receive a constant signal. By replacing the *stable* input variable with the corresponding *constant* value, the number of input variables is reduced. For each replaced input variable to a transition function, that functions input state space is reduced by a factor 1/2 and within this subspace it may be independent of yet other input variables. If in the end even this rule become a constant, the corresponding variable is stable (after a transient time), and can be replaced by a constant. Therefore, we have to repeat this procedure until no more stable variables are found. We summerize the method as follows:

1. For every transition function, f_i, remove all inputs it does not depend upon.
2. For those f_i with no inputs, clamp the variable σ_i to the corresponding constant value.
3. For every f_i, replace clamped inputs with the corresponding constants.
4. If any variable has been clamped, repeat from step 1.

It is clear that our method sometimes does not find all stable variables. We see an example of such a situation in Fig. 5.1. Here the inputs to a function are coupled logically and hereby confined to a subspace of possibilities. Within this subspace the otherwise unstable variable is stable. The figure illustrates just one of the possible couplings between inputs.

Once the stable variables are identified and removed from the network the non-regulating variables can be removed iteratively. Since our method keeps all

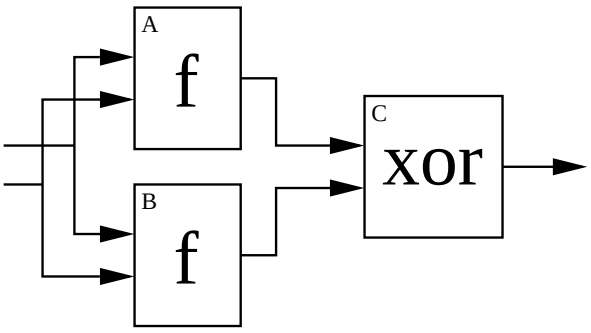

Figure 5.1: A and B have arbitrary but identical transition functions and C implements xor. Since A and B also have the same inputs their outputs will be identical. Thus, C will always output $false$, i.e. C is stable and can be removed. A and B are now non-regulating and can be removed too.

relevant variables the resulting network will have exactly the same asymptotic dynamics as the original network.

5.4 Results

Let us start by analyzing the size of the stable core as a function of the system size N. In Fig. 5.2 the the size of the stable core N^*, identified by the decimation procedure described above, is shown. Each data point is averaged over 10^4 instances of networks. For comparison the size of the stable core N^+ as estimated by the method used in [8] is also plotted. The latter procedure is based on observations of the dynamics of the full network and identification of nodes acquiring the same constant value independently of the start-configuration. Since only a small part of the state space can be probed in practice, the number N^+ is biased to overestimate the true size η of the stable core. On the other hand, our decimation procedure underestimates η because some configurations, like the one depicted in Fig. 5.1, which may lead to stable variables are not identified. Therefore, we have $N^+ \leq \eta \leq N^*$.

It is somewhat surprising to observe $N^+ \approx N^*$, which indicates that properties of the stable core, at least for $K = 2$, mostly can be understood by the comparatively trivial interactions detected in the decimation procedure. The probability s for a node to belong to the stable core can be estimated by using Eq. (2) in [7]

$$s(t+1) = \sum_{k=0}^{K} s(t)^{K-k}(1-s(t))^k \binom{K}{k} p_k, \tag{5.2}$$

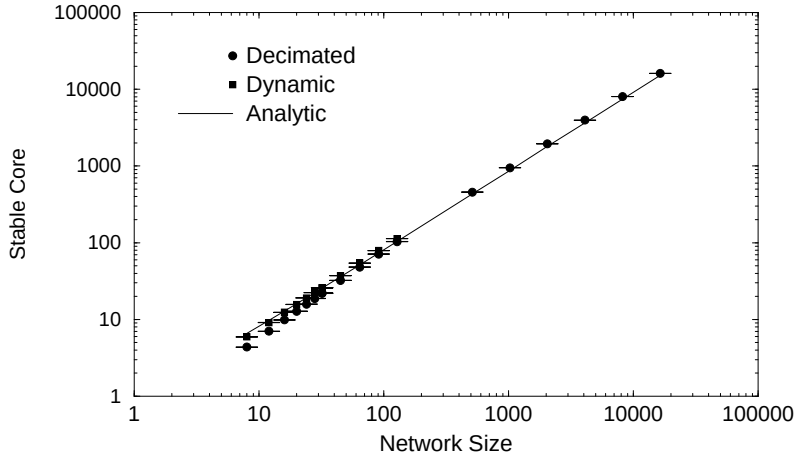

Figure 5.2: The size of the stable core – the number of variables going to the same constant value independently of the start-configuration – as a function of the network size for $K = 2$. The number of relevant variables estimated with our decimation procedure (circles), the observation of the dynamics [8] (squares) and Eq. (5.2) and (5.3) (full line) are in very good agreement.

