Hydrophobicity Patterns in Protein Folding

Frank Potthast

Department of Theoretical Physics
Complex Systems Group
Lund University
Sweden

Thesis for the degree of Doctor of Philosophy

Thesis Advisor: Anders Irbäck

Faculty Opponent: Michele Vendruscolo
Weizmann Institute of Science
Rehovot, Israel

To be presented, with the permission of the Faculty of Mathematics and Natural Sciences of Lund University, for public criticism in Lecture Hall B of the Department of Physics on Friday 4th September 1998, at 10.15 a.m.

DOKUMENTDATABLAD
enl SIS 61 41 21

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

Document Name
DOCTORAL DISSERTATION

Date of issue
June 1998

CODEN:
LUNFD6/(NFTF-1038)/1-20 (1998)

Author(s)
Frank Potthast

Sponsoring organization

Title and subtitle

Hydrophobicity Patterns in Protein Folding

Abstract

The protein folding problem is addressed focussing on the hydrophobicity patterns in the amino acid sequences and structures. Both real and model proteins are investigated.

The hydrophobicity pattern of real proteins is probed in two ways. First, it is asked which binary patterns are most conserved within groups of related proteins. Not unexpectedly, the most conserved patterns are strongly correlated with hydrophobicity.
Second, the hydrophobicity pattern is investigated using methods that are sensitive to long-range correlations along the chains. Solid statistical evidence is found that the hydrophobic amino acids are anticorrelated along protein sequences.

The considered models are coarse-grained in nature, representing all atoms of an amino acid as one site. There are only two types of amino acids in the models, hydrophobic and hydrophilic. The dynamical-parameter approach is used to explore the thermodynamic behavior of the models. The good folding sequences of the model show the same deviation from randomness as real proteins. A 3D off-lattice model is proposed, meant to model alpha-helical proteins. Finally, a novel Monte Carlo method for protein design, which is based upon the dynamical parameter method, is suggested.

Key words
protein folding, hydrophobicity, simulated tempering, multisequence, global optimization, Fourier analysis, alpha-helix, protein design.

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language
English

ISSN and key title

ISBN
91-628-3071-6

Recipient's notes

Number of pages 152

Price

Security classification

Distribution by (name and address)
Frank Potthast, Dept. of theoretical Physics,
Sölvegatan 14 A, SE-223 62 LUND

Signature ______________________ **Date** 1998-06-19

This thesis is based on the following papers:

1. Anders Irbäck and Frank Potthast,
Binary Assignments of Amino Acids from Pattern Conservation,
Protein Engineering **10**, 1013-1017 (1997).

2. Anders Irbäck, Carsten Peterson and Frank Potthast,
Evidence for Non-Random Hydrophobicity Structures in Protein Chains,
Proceedings of the National Academy of Sciences, USA **93**, 9533-9538 (1996).

3. Anders Irbäck and Frank Potthast,
Studies of an Off-Lattice Model for Protein Folding: Sequence Dependence and Improved Sampling at Finite Temperature,
Journal of Chemical Physics **103**, 10298-10305 (1995).

4. Anders Irbäck, Carsten Peterson and Frank Potthast,
Identification of Amino Acid Sequences with Good Folding Properties in an Off-Lattice Model,
Physical Review **E 55**, 860-867 (1997).

5. Anders Irbäck, Carsten Peterson, Frank Potthast and Ola Sommelius,
Local Interactions and Protein Folding: A 3D Off-lattice Approach,
Journal of Chemical Physics **107**, 273-282 (1997).

6. Frank Potthast,
A Minimal Model for α-Helical Proteins,
LU TP 98-13, submitted to *Journal of Computational Biology.*

7. Anders Irbäck, Carsten Peterson, Frank Potthast and Erik Sandelin,
A Novel Monte Carlo Procedure for Protein Design,
LU TP 97-33 submitted to *Physical Review* **E**.

Preface

Nature doesn't have a protein-folding problem - but we do [1].

In this thesis, the complex process of protein folding is under investigation. The approach taken is reductionistic in nature, focussing on hydrophobicity patterns in the sequences.

The thesis consists of an introduction, followed by the seven papers. The introduction starts with a brief discussion of protein structure and the hydrophobic effect. Problems encountered in thermodynamic studies of protein models are discussed next, followed by a schematic presentation of the models used. An outline of the Monte Carlo methods used, the dynamical-parameter algorithms, is then given. The introduction ends with a summary of the seven papers.

The seven papers are listed on page 3. The topics of these papers may be divided into three themes: sequence analysis (Papers I-IV), thermodynamic studies of minimal models (III-VI), and protein design (VII). Papers I and II contain analyses of real protein sequences. In Papers II and IV, the hydrophobicity patterns of good folding model sequences are compared to those of real proteins. Paper III also describes the dynamical-parameter algorithm, which is of central importance to our ability to investigate the thermodynamics. In Paper VII, the dynamical-parameter algorithm is used as the basis for a novel protein design method.

Introduction

Basics of Protein Structure

Proteins are essential molecular building blocks of life as we know it, fulfilling a vast number of functions in all living organisms. Proteins are linear chain molecules consisting of 20 possible subunits, the so-called amino acids. Therefore, a protein is utmostly simple to describe: it is just a word given in a 20-letter alphabet. Each of the 20 letters corresponds to one of the 20 amino acids, containing between 10 and 27 atoms each. The (word)length of a protein can vary from the tens to several hundreds. In their natural environment, i.e. correct temperature, pH, solvent conditions etc., many proteins fold into a remarkably well-defined structure, as depicted in Fig. 1. This structure, the so-called native structure, is encoded in the amino-acid sequence (the sequence of letters) [2].

This process of native structure formation is as essential as it is interesting and difficult to understand. It is called protein folding. Protein folding is

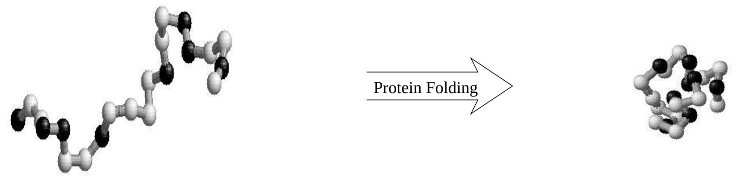

Figure 1: Schematic picture of protein folding. In their natural environment, most proteins fold into a compact structure with very well-defined shape, the so-called native structure.

essential to life since each protein is able to serve its function effectively in its native state only. Protein folding is interesting since it is the second half of the genetic code; the step from a piece of DNA to the sequence of amino acids is the first half, protein folding is the second half, leading from the amino-acid sequence to the functional structure. Finally, protein folding is difficult to understand (for the time being) in part simply due to the large number of degrees of freedom involved; moreover, these are far from independent because of the compactness of the native state.

In order to be a little bit more specific, let us consider the protein "pancreatic hormone" [3] [1]. Containing just 36 amino acids, pancreatic hormone is a very small protein. In comparison, a typical protein of yeast consists of around 300 amino acids [6]. The amino-acid sequence of pancreatic hormone is given by

G P S Q P T Y P G D D A P V E D L
I
R
Y R H R T V V N L Y Q Q L N D Y F

where each circle represents one amino acid. The letters denote the amino-acid types (see Table 1). The sequence of amino acids is called the primary structure, and is the blueprint for the protein.

---

[1] We consider pancreatic hormone found in turkey, PDB accession code: 1PPT [4, 5]. Pancreatic hormone may also be referred to as "avian pancreatic polypeptide."

All 20 amino acids in the alphabet, except proline, have the general structure

$$\mathrm{H_2N\text{-}}\overset{\sigma}{\mathrm{C}}_\alpha\mathrm{H\text{-}CO_2H} \quad ,$$

where σ denotes the amino-acid side chain containing between one and 18 atoms (see Table 1). Pancreatic hormone is assembled by connecting the amino acids in the given order, which is realized by polymerization reactions,

$$2\times\mathrm{H_2N\text{-}}\underset{\mathrm{H}}{\overset{\sigma}{\mathrm{C}}}_\alpha\mathrm{\text{-}CO_2H} \rightarrow \mathrm{H_2N\text{-}}\underset{\mathrm{H}}{\overset{\sigma}{\mathrm{C}}}_\alpha \overset{\text{Peptide Bond}}{\overset{\downarrow}{\boxed{\underset{\mathrm{O}}{\mathrm{C}}\text{-}\underset{\mathrm{H}}{\mathrm{N}}}}} \underset{\mathrm{H}}{\overset{\sigma}{\mathrm{C}}}_\alpha\mathrm{\text{-}CO_2H} + \mathrm{H_2O} \quad ,$$

where two amino acids react at the peptide bond. Hence, the 36 amino acids are held together by 35 peptide bonds.

As first noticed by Linus Pauling [11], the peptide bond is stiff, and the six atoms C_α-CO-NH-C_α (being those situated around the peptide bond) tend to be coplanar. Two possible orientations of the peptide bond exist, referred to as cis and trans. As pointed out by G.N. Ramachandran [12], the trans conformation dominates in real proteins. In this conformation, the distance between the two C_α atoms in the planar C_α-C-N-C_α unit is about 3.8Å.

The Native State

As stated above, it is of essential importance that a protein folds into its native state in order to serve its function. The geometrical structure of the native state has been determined experimentally for 7162 proteins [2], one of them being pancreatic hormone, which is displayed in Fig. 2.

One striking feature of proteins is the presence of repetitive local structures. This can be quantified by connecting the C_α atoms by virtual bond vectors, and studying the distribution of bend and torsional angles in the native state. In these distributions, as displayed in Fig. 5.2[3], one observes two preferred regions. The most dominant peak is a fingerprint of so-called α-helical structure, which is shown shematically in Fig. 6.7 . The second, less dominant and wider peak in these angle distributions corresponds to another type of structure, β-sheet. The distribution of these structural elements along the chain is called the "secondary structure".

[2] Counting released protein structures from the PDB [4,5] on May 27th 1998. The current release (35.0) of the SWISS-PROT database [13, 14] contains 69113 sequences.

[3] Fig. 5.2 is to be read as "Fig. 2 in Paper V."

Code		Name	Side chain	f[%]	h[kcal/mol]
			most hydrophilic		
R	Arg	Arginine	-CH2-CH2-CH2-NH-CNH2-NH2	5.7	-14.92
D	Asp	Aspartic Acid	-CH2-COO	5.3	-8.72
E	Glu	Glutamic Acid	-CH2-CH2-COO	6.2	-6.81
N	Asn	Asparigine	-CH2-CONH2	4.4	-6.64
K	Lys	Lysine	-CH2-CH2-CH2-CH2-NH3	5.7	-5.55
Q	Gln	Glutamine	-CH2-CH2-CONH2	4.0	-5.54
H	His	Histidine	-CH2-IMIDAZOLE	2.2	-4.66
S	Ser	Serine	-CH2-OH	6.9	-3.40
T	Thr	Threonine	-CH(CH3)-OH	5.8	-2.57
Y	Tyr	Tyrosine	-CH2-PHENOL	3.2	-0.14
G	Gly	Glycine	-H	7.2	0.94
P	Pro	Proline	-CH2-CH2-CH2-[N]	5.1	
C	Cys	Cysteine	-CH2-SH	1.7	1.28
A	Ala	Alanine	-CH3	8.3	1.81
W	Trp	Tryptophan	-CH2-INDOLE	1.3	2.33
M	Met	Methionine	-CH2-CH2-S-CH3	2.4	2.35
F	Phe	Phenylalanine	-CH2-PHENYL	3.9	2.98
V	Val	Valine	-CH-(CH3)2	6.6	4.04
I	Ile	Isoleucine	-CH(CH3)-CH2-CH3	5.2	4.92
L	Leu	Leucine	-CH2-CH(CH3)2	9.0	4.92
			most hydrophobic		

Table 1: The 20 amino acids: abbreviation codes [7], side-chain molecules [8], frequency of occurence in natural proteins, f [9], and hydrophobicities for the side-chain molecules, h [10]. The horizontal line between Alanine and Tryptophan indicates the division into hydrophobic and hydrophilic residues, as done in paper II.

Another important feature is a right/left asymmetry of the torsional angle distribution. This "chirality" results from an asymmetry of the four atoms bonded covalently to the C_α atom; only one out of two possible orientations is commonly found in amino acids [8]. A chemist calls this type of amino acid an "L isomer".

Guidance to the Native State

How can a protein, which in general contains many more amino acids than pancreatic hormone, find its native state among the astronomical number of possible conformations? Furthermore, the protein has to succeed with this task on biologically relevant timescales; indeed, proteins fold within seconds [8]. Simple estimates of the time that would be needed for a random search [15,16] show that some guidance must be involved in the folding process. To get an idea about possible guiding mechanisms, it is necessary to look at the physical

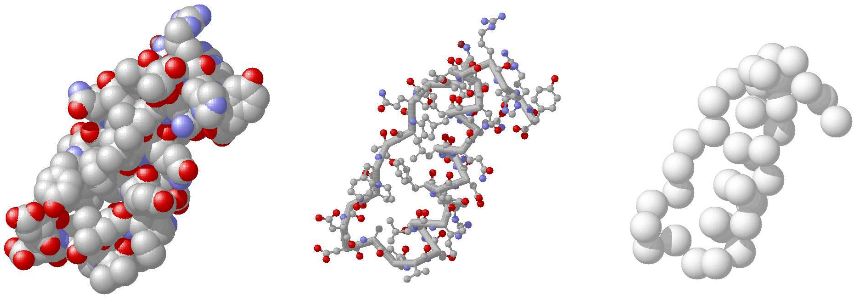

Figure 2: Illustration of protein architecture. Shown is the native state of pancreatic hormone, containing 36 amino acids. Left: All-atom representation (302 atoms, hydrogen atoms not included). Middle: All-atom representation with the backbone highlighted. Right: Only the C_α positions are displayed. The regular structure to the right is an α-helix.

interactions involved. These include electrostatic forces, hydrogen bonds, van der Waals interactions, to name some of the most important. In addition to this, one has to take into account that proteins in vivo are surrounded by a solvent, which is well described as salty water. The solvent gives rise to an effect which is widely believed to be very important in protein folding, the hydrophobic effect [8, 17]. Since water plays such an important role, a short discussion of the hydrophobic effect is given below.

Hydrophobic Core Formation

> The most essential informational property of the core is simply that it must remain hydrophobic [18].

As is well-known, oily (non-polar) liquids separate from water in the liquid phase. This is commonly denoted as the hydrophobic effect. It is widely believed that this is an entropic effect which can be understood in terms of the network of hydrogen bonds that exist between the polar water molecules [19,20]; as will be explained below. If a non-polar molecule is introduced into water, the hydrogen bonds tend to form an ordered iceberg-like structure (clathrate) around it [21]. Since the water gets more ordered in this process, the entropy decreases. If we add another non-polar molecule, it will stick to the first one. This is not primarily due to attraction between the non-polar molecules. Rather, it is based on favorable interactions between the water molecules and the fact that the entropy loss can be made smaller by forming just one big clathrate instead of two small ones. This is thought to be the physical mechanism behind the separation between the non-polar molecules and the polar water.

The 20 amino acids show a varying degree of hydrophobicity. This is illustrated in Table 1, where the amino acids have been ordered with respect to their hydrophobicity. This has important consequences for protein structure. In fact, it is well-documented that the core of proteins mainly consists of hydrophobic (non-polar) amino acids [17, 22]. This implies that the amino-acid sequences to some extent define an "inside-outside" pattern; the hydrophobic side chains are driven towards the inside by the hydrophobic effect. However, the division into inside/outside is not sharp; for example, it is not unusual to find hydrophobic residues at the surface [23].

Hydrophobic Patterns in the Sequences

If hydrophobicity is of such importance in protein folding, it should have left some "footprints" in the protein sequences themselves.

As has been mentioned above, two classes of repetetive local structure are found in proteins, α-helices and β-sheets. Taken together with hydrophobic core formation, it is reasonable to expect that these repetitive structures would lead to hydrophobicity patterns that are repetitive as well. This is indeed confirmed by a detailed analysis; both α-helices and β-sheets tend to show matching repetitive hydrophobicity patterns [24–27]. Furthermore, this signal is so strong that it survives even if one does not restrict the analysis to, say, α-helical regions. This can be seen from Paper II, where a pure sequence analysis, without structural input, is performed. In this paper, it is shown that the hydrophobic amino acids are anticorrelated along the chains for a large class of proteins.

Redundancy and the Importance of Hydrophobicity

Proteins with differing amino-acid sequences may fold to very similar conformations [8]. On a more quantitative level, it is estimated that pairs of protein sequences sharing 50% of their amino acids usually have very similar structures [28]. Surprisingly, similarity of protein structures has even been observed in cases where the amino-acid sequences lack apparent similarity [8]. This redundancy in the sequence-structure relationship suggests that there exists structurally superfluous information in the protein sequence code; the *structure* may be determined by a simpler code. To understand the nature of this redundancy is of great interest; it contains valuable information about what is essential to a protein's fold.

One possibility to study the redundancy is by mutation experiments. Here, one takes a protein, changes its amino-acid sequence in a controlled way, and monitors if the native state is unaffected. Such studies show that the core is of special importance [18, 29]; in particular, it has to remain hydrophobic. In fact, it has been argued that the hydrophobicity of a sequence contains more information about its potential acceptability in the core than does the side-chain volume [29].

Another possibility is to statistically compare closely related proteins. Such an analysis is performed in Paper I. As an example, Fig. 1.1 shows 16 versions of pancreatic hormone taken from different species. The first row in this figure gives the amino-acid sequence found in American Alligator, the second for Western Graylag Goose, and so on. It becomes clear at a glance that there is quite some similarity between the 16 sequences, which is evolutionary in origin. An intriguing question is what such closely related sequences have in common. In Paper I, it is shown that hydrophobicity patterns are highly conserved.

Thermodynamic Studies of Protein Folding

Papers III-VII present thermodynamic studies of simplified protein models based on hydrophobicity forces. In these minimal models, each amino acid is represented by a single site and the solvent is not explicitly included. There are only two types of amino acids, chosen to be hydrophobic and hydrophilic. The hydrophobic effect is modeled by assigning a stronger attraction to hydrophobic-hydrophobic pairs than to other pairs. In addition, the models contain sequence independent local interactions. Obviously, such minimal models are unrealistic in some respects. Nevertheless, there are a number of good reasons for investigating them:

- As discussed above, there is solid evidence for redundancy in the sequence code.
- Their reductionistic nature takes away distracting details, which in particular makes it easier to study the relative importance of different interaction types.
- Thermodynamic studies are feasible for minimal models only (for the time being).

Even for minimal protein models, studying the thermodynamics is generally very difficult due to the free-energy barriers present for self-avoiding, compact chains [30]. Therefore, past thermodynamic studies have mainly been performed using lattice models. Lattice models lend themselves to thermodynamic calculations since conformation space is discrete, enabling the enumeration of all conformations and therefore the calculation of exact results for short chains. For example, the minimal HP model of Lau and Dill [31] has been examined in quite some detail [32,33]. However, the approximations involved in lattice models are far from well-understood [34]. Furthermore, the energy barriers between the states may be poorly represented due to the discreteness of conformational space. In this thesis, all models studied are off-lattice, except the HP model, which is used in Paper VII to illustrate a new protein design algorithm.

It should be obvious that minimal models are not intended for detailed structure prediction; rather they are meant to provide generic scenarios for how proteins are guided to their native state.

The Models Explored in This Thesis

Thermodynamic studies of three different off-lattice models are discussed in this thesis, one is formulated in two dimensions (2D) and the others in three dimensions (3D). They all share the following properties:

1. Each amino acid is represented by a single site.
2. Two types of amino acids are considered, being hydrophobic and hydrophilic.
3. The hydrophobic effect is modeled by assigning stronger attraction to the hydrophobic-hydrophobic pairs than to the other pairs, thereby inducing hydrophobic core formation.
4. Sequence independent local interactions are present.

To give an overview, the energy functions of all three models are listed below. The geometry of a chain with N amino acids is described by (virtual) bond vectors of unit length, $\{\vec{b}_1, ..., \vec{b}_{N-1}\}$, which connect the amino acids. The distance between monomers i and j is denoted by r_{ij}. Finally, a binary string $\{\sigma_1, ..., \sigma_N\}$ specifies the sequence of hydrophobic ($\sigma_i = 0$) and hydrophilic ($\sigma_i = 1$) amino acids.

The energy functions of the models studied in Papers III and IV (2D), Paper V (3D), and Paper VI (3D), are given by

Paper	$E(b,\sigma)$
III, IV	$\sum_{i=1}^{N-2}\sum_{j=i+2}^{N} 4\left(\left(\frac{1}{r_{ij}}\right)^{12} - \epsilon(\sigma_i,\sigma_j)\left(\frac{1}{r_{ij}}\right)^{6}\right) - \frac{1}{4}\sum_{i=1}^{N-2}\vec{b}_i\cdot\vec{b}_{i+1}$
V	$\sum_{i=1}^{N-2}\sum_{j=i+2}^{N} 4\epsilon(\sigma_i,\sigma_j)\left(\left(\frac{1}{r_{ij}}\right)^{12} - \left(\frac{1}{r_{ij}}\right)^{6}\right) - \kappa_1\sum_{i=1}^{N-2}\vec{b}_i\cdot\vec{b}_{i+1} - \kappa_2\sum_{i=1}^{N-3}\vec{b}_i\cdot\vec{b}_{i+2}$
VI	$\sum_{i=1}^{N-2}\sum_{j=i+2}^{N} 4\epsilon(\sigma_i,\sigma_j)\left(\left(\frac{A_{ij}}{r_{ij}}\right)^{12} - \left(\frac{A_{ij}}{r_{ij}}\right)^{6}\right) - \sum_{i=1}^{N-3}\exp\left(-\frac{(\alpha_i-\hat{\alpha})^2+(\tau_i-\hat{\tau})^2+(\tau_{i+1}-\hat{\tau})^2}{2w^2}\right)$

Excluded volume effects of the protein chain are modelled by the $1/r_{ij}^{12}$-part of the Lennard-Jones potential. The attractivity of the Lennard-Jones terms is controled by the sequence dependent factors $\epsilon(\sigma_i, \sigma_j)$, which are chosen so as to favor the formation of a hydrophobic core. To this end, we set $\epsilon = 1$ for hydrophobic-hydrophobic pairs, and $\epsilon = 1/2$ for all other pairs. An exception is the 2D model of Papers III and IV, where the hydrophobic-hydrophilic pairs are chosen to repel each other, $\epsilon = -1/2$.

All three models combine sequence dependent global Lennard-Jones interactions with sequence independent local interactions. The importance of a proper balance between the global and local interactions is pointed out in Paper V; there, different strengths of the local interactions are tested by varying κ_1, κ_2. Further details of these models can be found in the papers.

Monte Carlo Integration

The thermodynamic behavior of the models is defined by the energy function $E(r)$, where r is the chain conformation. The probability to observe r, $p(r)$, is determined by the Boltzmann probability

$$p(r) = \frac{\exp(-\frac{E(r)}{T})}{\int d\tilde{r}\ \exp(-\frac{E(\tilde{r})}{T})} \quad , \tag{1}$$

where $\int dr$ symbolizes an integration over the whole conformation space. The thermal conformation average $\langle \Omega \rangle$ of an observable $\Omega(r)$ is "simply" given by

$$\langle \Omega \rangle = \int dr\ p(r) \times \Omega(r) = \int dr \frac{\exp(-\frac{E(r)}{T})}{\int d\tilde{r} \exp(-\frac{E(\tilde{r})}{T})} \times \Omega(r) \quad , \tag{2}$$

which, unfortunately, is impossible to calculate analytically for most systems. In this thesis, such integrals are evaluated using Monte Carlo methods. A fundamental contribution to this area was made in 1953 when the Metropolis algorithm was suggested [35]. This is a very general recipe for generating samples from the desired Boltzmann distribution. Given a set of S conformations obtained in this way, $\{r_1, r_2, ..., r_S\}$, one can approximate $\langle \Omega \rangle$ of Eq. 2 with

$$\langle \Omega \rangle \approx \frac{1}{S} \sum_{i=1}^{S} \Omega(r_i) \tag{3}$$

For this approximation to be useful, it is crucial that the samples r_i are as uncorrelated as possible. In protein models, simple Metropolis schemes become problematic at low temperatures. Here, it takes long times to cross free-energy barriers, resulting in strongly correlated samples. This is well-illustrated in Fig. 3.9, where the Monte Carlo dynamics of a system with two dominating states is displayed. In the upper plot of this figure it can be seen that the system tends to get stuck for long times in a standard Metropolis simulation. The samples obtained are therefore strongly correlated. Also shown in this figure are results obtained by the dynamical-parameter method. This method is described in the following section, and is of invaluable help in speeding up the simulations by obtaining less correlated samples.

The Dynamical-Parameter Method

As indicated above, sampling a representative part of conformation space is difficult for compact polymeric chains due to the presence of a rugged free-energy landscape. In the dynamical-parameter method [36, 37], one tries to overcome this problem by simulating several systems in parallel. The different systems simulated are indexed by an additional statistical variable, $k \in \{1, ..., K\}$, which is allowed to fluctuate, much as the coordinates of the system. The energy function as well as the temperature of the system may depend on k. In addition, a set of tunable parameters g_k is introduced, leading to a joint probability

$$p(r,k) = \frac{\exp(-\frac{E(r,k)}{T_k} - g_k)}{\sum_{\tilde{k}=1}^{K} \int d\tilde{r}\ \exp(-\frac{E(\tilde{r},\tilde{k})}{T_{\tilde{k}}} - g_{\tilde{k}})}. \quad (4)$$

This joint probability can be simulated using a standard Metropolis algorithm.

In this thesis, two different dynamical-parameter algorithms are discussed, differing in the choice of k dependence.

In simulated tempering [36,37], the energy function is chosen not to depend on k, while the temperature does. The set of temperatures is chosen to contain low and high temperatures. The main interest is in the low temperature behavior. Here, higher temperatures are included to decorrelate the samples.

In the multisequence method (see Paper III), the temperature is constant while the sequence depends on k. It turns out that this way of extending the configurational space also can lead to a significant speedup of the simulations. Other implementations are possible; for example, an implementation suitable for the Potts model can be found in [38].

The parameters g_k are crucial for the performance of the algorithm. In fact, without these parameters (i.e. $g_k = 0$), the marginal probability

$$p(k) = \int dr\ p(r,k) \quad (5)$$

is generally far from uniformly distributed, as illustrated in Fig. 3.8(a) for the multisequence method. This implies that the system tends to get stuck at a few k values. The tunable parameters g_k in Eq. 4 are needed to surmount this problem. It is easy to see that this can be achieved by taking $g_k = -F_k/T_k$, where F_k is the free energy; this choice makes the distribution $p(k)$ uniform. In practice, the free energy is, of course, unknown. However, it turns out that the g_k's can be adjusted by means of a number of trial runs. Details can be found in Paper III.

It is important to note that the conditional distribution

$$p(r|k) = \frac{p(r,k)}{p(k)} \tag{6}$$

is identical to the distribution $p(r)$ in Eq. 1 (where the index k is suppressed), independent of the choice of g_k; hence, the method is free from systematic errors for any choice of g_k.

It turns out that both simulated tempering and the multisequence method significantly speed up the simulations of the systems under consideration (see e.g. Fig. 3.9). The choice between these two methods depends on the problem at hand. For sequence design, as discussed in Paper VII, the multisequence method is found to be especially helpful.

The Papers

The research work presented in the papers may be divided into the three themes sequence analysis, minimal protein models and protein design. The papers are ordered, as far as possible, to follow these themes.

Paper I

The redundancy of the sequence-structure relationship is investigated by a statistical comparison of closely related sequences. A total of 607 groups of closely related sequences is used. These contain 1968 sequences in total and are taken from the SWISS-PROT [13] database. Each group consists of sequences with the same protein name and the same number of amino acids, but taken from different species. All possible binary assignments of the amino acids are considered, and the assignment which leads to the maximum degree of pattern conservation is determined. For an illustration, see Fig. 1.1. The optimal assignment is found to be strongly correlated with hydrophobicity. A cross-validation of the method is done on 14 subsets, and the results are found to be very stable. This method, which does not require any explicit structural input, provides a simple and clear demonstration of the importance of hydrophobicity.

Paper II

The binary hydrophobicity pattern of protein sequences is investigated, without using groups of related sequences or structural input. A similar approach was taken in [39], where the so-called run test was employed. On the level of single sequences, it turns out that the run test cannot distinguish protein sequences from random ones. The analysis presented in Paper II is based on a representation called hydrophobicity walk [40], which is sensitive to long-range correlations. Statistically significant deviations from randomness are found using a set of 36765 sequences taken from the SWISS-PROT [13, 14] database. The results are stable when restricting the sequences to a non-homologous [4] set. This evidence for non-randomness is found using sequence information only. An interesting possibility would be to maximize the signal for non-randomness by using the method of Paper I. A somewhat similar approach, but including structural input, was already taken in [41].

Paper III

Thermodynamic studies of a minimal model for protein folding are presented. The model is two-dimensional, and was first suggested in [42]. All chains of length eight and ten are examined. Most of the chains are thermodynamically unstable, while a few are stable. These "stable" sequences are of special interest from a biological point of view, since proteins are stable in their native state. The two dynamical-parameter methods discussed above are tested on this model, and are found to speed up the simulations by large factors.

Paper IV

300 sequences with 20 amino acids each are investigated with respect to their thermodynamic and kinetic behavior. In contrast to the thermodynamic properties, the kinetic behavior is found to be fairly sequence independent at fixed temperature. Hence criteria for chains with good folding properties are devised solely in terms of conformational stability. A statistical analysis is done for the sequences that fold well. Similar deviations from randomness are found as those reported in Paper II for real proteins. Using artificial neural network methods, it is tried to predict whether or not a sequence folds well, given its hydrophobicity pattern only.

Paper V

Paper V deals with an extension to three dimensions of the 2D model studied in Papers III and IV. Like the 2D model, it is based on sequence-independent local interactions and hydrophobicity forces only. Using simulated tempering, the thermodynamic behavior is investigated for varying strengths of the local interactions. In absence of local interactions, native structure formation does not occur at the temperatures considered. For proper strengths of the local interations, it is shown that it is possible to find sequences that have compact, well-defined structures. The local geometries of the model proteins are compared to those of real ones, and qualitative, but not quantitative, agreement is found.

[4] Closely related sequences, as those used in Paper I, are called homologous.

Paper VI

A minimal model for α-helical proteins, which is a modification of that introduced in Paper V, is presented. The modifications locally favor α-helical structure. The global structures resulting from the competition between global and local interactions are studied by simulated tempering. Two sequences, with hydrophobicity patterns consistent with two- and three-helix bundels, are investigated. The local structure is found to be very similar to those of real α-helical proteins, and the two sequences display the desired number of helices in the folded phase.

Paper VII

The protein folding problem amounts to: given a sequence, to which structure will it fold? Protein design is the "inverse" of this: given a target structure, which sequences will fold to it? The latter problem is of obvious interest in drug design. In this paper, this problem is addressed by using the multisequence method of Paper III to maximize conditional probabilities. This means that both the conformation and sequence degrees of freedom are allowed to fluctuate. The suggested algorithm is successfully tested on a 2D lattice model. Also, a bootstrap procedure for the sequence space search is devised, which makes large chains feasible.

Danksagung

Ich danke Anders und Carsten für die wundervolle Zusammenarbeit in den letzten Jahren. Besonderer Dank gilt jedoch meinen Eltern, Karola und Friedel Potthast.

Bibliography

[1] M. Groß and K.W. Plaxco, "Reading, Writing and Redesigning", *Nature* **388**, 419 (1997).

[2] C.B. Anfinsen, "Principles that Govern the Folding of Proteins", *Science* **181**, 223 (1973).

[3] T.L. Blundell, J.E. Pitts, I.J. Tickle, S.P.Wood and C.-W. WU, "X-Ray Analysis (1.4-Å Resolution) of Avian Pancreatic Polypeptide. Small Globular Protein Hormone", *Proc. Nat. Acad. Sci. USA* **78**, 4175 (1981).

[4] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, "The Protein Data Bank: A Computer Based Archival File for Macromolecular Structures", *J. Mol. Biol.* **112**, 535 (1977).

[5] Internet: *http://www.pdb.bnl.gov*

[6] S. Das, L.Y.C. Gaitatzes, R.R.J. Freeman, J.B.R.M. Adams and T.F. Smith, "Biology's new Rosetta Stone", *Nature* **385**, 29 (1997).

[7] IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN), "Nomenclature and Symbolism for Amino Acids and Peptides", *Eur. J. Biochem.* **138**, 9 (1984).

[8] T. Creighton, *Proteins: Structures and Molecular Properties* (Freeman, New York, 1993).

[9] P.McCaldon and P. Argos, "Frequency of occurrence of amino acid residue in primary structures of 1021 unrelated proteins of known sequence", *Proteins* **4**, 99 (1988).

[10] A. Radzicka and R. Wolfenden, "Comparing the Polarities of the Amino Acids: Side-Chain Distribution Coefficients between the Vapor Phase, Cyclohexane, 1-Octanol, and Neutral Aqueous Solution", *Biochemistry* **27**, 1664 (1988).

[11] L. Pauling, R.B. Corey and H.R. Branson, "The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain", *Proc. Nat. Acad. Sci. USA* **37**, 205 (1951).

[12] G.N. Ramachandran and V. Sassiekharan, "Conformation of Polypeptides and Proteins", *Adv. Prot. Chem.* **28**, 283 (1968).

[13] A. Bairoch and B. Boeckmann, "The SWISS-PROT Protein Sequence Data Bank: Current Status", *Nucleic Acids Res.* **22**, 3578 (1994).

[14] Internet: *http://expasy.hcuge.ch/sprot/*

[15] C. Levinthal, "Are there Pathways for Protein Folding?", *J. Chim. Phys.* **65**, 44 (1968).

[16] For a recent review, see K.A. Dill and H.S. Chan, "From Levinthal to Pathways to Funnels", *Nat. Struct. Biol.* **4**, 10 (1997).

[17] W. Kauzmann, "Some Factors in the Interpretation of Protein Denaturation", *Adv. Prot. Chem.* **14**, 1 (1959).

[18] W.A. Lim and R.T. Sauer, "Alternative Packing Arrangements in the Hydrophobic Core of λ Repressor", *Nature* **339**, 31 (1989).

[19] F.S. Henry, "The Structure of Ordinary Water", *Science* **169**, 635 (1970).

[20] F.H. Stillinger, "Water revisited", *Science* **209**, 451 (1980).

[21] See e.g. J.M. Goodfellow, "Computer Simulation of Hydration Networks around Amino Acids", *Int.J.Biol.Macromol.* **9**, 273 (1987).

[22] G.D. Rose and S. Roy, "Hydrophobic Basis of Packing in Globular Proteins", *Proc. Nat. Acad. Sci. USA* **77**, 4643 (1980).

[23] F.M. Richards, "The Interpretation of Protein Structures: Total Volume, Group Volume Distributions and Packing Density", *J. Mol. Biol.* **82**, 1 (1974).

[24] D. Eisenberg, R.M. Weiss and T.C. Terwilliger, "The Hydrophobic Moment Detects Periodicity in Protein Hydrophobicity", *Proc. Nat. Acad. Sci. USA* **81**, 140 (1984).

[25] E.T. Kaiser, F.J. Kezdy, "Secondary Structures of Proteins and Peptides in Amphiphilic Environments", *Proc. Nat. Acad. Sci. USA* **80**, 1137 (1983).

[26] H. Xiong, B.L. Buckwalter, H.M.Shieh and M.H. Hecht, "Periodicity of Polar nad Nonpolar Amino Acids is the Major Determinant of Secondary Structure in Self-assembling Oligomeric Peptides", *Proc. Nat. Acad. Sci. USA* **92**, 6349 (1992).

[27] M.W. West and M.H. Hecht, "Binary Patterning of Polar and Nonpolar Amino Acids in the Sequences and Structures of Native Proteins", *Protein Sci.* **4**, 2032 (1995).

[28] C. Chotia and A.M. Lesk, "The Relation between the Divergence of Sequence and Structure in Proteins", *EMBO J.* **5**, 823 (1986).

[29] J.U. Bowie, J.F. Reidhaar-Olson, W.A. Lim and R.T. Sauer, "Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitutions", *Science* **247**, 1306 (1990).

[30] See e.g. M. Karplus and A. Šali, "Theoretical Studies of Protein Folding and Unfolding", *Curr. Opin. Struct. Biol.* **5**, 58 (1995).

[31] K.F. Lau and K.A. Dill, "A Lattice Statistical Mechanics Model of the Conformational and Sequence Spaces of Proteins", *Macromolecules* **22**, 3986 (1989).

[32] C.J. Camacho and D. Thirumalai, "Minimum Energy Compact Structures of Random Sequences of Heteropolymers", *Phys. Rev. Lett.* **71**, 2505 (1993).

[33] K.A. Dill, S. Bromberg, K. Yue, K.M. Fiebig, D.P. Yee, P.D. Thomas and H.S. Chan, "Principles of Protein Folding — A Perspective from Simple Exact Models", *Protein Sci.* **4**, 561 (1995).

[34] A. Irbäck and E. Sandelin, "Local Interactions and Protein Folding: A Model Study on the Square and Triangular Lattices", *J. Chem. Phys.* **108**, 2245 (1998).

[35] N.A. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A. Teller and E. Teller, "Equation of State Calculations by Fast Computing Machines", *J. Chem. Phys.* **21**, 57 (1953).

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

[37] A.P. Lyubartsev, A.A. Martsinovski, S.V. Shevkunov and P.N. Vorontsov-Velyaminov, "New Method to Monte Carlo Calculation of the Free Energy: Method of Expanded Ensembles", *J. Chem. Phys.* **92**, 1776 (1992).

[38] W. Kerler and A. Weber, "Cluster Dynamics for First-order Phase Transitions in the Potts Model", *Phys. Rev. B* **47**, 11563 (1993).

[39] S.H. White and R.E. Jacobs, "Statistical Distribution of Hydrophobic Residues Along the Length of Protein Chains", *Biophys. J.* **57**, 911 (1990).

[40] V.S. Pande, A.Y. Grosberg and T. Tanaka, "Nonrandomness in Protein Sequences: Evidence for a Physically Driven Stage of Evolution", *Proc. Natl. Acad. Sci. USA* **91**, 12972 (1994).

[41] J.L. Cornette, K.B. Cease, H. Margalit, J.L. Spouge, J.A. Berzofsky and C. DeLisi, "Hydrophobicity Scales and Computational Techniques for Detecting Amphipathic Structures in Proteins", *J. Mol. Biol.* **195**, 659 (1987).

[42] F.H. Stillinger, T. Head-Gordon, and C.L. Hirshfeld, "Toy Model for Protein Folding", *Phys. Rev.* **E 48**, 1469 (1993).

Paper I

Binary Assignments of Amino Acids from Pattern Conservation

Anders Irbäck and Frank Potthast
Department of Theoretical Physics, University of Lund
Sölvegatan 14A, S-223 62 Lund, Sweden

Protein Engineering **10** 1013-1017 (1997)

Abstract:

We develop a simple optimization procedure for assigning binary values to the amino acids. The binary values are determined by a maximization of the degree of pattern conservation in groups of closely related protein sequences. The maximization is carried out at fixed composition. For compositions approximately corresponding to an equipartition of the residues, the optimal encoding is found to be strongly correlated with hydrophobicity. The stability of the procedure is demonstrated. Our calculations are based upon sequences in the SWISS-PROT database.

1.1 Introduction

An amino acid sequence is a message in 20-letter code that determines the shape and function of the protein. This message is degenerate; amino acids may be exchanged to a certain degree without affecting the functionality of the protein in a drastic way [1]. For example, it has been demonstrated that the function of λ repressor is very tolerant to exchanges of core residues as long as the pattern of hydrophobicity remains unchanged [2]. The nature of the degeneracy can be probed by analyzing the Dayhoff mutation matrix [3] with respect to conservation of different physico-chemical properties. Results of such studies convincingly show that hydrophobicity plays a central role in the formation of protein structure [4].

As a first-order approach to the structure of proteins, it may therefore be tempting to take a simple two-letter code where the residues are classified as either hydrophobic or hydrophilic. A two-letter code can contain much structural information, as shown by studies of simplified models such as the lattice-based HP model [5] or the off-lattice model of Refs. [6, 7]. However, there are 2^{20} possible ways to reduce a 20-letter code to a binary code, and it is not obvious that the most efficient encoding is closely related to a specific physico-chemical property such as hydrophobicity.

In this note we present a simple optimization procedure for assigning binary values, $\sigma_i = \pm 1$, to the amino acids. The quantity which is optimized is the degree of pattern conservation in groups of closely related proteins sequences. In order to avoid trivial solutions, the optimization is carried out for a fixed number of $\sigma_i = +1$, N_+. Depending on N_+, the optimal code may or may not have a simple physico-chemical interpretation. Here we shall focus on the results for $N_+ = 10$, which turn out to be strongly correlated with hydrophobicity. Although not unexpected (cf Ref. [4]), this finding underlines the importance of hydrophobicity as the method is free from physico-chemical inputs. Furthermore, it should be stressed that the method is global in the sense that all possible encodings, for fixed N_+, are considered.

Our calculations are based on groups of protein sequences extracted from the SWISS-PROT database [8], each group corresponding to a fixed length and a single protein but different species. Using a simple measure of similarity within these groups, the degree of pattern conservation is maximized. Although the procedure ignores problems due to insertions and deletions of residues, it turns out to be fairly robust. The robustness was tested by separately analyzing different parts of the data set. In this way one can also study the stability of individual binary values, and assign refined, non-binary, values to the residues.

Our method is somewhat related to the method of "optimal matching hydrophobicities" by Sweet and Eisenberg [9]. In this method hydrophobicity values are determined from mutation probabilities [3] by using an iterative procedure. Our method uses mutation frequencies rather than probabilities, and has the advantage that there is no need to specify initial values.

Another optimization procedure for determining hydrophobicity values has been proposed by Cornette et al. [10]. This method uses secondary-structure data rather than mutation data; sequence segments that form α-helices are analyzed using Fourier methods. The hydrophobicity scale is obtained by maximizing of the strength of the signal for the 3.6 residue period characteristic of the α-helix. A similar method has been used for detecting patterns in biologically related sequences [11].

The paper is organized as follows. In Sec. 1.2 we describe our method. In Sec. 1.3 we discuss the stability of the method and the results obtained. We end with a brief summary in Sec. 1.4.

1.2 Method

Forming groups of sequences

In our calculations we have used a set of sequence groups extracted from the SWISS-PROT database, release 31 [8]. Each group consisted of sequences with the same protein name and length N, but different biological sources. All possible groups of this type were formed for $N \leq 140$. As an example, one of the groups is shown in Fig. 1.1. In order to test the size dependence of the results, the data set was divided into two parts corresponding to $N \leq 100$ (380 groups containing 1251 sequences in total) and $100 < N \leq 140$ (227 groups, 717 sequences), respectively.

In forming these sequence groups we have ignored problems related to insertions and deletions of residues. This was done in order to keep the method as free as possible from physico-chemical inputs. It is possible that the method can be improved by incorporating some carefully chosen alignment technique. Also, the groups are small; typically, they contain three to four sequences. It is therefore important to check the stability of the procedure, which will be done in Sec. 1.3.

In our analysis we have removed uncertain sequences by ignoring all entries in the database containing the feature keys **UNSURE** (indicates uncertainties in the sequence), **NON_TER** (an extremity of the sequence is not the terminal residue) and **NON_CONS** (two residues in the sequence are not consecutive). Furthermore, we removed sequences containing the letters **B** (denoting Asp or Asn), **Z** (Gln or Glu) and **X** (any amino acid).

Comparing binary assignments

A binary encoding σ of the amino acids is a mapping from the original 20-letter alphabet to a two-letter alphabet, which we shall write as $\xi \to \sigma(\xi) \in \{+1, -1\}$ where ξ specifies the amino-acid type. Using the sequence groups described

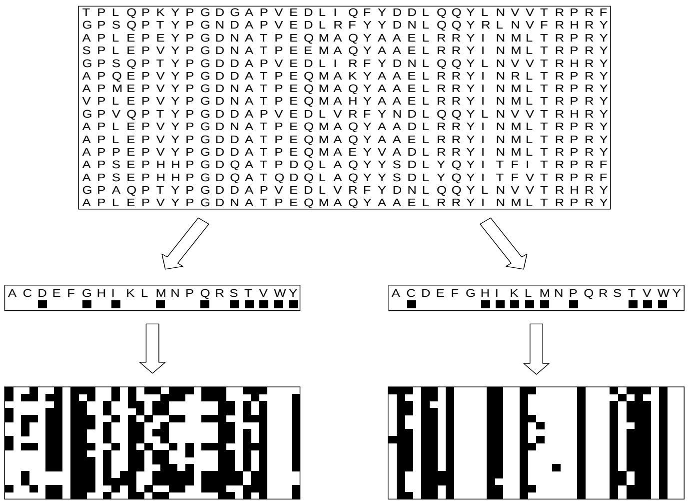

Figure 1.1: Illustration of the method for one of the sequence groups. The group consists of 16 sequences of pancreatic hormone with $N = 36$, which are shown in the upper box. Shown below are two binary assignments and the resulting patterns.

above, we shall assign a number Δ^σ to each encoding σ. The quantity Δ^σ provides a measure of pattern conservation and will be optimized. To define Δ^σ we proceed in three steps.

First we define the distance $D^\sigma(\xi^a, \xi^b)$ between two arbitrary amino-acid sequences $\xi^a = (\xi_1^a, \ldots, \xi_N^a)$ and $\xi^b = (\xi_1^b, \ldots, \xi_N^b)$ of the same length N,

$$D^\sigma(\xi^a, \xi^b) = \sum_{i=1}^{N} 1 - \delta_{\sigma(\xi_i^a),\sigma(\xi_i^b)} \qquad \delta_{st} = \begin{cases} 1 & \text{if } s = t \\ 0 & \text{otherwise} \end{cases} \tag{1.1}$$

$D^\sigma(\xi^a, \xi^b)$ is the usual Hamming distance between the binary strings $(\sigma(\xi_1^a), \ldots, \sigma(\xi_N^a))$ and $(\sigma(\xi_1^b), \ldots, \sigma(\xi_N^b))$ and takes integer values between 0 and N. Note that $D^\sigma(\xi^1, \xi^2)$ remains unchanged under a simultaneous change of the signs of all $\sigma(\xi)$'s, i.e., $D^\sigma(\xi^1, \xi^2) = D^{\sigma^-}(\xi^1, \xi^2)$ for all ξ^1 and ξ^2 if $\sigma^-(\xi) = -\sigma(\xi)$ for all ξ. This implies a twofold degeneracy in σ space.

Next we define a measure of the flucuations within one group of sequences. For a group labeled k consisting of the P_k sequences $\xi^1, \ldots, \xi^{P_k}$ (all having the same length) we put

$$D_k^\sigma = \frac{1}{P_k} \sum_{1 \le a < b \le P_k} D^\sigma(\xi^a, \xi^b) \tag{1.2}$$

where the normalization is chosen so as to make the scaling of D_k^σ linear in P_k (the sum has $P_k(P_k-1)/2$ terms). A low D_k^σ value signals high degree of similarity between the binary strings. The calculation of D_k^σ is illustrated in Fig. 1.1 for two different σ; D_k^σ is high to the left and low to the right.

The quantity D_k^σ will not have a unique minimum $\sigma = \sigma_{\min}$ unless the group labeled k is large. Finally, we therefore take a set of f different sequence groups and define the optimal encoding as

$$\sigma_{\min} = \min_{\sigma} \Delta^\sigma \tag{1.3}$$

where

$$\Delta^\sigma = \sum_{k=1}^{f} D_k^\sigma \tag{1.4}$$

It is not meaningful to minimize $\Delta^\sigma(\geq 0)$ over all possible σ, since Δ^σ vanishes if $\sigma(\xi)$ is a constant. In order to avoid this trivial solution, we have performed minimizations for different fixed numbers of positive $\sigma(\xi)$'s, N_+. The twofold degeneracy mentioned above disappears when N_+ is held fixed and $N_+ \neq 10$. Furthermore, the symmetry implies that it is sufficient to consider $N_+ \leq 10$. The number of distinguishable encodings varies from 20 for $N_+ = 1$ to 92378 for $N_+ = 10$.

1.3 RESULTS

Minimizing Δ^σ

We have performed the minimization described in the previous section for $N \leq 100$ and $100 < N \leq 140$ and for $N_+ = 1, \ldots, 10$. In this subsection we present the results and discuss the robustness of the procedure. In the next subsection we discuss the interpretation of the results.

The minimizations were carried out in a straightforward way by enumerating all possible encodings, which turned out to be feasible for all N_+. In Table 1.1 we show the minimum of Δ^σ, $\sigma_{\min}$, for different N and N_+. The interpretation of $\sigma_{\min}$ depends on N_+, as will be discussed below. Results for different N_+ are therefore not to be thought of as corresponding to some fixed property such as hydrophobicity (the results for $N_+ = 10$ are strongly correlated with hydrophobicity, as will be shown below).

From Table 1.1 it can be seen that the size dependence of the results is fairly weak, and that the structure of $\sigma_{\min}$ varies slowly with N_+. These observations indicate a certain degree of stability. In order to check the stability of the method in more detail, we divided the data set into blocks (see Table 1.2) and performed a separate optimization for each block. In Table 1.3 we show the results of these calculations for $N_+ = 10$. For each block the irrelevant overall

N_+	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−
2	−	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	+	−
3	−	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	+	+
4	−	−	−	−	−	−	+	−	−	−	+	−	−	−	−	−	−	−	+	+
5	−	+	−	−	+	−	+	−	−	−	−	−	−	−	−	−	−	−	+	+
6	−	+	−	−	+	−	+	−	−	−	+	−	−	−	−	−	−	−	+	+
7	−	+	−	−	+	−	+	−	−	−	+	−	+	−	−	−	−	−	+	+
8	−	−	−	−	+	−	+	+	−	+	+	−	−	−	−	−	−	+	+	+
9	−	+	−	−	+	−	+	+	−	+	+	−	−	−	−	−	−	+	+	+
10	−	+	−	−	+	−	+	+	−	+	+	−	+	−	−	−	−	+	+	+
1	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−
2	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−
3	−	+	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	+	−
4	−	+	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	+	+
5	−	+	−	−	+	−	+	−	−	−	−	−	−	−	−	−	−	−	+	+
6	−	+	−	−	+	−	+	−	−	−	+	−	−	−	−	−	−	−	+	+
7	−	+	−	−	+	−	+	−	−	−	+	−	−	+	−	−	−	−	+	+
8	−	+	−	−	+	−	−	+	−	+	+	−	−	−	−	−	−	+	+	+
9	−	+	−	−	+	−	+	+	−	+	+	−	−	−	−	−	−	+	+	+
10	−	+	−	−	+	−	+	+	−	+	+	−	+	−	−	−	−	+	+	+

Table 1.1: $\sigma_{\min}$ for different values of N_+. The results were obtained by using groups of length $N \leq 100$ (top) and $100 < N \leq 140$ (bottom). For $N_+ = 10$ there are two symmetry-related copies of $\sigma_{\min}$, and we have chosen the one with $\sigma(A) = -1$.

sign of $\sigma_{\min}$ (see Sec. 1.2) was chosen so as to minimize the Hamming distance to $\sigma_{\min}$ for the full data set. For the full data set the overall sign was fixed by setting $\sigma_{\min}(A) = -1$. The results shown in Table 1.3 clearly demonstrate the stability of the method; for example, it can be seen that twelve of the amino acids have been assigned the same binary value in twelve or more of the fourteen independent block calculations. Note that the stability of the assignment is amino-acid dependent. This will be used below to define a refined, non-binary, scale.

Another way to test the stability is to study the distribution of Δ^σ. In Fig. 1.2 we show two histograms of Δ^σ corresponding to $N \leq 100$ and $100 < N \leq 140$, respectively, for $N_+ = 10$. These histograms were obtained using the full data sets. Also using the full data set, we then computed Δ^σ for each $\sigma_{\min}$ from the block analysis. The positions of these fourteen values are shown in Fig. 1.2. They are all located in the low 0.1% tails of the histograms, which gives another demonstration of the stability of the method.

Interpretation

It is well-known that physico-chemical properties can be extracted from mutation probabilities, as, e.g., in the method of Sweet and Eisenberg [9]. Our

Block	First entry	Last entry	No. of sequences	No. of groups
B1	ACBP_BOVIN	CRBL_VESXA	158	51
B2	CSPA_ECOLI	FER_SYNY4	161	45
B3	FER_ARCLA	LYS3_SHISO	149	52
B4	MAST_POLJA	NXS2_NAJNI	158	45
B5	NXSB_LATCR	PYY_RANRI	169	40
B6	REV_SIVA1	S10A_HUMAN	155	47
B7	S10B_BOVIN	VA15_VARV	152	53
B8	VB09_VACCC	YVDC_VACCV	149	47
C1	ACP_BRACM	CYC_KLULA	117	33
C2	CYC_BRAOL	H3_VOLCA	119	39
C3	H3_ACRFO	NU3C_WHEAT	121	36
C4	NU3M_ASCSU	RL20_ECOLI	124	33
C5	RL20_MAIZE	VAL3_TYLCV	116	41
C6	VDBP_CAMVC	YV1_TYLCH	120	45

Table 1.2: The blocks used for $N \leq 100$ (top) and $100 < N \leq 140$ (bottom).

ξ	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Full	−	+	−	−	+	−	+	+	−	+	+	−	+	−	−	−	−	+	+	+
B1	+	−	−	−	+	+	−	+	−	+	+	−	+	−	−	+	+	+	−	−
B2	−	−	−	−	+	−	+	+	−	+	+	−	+	−	+	−	−	+	+	+
B3	+	−	−	−	+	−	−	+	−	+	+	−	+	−	−	−	+	+	+	+
B4	+	−	−	−	+	+	−	+	−	+	+	−	+	−	−	+	+	+	−	−
B5	−	+	−	−	+	−	+	+	−	+	+	−	+	−	−	−	−	+	+	+
B6	−	+	−	−	+	−	+	+	−	+	+	−	+	−	−	−	−	+	+	+
B7	−	+	−	−	+	−	+	+	−	+	+	−	+	−	−	−	−	+	+	+
B8	−	+	−	−	+	−	+	−	+	−	+	+	−	+	+	−	−	−	+	+
Full	−	+	−	−	+	−	+	+	−	+	+	−	+	−	−	−	−	+	+	+
C1	−	+	−	−	+	−	+	+	−	+	+	−	−	−	+	−	−	+	+	+
C2	−	+	−	−	+	−	+	+	−	+	+	−	−	+	−	−	−	+	+	+
C3	−	+	−	−	+	−	+	+	−	+	+	−	+	−	−	−	−	+	+	+
C4	+	−	−	−	+	−	−	+	−	+	+	−	+	−	−	+	+	+	+	−
C5	−	+	−	−	+	−	+	+	−	+	+	−	+	−	−	−	−	+	+	+
C6	+	+	−	−	+	−	−	+	−	+	+	−	−	−	−	−	+	+	+	+

Table 1.3: $\sigma_{\min}$ as obtained using the full data set and the different blocks in Table 1.2 ($N_+ = 10$). The results in the upper part are for $N \leq 100$, whereas those in the lower part are for $100 < N \leq 140$.

approach is different since it is based on mutation frequencies rather than probabilities. Somewhat surprisingly, it turns out that physico-chemical information can be obtained in this way also.

In our method the optimized quantity Δ^σ is defined as the total number of pattern-breaking bits, without reference to the frequency of occurrence of the amino acids. In this way emphasis is put on the overall degree of pattern conservation rather than the values assigned to individual amino acids. Whether the

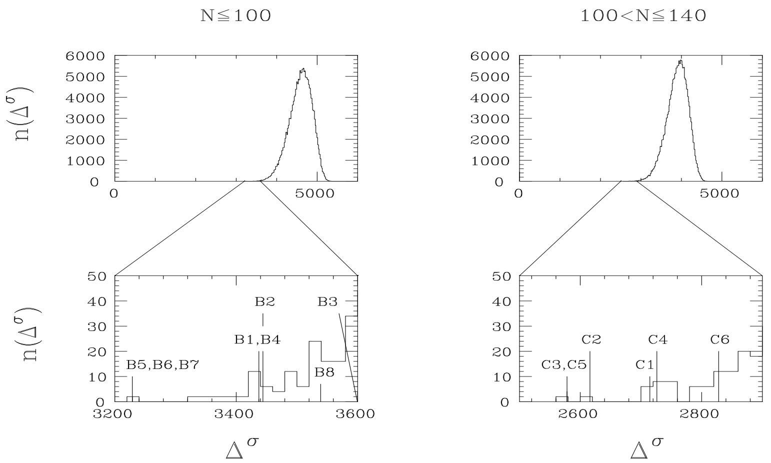

Figure 1.2: a) Histograms of Δ^σ for $N_+ = 10$, obtained using $N \leq 100$ (left) and $100 < N \leq 140$ (right). The lower figures are enlargements of the low-Δ^σ tails, in which we have indicated the values of Δ^σ for the different minima $\sigma_{\rm min}$ from the block analysis. The minima $\sigma_{\rm min}$ for the blocks B5, B6 and B7 coincide with the minimum obtained using the full data set. The same is true for the blocks C3 and C5.

individual amino-acid values still contain useful information is a priori unclear. A necessary condition for that is, of course, that the corresponding degree of pattern conservation is relatively high. If, on the other hand, the mutations are more random, then the assigned values tend to reflect the frequency of occurrence of the amino acids; the procedure then tends to put rare amino acids in the smallest group (corresponding to $\sigma = +1$ if $N_+ < 10$).

In Fig. 1.3 we show amino-acid frequency [15] against N_+ for the results corresponding to $100 < N \leq 140$. As can be seen from this figure, there is a strong correlation between frequency and binary value for $N_+ \leq 7$. This does not necessarily imply that the binary values solely reflect frequency; for example, W and C, the amino acids with $\sigma = +1$ for $N_+ = 2$ ($100 < N \leq 140$), are not only rare but have, in fact, been found to be the least mutable residues [16]. However, it is clear that in order to extract any physico-chemical information for $N_+ \leq 7$ a more detailed analysis is required, which is beyond the scope of the present paper.

For $N_+ \geq 8$ it is clear that the binary values contain non-trivial information since the correlation between binary value and frequency is fairly weak. In what follows we shall discuss the results for $N_+ = 10$ in some detail.

In order to take into account the fact that the stability of the binary value is amino-acid dependent, we begin by forming the average $a(\xi)$ of the results

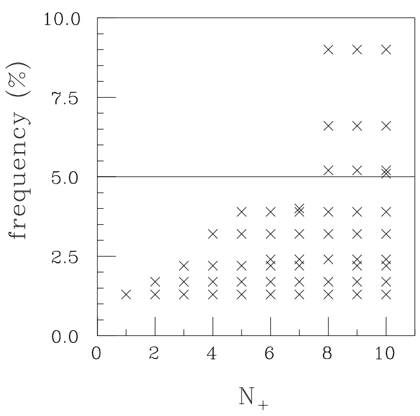

Figure 1.3: Frequency of occurrence for the amino acids with $\sigma = +1$ plotted against N_+, using data for $100 < N \leq 140$.

from the block analysis;

$$a(\xi) = \frac{1}{k} \sum_{i=1}^{k} \sigma_{\min}^{(i)}(\xi) \tag{1.5}$$

where $\sigma_{\min}^{(i)}$ denotes the result obtained using data block i. As the size dependence was found to be weak, the average has been taken over all the blocks in Table 1.2 ($k = 14$).

In Fig. 1.4a we have plotted $a(\xi)$ against hydrophobicity, using the scale of Fauchère and Pliska [12]. From this figure it is clear that $a(\xi)$ is strongly correlated with hydrophobicity. The correlation coefficient is 0.91, which, in fact, is a representative value for what Cornette *et al.* typically found when comparing various existing hydrophobicity scales [10]. As an example, we show in Fig. 1.4b the scale of Fauchère and Pliska against that of Roseman [13] (correlation coefficient 0.93). The correlation between $a(\xi)$ and hydrophobicity is much stronger than that between $a(\xi)$ and frequency of occurrence (correlation coefficient -0.32).