where p_k is the probability that a transition function for given values of $K-k$ of its input variables is independent of its other k inputs. This equation describes the growth of the stable core with the time. At $t = 0$ only nodes which happen to have a constant transition function are stable. At later times non-constant transition functions, which receive inputs from stable nodes, can acquire a constant value. In [7] Eq. (5.2) was used as a self-consistency equation for infinite systems, i.e. letting $t \to \infty$. In a finite system, the iteration has to stop at some time T, which reflects a characteristic length in the network, the maximal distance a signal can flow before it reaches all nodes. The length scale is set by the average distance (in number of links) a signal can travel. The signal pathway in a sparse directed random graph with only a few loops is approximately a branched polymer, where it is known (see e.g. [9]) that the average distance grows algebraically, i.e. $T \sim cN^{\gamma}$. We have fitted the constants c and γ numerically to our data for $K = 2$ and find $\gamma = 0.32(3)$.

After removing the $s(T)$ stable variables, the decimation procedure eliminates the leaves of the network, i.e. those nodes with out-degree $q(t = 0) = 0$. This changes the out-degree of the remaining variables. Therefore, this procedure is repeated until no more variables with $q(t) = 0$ are found. The fraction P_l of leafs, direct and indirect, can be estimated by the self consistent equation

$$P_l = \sum_{q=1}^{\infty} P(q|\tilde{N}, \tilde{K}) P_l^q, \tag{5.3}$$

where $\tilde{N} = N(1 - s)$ is the number of variables after removing the η stable

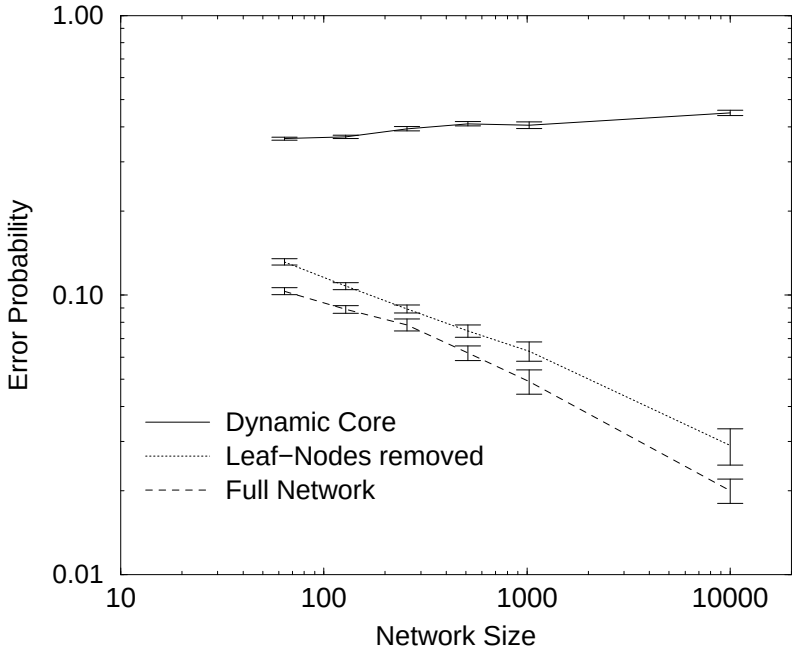

Figure 5.3: The probability to be pushed out from a limit cycle by the inversion of a randomly chosen variable. For the (full) Kauffman-network (dashed line) the error probability scales to zero for large lattices. This behavior is not changed if one does only a geometric reduction of leafs (dotted line). If the network is reduced to its relevant variables (full line) the tolerance against small fluctuations in the state space is completely lost.

ones, $\tilde{K}$ is the average in-degree, an $P(q|\tilde{N},\tilde{K})$ the distribution of the out-degree q given in Appendix B. We solved Eq. (5.2) and (5.3) numerically, the resulting graph is also shown in Fig. 5.2, which is in very good agreement with our numerical results.