The correlation between $a(\xi)$ and hydrophobicity becomes weaker as N_+ decreases, but remains fairly strong down to $N_+ = 8$.

1.4 Summary

We have developed a simple optimization procedure based on pattern conservation for assigning binary values to the amino acids. In this method the optimal encoding is determined by a global search at fixed composition, i.e., fixed value

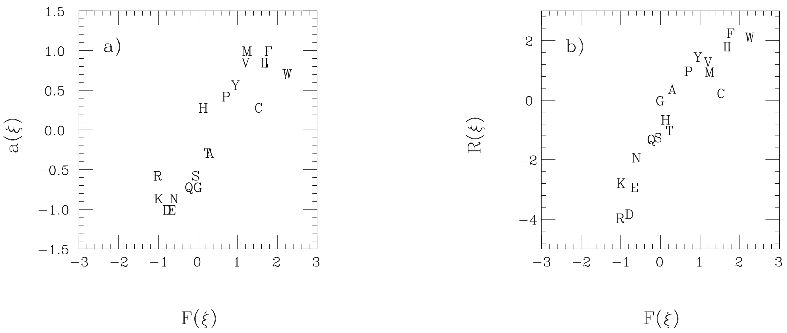

Figure 1.4: a) $a(\xi)$ against the hydrophobicity scale of Fauchère and Pliska [12], $F(\xi)$. b) $F(\xi)$ against the hydrophobicity scale of Roseman [13], $R(\xi)$.

of the parameter N_+. The interpretation of the optimal encoding depends on N_+. For $N_+ = 10$ we have shown that the results are strongly correlated with hydrophobicity. Since the method is global and free from physico-chemical inputs, this finding illustrates the importance of the hydrophobicity pattern.

The stability of the method was demonstrated by applying it to independent sets of protein sequences. The method can probably be improved by incorporating some sequence alignment technique. However, it is important that this is done in such a way that the amount of physico-chemical input is kept at a minimum. The method can easily be generalized to non-binary scales.

As an example of a possible application of our method, let us mention the question of how the statistical distribution of amino acids along functional protein sequences differs from a random distribution. This question has recently been addressed using binary assignments based on hydrophobicity [14]. By studying long-range correlations, it was demonstrated that protein sequences differ from random sequences in a statistically significant way. Alternative assignments such as those presented here may be useful in studying the nature of these deviations from randomness.

Acknowledgements

We thank Carsten Peterson, Roberto Ugoccioni and Christian Witt for useful discussions.

Bibliography

[1] See e.g. Bowie,J.U., Reidhaar-Olson,J.F, Lim,W.A. and Sauer,R.T. (1990), "Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitutions", *Science*, **247**, 1306–1310.

[2] Lim,W.A. and Sauer,R.T. (1989) "Alternative Packing Arrangements in the Hydrophobic Core of λ Repressor", *Nature*, **339**, 31–36.

[3] Dayhoff,M., Schwartz,R.M. and Orcutt,B.C. (1978) In Dayhoff,M. (ed.), *Atlas of Protein Sequences and Structures*, vol. 5, supplement 3, pp 345–352, National Biomedical Research Foundation, Silver Spring, Md.

[4] For a review, see Taylor,W.R. (1986) The classification of Amino Acid Conservation *J. Theor. Biol.*, **119**, 205–218.

[5] Lau,K.F. and Dill,K.A. (1989) "A Lattice Statistical Mechanics Model of the Conformational and Sequence Spaces of Proteins", *Macromolecules*, **22**, 3986–3997.

[6] Stillinger,F.H., Head-Gordon,T. and Hirshfeld,C.L. (1993) *Phys. Rev.*, **E 48**, 1469–1477.

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

[8] Bairoch,A. and Boeckmann,B. (1994) "The SWISS-PROT Protein Sequence Data Bank: Current Status", *Nucleic Acids Res.***22**, 3578–3580.

[9] Sweet,R.M. and Eisenberg,D. (1983) "Correlation of Sequence Hydrophobicties Measures Similarity in Three-dimensional Protein Structure", *J. Mol. Biol.*, **171**, 479–488.

[10] Cornette,J.L., Cease,K.B, Margalit,H., Spouge,J.L., Berzofsky,J.A. and DeLisi,C. (1987) "Hydrophobicity Scales and Computational Techniques for Detecting Amphipathic Structures in Proteins", *J. Mol. Biol.*, **195**, 659–685.

[11] Viari,A., Soldano,H. and Ollivier,E. (1990) "A Scale-Independent Signal Processing Method for Sequence Analysis", *CABIOS*, **6**, 71–80.

[12] Fauchère,J. and Pliska,V. (1983) "Hydrophobic parameters π of Amino-Acid Side Chains from the Partitioning of N-Acetyl-Amino-Acid Amides", *Eur. J. Med. Chem.* , **18**, 369–375.

[13] Roseman,M.A. (1988) "Hydrophilicity of Polar Amino Acid Side-Chains is Markedly Reduced by Flanking Peptide Bonds", *J. Mol. Biol.*, **200**, 513–522 (1988).

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

[15] Creighton,T.E. (1993) *Proteins: Structures and Molecular Properties.* W.H. Freeman and Company, New York.

[16] Jones,D.T, Taylor,W.R. and Thornton,J.M. (1992) *CABIOS*, **8**, 275–282.

Paper II

Evidence for Non-Random Hydrophobicity Structures in Protein Chains

Anders Irbäck, Carsten Peterson, and Frank Potthast

Department of Theoretical Physics, University of Lund
Sölvegatan 14A, S-223 62 Lund, Sweden

Proceedings of the National Academy of Sciences USA
93, 9533-9538 (1996)

Abstract:

The question of whether proteins originate from random sequences of amino acids is addressed. A statistical analysis is performed in terms of blocked and random walk values formed by binary hydrophobic assignments of the amino acids along the protein chains. Theoretical expectations of these variables from random distributions of hydrophobicities are compared with those obtained from functional proteins. The results, which are based upon proteins in the SWISS-PROT data base, convincingly show that the amino acid sequences in proteins differ from what is expected from random sequences in a statistically significant way. By performing Fourier transforms on the random walks one obtains additional evidence for non-randomness of the distributions.

We have also analyzed results from a synthetic model containing only two amino-acid types, hydrophobic and hydrophilic. With reasonable criteria on good folding properties in terms of thermodynamical and kinetic behavior, sequences that fold well are isolated. Performing the same statistical analysis on the sequences that fold well indicates similar deviations from randomness as for the functional proteins. The deviations from randomness can be interpreted as originating from anticorrelations in terms of an Ising spin model for the hydrophobicities.

Our results, which differ from some previous investigations using other methods, might have impact on how permissive with respect to sequence specificity the protein folding process is – only sequences with non-random hydrophobicity distributions fold well. Other distributions give rise to energy landscapes with poor folding properties and hence did not survive the evolution.

2.1 Introduction

Hydrophobicity is widely believed to play a central role in the formation of 3D protein structures. To understand the statistical distribution of hydrophobicity along proteins is therefore of utmost interest. This question has been addressed previously. In Ref. [1] the authors used binary hydrophobicity assignments, zero or one, and studied simultaneously the distribution of clumps of both zeros and ones by using the so-called run test. For the majority of the proteins examined it was found that the results could not be distinguished from those corresponding to completely random sequences. The same type of statistical test has also been applied to sequences stemming from a simplified protein model [2]. Here randomly selected sequences were compared with sequences that had been specially designed to have good folding properties. The statistical analysis did not reveal any difference between these two groups. These findings seem to indicate that the folding requirements on proteins are fairly permissive with little sequence specificity. A slightly different approach to analyze the same problem was pursued in Ref. [3], where by mapping the binary chains onto the trajectories of a random walk, deviations from random distributions are reported.

Also, recent work on simplified models suggest the non-randomness [4, 5]. In these studies a large number of randomly selected sequences were investigated,and it was found that only a small fraction of them folded easily into a thermodynamically stable state.

In this work we study the statistical distribution of hydrophobicity by using methods different from the run test in Ref. [1]. Along the same lines as in Ref. [3] rather than analyzing raw sequences of hydrophobicity, we focus on the corresponding random walk representation. In this way the analysis is more sensitive to long range correlations along the sequence. Our analysis has been carried out using two different methods, which differ substantially from what is used in Ref. [3] although the starting point is similar. First, we form block variables, and study how the behavior of these depends on the block size. When applied to the SWISS-PROT data base [6] of functional proteins, this method yields clear evidence for non-randomness. In addition, we have performed a Fourier analysis based on the random walk representation. In this analysis we find non-random behavior at the wave-length corresponding to α-helix structure, as one might have expected, but also at large wave-lengths.

In our analysis we have divided the sequences into groups corresponding to different fractions of hydrophobic residues. This division is important because the results for different groups deviate in different directions from those for random sequences. For sequences with a typical fraction of hydrophobic residues we find that the non-randomness can be interpreted as anticorrelations. This interpretation emerges from a simple Ising model of antiferromagnetic interactions among the residues.

Given the impact our results might have on the issue of how permissive with respect to sequence specificity the protein folding process is, we have carried out the same analysis for a toy model [7,8], for which unbiased samples of folding and non-folding sequences can be obtained. This model, hereafter denoted the **AB** model, consists of chains of two kinds of "amino acids" interacting with Lennard-Jones potentials. We have examined the behavior of 300 randomly selected chains of length 20 in this model [9]. Out of these only about 10% were found to have reasonable folding properties. Analyzing these sequences with the same methods as being used for the functional proteins, we obtain results that are qualitatively very similar to those for proteins with a typical fraction of hydrophobic residues. In particular, we again find deviations from random behavior that correspond to anticorrelations. One should keep in mind that the toy model chains are quite short and highly simplified as compared to functional proteins. Nevertheless, it is appealing to attempt an explanation for the observed similarity in behavior as originating from the fact that those amino acid sequences exhibiting this type of hydrophobicity distribution are the ones that fold well. Other distributions give rise to energy landscapes with poor folding properties and hence did not survive the evolution.

All our analysis concerns comparisons between distributions. The ultimate challenge is to decide whether a given sequence is non-random or not. This issue, which is beyond the scope of the paper, may be feasible when combining different cuts on the measures developed here.

This paper is organized as follows. In Section 2.2 we develop our two methods for analyzing binary hydrophobicity sequences. In Section 2.3 and 2.4 these methods are applied to real and toy model proteins respectively. Section 2.4 also contains the interpretation of deviations from randomness in terms of anticorrelations. Finally, a brief summary and outlook can be found in Section 2.5.

2.2 Methods

In this section we describe the variables and statistical methods employed. Two different, but not completely unrelated approaches are used – the blocking and Fourier transform methods. These will be described in some detail below.

Throughout the paper we consider sequences of N residues and denote by σ_i the hydrophobicity of residue i. We use a binary hydrophobicity scale: $\sigma_i = 1$ if residue i is hydrophobic and $\sigma_i = -1$ otherwise. The analysis can easily be extended to an arbitrary number of allowed hydrophobicity values and we do not expect our results to be affected by using such multivalued hydrophobicity assignments.

Hydrophobicities σ_i represent local properties of a chain. As in Ref. [3] in order to capture some long range correlation properties we consider a random walk representation

$$r_n = \sum_{i=1}^{n} \sigma_i \tag{2.1}$$

for $i = 1, \ldots, N$ and where $r_0 = 0$.

The Blocking Method

Analyzing the behavior of block variables is a widely used and fruitful technique in statistical mechanics, and our application will turn out to be no exception. For a block size s we define the variables

$$\sigma_i^{(s)} = \sum_{j=1}^{s} \sigma_{(i-1)s+j} = r_{is} - r_{(i-1)s} \qquad i = 1, \ldots, N/s \tag{2.2}$$

where it is assumed that N is a multiple of s. The scaling behavior of $\sigma_i^{(s)}$ with increasing s is determined by the correlations between σ_i and σ_j. If the σ_i's are independent random numbers drawn from the same distribution, the correlations between different σ_i and σ_j vanish and the variance of $\sigma_i^{(s)}$ scales linearly with s.

We need to be able to compare real proteins with a random distribution of hydrophobic residues. For this reason we average over all sequences with a fixed length and composition. These averages are denoted by $\langle \cdot \rangle_{N,N_+}$, where N is the total number of residues and N_+ is the number of hydrophobic residues.

In order to study the fluctuations of the block variables we introduce the normalized variables

$$\psi_i^{(s)} = \frac{1}{K}\Big(\sigma_i^{(s)} - \frac{s}{N}\sum_{j=1}^{N/s} \sigma_j^{(s)}\Big)^2 \qquad i = 1, \ldots, N/s \tag{2.3}$$

where

$$K = \frac{4N_+(N - N_+)}{N(N-1)}(1 - s/N) \tag{2.4}$$

The constant K is chosen such that $\langle \psi_i^{(s)} \rangle_{N,N_+} = s$ for all N and N_+. The fact that K depends on s implies that the variance of $\sigma_i^{(s)}$ is not linear in s, which is due to the fact that the average is taken at fixed composition. At fixed s this deviation from linearity disappears in the limit $N \to \infty$. If all the residues are of the same type, K vanishes. Such sequences are uninteresting in the present analysis and have therefore been excluded.

An important quantity is the (normalized) mean-square fluctuation of the block variables, defined by

$$\psi^{(s)} = \frac{1}{K}\frac{s}{N}\sum_{i=1}^{N/s}\Big(\sigma_i^{(s)} - \frac{s}{N}\sum_{j=1}^{N/s}\sigma_j^{(s)}\Big)^2 = \frac{s}{N}\sum_{i=1}^{N/s}\psi_i^{(s)} \tag{2.5}$$

Obviously, one has

$$\langle \psi^{(s)} \rangle_{N,N_+} = s \tag{2.6}$$

and that $\psi^{(1)} = 1$ is independent of configuration σ_i. It is also important to know the variance of $\psi^{(s)}$. The complete expression for this quantity is lengthy and can be found in Section 2.6.1. However, in the $N \to \infty$ limit it takes the simple form

$$\langle \psi^{(s)2} \rangle_{N,N_+} - \langle \psi^{(s)} \rangle^2_{N,N_+} \sim \frac{2s^2(s-1)}{N} \tag{2.7}$$

When studying proteins from the data base, we average over sequences with different length and composition. For a general quantity this requires some assumption about the probability of different values of N and N_+ in order to compare with random sequences. This problem is absent for $\psi^{(s)}$, since it is defined such that $\langle \psi^{(s)} \rangle_{N,N_+}$ is independent of N and N_+. The variance of $\psi^{(s)}$, on the other hand, does depend upon N and N_+. However, for an interval $N_1 < N < N_2$ with both N_1 and N_2 large and $(N_2 - N_1)$ not too large, it is still possible to use Eq. (2.7) for estimating the variance.

The Fourier Transform Method

The most direct way to detect periodicity in the distribution of hydrophobic residues is to use Fourier analysis. It is well-known that the Fourier component corresponding to a period of 3.6 residues tends to be strong for sequences that form α helices [11]. Also, sequences that form β sheets tend to exhibit a periodicity in the hydrophobicity of about 2.3 residues. In this paper we compare the full power spectrum for proteins with that for random sequences.

As a starting point for our Fourier analysis we take the random walk representation r_n. Since we want to compare with random sequences and since any permutation of the residues leaves the end point r_N unchanged, it is here convenient to introduce the modified random walk (see Section 2.6.2)

$$\rho_0 = r_0 = 0 \tag{2.8}$$

$$\rho_n = \sum_{i=1}^{n} \left(\sigma_i - \frac{2N_+ - N}{N} \right) = r_n - n\frac{2N_+ - N}{N} \qquad n = 1, \ldots, N \tag{2.9}$$

which is defined such that $\rho_0 = \rho_N = 0$. With these boundary conditions, we consider the sine transform

$$f_k = \sum_{n=1}^{N-1} \rho_n \sin \frac{\pi k n}{N} \qquad k = 1, \ldots, N-1 \tag{2.10}$$

where the kth component corresponds to a wave-length of $2N/k$ residues.

It is easy to see that the average of f_k over all sequences with a fixed length and composition vanishes, and for the squared amplitude we find

$$\langle f_k^2 \rangle_{N,N_+} = \frac{2N_+(N-N_+)}{N-1} \frac{1}{(2\sin\frac{\pi k}{2N})^2} \tag{2.11}$$

which shows that this quantity behaves as k^{-2} for small k. In Section 2.6.2 we also give the fourth moment of the f_k distribution.

In our calculations we have used the normalized squared amplitude

$$\tilde{f}_k^2 = \frac{f_k^2}{\langle f_k^2 \rangle_{N,N_+}} \tag{2.12}$$

which has an average $\langle \tilde{f}_k^2 \rangle_{N,N_+} = 1$ independent of N and N_+. By measuring $\tilde{f}_k^2$ one can, of course, only study the relative strength of the different components. In fact, it can easily be shown that

$$\sum_{k=1}^{N-1} \tilde{f}_k^2 = N - 1 \tag{2.13}$$

independent of configuration σ_i.

2.3 Real Proteins

Our analysis has been carried out using the SWISS-PROT data base, release 31 [6]. Some proteins were removed from this data base due to uncertainties (see Section 2.6.3 for details). Also, we limit the analysis to proteins with $N \geq 50$ after the endpoints have been removed according to a prescription to be dealt with below. To each residue we assigned a binary hydrophobicity value, which was taken to be $+1$ for Leu, Ile, Val, Phe, Met, and Trp and -1 for the others. This choice was done by picking the residues with strongest hydrophobic interactions down to a preset level. Alternative definitions, with 4 to 11[1] of the residues classified as hydrophobic, have also been tested, with qualitatively similar results.

Estimates of statistical errors on the measurements have been obtained by dividing the data into 20 groups and treating the corresponding averages as independent measurements. All statistical errors quoted are in σ error units.

Before starting our final analysis of hydrophobicity correlations, we need to deal with two important observations.

- The data originating from the ends of the sequences display a different behavior than the data from the rest of the sequences.

[1] This is the choice of Ref. [1].

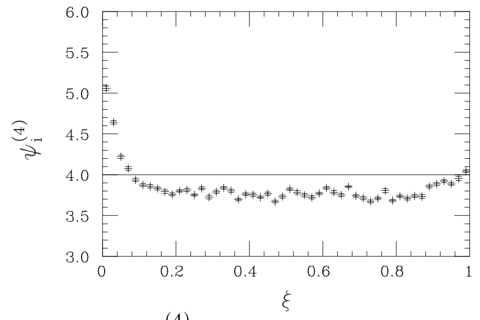

Figure 2.1: Average values of $\psi_i^{(4)}$ against the relative positions of the blocks along the sequence, ξ.

- Sequences with different fractions of hydrophobic residues tend to behave in different ways.

As a result, important effects can easily be missed out if averages are computed over the full data set, as will be seen below.

The Interior versus the Ends

We begin by examining whether the behavior of the block variable depends upon the position of the block along the sequence. This analysis is carried out for block sizes $s = 2, 3, 4, 6$, and 12. In order to obtain a sequence that can be divided into blocks for each of these sizes, we disregard up to eleven residues at the ends. In this way we form a sequence of length $N' = n \cdot 12$, where n is the largest integer such that $n \cdot 12 \leq N$, N being the length of the original sequence.

We study the block fluctuation $\psi_i^{(s)}$ as a function of the relative position ξ of the block center, ξ being 0 at the N-end and 1 at the C-end. The interval in ξ from 0 to 1 is divided into 50 bins and average values were computed for each of these bins, using all sequences in the data base with more than 50 residues. In Fig. 2.1 we show the result for block size $s = 4$. The results for other values of s are similar. The horizontal line in the figure represents random sequences; if the distribution of hydrophobic residues were random, the average of $\psi_i^{(s)}$ would be s, independent of i.

From Fig. 2.1 it is clear that the block fluctuations are roughly constant over a wide range in ξ. However, it is also evident that the fluctuations tend to increase in strength at the ends, in particular at the N-end. One also notices that the deviations from the random value tend to cancel if one averages over all positions.

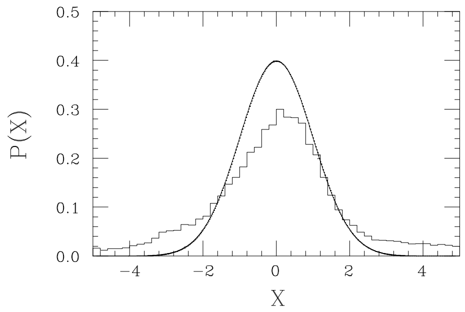

Figure 2.2: The distribution of X for the sequences in the data base. The curve is the Gaussian with zero mean and unit variance.

This shows that it is important to distinguish between the ends and the interior of the sequences when studying hydrophobicity correlations. In what follows we focus on the interior by ignoring 15% of the residues at each of the two ends, and analyze sequences containing the remaining 70% of the residues.

The Fraction of Hydrophobic Residues

Our main focus in this paper is on the distribution of hydrophobic residues along the sequence, and to what extent this distribution has random characteristics. One may also ask whether the total number of hydrophobic residues in a sequence follows a random pattern. This question can be addressed by studying the quantity

$$X = \frac{N_+ - Np}{\sqrt{Np(1-p)}} \tag{2.14}$$

where N_+ is the number of hydrophobic residues, N is the total number of residues, and p is the average of N_+/N over all sequences. If N hydrophobicity values are drawn randomly and independently with probability p for the value 1 and $(1-p)$ for -1, the distribution of X becomes approximately Gaussian with zero mean and unit variance for large N.

We have calculated X for the sequences in the data base, after eliminating 30% of the residues, as discussed above. The average fraction of hydrophobic residues was found to be $p \approx 0.291$. The distribution of X obtained is shown in Fig. 2.2, from which we see that the tails are larger than for the random distribution.

When studying correlations in hydrophobicity, we have divided the sequences into groups corresponding to different regions in X. This division need not have a simple interpretation in terms of standard groups of proteins,

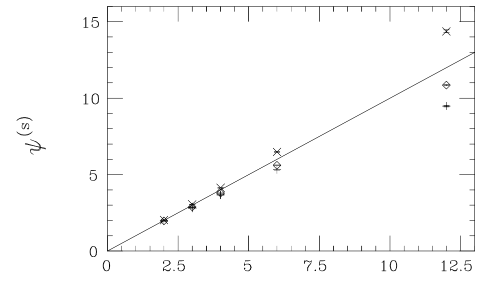

Figure 2.3: Mean-square fluctuation of the block variables, $\psi^{(s)}$, as a function of block size s for $|X| < 0.5$ (+ ;10154 qualifying proteins), $|X| > 3$ ($\times$; 4928), and all X ($\diamond$; 36765). The averages have been computed over sequences which contained more than 50 residues prior to the elimination of residues at the ends. The straight line is the result for random sequences.

but turns out to be useful. Indeed, we will find below that sequences with different X tend to display different types of correlations.

Results

We now turn to the results of our block and Fourier analysis. As discussed in the previous two subsections, we have chosen to consider the interior of the sequences and to study different regions in X.

First we consider the mean-square fluctuation of the block variables, $\psi^{(s)}$. In Fig. 2.3 results are shown corresponding to three different regions in X: $|X| < 0.5$, $|X| > 3$, and all X. The straight line represents random sequences. We see that the results for large X lie above this line, while the results for small X show the opposite behavior. The same pattern is observed when using alternative hydrophobicity assignments. Notice that $\psi^{(s)}$ cannot increase slower than linearly with s if the correlation between σ_i and σ_j is translationally invariant and non-negative. Therefore, these results suggest that there exists negative hydrophobicity correlations for small X.

We have also tested how these results depend on the length of the sequences by computing averages of $\psi^{(s)}$ corresponding to different intervals in N, where N is the length of the sequence prior to the elimination of residues at the ends. In Fig. 2.4 we show results obtained for $|X| < 0.5$ and three different intervals in N. It is clear from Fig. 2.4 that the size dependence is fairly weak. Another interesting feature is that the deviation from the result for random sequences grows with sequence length. Notice that the variance of $\psi^{(s)}$ scales as $N^{-1/2}$ for random sequences.

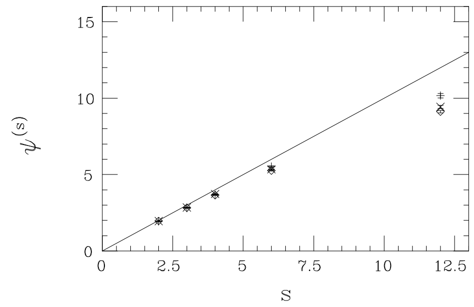

Figure 2.4: Mean-square fluctuation of the block variables, $\psi^{(s)}$, against block size s for $|X| < 0.5$ and $50 < N \leq 150$ (+; 2457 qualifying proteins), $150 < N \leq 250$ ($\times$; 2228), and $250 < N \leq 350$ ($\diamond$; 1642). The straight line is the result for random sequences.

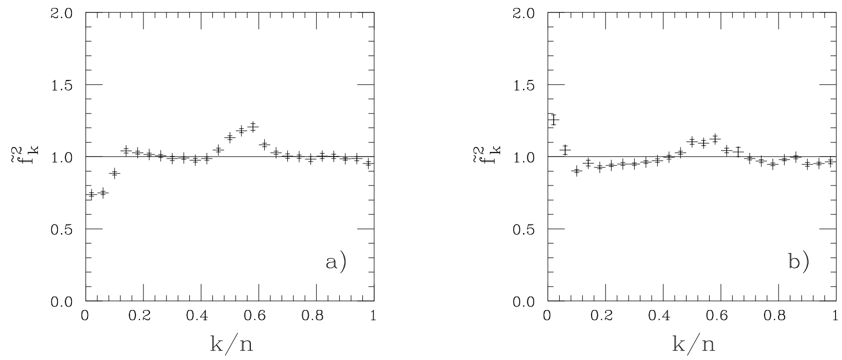

Figure 2.5: Normalized squared amplitude $\tilde{f}_k^2$ against k/N for a) $|X| < 0.5$ and b) $|X| > 3$. The sets of sequences considered are the same as in Fig. 3.

Next we compare the behavior of the Fourier components for small and large $|X|$. In Fig. 2.5 we have plotted the normalized squared amplitude $\tilde{f}_k^2$ against k/N for $|X| < 0.5$ and $|X| > 3$. Let us first consider the region of small and medium wave-length. Here the results for the two intervals in $|X|$ are similar. As one might have expected, there is a peak around the wave-length corresponding to α-helix structure, $2N/k = 3.6$. Away from this peak, the results are very close to those for random sequences.

At large wave-length, on the other hand, the results show a clear $|X|$ dependence, and they differ from the results for random sequences for both small and large $|X|$. As can be seen from the figure, these components are suppressed for small $|X|$ and strong for large $|X|$.

Tests on Non-Redundant Sets

A general problem in the statistical analysis of proteins is the presence of homologies since these may shift away distributions from an ideal set of independent samples.

In order to test for effects due to homologies, we redid the analysis above using a set of 486 selected sequences [12][2] from the Protein Data Bank [13]. This set was obtained by allowing for a maximum of 25% sequence similarity for aligned subsequences of more than 80 residues [14]. Within this set of minimally redundant sequences, 185 with $|X| < 0.5$ and 5 with $|X| > 3.0$ qualified for analysis. The results for $|X| < 0.5$ are within statistics identical to those described above. For $|X| > 3.0$ the results are not in conflict with the results above but quantitative comparisons are not meaningful due to the extremely small sample size.

The fact that our results survive when limiting ourselves to non-redundant proteins, implying a substantial cut in number of proteins involved in the analysis, makes the evidence of non-randomness even stronger.

2.4 A Simplified Synthetic Model

In this section we carry through the same hydrophobicity analysis as above for a simple toy model for proteins [7, 8] with binary amino acids - The **AB** model. Due to its simplicity and the relatively small sizes involved the folding properties of this model have been studied to quite some detail [5, 9]. The question we want to address here is whether the sequences, which have good folding properties in the AB model, deviate from the non-folding ones in a way qualitatively similar to what was found for the small-$|X|$ functional proteins above. As will be shown below this is indeed the case.

The AB Model

In this model, there are two kinds of residues with $\sigma_i = \pm 1$ (A and B) respectively. These are linked by rigid bonds to form linear chains in two dimensions. The interactions between the residues are given by σ_i-dependent Lennard-Jones potentials such that $(++)$ is strongly attractive, $(--)$ weakly attractive and $(+-)$ repulsive. In Ref. [5] the thermodynamics of this system at low temperature was studied using the hybrid Monte Carlo method. Fluctuations in the shape for a given chain were studied by measuring the mean-square distance δ^2 between pairs of configurations; the probability distribution of δ^2, for fixed temperature and sequence, describes the magnitude of the thermodynamically relevant fluctuations. It is suggestive to interpret a low average δ^2 as a signal

[2] The March 1996 edition was used.

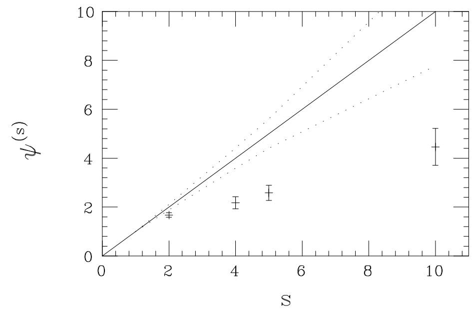

Figure 2.6: Mean-square fluctuation of the block variables, $\psi^{(s)}$, against block size s for good folding sequences in the AB model. Also shown are the mean s (full line) and the $s \pm \sigma$ band (bounded by dotted lines) for random sequences.

for good folding and stability properties. Recently an attempt to understand the systematics of how low δ^2 values relate to the σ_i sequence was pursued [9]. In this work 300 randomly selected sequences with 14 A and 6 B residues were studied, using an improved Monte Carlo method [5, 10]. The sequences were classified as having good folding properties if the average δ^2 was less than 0.3, or if the probability of $\delta^2 < 0.1$ was greater than 0.35. This yielded a total of 37 good folders (roughly 10 %).

Results

Using the 37 good folding sequences, we have repeated the analysis of the previous section. This set of sequences is fairly small, but has the advantage that it is has been generated in a bias-free way. Statistical errors given in this section have been obtained by taking the results for different sequences as independent measurements.

In Fig. 2.6 we show the mean-square fluctuation of the block variables, $\psi^{(s)}$. The average of $\psi^{(s)}$ over 37 random sequences has an approximately Gaussian distribution, with mean s and a standard deviation σ that can be obtained by using the results of Section 2.6.1. In the figure we have indicated the position of the $s \pm \sigma$ band. We see that the data points lie clearly below this band.

Our results for the squared Fourier amplitude are shown in Fig. 2.7. Although the statistical errors on this quantity are large, there are clear deviations from the result for random sequences at large wave-length. We see that components corresponding to large wave-lengths are suppressed.

These results show that good folding sequences in the AB model tend to exhibit small block fluctuations and weak Fourier components at large wave-length. Qualitatively, the results are very similar to those obtained in the

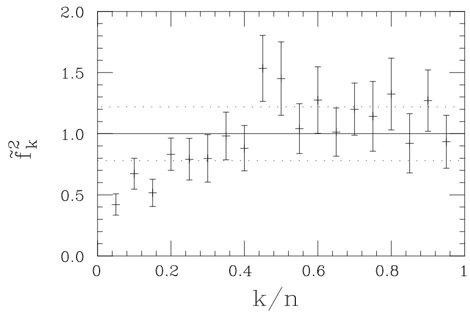

Figure 2.7: Normalized squared amplitude $\tilde{f}_k^2$ against k/N for good folding sequences in the AB model. The full line and dots are as in Fig. 6. The standard deviation for random sequences can be obtained by using the results of Section 2.6.2.

previous section for small $|X|$.

Interpretation of the Results

In this paper we have compared various results with those for random sequences. Random sequences correspond to a situation in which there is (essentially) no correlation between σ_i and σ_j for $i \neq j$. A simple but instructive way to introduce non-zero correlations into the system is to consider the one-dimensional Ising model. In this model there are N "spins" σ_i that take the values ± 1, and each configuration is given a statistical weight

$$P \propto \exp\Big(K \sum_{i=1}^{N-1} \sigma_i \sigma_{i+1}\Big) \tag{2.15}$$

As in our previous calculations, we consider configurations with a fixed number of positive spins, i.e., the magnetization $M = \sum_{i=1}^{N} \sigma_i = 2N_+ - N$ is held fixed. Also, as before free boundary conditions are assumed.

The properties of this system are determined by the parameter K. Neighboring spins tend to point in the same direction if $K > 0$ (ferromagnet), and in opposite directions if $K < 0$ (antiferromagnet). For $K = 0$ we recover the (random) system studied previously.

To illustrate the behavior of the system for non-zero K, we show in Fig. 2.8 results obtained at $K = \pm 0.25$. As in our AB calculations, we have taken $N = 20$ and $N_+ = 14$. At $K = 0.25$ we see that block fluctuations are large and that Fourier components with large wave-length are strong, while the behavior is the opposite at $K = -0.25$. This means that the results for this antiferromagnetic system are similar to those obtained for good folding

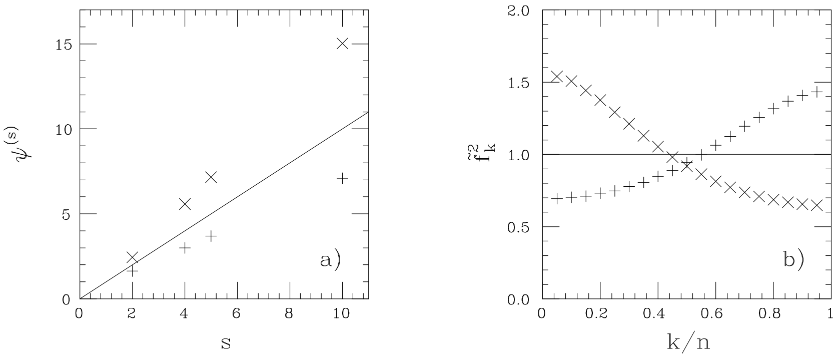

Figure 2.8: a) Mean-square fluctuation of the block variable, $\psi^{(s)}$, and b) normalized squared amplitude $\tilde{f}_k^2$ for the Ising model with $K = -0.25$ (+) and $K = 0.25$ ($\times$). The lines correspond to $K = 0$.

sequences in the AB model and for protein sequences with small $|X|$. On the other hand, the results for the ferromagnetic system resemble those for proteins with large $|X|$.

2.5 Summary

We have demonstrated that the statistical distribution of hydrophobic residues along chains of functional proteins are non-random. This result is in contrast with what was concluded in Ref. [1]. An important reason for this difference is probably that the blocking and Fourier analysis methods are able to capture long range correlations more efficiently than the method of Ref. [1]. In Ref. [3], on the other hand, a method more similar to ours was used and deviations from random behavior were observed, but the deviations may seem to differ in nature from what we have found. However, it is important to note that these authors focused on hydrophilicity rather than hydrophobicity, as they used a binary classification in which five strongly hydrophilic residues formed one group. Also, the interpretation of the results of Ref. [3] is somewhat unclear, as no distinction was made between the interior and the ends of the sequences. When limiting the data set to non-redundant protein chains the results from the analysis are unaffected. Hence we consider our evidence for non-randomness as being quite robust.

We have also applied our analysis method to a toy model data base (AB model) where chains with good folding properties were distinguished from the rest. The hydrophobicity distributions of the good folding sequences differ from random ones in qualitatively the same way as for the low-$|X|$ functional protein analysis. It is tempting to interpret this similarity as indicating that only those proteins with good folding properties have survived the evolution.

The deviation from randomness in the AB model case can be understood as originating from anticorrelations among the residues. The effects of correlations and anticorrelations on the observables considered were illustrated by using the simple one-dimensional Ising model.

Our analysis has been a statistical one in the sense that distributions are being compared. Given our encouraging result, it might be possible to reach the ultimate goal of being able to classify individual sequences in terms of belonging to one category or the other. This might be feasable by considering suitable cuts in the block and Fourier quantities. Very likely, one then needs to augment the method with additional discriminative variables and an automated procedure like artificial neural networks for setting the cuts.

Our analysis has been confined to binary hydrophobicity assignments. The results presented are insensitive to minor modifications of these assignments. We do not expect the results to change significantly if instead of binary assignments multivalued ones are used.

2.6 Appendix

2.6.1 Variance of $\psi^{(s)}$

In this Appendix we give the variance of $\psi^{(s)}$ (see Eq. 2.5). The average of $\psi^{(s)}$ over all sequences with fixed composition, N_+ and $N_- = N - N_+$, is given by

$$\langle \psi^{(s)} \rangle_{N,N_+} = \frac{1}{K} \sum_{i,j=1}^{s} c_{ij} \tag{2.16}$$

where c_{ij} is the connected correlation between σ_i and σ_j, for which one finds

$$c_{ij} = \langle \sigma_i \sigma_j \rangle_{N,N_+} - \langle \sigma_i \rangle_{N,N_+} \langle \sigma_j \rangle_{N,N_+} = \begin{cases} \frac{4N_+N_-}{N^2} & \text{if } i = j \\ -\frac{4N_+N_-}{N^2(N-1)} & \text{if } i \neq j \end{cases} \tag{2.17}$$

Using this, one obtains $\langle \psi^{(s)} \rangle_{N,N_+} = s$ (Eq. 2.6). The off-diagonal correlation, c_{ij} with $i \neq j$, has to be negative since $\sum_{i=1}^N \sum_{j=1}^N c_{ij} = 0$, but vanishes in the limit $N \to \infty$.

The variance can be computed in a similar way. In addition to c_{ij}, one then needs the correlation between four σ_i's. One finds

$$\langle \psi^{(s)2} \rangle_{N,N_+} - \langle \psi^{(s)} \rangle^2_{N,N_+} = 2s^2(s-1)N^{-1}K^{-2}G(s,N,N_+) \tag{2.18}$$

where

$$\begin{aligned} & G(s,N,N_+) = \qquad (2.19) \\ 1+ \quad & 2(s-2)\frac{(N_+-N_-)^2-N}{N(N-1)} \\ - \quad & (2s-3)\frac{(N_+-N_-)^4-6N(N_+-N_-)^2+3N^2+8(N_+-N_-)^2-6N}{N(N-1)(N-2)(N-3)} \\ + \quad & \frac{1}{2}(s-1)\cdot\Big(\frac{(4N-6)(N_+-N_-)^4}{N(N-1)^2(N-2)(N-3)} \\ + \quad & \frac{-6N(N_+-N_-)^2+3N^2+8(N_+-N_-)^2-6N}{(N-1)(N-2)(N-3)} + \frac{2(N_+-N_-)^2-N}{(N-1)^2}\Big) \end{aligned}$$

In the limit $N \to \infty$, for fixed s, this expression simplifies to Eq. 2.7.

2.6.2 Fourier Transforms of Random Walk Representations

In this Appendix Fourier transform moments of random walk representations are listed. The expressions are more general than what is required for binary hydrophobicity assignments. In order to do this we first list the following basic quantities for sequences of length N:

- Moment of order k, m_k.
  $m_k = \frac{1}{N}\sum_{i=1}^{N}\sigma_i^k$.
- Cumulant of order k, c_k.
  $c_2 = m_2 - m_1^2$
  $c_4 = m_4 - 4m_3m_1 - 3m_2^2 + 12m_2m_1^2 - 6m_1^4$
- Random walk, ρ_n.
  $\rho_0 = 0$
  $\rho_n = \sum_{i=1}^{n}\sigma_i - nm_1 \qquad i = 1,\ldots,N \qquad (\rho_N = 0)$
- Sine transform of ρ_n, f_k
  $f_k = \sum_{n=1}^{N-1}\rho_n \sin\frac{\pi kn}{N} \qquad k = 1,\ldots,N-1$

Averaging over all sequences with fixed composition, i.e., all permutations of σ_i, one obtains

$$\langle\rho_n\rangle = 0 \tag{2.20}$$

$$\langle\rho_n^2\rangle = \frac{N^2c_2}{N-1}\cdot\frac{n}{N}(1-\frac{n}{N}) \tag{2.21}$$

$$\langle f_k\rangle = 0 \tag{2.22}$$

$$\langle f_k^2\rangle = \frac{N^2c_2}{2(N-1)}\cdot\frac{1}{(2\sin\frac{\pi k}{2N})^2} \tag{2.23}$$

$$\begin{aligned}\langle f_k^4\rangle = {} & \frac{3N^4}{4(N-1)(N-2)}\Big[c_2^2 - \frac{1}{2N}(c_4+6c_2^2) \\ & -\delta_{2k,N}\frac{1}{24(N-3)}\Big(c_4 + \frac{1}{N}(c_4+6c_2^2)\Big)\Big]\cdot\frac{1}{(2\sin\frac{\pi k}{2N})^4}\end{aligned} \tag{2.24}$$

For the binary scale $\sigma_i = \pm 1$ one has $c_2 = 4N_+(N-N_+)/N^2$ and Eq. 2.23 becomes Eq. 2.11.

2.6.3 Removal of Uncertain Sequences

In our analysis we have removed "uncertain sequences" from the SWISS-PROT database by ignoring all entries containing the following feature keys in their feature key table:

- **UNSURE** - Indicates that there are uncertainties in the sequence.
- **NON_CONS** - Indicates that two residues in a sequence are not consecutive and that there are a number of unsequenced residues in between.
- **NON_TER** - The residue at an extremity of the sequence is not the terminal residue.

This reduces the size of the SWISS-PROT database from 43470 to 38050 protein entries.

Furthermore, when analyzing the interior parts of protein sequences, sequences containing the following letters

- **B** denoting Aspartic acid or Asparagine.
- **Z** denoting Glutamine or Glutamic acid.
- **X** denoting any amino acid.

within the interior are removed.

Bibliography

[1] S.H. White and R.E. Jacobs, "Statistical Distribution of Hydrophobic Residues Along the Length of Protein Chains", *Biophys. J.* **57**, 911 (1990).

[2] E.I. Shakhnovich and A.M. Gutin, "Engineering of Stable and Fast-Folding Sequences of Model Proteins", *Proc. Natl. Acad. Sci. USA* **90**, 7195 (1993).

[3] V.S. Pande, A.Y. Grosberg and T. Tanaka, "Nonrandomness in Protein Sequences: Evidence for a Physically Driven Stage of Evolution", *Proc. Natl. Acad. Sci. USA* **91**, 12972 (1994).

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

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

[6] A. Bairoch and B. Boeckmann, "The SWISS-PROT Protein Sequence Data Bank: Current Status", *Nucleic Acids Res.* **22**, 3578 (1994).

[7] F.H. Stillinger, T. Head-Gordon and C.L. Hirschfeld, "Toy Model for Protein Folding", *Phys. Rev.* **E48**, 1469 (1993).

[8] T. Head-Gordon and F.H. Stillinger, "Optimized Neural Network for Protein-Structure Prediction", *Phys. Rev.* **E48**, 1502 (1993).

[9] A. Irbäck, C. Peterson and F. Potthast, in preparation.

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

[11] D. Eisenberg, R.M. Weiss, and T.C. Terwilliger, "The Hydrophobic Moment Detects Periodicity in Protein Hydrophobicity", *Proc. Natl. Acad. Sci. USA* **81**, 140 (1984).

[12] U. Hobohm and C. Sander, "Enlarged Representative Set of Protein Structures", *Protein Science* **3**, 522 (1994).

[13] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, "The Protein Data Bank: A Computer Based Archival File for Macromolecular Structures", *J. Mol. Biol.* **112**, 535 (1977).

[14] U. Hobohm, M. Scharf, R. Schneider and C. Sander, "Selection of Representative Protein Data Sets", *Protein Science* **1**, 409 (1992).

Paper III

Studies of an off-lattice model for protein folding: Sequence dependence and improved sampling at finite temperature

Anders Irbäck and Frank Potthast
Department of Theoretical Physics, University of Lund
Sölvegatan 14A, S-223 62 Lund, Sweden

Journal of Chemical Physics **103**, 10298–10305 (1995)

Abstract:

We study the thermodynamic behavior of a simple off-lattice model for protein folding. The model is two-dimensional and has two different "amino acids". Using numerical simulations of all chains containing eight or ten monomers, we examine the sequence dependence at a fixed temperature. It is shown that only a few of the chains exist in unique folded state at this temperature, and the energy level spectra of chains with different types of behavior are compared. Furthermore, we use this model as a testbed for two improved Monte Carlo algorithms. Both algorithms are based on letting some parameter of the model become a dynamical variable; one of the algorithms uses a fluctuating temperature and the other a fluctuating monomer sequence. We find that by these algorithms one gains large factors in efficiency in comparison with conventional methods.

3.1 Introduction

The protein molecule is a very complex system, and understanding the folding of natural proteins remains one of the most challenging problems in biophysics. In recent years there has been an increasing interest in understanding the relevant mechanisms of the folding process by studying simplified models. Several authors have used ideas from spin-glass theory to gain useful insights into the behavior of self-interacting random chains [1–5]. The picture emerging from these studies is that the phase diagram consists of coil, globule and frozen phases. In the frozen phase the system exists in one of a number of different states, each corresponding to a fairly well-defined shape. This phase is interesting from the viewpoint of proteins, but for a generic random sequence it appears impossible to identify a state corresponding to the unique native state of a protein.

The sequence dependence of the folding properties has been studied numerically in various models [6–14]. An extensive study was reported recently by Šali *et al.* [11], who examined the behavior of 200 randomly selected sequences in a lattice model with contact interactions. It was found that 30 of these exhibited a folded state which was both thermodynamically dominant and kinetically accessible in a reasonable time. The existence of folding and nonfolding sequences has been observed also in off-lattice models [10], but much less is known about the behavior of such models. In this paper we study the thermodynamic behavior of different sequences in a very simple off-lattice model.

The model studied has been proposed by Stillinger *et al.* [15]. It is two-dimensional and has only two kinds of "amino acids". In Refs. [15, 16] the energy and structure of the ground state was determined for all possible chains with seven or fewer monomers, and an interesting interpretation of the results was obtained by employing neural-network techniques. In this paper we study the behavior of the model at finite temperature. To study the sequence dependence, we have carried out numerical simulations of all possible chains containing eight or ten monomers at a fixed temperature. We find that only a few of the chains exist in a unique folded state at this temperature. These chains have a relatively high folding temperature T_f. This property is important because the dynamics tends to be very slow at low temperature. Chains with a high T_f are therefore more likely to satisfy both the thermodynamic and kinetic requirements for folding. We have also determined the low-lying energy levels of a few different sequences. Sequences with good folding properties are expected to exhibit a large stability gap [14], which is defined as the energy gap between the native state and the lowest of all states with little structural similarity to the native state. The stability gaps of the sequences we have studied show a wide variation, and, as expected, it is large when T_f is high.

This first part of our study shows that the model displays interesting features for relatively short chains. It has been performed by using the well established hybrid Monte Carlo algorithm [17]. In the second part of the paper, we use the model as a testbed for two algorithms that are meant to facilitate the study of longer chains.

Simulations of heteropolymeric chains at low temperature are notoriously difficult, due to the presence of high free-energy barriers separating different folded states. The problem is that any local algorithm requires the system to pass through these barriers, which leads to a suppression of transitions between different free-energy valleys. A method designed to overcome this difficulty is the multicanonical Monte Carlo algorithm [18], which is closely related to the umbrella-sampling method [19]. The trick used here is to simulate a modified energy function which one tries to choose so as to eliminate the free-energy barriers. The canonical distribution is then extracted by means of the reweighting technique [20].

An alternative approach is provided by the method of simulated tempering [21]. In this method the temperature becomes a dynamical variable which takes values ranging over a definite set. In this way one tries to utilize the fact that at higher temperature the free-energy barriers are lower. In Ref. [21] this method was successfully applied to the random-field Ising model. The idea to let some parameter of the model become a dynamical variable has subsequently been used to accelerate simulations of other systems too [22–25]. It had earlier been shown by Lyubartsev *et al.* [26] that this is a useful method for calculating the free energy.

In this paper we investigate the use of the dynamical-parameter method in simulating heteropolymeric chains. We consider the simulated-tempering algorithm, and also another algorithm of the same type which we call the multisequence algorithm. In this algorithm the sequence degrees of freedom become a dynamical variable, which means that a number of different monomer sequences are simulated in parallel. We compare the performance of these two algorithms with that of hybrid Monte Carlo. Our results show that by the dynamical-parameter algorithms one greatly reduces the amount of computer time required for a representative sampling of the different folded states.

The plan of this paper is as follows. The model is described in Sec. 3.2 and the algorithms in Sec. 3.3. Sec. 3.4 deals with homopolymers, whereas general sequences are studied in Sec. 3.5. In Sec. 3.6 we present the results of our tests of the dynamical-parameter algorithms. Sec. 3.7 is a summary.

3.2 The Model

The model studied has two kinds of monomers, to be called A and B. The monomers are linked by rigid bonds of unit length to form linear chains living in two dimensions. For an N-mer we specify the sequence of monomers by the binary variables $\xi_1, \ldots, \xi_N$ and the configuration by the angles $\theta_2, \ldots, \theta_{N-1}$, where θ_i denotes the bend angle at site i and is taken to satisfy $|\theta_i| \leq \pi$. The energy function is

$$E(\theta, \xi) = \sum_{i=2}^{N-1} E_1(\theta_i) + \sum_{i=1}^{N-2} \sum_{j=i+2}^{N} E_2(r_{ij}, \xi_i, \xi_j) \tag{3.1}$$

where

$$\begin{aligned} E_1(\theta_i) &= \frac{1}{4}(1 - \cos\theta_i) \\ E_2(r_{ij}, \xi_i, \xi_j) &= 4(r_{ij}^{-12} - C(\xi_i, \xi_j) r_{ij}^{-6}) \end{aligned} \tag{3.2}$$

and $r_{ij} = r_{ij}(\theta_{i+1}, \ldots, \theta_{j-1})$ denotes the distance between sites i and j of the chain. The term $E_1(\theta_i)$ favors alignment of the three successive sites $i-1$, i and $i+1$. The nonbonded interactions E_2 are Lennard-Jones potentials with a species-dependent coefficient $C(\xi_i, \xi_j)$, which is taken to be 1 for an AA pair (strong attraction), 1/2 for a BB pair (weak attraction) and -1/2 for an AB pair (repulsion). Consequently, there is an energetic preference for separation between the two kinds of monomers. In fact, it was demonstrated in Ref. [15] that ground-state configurations tend to have a core consisting mainly of A monomers, which shows that A and B monomers behave respectively as hydrophobic and polar residues.

The energy is, for fixed sequence, a function only of the angles θ_i, and is therefore translationally and rotationally invariant. In addition, it is invariant under reflection ($\theta_i \to -\theta_i$ for all i) and change of the orientation ($\theta_i \to -\theta_{N-i+1}$ and $\xi_i \to \xi_{N-i+1}$ for all i). This implies that a generic energy level is fourfold degenerate for symmetric sequences, and twofold degenerate for asymmetric sequences. In numerical simulations these symmetries provide useful checks on effective ergodicity breaking.

The behavior of the model at finite temperature T is defined by the partition function

$$Z(T, \xi) = \int \left[\prod_{i=2}^{N-1} d\theta_i \right] \exp(-E(\theta, \xi)/T) \ . \tag{3.3}$$

That the T dependence of this simple model exhibits an interesting phase structure will become clear in Secs. 4 and 5.

3.3 The Algorithms

Hybrid Monte Carlo

We give a brief description of the hybrid Monte Carlo algorithm [17], which is a general method for simulating systems with continuous degrees of freedom at constant temperature. More about the use of this method in simulating polymers and proteins can be found in Refs. [27–30].

The algorithm is based on the evolution arising from a fictitious Hamiltonian, which in our calculations was taken as

$$H_{\mathrm{MC}}(\pi,\theta) = \frac{1}{2}\sum_{i=2}^{N-1} \pi_i^2 + E(\theta,\xi)/T \tag{3.4}$$

where π_i is an auxiliary momentum variable conjugate to θ_i. The first step in the algorithm is to generate a new set of momenta π_i from the equilibrium distribution $P(\pi_i) \propto \exp(-\frac{1}{2}\pi_i^2)$. Starting from these momenta and the old configuration, the system is evolved through a finite-step approximation of the equations of motion. The configuration generated in such a trajectory is finally subject to an accept-or-reject question. The probability of acceptance in this global Metropolis step is $\min(1, \exp(-\Delta H_{\mathrm{MC}}))$, where ΔH_{MC} is the energy change in the trajectory. This accept-or-reject step removes errors due to the discretization of the equations of motion.

When integrating the discretized equations of motion, two sites of the chain can come so close to each other that the hard r^{-12} repulsion causes numerical instabilities. To avoid this problem, we have used a modified, linear potential for very small monomer-monomer separations. The effect of this on the sampling distribution was negligible, as the original and modified Boltzmann weights were both extremely small in this part of configuration space. In our simulations none of the accepted configurations belonged to the region with modified potential.

The algorithm described has two tunable parameters, namely the step size, ϵ, and the number of steps in each trajectory, n. In our calculations we have used trajectories of length $n\epsilon = 1$ and values of n between 50 and 125.

It has been shown how hybrid Monte Carlo can be used to simulate efficiently various homopolymers with self-repulsion [29]. However, in the application considered here, we expect the efficiency of hybrid Monte Carlo to be similar to that of conventional methods.

The dynamical-parameter method

A simulation method meant for systems with a rough free-energy landscape is the dynamical-parameter method. The basic step in this approach is to create an enlarged configuration space by introducing a new variable k. This variable can take K different values, $k = 1, \ldots, K$, which correspond to different values of the parameters of the model. In this way one tries to circumvent the barriers in the original free-energy landscape.

We use the dynamical-parameter method to simulate the Boltzmann distribution $P_{\xi,T}(\theta) \propto \exp(-E(\theta,\xi)/T)$ for fixed sequence ξ and temperature T. To do that, we assign one sequence, $\xi^{(k)}$, and one temperature, $T^{(k)}$, to each k, taking $\xi^{(k)} = \xi$ and $T^{(k)} = T$ for some k. Using standard methods, we then simulate the joint probability distribution $P(\theta, k) \propto \exp(-g_k - E(\theta, \xi^{(k)})/T^{(k)})$, where the g_k's are constants that will be discussed below. Finally, knowing $P(\theta, k)$, one can obtain the distribution $P_{\xi^{(k_0)},T^{(k_0)}}(\theta)$, for any k_0, by simply restricting $P(\theta, k)$ to the subspace of fixed $k = k_0$.

We study two different algorithms of this type, simulated tempering [21] and the multisequence algorithm. The difference between these algorithms lies in the choice of the sets $\{\xi^{(k)}\}$ and $\{T^{(k)}\}$. In simulated tempering the temperature fluctuates, while the sequence is held fixed, i.e., $\xi^{(1)} = \ldots = \xi^{(K)}$. The multisequence algorithm, on the other hand, uses a set of different sequences and a fixed temperature, i.e., $T^{(1)} = \ldots = T^{(K)}$.

The g_k's are free parameters that determine the weights p_k of the different k values, which are given by

$$p_k = \frac{\exp(-g_k) Z_k}{\sum_{k'=1}^{K} \exp(-g_{k'}) Z_{k'}} \tag{3.5}$$

where $Z_k = Z(T^{(k)}, \xi^{(k)})$. The method is free from systematic errors for any choice of the g_k's, but the efficiency depends strongly on these parameters. In fact, if they are not chosen carefully, it can easily happen that the system gets trapped at a fixed value of k. In our calculations the g_k's have been chosen so as to obtain a roughly uniform k distribution. To that end, we used a number of trial runs, as will be described in Sec. 3.6. If the p_k's corresponding to some choice of the g_k's are known, the uniform distribution is obtained by replacing g_k by $g_k + \ln p_k$.

Next we turn to our simulations of the joint distribution $P(\theta, k)$. For the θ update we have employed the hybrid Monte Carlo method described above. To facilitate comparisons, we used the same values of n and ϵ as in our hybrid Monte Carlo runs. Our updates of k were ordinary Metropolis steps [31]. In our simulated-tempering runs we used an ordered set of allowed temperature values and changed k in steps of ± 1. Each hybrid Monte Carlo trajectory was followed by one Metropolis step in k, and the average acceptance rate was 79% for the k update. In our multisequence runs we chose to work with all possible sequences with a fixed composition, i.e., with fixed numbers of A and B monomers, and the updates of k corresponded to interchanges of the monomers at two randomly chosen sites. Here one update cycle consisted of one hybrid Monte Carlo trajectory followed by 100 Metropolis steps in k. In the final run, after adjusting the g_k's, about 4% of the cycles led to a change of k.

The acceptance rate for the θ update depends on k and was typically around 90%. The fraction of the total computer time spent on k updates was about 12% in our multisequence calculations and much less than 1% in our simulated-

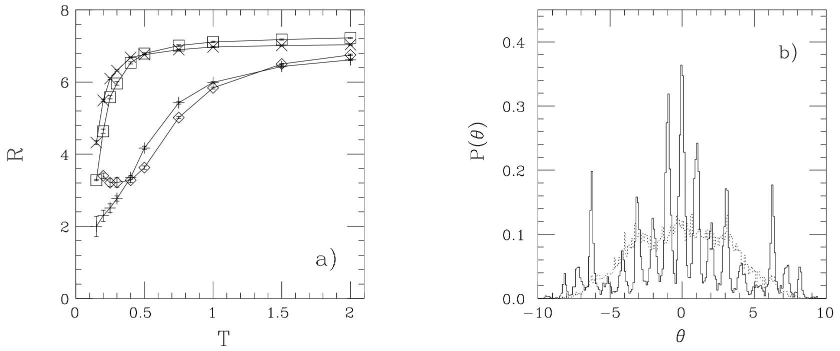

Figure 3.1: a) The temperature dependence of the ratio $R = \langle r_{ee}^2 \rangle / \langle r_{gyr}^2 \rangle$ for A (pluses $N = 8$, diamonds $N = 12$) and B (crosses $N = 8$, squares $N = 12$) homopolymers. b) Histogram of the total bend angle θ for A (solid line) and B (dotted line) homopolymers with $N = 12$ at $T = 0.2$.

tempering calculations. Notice that the k update leaves the energy unchanged in simulated tempering.

In simulated tempering it is possible to avoid a low acceptance rate for the k update by choosing a sufficiently dense (and large) set of $T^{(k)}$'s. The situation is different for the multisequence algorithm, since the sequences $\xi^{(k)}$ are picked from a discrete set. If the acceptance rate is low in a multisequence simulation, it may be useful to introduce auxiliary sequences that interpolate between those of the model studied.

3.4 Homopolymers

We begin our study of the model defined in Sec. 3.2 by investigating the behavior of homopolymers and how that depends on the temperature. The numerical simulations discussed in this and the next section have been performed by using the hybrid Monte Carlo method.

It is well-known that the importance of the r^{-6} terms in the potential varies with the temperature. At high temperature the influence of these terms is weak and the chains are expected to behave essentially as self-avoiding walks. At low temperature, on the other hand, they give rise to a strong effective attraction between monomers of the same type, which leads to globular chains.

To study how the extent of the homopolymers varies with the temperature, we carried out numerical simulations for $N = 8$ and 12. In Fig. 3.1a we show the results for the ratio $R = \langle r_{\mathrm{ee}}^2 \rangle / \langle r_{\mathrm{gyr}}^2 \rangle$, where r_{ee} denotes the end-to-end distance and r_{gyr} the radius of gyration. At high temperature the results are fairly close to $R \approx 7.13$, which is the value expected for two-dimensional self-avoiding walks

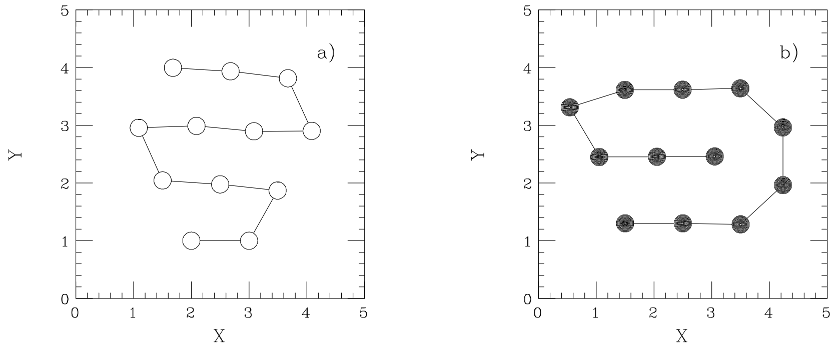

Figure 3.2: Ground-state configurations for the a) A and b) B homopolymers with $N = 12$.

in the limit $N \to \infty$ [32,33]. The corresponding value for ideal and quasi-ideal chains is $R = 6$. Values of R less than 6 indicate that the system is in the globule phase. From the figure we estimate that both the homopolymers are in the globule phase if $T < 0.3$. Our study of general sequences, which will be discussed in the next section, has been carried out using $T = 0.2$.