One of the important features of Kauffman's model is the intrinsic stability of critical Boolean networks. How does decimation affect this behavior? While the network decimation does not change the number and the length of limit cycles, the size of the basins of attraction has to be reduced because the state space is shrunken by orders of magnitude. To get a quantitative picture, we analyze the network stability with respect to the inversion of one randomly chosen variable, after the state trajectory has reached a limit cycle. The *error-probability*, i.e. the probability to end up in a *different* limit cycle compared to the undisturbed system, is shown in Fig. 5.3. If a limit cycle has not been found within 10^5 steps the network is discarded. For the full network we observe the well known stability, the probability to end up in a a different limit cycle asymptotically approaches zero for large lattices. By contrast, for the reduced network the error-probability grows slowly with the network size and

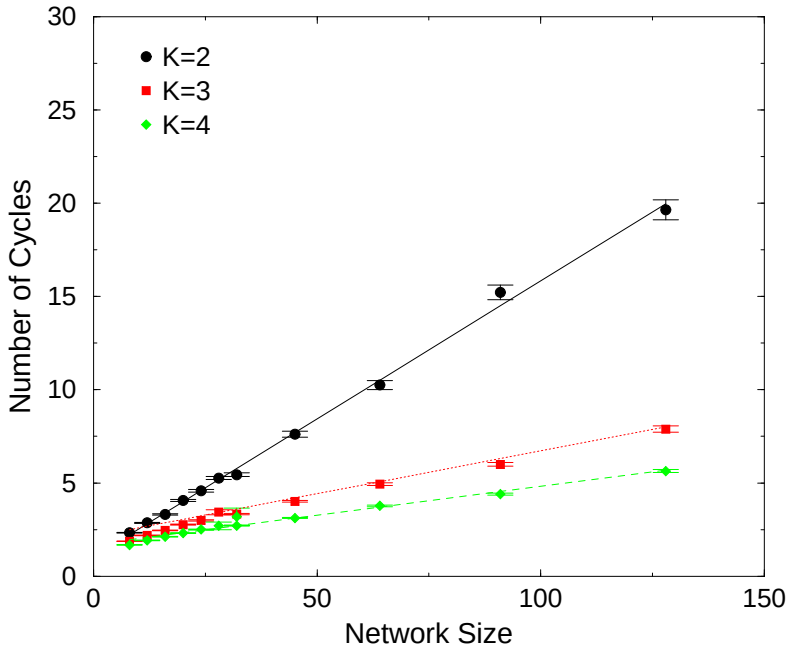

Figure 5.4: The number of limit cycles as a function of the network size for critical Boolean networks with $K = 2, 3, 4$ inputs. The solid lines connecting the data points are *linear* interpolations with $\chi^2/\text{D.O.F} = 1.5$ $(K = 2), 1.6$ $(K = 3), 1.4$ $(K = 4)$. A $\sqrt{N}$ behavior fits the data much worse with $\chi^2/\text{D.O.F} = 11.5$ $(K = 2), 8.6$ $(K = 3), 8.5$ $(K = 4)$. The dotted line represents the the best $\sqrt{N}$ fit for $K = 2$.

the stability is essentially lost. This means, the tolerance against perturbations observed for Kauffman networks is mostly generated by the stable core: in most cases the perturbed signal is "lost" in the stable core and the full network remains unaffected. It has recently been argued [10] that the in-homogeneous, for example scale-free, *geometry* of real world networks is underlying the stability of these systems. Here we find the opposite: stability is primarily generated dynamically by the propagation (Eq. 5.2) of the stable core in the network *logic*. The homogeneous geometry plays only a secondary role: if just a geometric reduction of the network is performed, i.e. the leafs are removed (see the discussion of of Eq. (5.3)), the error-sensitivity is almost unchanged compared to the full network.

The decimation of constant variables from the network enables us to probe a much larger fraction of the state space for a given network. Therefore one may expect to get an better estimate for the number of limit cycles n_c, which with the commonly used method tends to be underestimated, because some limit cycles may have been missed due to the huge state space. By decimating the networks we can fully enumerate the state space for $N \leq 32$ and hereby get

an unbiased estimate. For larger systems we use the standard method with 1000 restarts on each of the reduced networks. Not unexpectedly we observe a small discontinuity in the curve at the point were the simulation scheme is changed. In Fig. 5.4 we plot n_c as a function of the network size N. We do not find the quite often cited $\sqrt{N}$ behavior for this observable. Rather, we find a linear growth with N. A possible explanation for the different results obtained in some earlier works may be the bias introduced by the standard method in combination with lower computational power.

5.5 Summary

The source of the remarkable error tolerance of critical Kauffman model is identified as the "dynamics of the stable core". While this seems to be a contradiction in terms, it quite nicely describes the percolation-like process, which underlies the propagation of the "stability" signal. Starting from the relatively few nodes with transition functions which do not at all depend on their inputs, the islands of frozen states grow in time by the interaction with the already stable nodes. This process is only limited by the finite size of the system. A small fluctuation in the state of the system will most probably not propagate through the stable core and therefore in most cases has no effect. We demonstrate this by studying reduced networks , where most of the stable, irrelevant variables have been removed. The stability against small fluctuations for these networks is reduced by orders of magnitude and will probably go to zero for infinite networks. It is interesting to observe that these effects are mostly driven by the network *logic* and not by the network geometry.

For the identification of the relevant variables we have developed a decimation procedure, which is based on inspection of the networks connectivity and logic. The relatively simple procedure works surprisingly well. The results for the size of the stable core are in very good agreement with the values obtained by observing the dynamics of state-space trajectories in the full network [8].

As a by-product we use the reduced networks to get an improved estimate for the number of limit cycles as a function of the network size. We find that the number of limit cycles grows linearly with N, which is in sharp contrast to the square-root behavior reported by other groups. Even though this $\sqrt{N}$ behavior was an interesting analogy with multi-cellular organisms (with approximately $\sqrt{N}$ different cell types for genomes with genome size N), our result does in no way reduce the importance of Kauffman networks as an example of self-organized order.

Acknowledgments: We have benefitted from discussions with C. Peterson. This work was in part supported by the Swedish Foundation for Strategic Research and the Knut and Alice Wallenberg Foundation through the SWEGENE consortium.

5.A In- and Out-degree distribution

The reduced number $\tilde{K}$ of inputs after the decimation described in Eq. (5.2) is the expectation value of the in-degree for the number of inputs from a non-stable variable. For two inputs in the original network we get:

$$\tilde{K} = \frac{1(1-s(T)) + 2(1-s(T))^2}{1-s(T)} = 1 + 2(1-s(T)). \tag{5.4}$$

The out-degree distribution for a node in the random network can be understood by enumerating the number of ways the NK links can be distributed over this node and the $N-1$ remaining nodes, weighted by the corresponding probabilities to choose the nodes:

$$P(q|N,K) = \left(\frac{1}{N}\right)^q \left(\frac{N-1}{N}\right)^{NK-q} \binom{NK}{q}. \tag{5.5}$$

References

[1] S. A. Kauffman, J. Theor. Biol. **22**, 437 (1969).

[2] S. A. Kauffman, *The Origins of Order* (Oxford University Press, 1993).

[3] B. Derrida and H. Flyvbjerg, J. Phys. A: Math. Gen. **19**, L1003 (1986).

[4] M. Mezard, G. Parisi and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).

[5] B. Derrida and Y. Pomeau, Europhys. Lett. **1**, 45 (1986).

[6] U. Bastolla and G. Parisi, Physica D **98**, 1 (1996).

[7] H. Flyvbjerg, J. Phys. A: Math. Gen. **21**, L955 (1988).

[8] U. Bastolla and G. Parisi, Physica D **115**, 203 (1998).

[9] J. Ambjørn, B. Durhuus, T. Jonsson, *Quantum Geometry* (Cambridge University Press, 1997).

[10] H. Jeong, S. P. Mason, Z. N. Oltvai, A.-L. Barabasi, Nature **407**, 651 (2000).