Even though A and B homopolymers are both compact at $T = 0.2$, they behave slightly differently. This can be seen from the probability distributions of the total bend angle $\theta = \sum_{i=2}^{N-1} \theta_i$ which are shown in Fig. 3.1b. The distribution for the A chain has many narrow peaks, while that for the B chain is dominated by a single, broad peak. The origin of this difference becomes clear when studying the low-lying energy minima. To determine local energy minima we have employed a quenching procedure; during a Monte Carlo simulation at fixed temperature, the system was quenched to zero temperature at regular intervals by using a conjugate gradient method. In Fig. 3.2 we show the minimum energy configurations for $N = 12$. After checking the results by repeated use of simulated annealing, we feel confident that indeed these are ground-state configurations. Also, we checked that all the four symmetry related copies of these states were visited in the simulations.

The structure of the ground state is the result of an interplay between the bend energy and the Lennard-Jones terms. In Fig. 3.2 the sites of the A chain approximately reside on a regular triangular lattice, which is not true for the B chain. This reflects the fact that the bend energy plays a more important role for the B than for the A chain, which has deeper Lennard-Jones potentials. This lattice structure can also be seen in other low-lying states of the A homopolymer, and is responsible for the many distinct peaks in the θ distribution (see Fig. 3.1b).

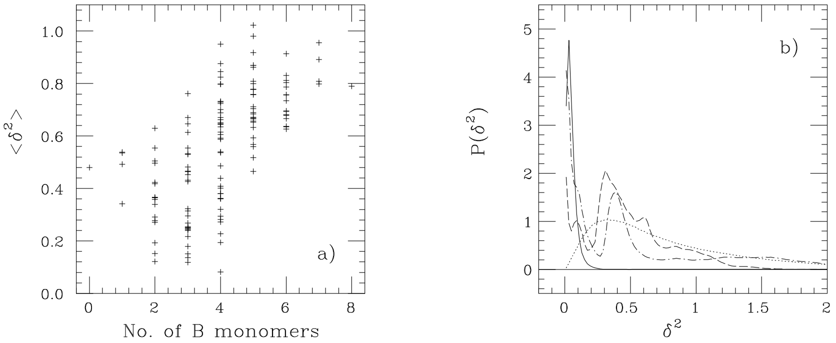

Figure 3.3: a) Average values $\langle\delta^2\rangle$ for the 136 chains with $N = 8$, plotted against the number of B monomers. The statistical errors are roughly of the size of the plot symbols. b) Histograms of δ^2 for the sequences S1 (dashes), S2 (dots), S3 (solid) and S4 (dot-dash) (see Table 3.1). The distribution for S3 has been normalized to 1/3.

3.5 General Sequences

Our study of general sequences has been carried out using $N = 8$ and 10 and a fixed temperature $T = 0.2$. At this temperature the homopolymers have a compact form, but large fluctuations take place in the positions of the individual monomers. As will be shown in this section, there are mixed sequences which, by contrast, exist in a unique state of fairly well-defined shape.

To get a measure of the fluctuations in shape, we introduce the usual mean-square distance between configurations. For two configurations a and b we define

$$\delta^2_{ab} = \min \frac{1}{N}\sum_{i=1}^{N} |\bar{x}_i^{(a)} - \bar{x}_i^{(b)}|^2 \tag{3.6}$$

where $|\bar{x}_i^{(a)} - \bar{x}_i^{(b)}|$ denotes the distance between the sites $\bar{x}_i^{(a)}$ and $\bar{x}_i^{(b)}$ ($\bar{x}_i^{(a)}, \bar{x}_i^{(b)} \in R^2$), and where the minimum is taken over translations, rotations and the discrete symmetries discussed in Sec. 3.2. The probability distribution of δ^2 for fixed temperature and sequence, $P(\delta^2)$, can be obtained numerically and is very informative [34], as it describes the magnitude of the thermodynamically relevant fluctuations. To determine $P(\delta^2)$, we computed δ^2 for every possible pair in a set of 2000 configurations, recorded at intervals of 100 trajectories or more.

We have calculated $P(\delta^2)$ for all the 136 distinguishable chains with $N = 8$ (the number of distinguishable $2n$-mers is $2^{n-1}(2^n + 1)$). In Fig. 3.3a we show the average values $\langle\delta^2\rangle$ of these distributions, plotted against the number of B monomers. This figure shows that $\langle\delta^2\rangle$ varies considerably, and that values occur which are much smaller than those for the homopolymers. In Fig. 3.3b we show $P(\delta^2)$ for the sequences S1–4 in Table 3.1. S1 and S2 are the two homopolymers. S3 has the smallest average value $\langle\delta^2\rangle$ of all the 136 sequences. S4 has the same composition as S3 but a more typical $\langle\delta^2\rangle$. That the shape of $P(\delta^2)$ is irregular for the A homopolymer is due to the lattice structure discussed above. It should become smoother with increasing N. For S3 the distribution is dominated by a single, narrow peak located at low δ^2, which shows that this chain exists in a state of fairly precise shape. There are other $N = 8$ chains which exhibit the same type of behavior. However, the number of such sequences is small, as can be seen from the average values in Fig. 3.3a.

The folding temperature T_f may be defined as the lowest temperature at which $\langle\delta^2\rangle$ takes some suitably chosen value c. For reasonable choices of c, it is clear from Fig. 3.3a that the number of $N = 8$ sequences with $T_f > 0.2$ is small.

To test the size dependence of these results, we have also studied the chains containing ten monomers. First we performed short simulations of all the 528 chains of this length. Even though some of these simulations were too short, the results clearly showed that most of the chains do not exhibit a well-defined shape at this temperature. Notice that too short runs tend to give underestimates of $\langle\delta^2\rangle$. After these preliminary runs, we picked out the four sequences S5–8 in Table 3.1 and performed longer simulations of these. S7 is the $N = 10$ sequence with smallest $\langle\delta^2\rangle$, according to the preliminary runs, and S8 is a sequence with the same composition as S7. In Fig. 3.4a we show $P(\delta^2)$ for S5–8. The result of the longer run confirms that S7 exists in a unique folded state. Since the sequence is asymmetric, there are two copies of this state. From Fig. 3.4b it can be seen that these symmetry related states were both visited in the simulation, but transitions from one state to the other are rare.

	N	sequence	$\langle\delta^2\rangle$	$\langle r^2_{\text{gyr}}\rangle$
S1	8	AAAAAAAA	0.48(2)	1.262(3)
S2		BBBBBBBB	0.79(2)	2.31(2)
S3		AABBBBAA	0.081(6)	1.555(5)
S4		ABAABBBA	0.57(2)	1.919(2)
S5	10	AAAAAAAAAA	0.778(6)	1.633(2)
S6		BBBBBBBBBB	1.164(12)	2.941(11)
S7		AABAABAABA	0.090(8)	1.707(2)
S8		ABAAAABABA	0.510(12)	1.827(2)

Table 3.1: Sequences referred to in the text. The average values were obtained at $T = 0.2$.

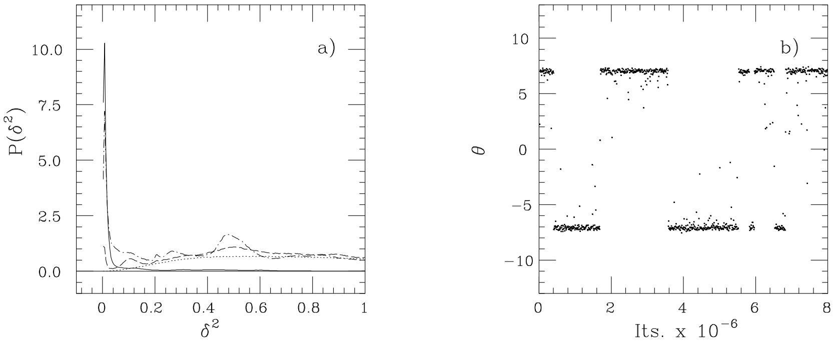

Figure 3.4: a) Histogram of δ^2 for the sequences S5 (dashes), S6 (dots), S7 (solid) and S8 (dot-dash) (see Table 3.1). The distribution for S7 has been normalized to 1/5. b) Monte Carlo evolution of the total bend angle θ for the chain S7. Data have been taken every 10,000 trajectories.

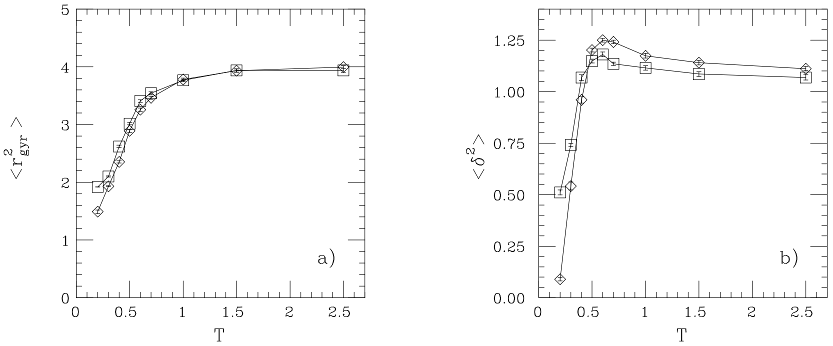

Figure 3.5: Temperature dependence of a) $\langle r^2_{gyr} \rangle$ and b) $\langle \delta^2 \rangle$ for the sequences S7 (diamonds) and S8 (squares) (see Table 3.1).

The results presented so far in this section were obtained at a fixed temperature. Let us also briefly discuss the transition to the folded low-temperature phase for the sequence S7. We compare the behavior of S7 with that of S8, which has a folding temperature lower than 0.2. In Fig. 3.5 we show the temperature dependence of $\langle r^2_{\rm gyr} \rangle$ and $\langle \delta^2 \rangle$. The radius of gyration is similar for the two sequences, which is expected since they have the same composition, and decreases gradually as the temperature is decreased. The temperature dependence of $\langle \delta^2 \rangle$ is more dramatic. We see that the transition from a high-temperature state with large fluctuations in shape to a frozen state of well-defined shape is fairly abrupt for S7.

Finally, let us compare the energy level spectra of the sequences S3 and

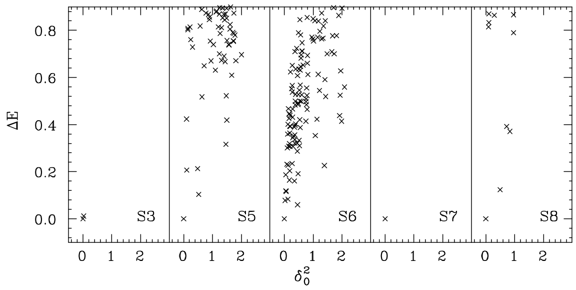

Figure 3.6: Local energy minima for the sequences S3 and S5–8 (see Table 3.1).

S5-8. We have seen that S3 and S7, unlike the other three of these sequences, have a folding temperature higher than 0.2. In Fig. 3.6 we show how the low-lying local energy minima are distributed in the $(\delta_0^2, \Delta E)$ plane, where δ_0^2 is the mean-square distance to the lowest minimum found (cf Eq. 3.6) and ΔE is the energy difference to this minimum. The minima were determined by using the quenching procedure mentioned in Sec. 3.4, and the figure shows all the minima found with $\Delta E < 0.9$. From the figure it can be seen that for S7 there is only with minimum with $\Delta E < 0.9$. For S3 there are two such minima which are structurally very similar. The stability gap defined in Sec. 3.1 is therefore large for both the sequences with high T_f, as expected. The behavior of S3 illustrates that it is important to distinguish between this gap and the energy gap between the two lowest of all minima.

3.6 Tests of the Dynamical-Parameter Algorithms

We have tested performance of the two dynamical-parameter algorithms described in Sec. 3.3 using several different sequences. The results obtained for the different sequences were qualitatively very similar. For clarity we focus here on the results for a specific sequence, namely $AABAAABBAA$, at $T = 0.17$.

In Fig. 3.7a we show the probability distribution of the total bend angle $\theta = \sum_{i=2}^{N-1} \theta_i$, which is dominated by three narrow and well separated peaks. By measuring mean-square deviations (cf Eq. 3.6), we checked that the central peak represents small fluctuations around the minimum energy configuration, which is shown in Fig. 3.7b. The next-lowest energy minimum we found is shown in Fig. 3.7c, and is very similar to the lowest minimum if one neglects the chain structure. However, for the next-lowest minimum we have $\theta \approx 6.1$, which coincides with the position of the rightmost peak in $P(\theta)$. The third peak, at $\theta \approx -6.1$, is related by symmetry to the one at $\theta \approx 6.1$.

When simulating this system by standard methods, the major difficulty

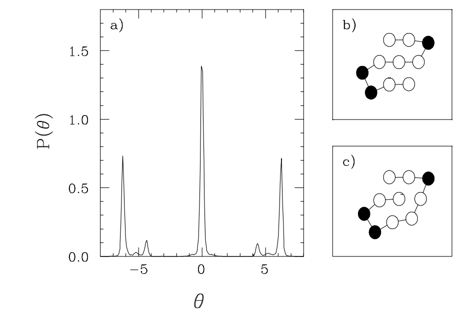

Figure 3.7: a) Histogram of the total bend angle θ. b) Ground-state configuration. c) The next-lowest local energy minimum. Open and filled circles represent A and B monomers, respectively.

is that transitions between the three different regions corresponding to these peaks are strongly suppressed. By the dynamical-parameter method one can greatly improve the frequency of these transitions, as will be seen below. Frequent transitions between the two symmetry related regions can, of course, also be obtained by incorporating global, random flips of the signs of the θ_i's into the update scheme. The important and nontrivial problem is to obtain transitions that connect these two regions with the $\theta \approx 0$ region.

Before running the dynamical-parameter algorithms, one has to determine the parameters g_k. We have done that in slightly different ways for the two algorithms. In our multisequence runs the set of sequences $\xi^{(k)}$ consisted of all the 60 sequences containing seven A and three B monomers. To determine the corresponding g_k's, we carried out four trial runs, covering a total of 6×10^5 update cycles. After each of these runs the g_k's were adjusted, using the measured weights p_k as described in Sec. 3.3. In Fig. 3.8 we show the 60 different weights before and after the tuning procedure. None of the final p_k's is smaller than $1/2 \times 1/60 \approx 0.008$.

In our simulated-tempering runs there were 12 allowed values of the temperature, $0.17 = T^{(1)} < \ldots < T^{(12)} = 0.47$, which were equidistant in $1/T$. Here we first performed short hybrid Monte Carlo simulations, each consisting of 1000 trajectories, to get rough estimates of the average energy at these temperature values, $\langle E\rangle_k$. We then put $g_{12} = 0$ and $g_{k-1} = g_k - \langle E\rangle_k(1/T^{(k-1)} - 1/T^{(k)})$ for $k < 12$. This relation between g_{k-1} and g_k is obtained by requiring that $p_{k-1} = p_k$ and neglecting terms that are of order two or higher in $1/T^{(k-1)} - 1/T^{(k)}$. One of the g_k's can be chosen freely, since only the differences between them are relevant. To fine-tune the g_k's, we finally carried out

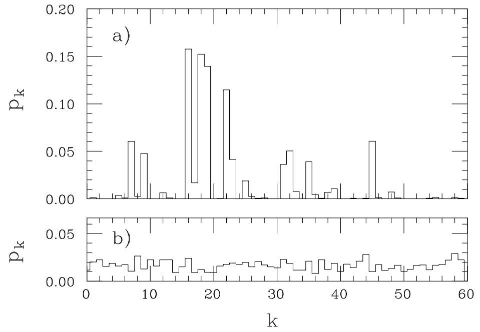

Figure 3.8: The weights p_k in the multisequence calculation: a) for all $g_k = 0$ and b) for our final choice the g_k's.

two simulated-tempering runs, each consisting of 2×10^4 update cycles.

Having chosen the g_k's, we performed one long production run for each of the two algorithms. In Fig. 3.9 we show the evolution of θ in these simulations. Only data corresponding to $T = 0.17$ and sequence $AABAAABBAA$ are shown, and they are plotted against the total number of update cycles. Also shown in Fig. 3.9 is the result of a hybrid Monte Carlo run. The same values of the simulation parameters n and ϵ have been used in the three different runs. The CPU time per iteration is therefore very similar for the hybrid Monte Carlo and simulated-tempering runs, and 12% higher for the multisequence run.

From Fig. 3.9 it is evident that transitions between the three different regions mentioned above are indeed much more frequent in the dynamical-parameter simulations than in the hybrid Monte Carlo simulation. In the hybrid Monte Carlo run there are only six transitions (three of which are difficult to see in the figure) in 8×10^6 trajectories, and one of the three regions is, in fact, never visited. By contrast, every region is visited many times in the two other runs, even though these are shorter. The frequency of transitions to and from the region with $\theta \approx 0$ is respectively 9.4×10^{-4} and 5.7×10^{-5} per update cycle in the simulated-tempering and multisequence runs.

The improved transition frequency is the major advantage of the dynamical-parameter simulations. In addition, these simulations can be used to study several different temperature values or sequences. The disadvantage of the method is that a number of trial runs are needed in order to determine the g_k's. In our simulated-tempering and multisequence calculations we have spent respectively 12% and 30% of the total computer time on trial runs. Let us stress that the total cost of each of these two calculations is still considerably lower than that of our hybrid Monte Carlo run (cf Fig. 3.9).

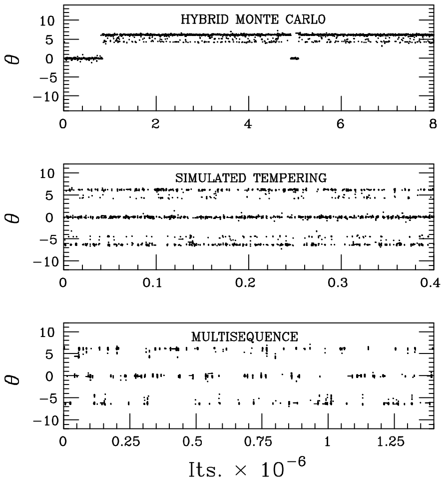

Figure 3.9: Evolution of the total bend angle θ in three different simulations.

3.7 Summary

We have studied numerically the finite-temperature behavior of a simple off-lattice model for protein folding. All possible chains containing eight or ten monomers were studied at a fixed temperature, which was chosen low enough for all the chains to be fairly compact. We showed that most of the chains undergo large fluctuations in shape at this temperature, while a few exist in a unique state of well-defined shape. We determined the low-lying local energy minima for a few different sequences, and found that the stability gaps varied considerably. This suggests that there are important differences in folding temperature, and, therefore, in the ability to fold fast, between different sequences.

Although the relatively short chains studied here exhibit interesting properties, it would, of course, be desirable to extend the study to considerably longer chains. To be able to do that, it is of great interest to find improved Monte Carlo methods. We have in this paper studied two dynamical-parameter algorithms, simulated tempering and the multisequence algorithm, which were found to be much more efficient than conventional simulation methods. The two algorithms are also useful for energy minimization, if combined with a suitable local optimization method.

It would be interesting to compare the performance of these algorithms with that of the multicanonical Monte Carlo algorithm which was mentioned in the introduction. Multicanonical heteropolymer simulations have recently been reported [35–37].

Bibliography

[1] D. L. Stein, Proc. Natl. Acad. Sci. USA **82**, 3670 (1985).

[2] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **84**, 7524 (1987).

[3] T. Garel and H. Orland, Europhys. Lett. **6**, 307 (1988).

[4] E. I. Shakhnovich and A.M. Gutin, Biophys. Chem. **34**, 187 (1989).

[5] G. Parisi and M. Mézard, J. Phys. (Paris) **I1**, 809 (1991).

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

[7] E. Shakhnovich, G. Farztdinov, A. M. Gutin, and M. Karplus, Phys. Rev. Lett. **67**, 1665 (1991).

[8] P. E. Leopold, M. Montal, and J. C. Onuchic, Proc. Natl. Acad. Sci. USA **89**, 8721 (1992).

[9] R. Miller, C. A. Danko, M. J. Masolka, A. C. Balazs, H. S. Chan, and K. A. Dill, J. Chem. Phys. **96**, 768 (1992).

[10] M. Fukugita, D. Lancaster, and M. G. Mitchard, Proc. Natl. Acad. Sci. USA **90**, 6365 (1993).

[11] A. Šali, E. Shakhnovich, and M. Karplus, J. Mol. Biol. **235**, 1614 (1994).

[12] E. I. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994).

[13] H. S. Chan and K. A. Dill, J. Chem. Phys. **100**, 9238 (1994).

[14] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins: Struct. Funct. Genet. **21**, 167 (1995).

[15] F. H. Stillinger, T. Head-Gordon, and C. L. Hirshfeld, Phys. Rev. **E48**, 1469 (1993).

[16] T. Head-Gordon and F. H. Stillinger, Phys. Rev. **E48**, 1502 (1993).

[17] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth, Phys. Lett. **B195**, 216 (1987).

[18] B. A. Berg and T. Neuhaus, Phys. Lett. **B267**, 249 (1991); Phys. Rev. Lett. **68**, 9 (1992).

[19] G. M. Torrie and J. P. Valleau, *J.Comput.Phys.* **23**, 187 (1977).

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

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

[22] W. Kerler and A. Weber, Phys. Rev. **B47**, 11563 (1993).

[23] Z. Liu and B. J. Berne, J. Chem. Phys. **99**, 6071 (1993).

[24] W. Kerler and P. Rehberg, Phys. Rev. **E50**, 4220 (1994).

[25] W. Kerler, C. Rebbi, and A. Weber, Nucl. Phys.**B450**, 452 (1995).

[26] A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov, and P. N. Vorontsov-Velyaminov, J. Chem. Phys. **96**, 1776 (1991).

[27] A. Brass, B. J. Pendleton, Y. Chen, and B. Robson, Biopolymers **33**, 1307 (1993).

[28] D. W. Heermann and L. Yixue, Makromol. Chem., Theory Simul. **2**, 299 (1993).

[29] A. Irbäck, J. Chem. Phys. **101**, 1661 (1994).

[30] B. M. Forrest and U. W. Suter, Molec. Phys. **82**, 393 (1994); J. Chem. Phys. **101**, 2616 (1994).

[31] N. A. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).

[32] J. Des Cloizeaux and G. Jannink, *Polymers in Solution: Their Modelling and Structure* (Clarendon Press, Oxford, 1990).

[33] B. Li, N. Madras and A. Sokal, J. Stat. Phys. **80**, 661 (1995).

[34] G. Iori, E. Marinari, and G.Parisi, J. Phys. **A24**, 5349 (1991).

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

[36] U. H. E. Hansmann and Y. Okamoto, Physica **A212**, 415 (1994).

[37] M.-H. Hao and H. A. Scheraga, J. Phys. Chem. **98**, 4940 (1994).

Paper IV

Identification of Amino Acid Sequences with Good Folding Properties in an Off-Lattice Model

Anders Irbäck, Carsten Peterson and Frank Potthast

Complex Systems Group, Department of Theoretical Physics
University of Lund, Sölvegatan 14A, S-223 62 Lund, Sweden

Physical Review **E 55**, 860-867 (1997)

Abstract:

Folding properties of a two-dimensional toy protein model containing only two amino-acid types, hydrophobic and hydrophilic respectively, are analyzed. An efficient Monte Carlo procedure is employed to ensure that the ground states are found. The thermodynamic properties are found to be strongly sequence dependent in contrast to the kinetic ones. Hence, criteria for good folders are defined entirely in terms of thermodynamic fluctuations. With these criteria sequence patterns that fold well are isolated. For 300 chains with 20 randomly chosen binary residues approximately 10% meet these criteria. Also, an analysis is performed by means of statistical and artificial neural network methods from which it is concluded that the folding properties can be predicted to a certain degree given the binary numbers characterizing the sequences.

4.1 Introduction

The protein folding problem is not merely an engineering task – given sequences of amino acid residues, compute its 3D structure by minimizing an appropriately chosen energy function. Since for such models the energy landscape is often rugged, the resulting 3D configurations may be hard to reach and may furthermore not be thermodynamically stable. It has therefore been argued that only those sequences with "nice" energy landscapes have survived the evolution [1].

A proper understanding of the thermodynamics and kinetics of protein folding requires studies of simplified toy models where the conditions can be somewhat controlled. For the choice of such models two major pathways exist. The currently most popular choice is lattice models with contact term interactions, see e.g. Refs. [1–4]. This approach has the advantage that the ground states are known, but at the same time it has the potential danger that the energy landscape contains artifacts from the discrete description of space. Alternatively, one may use a continuum model with simplified interactions, see e.g. Refs. [5–7], in which case substantial simulations are needed to map out the ground states. On the other hand, in this case properties of the energy landscape should be closer to those of the real world.

The aim of this paper is twofold - to map out the folding properties of the two-dimensional continuum model of [7, 8], hereafter denoted the AB model, and to analyze how the folding properties depend upon the sequences using statistical and state-of-the-art regression methods.

The folding properties of the AB model are investigated with respect to thermodynamics and kinetics given a set of thoroughly simulated sequences. In total 300 sequences of 20 hydrophobic and hydrophilic residues ($+1$ and -1 respectively) are studied using an efficient dynamical-parameter algorithm (see Ref. [7] and references therein). The thermodynamic properties are studied using the mean-square distance δ^2 between different configurations. A low average value $\langle\delta^2\rangle$ signals that the chain exists in a state with well-defined structure. It turns out that $\langle\delta^2\rangle$ exhibits very strong sequence dependence in contrast to the kinetic properties. Based on this we formulate criteria for good folders entirely based on the distribution of δ^2. Using these criteria roughly 10% of the 300 generated and studied sequences survive as good folders.

Next we pose the question of what characterizes the good folders in terms of sequence patterns. Rather than analyzing the "bare" binary sequences of hydrophobicity, we focus on effective variables like random walk representations, block fluctuations and the number of $+1$ embedded between two -1 . This has the virtue that the analysis will capture long range correlations in addition to the local ones. We investigate how $\langle\delta^2\rangle$ depends upon these quantities. This is done using tools of varying sophistication – covariance matrix and feedforward Artificial Neural Networks (ANN). Using ANN we predict $\langle\delta^2\rangle$ given the sequence. With our limited data set the results look very promising. Indeed, the folding properties strongly depend upon sequence patterns. These findings give further evidence of the non-randomness reported in Ref. [10].

Hydrophobicity is widely believed to play a central role in the formation of 3D protein structures. In Ref. [10] the question of whether proteins originate from random sequences of amino acids was addressed by means of a statistical analysis in terms of blocked and random walk values formed by binary hydrophobic assignments of the amino acids along the protein chains. The results, which were based upon proteins in the SWISS-PROT data base [11], convincingly demonstrated that the amino acid sequences in proteins differ from what is expected from random sequences in a statistical significant way. In Ref. [10] also preliminary results from the AB model using the same data as in this work were subject to the same statistical analysis. The interesting observation was made that the AB model sequences that fold well according to low $\langle\delta^2\rangle$-value criteria exhibit similar deviations from randomness as for the functional proteins. The deviations from randomness can be interpreted as originating from anticorrelations in terms of an Ising spin model for the hydrophobicities.

Our studies of AB model are limited to two dimensions in order to be able to analyze many sequences within limited CPU resources. How realistic this approximation is, can of course be questioned. The system may be "stiffer" than a three dimensional one when it comes to swapping monomer positions.

This paper is organized as follows. In Sec. 4.2 we briefly describe the model and generation of sequences. The Monte Carlo method and what is being measured can be found in Sec. 4.3. The thermodynamics and kinetics of the system are described in Sec. 4.4, whereas Sec. 4.5 contains our statistical and ANN analysis. In Sec. 4.6 we briefly review the results from Ref. [10] comparing deviations from non-randomness in the two-dimensional AB model with those of functional proteins. A brief summary can be found in Sec. 4.7.

4.2 The Model

General Formulation

The AB model consists of two kinds of monomers, A and B respectively. These are linked by rigid bonds of unit length to form linear chains living in two dimensions. For an N-mer the sequence of monomers is described by the binary variables $\sigma_1, \ldots, \sigma_N$ and the configuration by the angles $\theta_2, \ldots, \theta_{N-1}$, where θ_i denotes the bend angle at site i and is taken to satisfy $|\theta_i| \leq \pi$. The energy function is given by

$$E(\theta, \sigma) = \sum_{i=2}^{N-1} E_1(\theta_i) + \sum_{i=1}^{N-2} \sum_{j=i+2}^{N} E_2(r_{ij}, \sigma_i, \sigma_j) \tag{4.1}$$

where

$$\begin{aligned} E_1(\theta_i) &= \frac{1}{4}(1 - \cos\theta_i) \\ E_2(r_{ij}, \sigma_i, \sigma_j) &= 4(r_{ij}^{-12} - C(\sigma_i, \sigma_j) r_{ij}^{-6}) \end{aligned} \tag{4.2}$$

and $r_{ij} = r_{ij}(\theta_{i+1}, \ldots, \theta_{j-1})$ denote the distances between sites i and j. The term $E_1(\theta_i)$ favors alignment of three successive sites; $i-1$, i and $i+1$. The non-bonded interactions E_2 are Lennard-Jones potentials with a species-dependent coefficient $C(\sigma_i, \sigma_j)$, which is taken to be 1 for an AA pair (strong attraction), 1/2 for a BB pair (weak attraction) and -1/2 for an AB pair (repulsion). Consequently, there is an energetic preference for separation between the two kinds of monomers. In fact, it was demonstrated in [8] that ground-state configurations tend to have a core consisting mainly of A monomers, which shows that A and B monomers behave as hydrophobic and polar residues respectively. The behavior of the model at finite temperature T is defined by the partition function

$$Z(T, \sigma) = \int \left[\prod_{i=2}^{N-1} d\theta_i\right] \exp(-E(\theta, \sigma)/T) \ . \tag{4.3}$$

The Sequences

In total 300 sequences were drawn randomly from the set of all distinguishable chains with 14 A and 6 B monomers. Our motivation for this somewhat arbitrary choice of A/B ratio is that there are thermodynamically stable structures at relatively high temperatures for this ratio [7]. This set contains 19980 sequences, whereas the total number of sequences with the same composition is 38760. Among the 300 sequences 4 are symmetric. The 300 distinguishable sequences can be taken as 300 independent sequences drawn from the distribution of all sequences with double weight for every asymmetric sequence.

4.3 Simulations

Methods

We have performed numerical simulations of both the thermodynamic and kinetic behavior of the 300 randomly selected sequences. At low temperature the system is in a folded phase with high free-energy barriers, which makes conventional simulation methods very time-consuming. As in Ref. [7], we therefore employ the dynamical-parameter method for the thermodynamic simulations. In this approach one tries to accelerate the simulation by letting some parameter of the model become a dynamical variable, which takes values ranging over a definite set. In the present work we have taken the temperature as a dynamical parameter ("simulated tempering" [9]) which means that we simulate the joint probability distribution

$$P(\theta, k) \propto \exp(-g_k - E(\theta, \sigma)/T_k) \tag{4.4}$$

where T_k, $k = 1, \ldots, K$, are the allowed values of the temperature. The g_k's are tunable parameters which must be chosen carefully for each sequence. The determination of these parameters has been carried out by the same methods as in Ref. [7].

The joint distribution $P(\theta, k)$ is simulated by using an ordinary Metropolis step [12] in k and a hybrid Monte Carlo update [13] of θ. The hybrid Monte Carlo update is based on the evolution arising from the fictitious Hamiltonian

$$H_{\mathrm{MC}}(\pi, \theta) = \frac{1}{2} \sum_{i=2}^{N-1} \pi_i^2 + E(\theta, \sigma)/T \tag{4.5}$$

where π_i is an auxiliary momentum variable conjugate to θ_i. The first step in the update is to generate a new set of momenta π_i from the equilibrium distribution $P(\pi_i) \propto \exp(-\pi_i^2/2)$. Starting from these momenta and the old configuration, the system is evolved through a finite-step approximation of the equations of motion. The configuration generated in such a trajectory is finally subject to an accept-or-reject question, which removes errors due to the discretization of the equations of motion. The hybrid Monte Carlo update has two tunable parameters, the step size ϵ and the number of steps n in each trajectory.

The dynamical-parameter method greatly improves the frequency of transitions between different free-energy valleys, as compared to plain hybrid Monte Carlo. In Ref. [7] a speed up factor of almost 10^3 was observed for system size $N = 10$. For $N = 20$ we expect the gain to be even larger.

In our simulations we used a set of $K = 13$ allowed temperature values, which were equidistant in $1/T$ and ranging from 0.15 to 0.60. Each hybrid Monte Carlo trajectory was followed by one Metropolis step in k. The step size parameter ϵ was taken to vary with T, from 0.0025 at $T = 0.15$ to 0.005 at $T = 0.6$, while $n = 100$ was held fixed. For the typical sequence the average acceptance rate was around 95% for the θ update and 65% for the k update. For each sequence a total of 440000 update cycles were carried out, which requires around 4 CPU hours on a DEC Alpha 2000.

In order to study the kinetic behavior of the model, we have performed hybrid Monte Carlo simulations at different fixed values of T. Starting from random coils, we study the rate of the subsequent relaxation process. While our hybrid Monte Carlo dynamics is certainly different from any real dynamics, it is still a small-step evolution, so the system has to pass through the free-energy barriers. Hence, we expect relaxation times obtained in this way to reflect the actual kinetic properties of the system.

Our simulations of the kinetics have been performed for five different T. The trajectory length $n\epsilon = 0.25$ was held fixed, and the average acceptance rate was typically 85% or higher.

Measurements

In our thermodynamic simulations the main goal is to find out whether the system exists in a state with well-defined shape. To address this question, we introduce the usual mean-square distance between configurations. For two configurations a and b we define

$$\delta_{ab}^2 = \min \frac{1}{N} \sum_{i=1}^{N} |\bar{x}_i^{(a)} - \bar{x}_i^{(b)}|^2 \tag{4.6}$$

where $|\bar{x}_i^{(a)} - \bar{x}_i^{(b)}|$ denotes the distance between the sites $\bar{x}_i^{(a)}$ and $\bar{x}_i^{(b)}$ $(\bar{x}_i^{(a)}, \bar{x}_i^{(b)} \in R^2)$, and where the minimum is taken over translations, rotations, reflections and orientations. The probability distribution of δ^2 for fixed temperature T and sequence σ is given by

$$P(\delta^2) = \frac{1}{Z(T,\sigma)^2} \int d\theta^{(a)} d\theta^{(b)} \delta(\delta^2 - \delta_{ab}^2) \mathrm{e}^{-E(\theta^{(a)},\sigma)/T} \mathrm{e}^{-E(\theta^{(b)},\sigma)/T}, \tag{4.7}$$

where $\delta(..)$ denotes the Dirac delta function. $P(\delta^2)$ is a very useful quantity [5] for describing the magnitude of the relevant thermodynamic fluctuations, and can be determined numerically. In Fig. 4.1 we show three examples of δ^2-distribution at two different temperatures. In what follows we will also frequently use the mean of $P(\delta^2)$,

$$\langle \delta^2 \rangle = \int d\delta'^2 P(\delta'^2) \delta'^2 \tag{4.8}$$

The energy level spectrum can be studied by using a quenching procedure, where whenever the lowest allowed temperature value is visited, the system is quenched to zero temperature by means of a conjugate gradient minimization. With this method the ground states are found for most of the sequences. One reason for believing this is that the two or four symmetry-related copies of

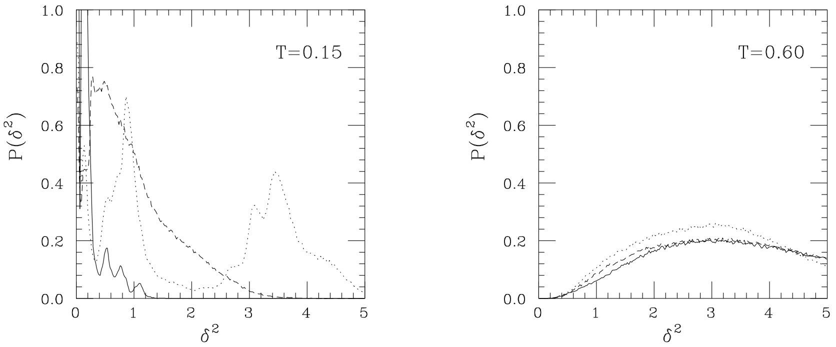

Figure 4.1: $P(\delta^2)$ for $T = 0.15$ and $T = 0.60$ respectively for the sequences in Table 4.1; 81 (solid line), 10 (dashed line) and 50 (dotted line). For $T = 0.15$ the distribution for sequence 81 is dominated by two narrow peaks at small δ^2 which extend outside the figure with a maximum value of around 20.

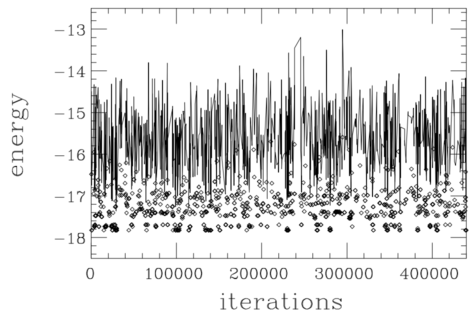

Figure 4.2: Evolution of the quenched (diamonds) and unquenched (line) energies in the simulation of sequence 10 (see Table 4.1). Measurements were taken every 10 iterations. Shown are the data corresponding to the lowest allowed temperature.

the lowest-lying minimum were all visited in the simulations. In Fig. 4.2 we show the evolution of the quenched and unquenched energies in one of the simulations.

4.4 Thermodynamic and Kinetic Properties

In this section we present the results of our thermodynamic and kinetic simulations. Based on these results, we then formulate criteria for good folding sequences.

81	$AAABAAABAAABBAABBAAA$
10	$ABAAABBAABAAAAAAAABB$
50	$BABAAAAAAABAAAABAABB$

Table 4.1: Examples of three sequences.

Thermodynamics

The thermodynamic behavior of the AB model has been studied previously for chain length $N = 8$ and 10 [7]. This study showed that whether or not the chain exhibits a well-defined structure depends strongly on the sequence, at fixed temperature. Using the dynamical-parameter method, we have now been able to extend these calculations to $N = 20$. The results obtained closely resemble those for $N = 8$ and 10 [7]; in particular, they show that the sequence dependence of the thermodynamic behavior remains strong for $N = 20$.

In order to illustrate this, we show in Figs. 4.1 and 4.3 results for the three sequences in Table 4.1. In Fig. 4.1 the δ^2 distributions are shown for the lowest and highest temperature studied, $T = 0.15$ and 0.60. At $T = 0.60$ the distributions are similar, and show that the fluctuations in shape are large. At $T = 0.15$ $P(\delta^2)$ is, by contrast, strongly sequence dependent. We see that one chain exists in a state with very well-defined shape, while the other two still undergo large fluctuations. The differences in $P(\delta^2)$ are clearly reflected in the mean value $\langle\delta^2\rangle$, which is shown as a function of T in Fig. 4.3b.

Although these three chains behave in very different ways, they have similar extension. In fact, from Fig. 4.3a it can be seen that the differences in radius of gyration are 10% or smaller at all the temperatures studied. Also, we note that the radius of gyration decreases gradually with decreasing temperature. No abrupt changes can be seen.

Kinetics

Next we study the time needed by the system to find the minimum energy configuration. Using hybrid Monte Carlo dynamics, we monitor the mean-square distance δ_0^2 to this configuration (see Eq. 4.6). The simulations are started from random coils, and as a criterion of successful folding we use the condition $\delta_0^2 < 0.3$. At low temperatures the folding time fluctuates widely, which makes the average folding time difficult to measure. For this reason we have chosen to measure the probability of successful folding within a given time [1, 15]. Following Ref. [1], this quantity will be called the foldicity. In our simulations the maximum allowed folding time is set to 5000 trajectories. For each of the 300 sequences we studied five different temperatures, $T = 0.15, 0.18, 0.24, 0.34$ and 0.60. For each T we carried out 25 simulations for

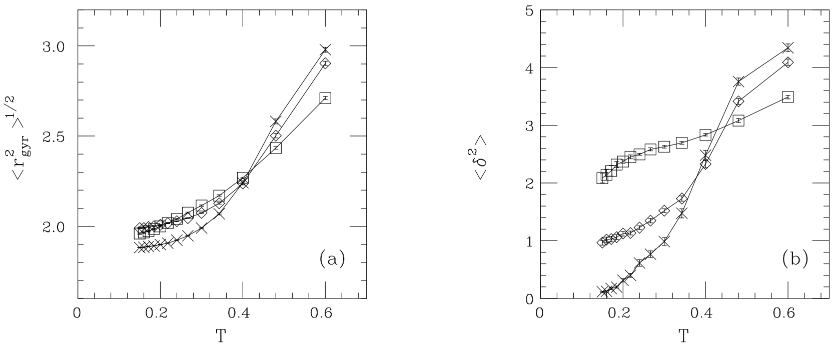

Figure 4.3: The temperature dependence of **(a)** the radius of gyration $\langle r_{gyr}^2 \rangle^{1/2}$ and **(b)** $\langle \delta^2 \rangle$ for the sequences in Table 4.1: 81 ($\times$), 10 ($\diamond$) and 50 ($\Box$).

different initial configurations.

The foldicity is expected to be low both at high and low temperature. At low temperature the suppression is due to the ruggedness of the free-energy landscape. At high temperature folding is slow because the search is random.

For a majority of the sequences studied we find that the foldicity exhibits a peak in the interval $0.15 \leq T \leq 0.60$. In order to get precise estimates of the height and location of the peak, one would clearly need more data points. However, the available data demonstrate that the position of the peak is fairly sequence independent. In fact, for 243 of the 300 sequences we obtained a higher foldicity at $T = 0.34$ than at the other four temperatures. Also, we note that all sequences have a foldicity of 12% or higher at $T = 0.34$, as can be seen from Fig. 4.4. Therefore, it appears that the kinetic behavior has a relatively weak sequence dependence.

In order to illustrate the implications of this, we have in Fig. 4.5 plotted the foldicity in two different ways. In Fig. 4.5a foldicity is plotted against temperature, and in Fig. 4.5b it is plotted against $\langle \delta^2 \rangle$. Again, we use the three sequences in Table 4.1 as examples. From a) it can be seen that, at a given temperature, the foldicity is roughly similar for the three sequences. Nevertheless, it follows from b) that their folding properties are very different; sequence 10 has a well-defined shape at the point where the system freezes, which is not true for the other two sequences.

Folding Criteria

A good folder is a sequence which, at some value of the parameter T, exists in a unique and kinetically accessible state with well-defined shape. One way to find out whether or not a given sequence meets this requirement is to first locate the lowest temperature at which folding is fast, $\tilde{T}$, and then study the thermo-

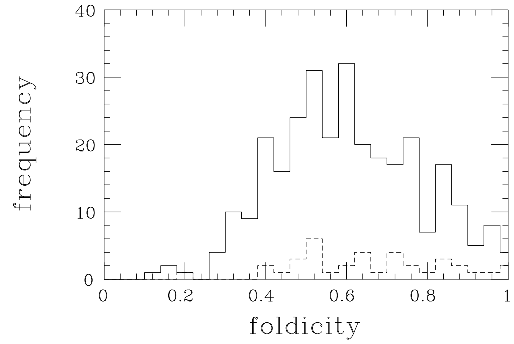

Figure 4.4: Histogram of the foldicity at $T = 0.34$ for the full set of 300 sequences (solid line) and for a subset of 37 sequences that satisfy the thermodynamic stability condition in Eq. 4.9 (dashed line).

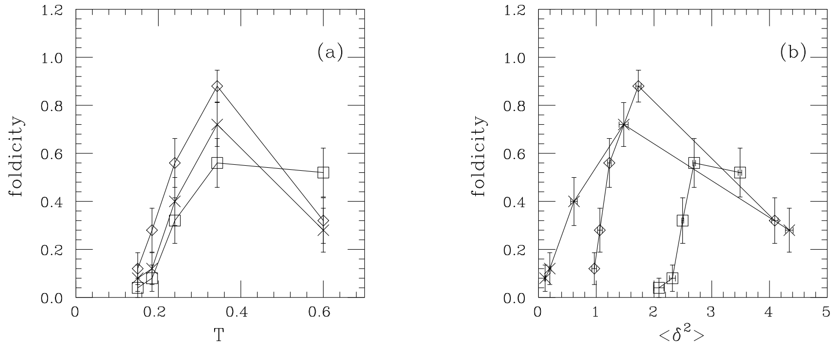

Figure 4.5: Foldicity against **(a)** T and **(b)** $\langle\delta^2\rangle$ for the sequences 81 ($\times$), 10 ($\diamond$) and 50 ($\Box$), see Table 4.1.

dynamic behavior at this temperature. If the sequence is found to exhibit a well-defined structure at $\tilde{T}$, then it is a good folder. Alternatively, this can be formulated in terms of the folding temperature T_f, defined as the temperature where the dominance of a single state sets in; a good folder is a sequence with $T_f > \tilde{T}$.

In our determination of good folders we take the kinetic quantity $\tilde{T}$ to be same for all sequences. This means that our classification is entirely determined by the thermodynamic behavior at a fixed temperature, which we take as $\tilde{T} = 0.15$ (cf Fig. 4.5a). Our motivation for this simplifying approximation is that the kinetic behavior has a relatively weak sequence dependence, as discussed in the previous subsection.

With this approximation, a natural criterion for good folders would be to require that $\langle\delta^2\rangle < \tilde{\delta}^2$ for some $\tilde{\delta}$. Such a cut is appropriate for most sequences. However, some care has to be exercised since one might encounter situations where $P(\delta^2)$ has a tiny but distant outlier bump, which can make $\langle\delta^2\rangle$ large even though the system spends a large fraction of the time very near one particular configuration. Taking this into account we define a sequence to be a good folder if

$$\langle\delta^2\rangle < \tilde{\delta}^2 \qquad \text{or} \qquad P(\delta^2 < 0.1) = \int_0^{0.1} d\delta'^2 P(\delta'^2) > \tilde{P} \tag{4.9}$$

with $\tilde{\delta}^2 = 0.3$ and $\tilde{P} = 0.35$.

With this choice of parameters, we find that 24 of our sequences satisfy $\langle\delta^2\rangle < \tilde{\delta}^2$ whereas 30 satisfy $P(\delta^2 < 0.1) > \tilde{P}$. Our set of good folders, which satisfy one or both of these conditions, contains 37 of the 300 sequences, or 12%.

The precise number of sequences classified as good folders depends, of course, on the choice of $\tilde{\delta}$ and $\tilde{P}$, which to some extent is arbitrary. However, it should be stressed that for most of the sequences the classification is unambigiuous in the sense that it is insensitive to small changes of $\tilde{\delta}$ and $\tilde{P}$. Furthermore, we note that the foldicity distribution for good folders is similar to that for all sequences, as can be seen from Fig. 4.4. If these distributions had been different, the use of a sequence independent $\tilde{T}$ would have been unjustified.

4.5 Sequence Characteristics of Good Folders

In this section we analyze how $\langle\delta^2\rangle$ depends upon the binary patterns of the sequences. This is done in two steps. First we make a statistical analysis in terms of correlations. Second, we employ a feedforward Artificial Neural Network (ANN) to predict $\langle\delta^2\rangle$ given the sequence as input.

Choice of Variables

It turns out to be enlightening and profitable to transform the original binary patterns into more global variables prior to performing the statistical and ANN analysis. The following variables are formed:

Random Walk Representations – r_n. In order to build in some long range correlation properties we consider random walk representations

$$r_n = \sum_{i=1}^{n} \sigma_i \qquad n = 1, \ldots, N \tag{4.10}$$

Block Fluctuations – $\psi_i^{(s)}$ **and** $\psi^{(s)}$**.** For a block size s we define the variables [10]

$$\sigma_i^{(s)} = \sum_{j=1}^{s} \sigma_{(i-1)s+j} = r_{is} - r_{(i-1)s} \qquad i = 1, \ldots, N/s \tag{4.11}$$

In order to efficiently capture the fluctuations of the block variables we introduce the normalized variables

$$\psi_i^{(s)} = \frac{1}{K}\Big(\sigma_i^{(s)} - \frac{s}{N}\sum_{j=1}^{N/s}\sigma_j^{(s)}\Big)^2 \qquad i = 1, \ldots, N/s \tag{4.12}$$

and the (normalized) mean-square fluctuation of the block variables

$$\psi^{(s)} = \frac{s}{N}\sum_{i=1}^{N/s}\psi_i^{(s)} \tag{4.13}$$

where the constant K can be found in Ref. [10]

Number of Internal A**'s –** N_{IA}**.** This is the number of σ=+1 surrounded on both sides by σ=+1. For boundary residues, a single adjacent σ=+1 is sufficient for giving a count. The reason for this choice of variable is that in the homopolymer limit, $AAAA...A$, the energy landscape is degenerate and, hence, the fluctuations are large. Therefore, one expects long stretches of A's (or B's) to be rare in good folders, and that N_{IA} tends to be low for such sequences.

Number of Clumps – N_C**.** This quantity is defined as the number of clumps of σ=±1. The reason for including this variable is similar to what was argued for N_{IA} above; it seems natural to expect a high N_C for good folders.

We expect these preprocessed variables, which of course are not independent of each other, to shed more light on the structure than the "raw" σ=±1 ones, when it comes to relate the sequences to $\langle\delta^2\rangle$.

In this section we will make use of all patterns, even those generated by symmetry giving rise to the same $\langle\delta^2\rangle$. As mentioned in Sec. 4.2, 4 of the 300 original sequences are symmetric, which means that we have a total of 596 patterns at our disposal.

Correlations

To what extent is $\langle\delta^2\rangle$ correlated with r_n, $\psi_i^{(4)}$, $\psi^{(4)}$, N_{IA} and N_C? In Fig. 4.6 the correlation between $\langle\delta^2\rangle$ and these variables are shown.

As can be seen from Fig. 4.6 the correlations are substantial. It should also be mentioned that the same correlation patterns emerge when reducing the data set by a factor four. The symmetry for r_n around n=10 is inherent

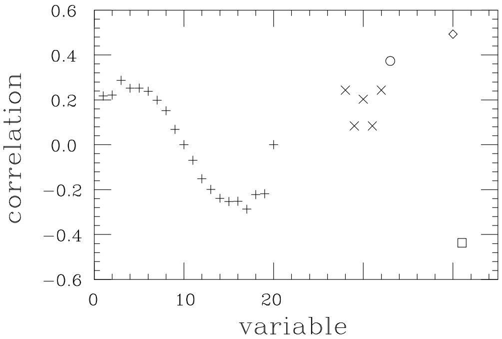

Figure 4.6: Correlations of $\langle\delta^2\rangle$ against r_n [n=1,…,20] (+), $\psi_i^{(4)}$ [i=1,…,5] ($\times$), $\psi^{(4)}$ ($\circ$), N_{IA} ($\diamond$) and N_C ($\square$).

from the way we have generated the data, and leaves us with nine independent measurements, corresponding to e.g. r_1 through r_9. The following general observations can be made:

- $\langle\delta^2\rangle$ and r_n exhibit sizable positive correlations near the endpoints, implying that proteins with many A's near the ends do not fold well.
- $\langle\delta^2\rangle$ and the block fluctuations, $\psi_i^{(4)}$ and $\psi^{(4)}$, also show strong positive correlations. This is consistent with what was observed for real proteins with limited net hydrophobicity in Ref. [10], where these variables are anticorrelated as compared to what is expected from random hydrophobicity distributions.
- The strongest positive correlation is between $\langle\delta^2\rangle$ and the number of internal A's (N_{IA}). This is in line with what is expected according to the motivation when introducing N_{IA} above.
- Related to positive correlation of N_{IA} is the strong anticorrelation between $\langle\delta^2\rangle$ and the number of clumps N_C.

Artificial Neural Networks

Given the substantial correlations between $\langle\delta^2\rangle$ and the various quantities formed out of the sequence patterns, it should be possible to make a regression model. If enough data are available for an efficient and reliable fit one should be able to predict the folding properties given a binary sequence. The state-of-the-art technique for such modeling are feedforward ANN (see e.g. Ref. [14]), which will be used here. This method has the advantage of capturing non-linear dependencies in a generic way in contrast to standard linear regression

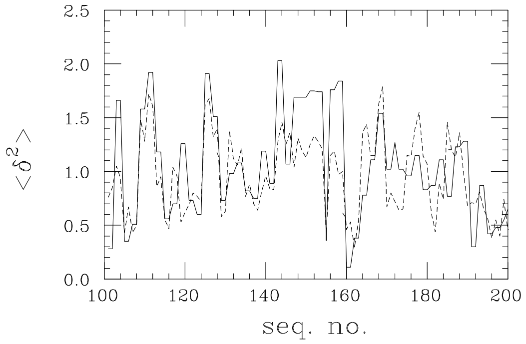

Figure 4.7: Comparison of predicted and true values, $\langle\hat{\delta^2}\rangle$ and $\langle\delta^2\rangle$ respectively, using K=3 for sequences 100 to 200. In order to guide the eye, lines connecting the points are drawn. The full line represents $\langle\delta^2\rangle$ and the dashed line $\langle\hat{\delta^2}\rangle$.

approaches. In our case the feedforward ANN consists of an input layer representing the variables defined in Sec. 4.5 above, an output unit for $\langle\delta^2\rangle$ and a set of hidden units in order to model nonlinearities. The weights connecting the nodes are the parameters of the system. In order to avoid overfitting, the number of weights (parameters) should be less than the number of "training" patterns. Also, some of the patterns should be set aside for "testing". Using all the quantities (only r_1 through r_9 due to symmetries) defined in Sec. 4.5 with 5 hidden units implies $17 \cdot 5 + 5 = 90$ parameters.

The network was trained using the `JETNET 3.0` package [16] with an initial learning rate η_0=0.5, which decreases according to $\eta_k = 0.998\eta_{k-1}$ and momentum α=0.7. In order to obtain as reliable performance as possible, the K-fold cross validation procedure was used, where the data set was randomly divided into K equal parts. Each of the the K different parts were once used as test sets, while the remaining $K-1$ sets were used for training. In our problem sets with K=2, 3, and 4 were used. In Fig. 4.7 the resulting prediction, $\langle\hat{\delta^2}\rangle$, is compared with the true values, $\langle\delta^2\rangle$, for a representative subset of the sequences using K=3 is shown. As can be seen from Fig. 4.7 the predicted values follow the true values pretty well even though sometimes quantitative agreement is lacking. The degree of success can be quantified using the standard error of prediction,

$$R = \frac{1}{\sigma^2}\frac{1}{N}\sum(\langle\hat{\delta^2}\rangle - \langle\delta^2\rangle)^2 \tag{4.14}$$

where σ is the standard deviation of $\langle\delta^2\rangle$ for all 596 patterns. For K=3 and 4 we find R=0.69±0.02. This value deteriorates somewhat if a smaller subset of the data is used, which indicates that with a larger available data set the performance is likely to improve. In the long term one would like to predict a classification (folder/non-folder) but the limited folding data at our disposal do not yet allow for that.

4.6 Non-Randomness and Folding

In Sec. 4.5 we studied the difference between folding and nonfolding sequences in the AB model. This could be done in a controlled way due to the fact that we have unbiased samples of folding and nonfolding sequences at our disposal. To assess the applicability of our methods to real amino acid sequences is more difficult, and we shall not deal here with the problem of predicting the behavior of individual amino acid sequences. However, we would like to stress that the binary hydrophobicity patterns corresponding to a large class of functional proteins do exhibit interesting similarities with folding sequences in the AB model, as was demonstrated in Ref. [10]. The proteins sequences considered in Ref. [10] have a limited net hydrophobicity

$$X = \frac{N_+ - Np}{\sqrt{Np(1-p)}} \tag{4.15}$$

where N_+ is the number of hydrophobic residues, N is the total number of residues, and p is the average of N_+/N over all sequences. In the following, we consider sequences with $|X| < 0.5$.

In order to examine these two groups of binary strings, it is instructive to consider the scaling of the fluctuation variable $\psi^{(s)}$ with block size s [10]. In Fig. 4.8 we show results for this quantity. For both groups, it can be seen that $\psi^{(s)}$ grows significantly slower with s than for random sequences. This behavior implies that the sequence variables σ_i exhibit anticorrelations. A simple way to see that is to consider the one-dimensional Ising model, where each configuration is given a statistical weight

$$P \propto \exp\Big(K \sum_{i=1}^{N-1} \sigma_i \sigma_{i+1}\Big) \tag{4.16}$$

For $K = 0$ the σ_i's are completely random, while $K > 0$ and $K < 0$ correspond to ferromagnetic and antiferromagnetic behavior respectively. It turns out that one obtains results similar to those seen in Figs. 4.8a and b for $K = -0.25$.

4.7 Summary

Fairly detailed studies of the folding properties of a toy model for proteins, containing effective hydrophobicity interactions only, have been performed. The thermodynamic quantity $\langle \delta^2 \rangle$ exhibits a strong sequence dependence in contrast

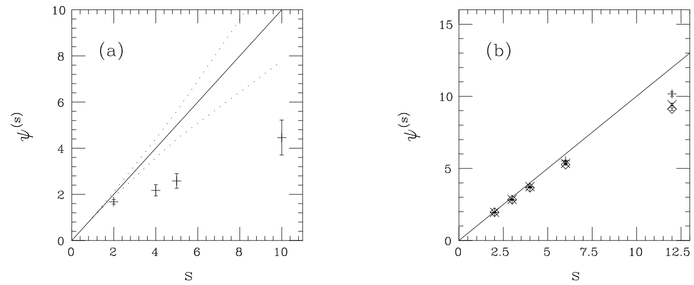

Figure 4.8: Mean-square fluctuation of the block variables, $\psi^{(s)}$, against block size s. **(a)** Good folding sequences in the AB model. Also shown are the mean s (full line) and the $s \pm \sigma$ band (bounded by dotted lines) for random sequences [10]. **(b)** Functional proteins for $|X| < 0.5$, $50 < N \leq 150$ (+; 2457 qualifying proteins), $150 < N \leq 250$ ($\times$; 2228), and $250 < N \leq 350$ ($\diamond$; 1642). All data are from the SWISS-PROT data base [11]. The straight line is the result for random sequences [10].

to foldicity, which measure the kinetic properties. Hence criteria for chains with good folding properties have been devised solely in terms of $\langle\delta^2\rangle$. With these criteria, approximately 10% of the 300 generated sequences are classified as good folders. These conclusions have been possible due to extensive searches using an efficient dynamical-parameter algorithm, which with very large probability visits the ground states. A similar fraction of good folders was obtained in the lattice model study of Ref. [1], where 30 out of 200 randomly chosen sequences were classified as good folders.

Our conclusion that the thermodynamic properties are more important for the classification of folders/non-folders than the kinetic ones is in line with the lattice model results of Ref. [4]. These authors introduced a kinetic glass transition temperature, which was found to be nearly sequence independent. Although the kinetic studies of Ref. [4] are more elaborate than ours, it should be stressed, however, that our data provide no justification for introducing this transition temperature, which may indicate a difference between the models.

Using statistical and artificial neural network methods, substantial functional dependencies between sequence patterns and $\langle\delta^2\rangle$ are revealed. With larger statistics it should be possible given the sequence pattern to predict $\langle\delta^2\rangle$ within a reasonable confidence level; in other words to predict whether a given sequence folds or not.

Related to this strong correlation is the observed pattern of the non-randomness for the folders, which show similar qualitative behavior with what is observed for real proteins [10].

Bibliography

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

[2] C.J. Camacho and D. Thirumalai, "Kinetics and Thermodynamics of Folding in Model Proteins", *Proc. Natl. Acad. Sci. USA* **90**, 6369 (1993).

[3] H.S. Chan and K.A. Dill, "Transition States and Folding Dynamics of Proteins and Heteropolymers", *J. Chem. Phys.* **100**, 9238 (1994).

[4] J.D. Bryngelson, J.N. Onuchic, N.D. Socci and P.G. Wolynes, "Funnels, Pathways, and the Energy Landscape of Protein Folding", *Proteins* **21**, 167 (1995).

[5] G. Iori, E. Marinari and G. Parisi, "Random Self-Interacting Chains: A Mechanism for Protein Folding", *Journal of Physics* **A24**, 5349 (1991).

[6] M. Fukugita, D. Lancaster and M.G. Mitchard, "Kinetmatics and Thermodynamics of a Folding Heteropolymer", *Proc. Natl. Acad. Sci. USA* **90**, 6365 (1992).

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

[8] F.H Stillinger, T. Head-Gordon and C.L. Hirschfeld, "Toy Model for Protein Folding", *Phys. Rev.* **E48**, 1469 (1993).

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

[10] A. Irbäck, C. Peterson and F. Potthast, "Evidence for Non-Random Hydrophobicity Structures in Protein Chains", *Proc. Natl. Acad. Sci. USA* **93**, 9533 (1996).

[11] A. Bairoch and B. Boeckmann, “The SWISS-PROT protein sequence data bank: release 31”, *Nucleic Acids Res.* **22**, 3578 (1994).

[12] N.A. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A. Teller and E. Teller, “Equation of State Calculations by Fast Computing Machines”, *J. Chem. Phys.* **21**, 1087 (1953).

[13] S. Duane, A.D. Kennedy, B.J. Pendleton and D. Roweth, “Hybrid Monte Carlo”, *Phys. Lett.* **B 195**, 216 (1987).

[14] J. Hertz, A. Krogh and R.G. Palmer, *Introduction to the Theory of Neural Computation*, Addison-Wesley, Redwood City, Ca. (1991).

[15] N.D. Socci and J. N. Onuchic, ”Folding Kinetics of Proteinlike Heteropolymers”, *J. Chem. Phys.* , **101** 1519 (1994).

[16] C. Peterson, T. Rögnvaldsson and L. Lönnblad, “JETNET 3.0 - A Versatile Artificial Neural Network Package”, *Computer Physics Communications* **81**, 1 (1994)

Paper V

V

Local Interactions and Protein Folding: A 3D Off-Lattice Approach

Anders Irbäck, Carsten Peterson,
Frank Potthast and Ola Sommelius

Complex Systems Group, Department of Theoretical Physics
University of Lund, Sölvegatan 14A, S-223 62 Lund, Sweden

Journal of Chemical Physics **107**, 273-282 (1997)

Abstract:

The thermodynamic behavior of a three-dimensional off-lattice model for protein folding is probed. The model has only two types of residues, hydrophobic and hydrophilic. In absence of local interactions, native structure formation does not occur for the temperatures considered. By including sequence independent local interactions, which qualitatively reproduce local properties of functional proteins, the dominance of a native state for many sequences is observed. As in lattice model approaches, folding takes place by gradual compactification, followed by a sequence dependent folding transition. Our results differ from lattice approaches in that bimodal energy distributions are not observed and that high folding temperatures are accompanied by relatively low temperatures for the peak of the specific heat. Also, in contrast to earlier studies using lattice models, our results convincingly demonstrate that one does not need more than two types of residues to generate sequences with good thermodynamic folding properties in three dimensions.

5.1 Introduction

In the process of unveiling central issues in the thermodynamics and kinetics of protein folding, simplified models where the amino acid residues constitute the basic entities seem to exhibit many non-trivial and interesting properties [1]. In particular lattice model approaches with contact interactions only have become increasingly popular. The lattice and contact term approximations may seem drastic. Nevertheless, it turns out that such simple models are able to catch non-trivial aspects of the folding problem. This is interesting and encouraging. However, it does not imply that the approximations involved are understood. It is therefore crucial to pursue the study of alternative models, e.g. exploring off-lattice models with Lennard-Jones interactions.

The major advantage of lattice models is computational efficiency, at least for small chain sizes. With improved algorithms and faster computers this advantage is losing in importance. In this paper we propose and study an extension to three dimensions (3D) of a two-dimensional (2D) off-lattice model [2] that was successfully studied previously by Irbäck *et al.* [3, 4] This model contains only two types of amino acids, hydrophobic and hydrophilic, and the key part of the energy function is a species-dependent term that favors the formation of a hydrophobic core. However, as will be demonstrated below, this term alone is not sufficient in order to have thermodynamically dominant native states. This observation is reminiscent of existing lattice model results [5, 6], and could be taken as an indication that it is essential to have more than two different types of amino acids. In this paper we explore another possibility; do the folding properties depend on species-independent local interactions sticking to two amino acid types only? It should be noted that both lattice and 2D off-lattice models implicitly contain local interactions. In the lattice case the mere presence of a discretized space "stiffens" the dynamics locally, and in two dimensions the continuum movements are hampered by compressing one dimension.

The purpose of this work is to construct and numerically study a 3D generalization of the 2D model. [3, 4] As design criteria for such a model we have:

- The model should give rise to thermodynamically dominant states – i.e. be biologically plausible from a stability point of view.
- The local interactions should be chosen to at least qualitatively reproduce bond and torsional angle distributions and local correlations found in functional proteins.

We propose a simple form for the local interactions, which are found to play an important role. Without any local interaction in the model, the local structure of the chains is much more irregular than for proteins. It turns out that one can obtain a local structure qualitatively similar to that of proteins, by adjusting the strength of the local interactions. Having chosen the local interactions in this way, we reexamine the overall thermodynamic behavior. We find that not only have the local properties improved, but the native states have also become more stable.

Furthermore, we examine the structure formation and the properties of the folding transition by studying the temperature dependence of various thermodynamic and structural variables, and the distributions of these quantities at the folding temperature. These results can be understood in terms of a gradual collapse to compact structures where the (sequence dependent) folding transition occurs. The qualitative aspects of this picture confirm what has been found in lattice model calculations. [1, 8] However, it turns out that our results differ from lattice approaches in that bimodal energy distributions are not observed and that high folding temperatures are accompanied by relatively low temperatures for the peak of the specific heat.

The paper is organized as follows. In Sec. 5.2 the 3D model is defined and in Sec. 5.3 we extract local properties and correlations for functional proteins from the PDB data base [9]. Monte Carlo (MC) methods and measurements are described in Sec. 5.4. The results and the summary are found in Sec. 5.5 and 5.6 respectively.

5.2 The Model

We start by defining the simplified geometry of the model. Each residue is represented by a single site located at the C^{α} position. These sites are linked by rigid bonds of unit length, $\hat{b}_i$, to form linear chains living in three dimensions. The shape of an N-mer is either specified by the $N-1$ bond vectors $\hat{b}_i$ or by $N-2$ bond angles τ_i and $N-3$ torsional angles α_i (see Fig. 5.1).

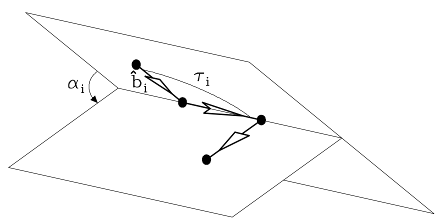

Figure 5.1: Definition of $\hat{b}_i$, τ_i and α_i.

The model contains two kinds of residues, A and B, which behave as hydrophobic (σ_i=+1) and hydrophilic ($\sigma_i = -1$) residues respectively. The energy function is given by

$$E(\hat{b};\sigma) = -\kappa_1 \sum_{i=1}^{N-2} \hat{b}_i \cdot \hat{b}_{i+1} - \kappa_2 \sum_{i=1}^{N-3} \hat{b}_i \cdot \hat{b}_{i+2} + \sum_{i=1}^{N-2} \sum_{j=i+2}^{N} E_{\mathrm{LJ}}(r_{ij};\sigma_i,\sigma_j) \quad (5.1)$$

where $r_{ij} = r_{ij}(\tau_{i+1},\ldots,\tau_{j-1};\alpha_{i+1},\ldots,\alpha_{j-2})$ denotes the distance between residues i and j, and $\sigma_1,\ldots,\sigma_N$ is a binary string that specifies the primary sequence. The *species-dependent* global interactions are given by the Lennard-Jones potential,

$$E_{\mathrm{LJ}}(r_{ij};\sigma_i,\sigma_j) = 4\epsilon(\sigma_i,\sigma_j)\Big(\frac{1}{r_{ij}^{12}} - \frac{1}{r_{ij}^{6}}\Big) \ . \quad (5.2)$$

The depth of the minimum of this potential, $\epsilon(\sigma_i,\sigma_j)$, is chosen to favor the formation of a core of A residues

$$\epsilon(\sigma_i,\sigma_j) = \begin{cases} 1 & \mathrm{AA} \\ \\ \frac{1}{2} & \mathrm{BB,\ AB} \end{cases} \quad (5.3)$$

The two parameters of the energy function, κ_1 and κ_2, determine the strength of *species-independent* local interactions. The model will be explored for different values of these two parameters, and our final choice will be $\kappa_1 = -1$ and $\kappa_2 = 0.5$.

The behavior of the model at finite temperature T is given by the partition function

$$Z(T;\sigma) = \int \Big[\prod_{i=1}^{N-1} d\hat{b}_i\Big] \exp(-E(\hat{b};\sigma)/T) \ . \quad (5.4)$$

Let us stress that the interactions in Eq. 5.1 are not chosen so as to stabilize the native state of some particular sequence. Rather, our goal is to study general sequences for a given energy function. We have attempted to choose this energy function as simple as possible. Anticipating some of the results to be presented in Sec. 5.5, let us here briefly discuss the relevance of the different interaction terms.

Species-dependent interactions. It is obvious that the global interactions play a key role in the model; these interactions are responsible for the compactification of the chain, and for the formation of a hydrophobic core. For functional proteins, we find that probability distributions of bond and torsional angles depend only weakly on the hydrophobicity pattern for the residues involved. In our model, the form of these distributions is very sensitive to the choice of global interactions, and it can be strongly sequence dependent. In order to avoid this, we have chosen the potentials for the three different types of residue pairs to be fairly similar; they are all attractive at large separations, and the location of the minimum is the same.

Species-independent interactions. It is less clear how important the local interactions are. When studying the behavior of several different sequences in the absence of the local interactions ($\kappa_1 = \kappa_2 = 0$), the stability of the native state tends to be very low. Furthermore, we find that local correlations along such chains are weak, which is in line with the findings of Socci *et al.* [10] For functional proteins the corresponding correlations are fairly strong, which is a manifestation of the presence of secondary structure. For these reasons we decided to incorporate the local interactions in the model. In this way one gets stiffer chains, which implies stronger local correlations. In addition, the stability of the native states tends to improve, as we will see below.

We have studied the model for many different choices of κ_1 and κ_2. The general behavior of the system is fairly insensitive to small changes of these parameters. Below we will focus on the results obtained for the three choices (κ_1, κ_2)=(0,0),(-1,0) and (-1,0.5).

5.3 Local Structures in Functional Proteins

In a 3D off-lattice model, it is possible to check the local properties against those for functional proteins in a direct way. We have probed the local structure in two different ways. First we consider the distributions of bond and torsional angles. Second we study local correlations along the chain. In this section we describe the results one finds for functional proteins. These results will be used in Sec. 5.5 to compare with the qualitative local behavior of our model.

Distributions of bond and torsional angles. These distributions for functional proteins are well-known [11, 12] and are included here for completeness. We consider the structure defined by the backbone of C^α atoms. The results were obtained using a set of 505 selected sequences [13, 14] from the Protein Data Bank. [9] This set was obtained by allowing for a maximum of 25% sequence similarity for aligned subsequences of more than 80 residues. [15] Within this set of 505 minimally redundant sequences, 491 contained complete backbone information; the others were excluded from our analysis.

In Fig. 5.2 the distributions of bond and torsional angles are shown together with a scatter plot for these two quantities. Note that the calculation of the torsional angle α_i requires four consecutive C^α atoms, while the bond angle τ_i requires only three. The additional fourth C^α atom needed for α_i is taken in the N-terminus direction.

From Fig. 5.2 it is evident that there are strong regularities in the local structure. The τ_i and α_i distributions both exhibit a clear two-peak structure, which can be associated with two well-populated regions in the (τ_i, α_i)-plane. One of these regions, $\tau_i \in$ [85°,100°] and $\alpha_i \in$ [35°,70°], corresponds to right-handed

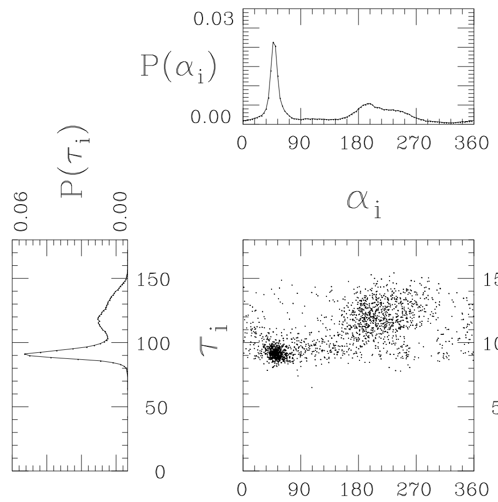

Figure 5.2: Bond (τ_i) and torsional (α_i) angle distributions for functional proteins.

α-helix and the other, $\tau_i \in$ [105°,145°] and $\alpha_i \in$ [170°,250°], corresponds to β-sheet.

Local correlations. We study local correlations using the function

$$C_b(d) = \frac{1}{N-d-1} \sum_{i=1}^{N-d-1} \hat{b}_i \cdot \hat{b}_{i+d} \tag{5.5}$$

where $\hat{b}_i$ are normalized (virtual) bond vectors. Local correlations along protein chains have been studied by Socci *et al.* [10], using a correlation function slightly different from that in Eq. 5.5. In Fig. 5.3 we show the correlation function $C_b(d)$ for functional proteins. As can be seen from this figure, there are significant correlations at least out to separations of about eight residues. The oscillations present can be related to the presence of α-helix structure, which has a period of 3.6.

5.4 Methods

Monte Carlo Method

Numerical simulations have been performed for a variety of different choices of (κ_1, κ_2) and different sequences and temperatures. At low temperatures conventional Monte Carlo methods tend to become extremely time-consuming,

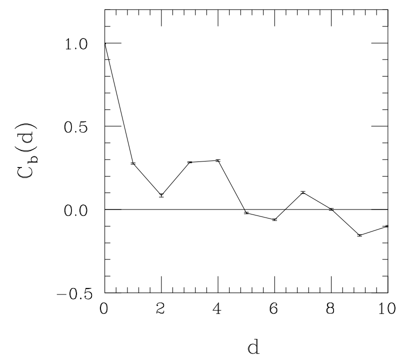

Figure 5.3: The correlation function $C_b(d)$ for functional proteins (see Eq. 5.5).

due to the presence of high free-energy barriers. As in previous work [3, 4] we have therefore chosen to employ the dynamical-parameter method, which means that some parameter of the model is taken as a dynamical variable which takes values ranging over a definite set. In this way it is possible to greatly improve the frequency of transitions between the different free-energy valleys; for the 2D model speedup factors of 10^3–10^4 were observed [3], compared with conventional methods.

In the present work the temperature is treated as a dynamical variable ("simulated tempering" [16]). More precisely, the joint probability distribution

$$P(\hat{b}, k) \propto \exp(-g_k - E(\hat{b}, \sigma)/T_k) \tag{5.6}$$

is simulated, where T_k, $k = 1, \ldots, K$, are the allowed values of the temperature. The g_k's are tunable parameters which must be chosen carefully for each sequence. Details on how to determine these parameters can be found elsewhere. [3, 4] The joint distribution $P(\hat{b}, k)$ is simulated by using separate Metropolis steps [17] in k and $\hat{b}$. For $\hat{b}$ we use two types of elementary moves: rotations of single bonds and moves of pivot type. [18]

In our simulations we used a set of $K = 25$ allowed temperatures, which are equidistant in $1/T$ and ranging from 0.15 to 3.0.

In order to study the energy level spectrum we use a quenching procedure; in the course of the simulations the system is quenched to zero temperature at regular intervals. For this purpose we employ a conjugate gradient method. We found this method more efficient than using a Monte Carlo algorithm with $T = 0$. Also, we tested two different conjugate gradient algorithms. In the conjugate gradient method successive minimizations are carried out along different lines. Information about the gradient is needed to define the lines, and may or may not be utilized for the minimizations. For the present problem we found that

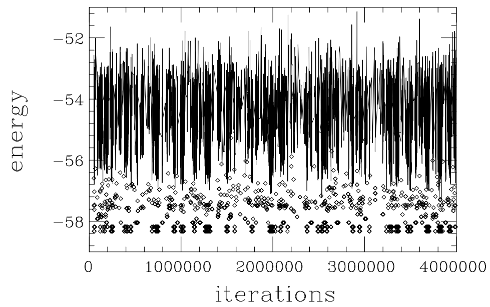

Figure 5.4: Evolution of the quenched (diamonds) and unquenched (line) energies in the simulation of sequence 1 (see Table 5.1). Measurements were taken every 10 iterations. Shown are the data corresponding to the lowest allowed temperature. The thermalization phase of 50000 sweeps is not shown.

minimization without gradient calculation is faster. The quenching procedure only accounts for a small portion of the total computing time.

In Fig. 5.4 we show the evolution of the quenched and unquenched energies in a typical simulation of a $N = 20$ chain.

Measurements

In Sec. 5.3 we discussed measurements of local properties of the chains. In order to study the stability of the full native structure, further information is needed. To this end we have studied the distribution of the mean-square distance between configurations, δ^2_{ab}. For two configurations a and b, δ^2_{ab} is defined as

$$\delta^2_{ab} = \min \frac{1}{N} \sum_{i=1}^{N} |\bar{x}_i^{(a)} - \bar{x}_i^{(b)}|^2 \tag{5.7}$$

where $|\bar{x}_i^{(a)} - \bar{x}_i^{(b)}|$ denotes the distance between the sites $\bar{x}_i^{(a)}$ and $\bar{x}_i^{(b)}$, and where the minimum is taken over translations, rotations and reflections. The corresponding distribution, $P(\delta^2)$, for fixed temperature and sequence, reads

$$P(\delta^2) = \frac{1}{Z(T;\sigma)^2} \int d\hat{b}^{(a)} d\hat{b}^{(b)} \delta(\delta^2 - \delta^2_{ab}) \mathrm{e}^{-E(\hat{b}^{(a)};\sigma)/T} \mathrm{e}^{-E(\hat{b}^{(b)};\sigma)/T} \tag{5.8}$$

where $\delta(\cdot)$ denotes the Dirac delta function. For convenience, we often use the mean

$$\langle \delta^2 \rangle = \int d\delta'^2 P(\delta'^2) \delta'^2 \tag{5.9}$$

rather than the full distribution $P(\delta^2)$.

We have also measured the specific heat C_V and gyration radius $r_{\rm gyr}$, defined by

$$C_V = \frac{1}{T^2}\left(\langle E^2\rangle - \langle E\rangle^2\right) \tag{5.10}$$

$$r_{\rm gyr}^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\langle\bar{x}_i^2\rangle - \langle\bar{x}_i\rangle^2\right) \tag{5.11}$$

The specific heat has a maximum in the vicinity of the folding transition. To accurately determine the height and location of this maximum we use the multihistogram method. [19, 20]

In addition to these measurements, we employ the quenching procedure described above. Removing the thermal noise in this way is of great help in monitoring the evolution of the system, but requires a substantial amount of computer time. Therefore, we have performed these calculations at larger intervals than other measurements, and only at the lowest of the temperatures studied.

The quenching procedure provides us with a set of low-lying local energy minima. The mean-square distance to the global minimum will be denoted by δ_0^2 (cf Eq. 5.7). Using the corresponding distribution, $P(\delta_0^2)$, we define a probability p_0 of finding the system within a small neighborhood of the global minimum. We take

$$p_0 = \int_0^{\Delta} P(\delta_0^2)\,d\delta_0^2 \tag{5.12}$$

with the parameter $\Delta = 0.04$; this choice of Δ is motivated by the δ_0^2 distributions (cf Figs. 5.12 a and b). Let us stress that the distribution $P(\delta_0^2)$ is different from the distribution $P(\delta^2)$ introduced earlier; $P(\delta^2)$ measures general fluctuations rather than deviations from a given state.

The degree of folding may be defined in terms of the quantity p_0. Alternatively, it may be defined as

$$Q = \frac{n^o}{n} \tag{5.13}$$

where n^o is the number of occupied native contacts, and n is the total number of native contacts. Two monomers i and j are taken to be in contact if $r_{ij}^2 < 1.75$; this cutoff is motivated by the distribution $P(r_{ij}^2)$ (not displayed in this paper). We have $Q = 1$ for the native structure. Also, it is useful to divide the set of native contacts into local and global contacts. A contact between monomers i and j will be called local if $2 \le |i-j| \le 4$ and global if $|i-j| > 4$. We set $Q_l = n_l^o/n_l$, where n_l^o is the number of occupied local native contacts and n_l is the total number of local native contacts. Similarly, we define Q_g as the fraction of occupied global native contacts.

no.	sequence	T_s	$C_V^{\max}$	T_f
1	$BAAA\ AAAB\ AAAA\ BAAB\ AABB$	0.361	51.4	< 0.15
2	$BAAB\ AAAA\ BABA\ ABAA\ AAAB$	0.319	55.9	< 0.15
3	$AAAA\ BBAA\ AABA\ ABAA\ ABBA$	0.298	65.8	0.23
4	$AAAA\ BAAB\ ABAA\ BBAA\ ABAA$	0.273	61.5	0.22
5	$BAAB\ BAAA\ BBBA\ BABA\ ABAB$	0.327	49.8	< 0.15
6	$AAAB\ BABB\ ABAB\ BABA\ BABA$	0.257	62.6	0.15

Table 5.1: The six sequences studied. The errors in T_s and $C_V^{\max}$ are less than 0.005 and 0.3 respectively. The errors in T_f are approximately 0.02.

5.5 Results

The model is defined by two parameters, κ_1 and κ_2, which set the strengths of the local interactions. In order to investigate the importance of these interactions, we first performed preliminary runs for a number of different (κ_1, κ_2) values. In particular, these explorations aimed at establishing the balance between the local and global interactions — the overall scale of (κ_1, κ_2). More extensive simulations were then carried out for three different choices

$$(\kappa_1, \kappa_2) = \begin{cases} (0,0) \\ (-1,0) \\ (-1,0.5) \end{cases} \tag{5.14}$$

using sequences of length $N = 20$. It turns out that shorter chains exhibit less interesting and discriminative behavior. The sequences, which were deliberately chosen to represent a variety of behavior, are listed in Table 5.1.

In this section we first compare the behavior of the model at the three values of (κ_1, κ_2) with respect to the local interactions and low T properties in order to single out one set of (κ_1, κ_2) values. We then examine the folding properties and thermodynamic behavior for $(\kappa_1, \kappa_2) = (-1, 0.5)$ in some detail.

The Local Interactions

Prior to investigating folding properties for different choices of (κ_1, κ_2), we compare the bond and torsional angle distributions and the local correlations of the chains with those of functional proteins. In Figs. 5.5 and 5.6 the model counterparts of Figs. 5.2 and 5.3 are shown. The data in Fig. 5.5 were obtained at a fixed temperature, $T = 0.15$, while those in Fig. 5.6 were obtained using

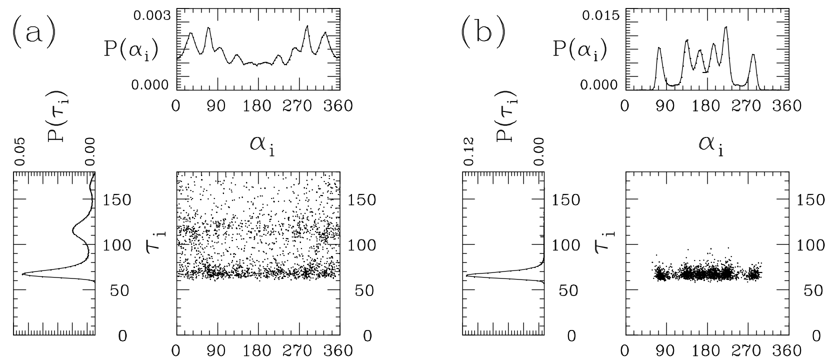

Figure 5.5: Bond (τ_i) and torsional (α_i) angle distributions at $T = 0.15$ for the six simulated sequences, **(a)** pure Lennard-Jones potential $(\kappa_1, \kappa_2) = (0, 0)$ and **(b)** $(\kappa_1, \kappa_2) = (-1, 0.5)$. The potential corresponding to $(\kappa_1, \kappa_2) = (-1, 0)$ yields a similar distribution as in (b).

quenched low energy structures. In both cases we expect the data to reflect the behavior for a wide range of not too high temperatures.

The results strongly indicate the need for local interactions when it comes to mimicking functional proteins; the strong regularities in the local structure observed for proteins are to a large extent missing for a pure Lennard-Jones potential. This can be seen from the torsional angle distribution and the correlation function $C_b(d)$. The torsional angle α_i varies over the whole range of 360°, without any strongly suppressed regions, and the correlation $C_b(d)$ is very weak for all $d > 0$. The conclusion that local correlations are weak for a pure Lennard-Jones potential is in good agreement with the findings of Socci *et al.* [10]

From the Figs. 5.5 and 5.6 it is also clear that changing κ_1 from 0 to -1 leads to a significant improvement; the torsional angle distribution becomes concentrated to a few relatively small regions, and local correlations become stronger. The range of the correlation $C_b(d)$ is for $\kappa_1 = -1$ comparable to what it is for proteins. Needless to say, the model is not intended to reproduce the precise form of the correlations. However, it is encouraging that the qualitative behavior of this very simple model is consistent with the one from functional proteins.

The remaining question is whether a non-zero κ_2 is called for. No conclusive evidence can be drawn from Figs. 5.5 and 5.6 alone in this respect, although one may argue that the range of the correlation $C_b(d)$ is still somewhat short for $(\kappa_1, \kappa_2) = (-1, 0)$.

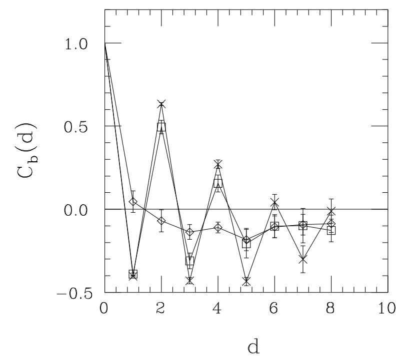

Figure 5.6: The correlation function $C_b(d)$ (Eq. 5.5) averraged over the six sequences in Table 5.1 using (κ_1,κ_2)=(0,0) ($\diamond$), (-1,0) ($\square$) and (-1,0.5) ($\times$).

Next we investigate how the choice of parameters affects the key folding property [4] — the δ^2 distribution should be peaked for low δ^2. In Fig. 5.8, $P(\delta^2)$ is shown for sequence 4 in Table 5.1. In contrast to the local properties discussed above, $P(\delta^2)$ is strongly sequence dependent. Nevertheless, Fig. 5.8 illustrates some general trends seen in the data. First, the pure Lennard-Jones interaction yields a very broad distribution of δ^2, implying that folding properties are poor. This is true for all the six sequences studied. Second, although the behavior is different for one of the six sequences, the $\kappa_2 \neq 0$ choice appears to have a distinct edge over the one ignoring the torsional angle interaction.

In summary, our results show that local interactions are essential in order to get a regular local structure and structural stability. When comparing the results for $(\kappa_1,\kappa_2) = (-1,0)$ and (-1,0.5), we find that the structural stability tends to be highest for $(\kappa_1,\kappa_2) = (-1,0.5)$. In what follows we will focus on this choice of parameters for the six sequences in Table 5.1.

In Fig. 5.7 a representative example of ground state structure for $(\kappa_1,\kappa_2) = (-1,0.5)$ is shown — the local structure is regular and the global structure is such that the A monomers form a core. Such formation of a hydrophobic core is observed for all sequences studied.

2D Revisited

In the work by Irbäck *et al.* [3,4], a similar model [2] was studied in 2D using somewhat different Lennard-Jones parameters and local interactions corresponding to (κ_1,κ_2)=($\frac{1}{4}$,0) (cf Eq. 5.1). Here, we briefly discuss the importance of local interactions for this 2D model.

For this purpose, we leave out the local interaction term, i.e. we set

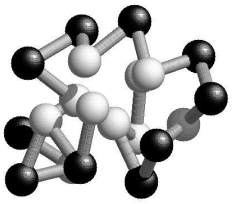

Figure 5.7: A representative native state (seq. 6) for $(\kappa_1, \kappa_2) = (-1, 0.5)$ illustrating the formation of a hydrophobic core. The hydrophobic (A) and hydrophilic (B) monomers are represented by white and black spheres respectively.

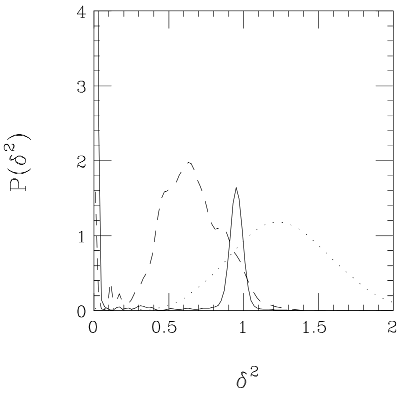

Figure 5.8: $P(\delta^2)$ at $T = 0.15$ for sequence 4 (cf Table 5.1); (κ_1,κ_2)=(0,0) (dots), (-1,0) (dashes), and (-1,0.5) (solid). The low-δ^2 peak for (κ_1,κ_2) = $(-1, 0.5)$ extends to 37.

(κ_1, κ_2)=$(0, 0)$. Using this pure Lennard-Jones potential, we redo the simulations for 15 sequences. At $T = 0.15$, we find a strong correlation between the $\langle\delta^2\rangle$'s for the two types of potentials; thus $\langle\delta^2\rangle$ varies widely with sequence even for the (κ_1, κ_2)=$(0, 0)$ potential. This is in contrast to our findings in 3D, where all the sequences studied have a large $\langle\delta^2\rangle$ for (κ_1, κ_2)=$(0, 0)$.

Folding Properties

When investigating the folding properties of the chains we focus on the thermodynamics. This is in part inspired by earlier results [4], where the thermodynamic properties exhibited strong sequence dependence. Initially we examine the various thermodynamic quantities defined in Sec. 5.4 over the entire probed T range. Then we proceed with the "magnifying glass" to the 0.15< T <0.50

region, where the different chains exhibit strongest difference in behavior and the folding properties can be studied in some detail.

Thermodynamic Behavior

In Fig. 5.9 the behavior of the different thermodynamic quantities are shown over the entire T range. The overall size of the molecule, as measured by $r^2_{\rm gyr}$, decreases substantially when T decreases from 3 to 0.15, as can be seen from Fig. 5.9. This decrease is gradual. The data points essentially fall onto two different curves. The upper curve corresponds to the sequences 5 and 6 with composition 10A+10B, and the lower curve to the sequences 1–4 with composition 14A+6B.

Next we turn to $\langle\delta^2\rangle$, which measures the size of the fluctuations. In Fig. 5.9 we show the relative magnitude $\langle\delta^2\rangle/r^2_{\rm gyr}$. This ratio exhibits a peak between $T = 0.6$ and 0.9, and is fairly sequence independent down to $T = 0.4$. At the lowest temperatures studied the situation is different. Here $\langle\delta^2\rangle$ and $\langle\delta^2\rangle/r^2_{\rm gyr}$ are strongly sequence dependent. In Fig. 5.10 the results are shown for $\langle\delta^2\rangle$ in the low-T region.

In order to understand the low-T behavior it is useful to consider p_0, which measures the relative population of the lowest energy state (Eq. 5.12). A comparison of the data for $\langle\delta^2\rangle$ and p_0 shows that small $\langle\delta^2\rangle$ values are associated with large p_0 values. Therefore, it is reasonable to define the folding temperature T_f as the temperature where $p_0 = 1/2$. Estimates of T_f are given in Table 5.1. For three of the six sequences T_f is smaller than 0.15, as can be seen from Fig. 5.11a. It should be stressed that the shape of the molecule is very compact already above T_f. This will be discussed in more detail below for one of the sequences.

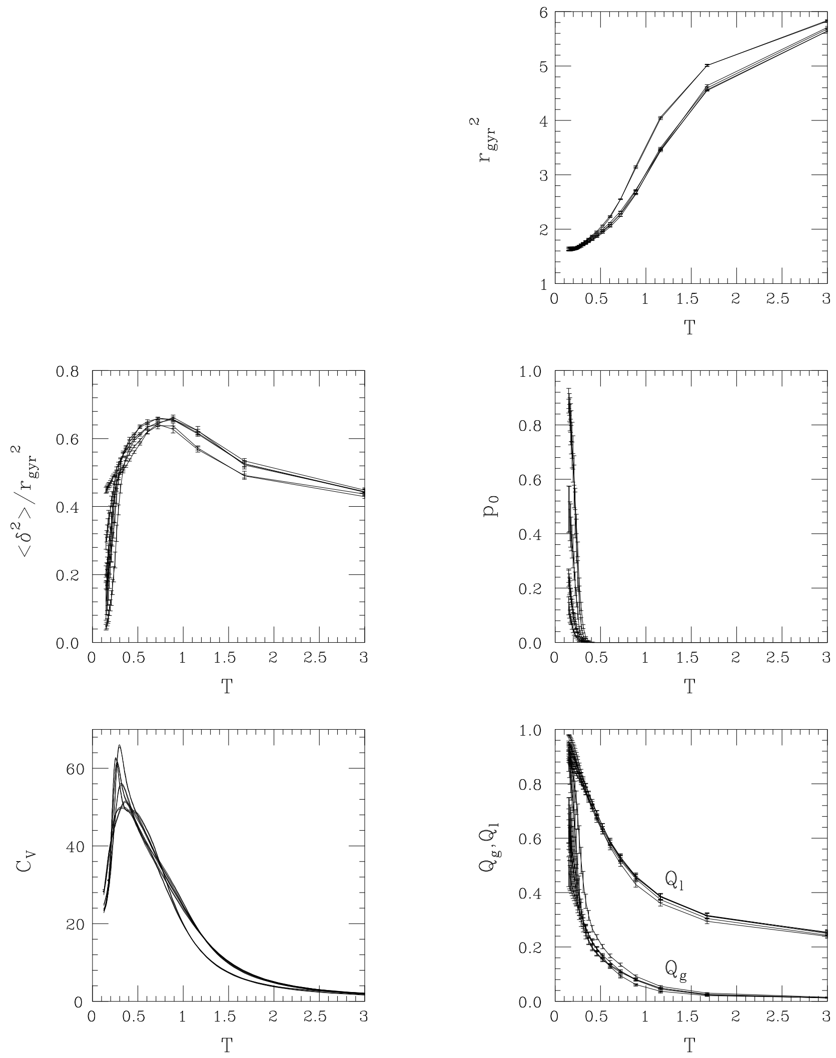

Figure 5.9: Thermodynamic properties for the six chains in Table 5.1 using (κ_1, κ_2)=(-1,0.5).

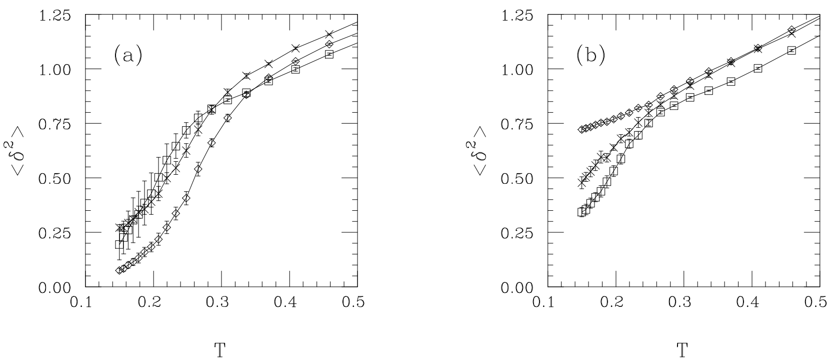

Figure 5.10: $\langle \delta^2 \rangle$ as a function of T for the sequences of Table 5.1 using (κ_1, κ_2)=(-1,0.5) : **(a)** 4 ($\Box$), 2 ($\times$) and 3 ($\diamond$). **(b)** 6 ($\Box$), 1 ($\times$) and 5 ($\diamond$).

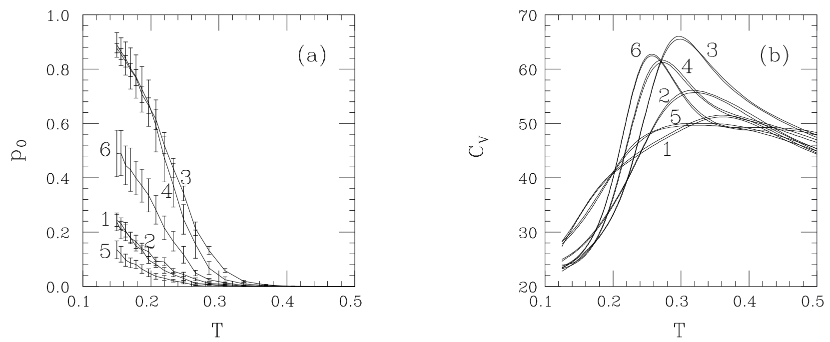

Figure 5.11: Enlargements of **(a)** p_0 and **(b)** C_V in Fig. 5.9, (κ_1, κ_2)=(-1,0.5). The bands in (b) correspond to $C_V \pm \sigma$, where σ is the standard deviation.

From Figs. 5.9 and 5.11b it can be seen that the specific heat exhibits a peak in the low-T region, slightly above T_f. The height, C_V^{max}, and location, T_s, of the peak are given in Table 5.1. We find that C_V^{max} is large for the sequences with high T_f, which is in accordance with the results from lattice model studies [8]. However, our results for T_s are somewhat different from these [8]. In our model T_s is relatively low for sequences with high T_f, while the lattice results [8] show the opposite behavior. Thus, the separation $T_s - T_f$ exhibits a stronger sequence dependence in our model.

Also shown in Fig. 5.9 is the T dependence for the occupancy of local and global native contacts. In the vicinity of T_f, Q_l is large and varies slowly, whereas Q_g changes rapidly. In particular, this demonstrates that the formation of local native structure, like the compactification, is a gradual process that to a large extent takes place above T_f.

The Folding Transition Region

In this subsection we study the behavior at the folding temperature T_f, where $p_0 = 1/2$, in some more detail. Two different sequences are considered, 4 and 6. Sequence 4 has a fairly high T_f value, whereas sequence 6 represents a more typical T_f.

In Figs. 5.12 a and b we show histograms of δ_0^2, the mean-square distance to the lowest energy state, near $T = T_f$. Both histograms exhibit a multi-peak structure, with a narrow peak at low δ_0^2 that corresponds to the native state. The thermodynamic weight of the native state is, by definition, approximately 50% in both cases. However, the distributions differ in shape, which is important from the viewpoint of kinetics. From Figs. 5.12 a and b one concludes that the δ_0^2 distribution is much more rugged for sequence 6 than for sequence 4. This suggests that, at $T = T_f$, folding is fastest for sequence 4, the sequence with highest T_f. We also explored using Q rather than δ_0^2 as reaction coordinate, with similar results.

The Q distribution at T_f has been studied previously, using a lattice model [7]. Here the distributions for a folding and a non-folding sequence were compared, and these were found to be almost identical. This may seem to contradict our findings, and suggest that there is a difference between the two models. However, it should be remembered that Šali *et al.* [7] did not study the full Q distribution, but rather the Q distribution corresponding to maximally compact structures only.

In Figs. 5.12 c and d we show the probability distributions of E and r_{gyr}^2 at $T \approx T_f$ for sequence 4. Also shown are the contributions to these distributions from conformations with $\delta_0^2 < 0.04$ and $\delta_0^2 > 0.04$, respectively. This corresponds to a simplified two-state picture where each conformation is classified as either native-like or not. The shape of the δ_0^2 distribution shows that such a classification is feasible in an essentially unambiguous way.

As one might have expected from the sharpness of the peak in the specific heat, Fig. 5.12 c shows that these two "states" differ significantly in energy, although the total E distribution is unimodal. However, the average size of the molecule is very similar for the two states. This fact clearly demonstrates that the compactification occurs prior to the dominance of the native state setting in.

These results may be compared with those of Socci *et al.* [8], where a detailed study of the behavior at T_f was carried out for a lattice model. Here the total number of contacts, C, rather than r_{gyr}^2 was used as a measure of compactness. In contrast to the results shown in Figs. 5.12 c and d, the probability distributions of both E and C were found to be bimodal for a sequence with high T_f. This may reflect a genuine difference between lattice and off-lattice models.

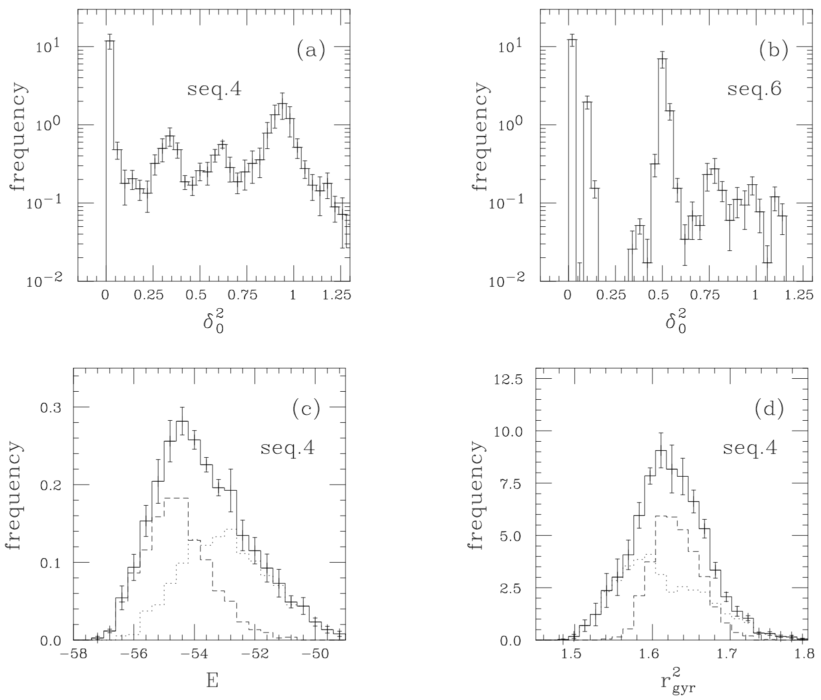

Figure 5.12: Histograms at $T \approx T_f$ for sequence 4 ($T = 0.22$) and 6 ($T = 0.15$), $(\kappa_1, \kappa_2) = (-1, 0.5)$. **(a)** δ_0^2 for seq. 4. **(b)** δ_0^2 for seq. 6. **(c)** E for seq. 4. **(d)** r^2_{gyr} for seq. 4. The dashed and dotted lines in (c) and (d) represent the contributions corresponding to $\delta_0^2 < 0.04$ and $\delta_0^2 > 0.04$, respectively.

5.6 Summary

We have extended the off-lattice protein model of Irbäck *et al.* [4] to three dimensions. In doing so, one key issue has been the choice of species-independent local interactions that balance the species-dependent non-bonded Lennard-Jones interactions. The local bond and torsional angle interactions were chosen to satisfy two criteria:

- The resulting low temperature configurations should at least qualitatively reproduce the local angle distributions and correlations of functional proteins, in particular the range over which structural correlations exist. In order to reproduce the precise shape of the correlation function, and that of bond and torsional angle distributions, one would have to work at a different level of detail, which is beyond the scope of this paper.

- It should be possible to produce good folders; i.e. there should be sequences with thermodynamically stable structures at not too low temperatures.

It turns out that the presence of local interactions is necessary for satisfying these criteria. Among the two local interactions, the bond angle one is the most crucial one in this respect.

After having specified the interaction to meet these requirements we study the thermodynamic behavior for six different sequences. The following generic behavior emerges:

- As the temperature is decreased, a gradual compactification takes place.
- In the compact state a sequence dependent folding transition occurs, where the good folding sequences are characterized by a higher folding temperature. In terms of the specific heat, these good folders also have more pronounced peaks.
- In the state from which the transition to the native state occurs, a large fraction of the native contacts are already formed. The contacts missing are mainly those corresponding to large topological distance along the chain.

This picture is consistent with what is observed for lattice models. [1,8] A few minor, but significant, differences should be mentioned though.

- We do not observe bimodal distributions of energy or compactness as in the work by Socci *et al.* [8] In a simplified two-state picture, the two coexisting states at the folding temperature do correspond to different energy distributions, but the overlap between the two is large enough to blur this effect.
- For the sequences studied, high folding temperature is accompanied by a relatively low temperature for the peak of the specific heat. This is in contrast to what has been previously reported. [8]

Finally, we stress the fact that we have studied sequences with only two types of residues. In the work by Shakhnovich *et al.* [5,6] a number of binary (two-letter code) sequences were studied in a 3D lattice model. The two-letter code was found to be insufficient in the sense that these sequences did not have unique native structures. In our model, many binary sequences do have unique native structures.

Our ability to map out the thermodynamics of the 3D off-lattice model, relies heavily upon the efficiency of the dynamical-parameter algorithm. [3,4] As it stands, the results for each $N = 20$ chain require 70 CPU hours on an Alpha DecStation 200. We feel confident that additional algorithmic efficiency improvements can be made, which will enable us to probe longer and more 3D chains than reported in this work.

Bibliography

[1] For a review, see M. Karplus and A. Šali, "Theoretical Studies of Protein Folding and Unfolding", *Curr. Opin. Struct. Biol.* **5**, 58 (1995).

[2] F.H. Stillinger, T. Head-Gordon and C.L. Hirshfeld, "Toy Model for Protein Folding", *Phys. Rev.* **E 48**, 1469 (1993).

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

[4] A. Irbäck, C. Peterson and F. Potthast, "Identification of Amino Acid Sequences with Good Folding Properties", *Phys. Rev.* **E 55**, 860 (1997).

[5] E.I. Shakhnovich, "Proteins with Selected Sequences Fold into Unique Native Conformation", *Phys. Rev. Lett.* **72**, 3907 (1994).

[6] K. Yue, K.M. Fiebig, P.D. Thomas, H.S. Chan, E.I. Shakhnovich and K.A. Dill, "A Test of Lattice Protein Folding Algorithms", *Proc. Natl. Acad. Sci. USA* **92**, 325 (1995).

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

[8] N.D. Socci and J.N. Onuchic, "Kinetic and Thermodynamic Analysis of Proteinlike Heteropolymers: Monte Carlo Histogram Technique", *J. Chem. Phys.* **103**, 4732 (1995).

[9] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, "The Protein Data Bank: A Computer Based Archival File for Macromolecular Structures", *J. Mol. Biol.* **112**, 535 (1977).

[10] N.D. Socci, W.S. Bialek and J.N. Onuchic, "Properties and Origins of Protein Secondary Structure", *Phys. Rev.* **E 49**, 3440 (1994).

[11] M. Levitt, "A Simplified Representation of Protein Conformations for Rapid Simulation of Protein Folding", *J. Mol. Biol.* **104**, 59 (1976).

[12] A. Godzik, A. Kolinski and J. Skolnick, "Lattice Representations of Globular Proteins: How Good are They?", *J. Comp. Chem.* **14**, 1194 (1993).

[13] U. Hobohm and C. Sander, "Enlarged Representative Set of Protein Structures", *Protein Sci.* **3**, 522 (1994).

[14] The May 1996 edition was used

[15] U. Hobohm, M. Scharf, R. Schneider and C. Sander, "Selection of a Representative Set of Structures from the Brookhaven Protein Data Bank", *Protein Sci.* **1**, 409 (1992).

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

[17] N.A. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A. Teller and E. Teller, "Equation of State Calculations by Fast Computing Machines", *J. Chem. Phys.* **21**, 1087 (1953).

[18] M. Lal, "Monte Carlo Computer Simulations of Chain Molecules", *Molec. Phys.* **17**, 57 (1969).

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

[20] A.M. Ferrenberg and R.H. Swendsen, "Optimized Monte Carlo Data Analysis", *Phys. Rev. Lett.* **63**, 1195 (1989).

Paper VI

A Minimal Off-Lattice Model for α-Helical Proteins

Frank Potthast

Complex Systems Group, Department of Theoretical Physics
University of Lund, Sölvegatan 14A, S-223 62 Lund, Sweden

LU TP 98-13
Submitted to *Journal of Computational Biology*

Abstract:

A minimal off-lattice model for α-helical proteins is presented. It is based on hydrophobicity forces and sequence independent local interactions. The latter are chosen so as to favor the formation of α-helical structure. They model chirality and α-helical hydrogen bonding. The global structures resulting from the competition between these forces are studied by means of an efficient Monte Carlo method. The model is tested on two sequences of length $N=21$ and 33 which are intended to form 2- and 3-helix bundles, respectively. The local structure of our model proteins is compared to that of real α-helical proteins, and is found to be very similar. The two sequences display the desired numbers of helices in the folded phase. Only a few different relative orientations of the helices are thermodynamically allowed. Our ability to investigate the thermodynamics relies heavily upon the efficiency of the used algorithm, simulated tempering; in this Monte Carlo approach, the temperature becomes a fluctuating variable, enabling the crossing of free-energy barriers.

Key words: protein folding, hydrophobicity, simulated tempering, global optimization, helical bundle, alpha-helix

6.1 Introduction

Protein folding may be described on all levels of complexity, ranging from most simplistic concepts that do not even incorporate the concept of geometry, to all-atom representations including solvent. Evidently, different levels of descriptive complexity address different aspects of the protein folding problem.

Addressing the thermodynamics of a given protein model is generally very difficult due to the free energy barriers present for compact chains [1]. Therefore, past thermodynamic studies of proteins have mainly been performed using lattice models. Lattice models lend themselves to thermodynamic calculations since the conformation space is discrete, enabling the enumeration of all conformations and therefore the calculation of exact results for short chains. For example, the minimal HP model of Lau and Dill [2] has been examined in quite some detail [3, 4]. However, the approximations involved in lattice models are far from well understood [5], and their geometry is clearly not that of real proteins. Furthermore, the energy barriers between the states may be poorly represented due to the discreteness of the conformation space.

In this paper, we study the thermodynamic behavior of a simple off-lattice model for protein folding. Studies of similar, minimal, off-lattice models have been performed before, see e.g. [6–11]. The main difference between our model and those studied previously is that it is deliberately constructed so as to model the C_α backbone geometry of α-helical proteins. A somewhat similar approach was taken in [12]. However, these authors used a conventional Monte Carlo method, and reliable thermodynamic averages for global quantities were not obtained.

Our starting point is the model recently suggested by Irbäck et al. [11]. The major change is a modification of the local interactions which we here choose so as to model α-helical structure. We stick to the original concept of presenting each amino acid as a single site. Furthermore, only two types of residues are considered, hydrophobic and hydrophilic. The residues are linked by rigid bonds. The local interactions in our model are meant to model the chirality of the amino acids and the hydrogen bonds found in real α-helices.

It has been shown by Kamtekar et al. [13] that the binary sequence pattern of hydrophobic and hydrophilic amino acids is of central importance in the design of de novo proteins. The model presented here has some potential to be of help when pursuing this type of protein design. In particular, it may give valuable hints on different thermodynamically stable structures of a given hydrophobic/hydrophilic sequence; such information would certainly be useful in this design approach.

6.2 Methods

The Model

Our starting point is the model proposed in [11], to be referred to as the AB model, which is briefly described in the appendix. In the AB model, there are sequences which are thermodynamically stable at kinetically acceptable temperatures [11]. Also, there are strong regularities in the local structure of the chains, qualitatively similar to those for real proteins. However, at a more quantitative level, the local structure observed in [11] is very different from that of real proteins - partly because it is insensitive to space reflections (no chirality). The modifications of the AB model presented in this paper are meant to model α-helical local structure; in particular, chirality is introduced. Globally, the modifications lead to sizes similar to those of real proteins.

Let us call the right-handed α-helical structure found in real proteins an **ideal α-helix**; this helix has 3.6 amino acids per turn and a translation of 5.4Å per turn. An ideal α-helix is completely described by the geometry of three successive (virtual) C_α-C_α bonds; that is by one torsional (α_i) and two bend angles (τ_i, τ_{i+1}). Equivalently, one may describe an ideal α-helix by the mutual distances r_{ij}^α of the four participating C_α atoms (plus the correct chirality).

In our model, each residue is represented by a single site corresponding to the C_α position in the polypeptide backbone. These sites are linked by rigid bonds, $\vec{b}_i$, of length 3.8Å. The shape and energy of an N-mer is specified by the $N-1$ bond vectors $\vec{b}_i$. r_{ij} denotes the distance between residues i and j. $\sigma_1, \ldots, \sigma_N$ is a binary string that specifies the primary sequence; we consider hydrophobic ($\sigma = 0$) and hydrophilic ($\sigma = 1$) monomers only. The energy function is defined to be

$$E(\vec{b};\sigma) = \sum_{i=1}^{N-2} \sum_{j=i+2}^{N} 4\epsilon(\sigma_i,\sigma_j) \left(\left(\frac{A_{ij}}{r_{ij}}\right)^{12} - \left(\frac{A_{ij}}{r_{ij}}\right)^{6} \right) + \sum_{i=1}^{N-3} E_l(i) \tag{6.1}$$

The first term in this equation consists of Lennard-Jones (LJ) interactions. The LJ parameters A_{ij} are chosen as $A_{ij} = 2^{-1/6} r_{ij}^\alpha$ for $|j-i| \leq 4$, where r_{ij}^α denotes the distance in an ideal α-helix. This choice of A_{ij} gives the corresponding terms in eq. 6.1 a minimum at r_{ij}^α. For all other (i,j), we set $A_{ij} \approx 6.01$Å, which is calculated from the average volume, 161Å^3, of an amino acid [14]. Hence there are four different A_{ij} values, three of which are determined by the geometry of an ideal α-helix, the fourth being defined by the average size of an amino acid. Finally, the depth of the LJ potential is chosen so as to favor the formation of a hydrophobic core; $\epsilon(0,0) = 1$ for a hydrophobic-hydrophobic pair, $\epsilon = 1/2$ for all other pairs.

The sequence-independent local interaction $E_l(i)$ depends on the torsional and bend angles of three successive bond vectors $(\alpha_i, \tau_i, \tau_{i+1})$, and is given by

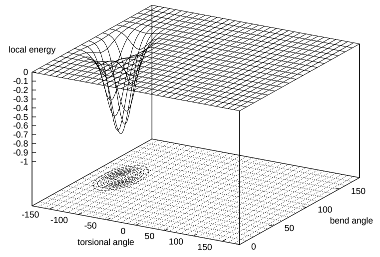

Figure 6.1: Local energy $E_l(i)$ (Eq. 6.2) displayed in one bond and one torsional angle. The second bend angle is set to $\hat{\tau}$, the bend angle of an ideal α-helix. $E_l(i)$ has a single minimum of depth one around the α-helical conformation and is close to zero elsewhere. Equipotential lines for $E_l(i) \in \{-0.1, -0.2, ..., -0.9\}$ are displayed with circles.

a negative gaussian with unit depth and width w;

$$E_l(i) = -\exp\Big(-\frac{(\alpha_i - \hat{\alpha})^2 + (\tau_i - \hat{\tau})^2 + (\tau_{i+1} - \hat{\tau})^2}{2w^2}\Big) \qquad E_l(i) \in [-1, 0] \quad (6.2)$$

where $\hat{\alpha}$ and $\hat{\tau}$ are the torsional and bend angles in an ideal α-helix. $E_l(i)$ is close to minus one if and only if the three participating bond vectors are close to α-helical structure, which is illustrated in Fig. 6.1. The width is taken to be $w = 1/4$ rad $\approx 14.32°$; this choice is justified by the resulting angle distributions, which will be presented in section 6.3. The aim of the local interaction is twofold. First, it introduces a right-handed chirality into the model that is not present in the LJ term alone. Second, it is meant to model the hydrogen bonds of the α-helical conformation. Observe that the depth of the local interaction $E_l(i)$ equals the depth of a hydrophobic-hydrophobic contact. Exact values of the parameters and technicalities, as well as a comparison to the AB model, can be found in the appendix.

In section 6.3, it will be shown that, not surprisingly, the model indeed gives rise to α-helical structure. This is done by comparing the local structure of our model to that of real α-helical proteins. Quantitatively very similar results are found. Furthermore, we examine the overall topology of the simulated sequences. As will be shown in section 6.3, we find topologies corresponding to 2- and 3-helix bundles. By looking at the C_α-C_α distance distribution functions, it can be seen that even the linear dimensions of our model proteins are similar to those of real ones.

Sequences

We have studied the behavior of two sequences in this model. These were deliberately chosen so as to be consistent with α-helical structure. Specifically, we start from a sequence segment of length 14, which is known to occur in amphiphilic α-helices [15]. This segment has been used earlier for the design of (real) de novo α-helical proteins [13]. We take the first 9 positions of this $N = 14$ segment as our basic building block. Our two sequences are then obtained by taking two (three) copies of this building block, and connecting them by hydrophilic segments of length three. These sequences have length $N = 21$ and 33 respectively, and are given in Table 6.1, where the spaces are meant to clarify the construction of the sequences. It should be obvious that these are meant as templates for 2- and 3-helix bundles, respectively.

Length N	sequence $\sigma_1, ..., \sigma_N$
21	101100110 111 101100110
33	101100110 111 101100110 111 101100110

Table 6.1: The two sequences studied. Basic building blocks of length 9 are connected by hydrophilic segments of length 3.

Monte Carlo Methods: Simulated Tempering

We calculate thermodynamic properties by using the method of simulated tempering [10, 16, 17]. To simulate proteins is notoriously difficult, due to the presence of a rugged energy landscape. One tries to overcome this problem in simulated tempering by treating the temperature as a fluctuating variable. This means that one simulates a joint distribution in conformation and temperature, which is taken to be

$$P(\vec{b}, k) \propto \exp(-g_k - E(\vec{b}, \sigma)/T_k)\,, \tag{6.3}$$

where T_k, $k = 1, \ldots, K$, are the allowed values of the temperature. The g_k's are free parameters that determine the probabilities $P(T_k)$ of visiting the different temperatures. Although the method is free from systematic errors for any set of g_k, it is crucial for the performance of the algorithm to make a careful choice of these parameters. In our simulations, the parameters g_k were adjusted so as to have a roughly uniform distribution $P(T_k)$. This was done by means of trial runs. Details may be found in [10].

The simulations reported in this paper were carried out using a set of $K = 20$ temperatures, ranging from $T_1 = 0.15$ to $T_{20} = 1.0$ with $1/T_i - 1/T_{i+1}$ constant. On a DEC ALPHA 200 (266 MHz), they took around 30 hours for $N = 21$ and around 500 hours for $N = 33$. In both cases, around 20 percent of the total computing time was spent on tuning of the weights g_k. We also simulated a sequence of length $N = 45$, whose construction is analogous to that of the other two sequences. We spent 1000 CPU hours on this sequence, and believe that we obtained reliable estimates for its local properties. However, the results for global properties must be interpreted with care for this sequence. By contrast, we feel very confident that the simulations for $N = 21$ and $N = 33$ are under control.

6.3 Results

The overall thermodynamic behavior of our model turns out to closely resemble that of the AB model [11]. In particular, the chains compactify gradually with decreasing temperature, and the folding transition takes place in the compact phase. Furthermore, we do not observe bimodal distributions in energy, compactness or the local interactions $E_l(i)$. All these observations are in agreement with earlier results for the AB model. By contrast, our model differs significantly from the AB model when it comes to structural properties. We shall therefore focus on these. Local structure as well as global topology and stability will be discussed in some detail.

Local Structure: Comparison with Real Proteins

In this section we examine the local structure of the chains in our model, and compare it to that of real α-helical proteins. The protein data were extracted using the C_α coordinates for 27 proteins containing mainly α-helices and almost no β-sheets. All these 27 proteins are listed in the all-α-helical class of the SCOP database [18]. The corresponding structures were taken from the Brookhaven Protein Data Bank (PDB) [19] [1]. One of these proteins, the $N = 54$ DNA-binding protein 1enh, which contains three α-helices [20], will be considered in more detail.

The local structure will be probed in two ways, by bend and torsional angle distributions and by bond-bond correlations.

In Fig. 6.2, left plot, bend and torsional angle distributions are displayed for the set of 27 α-helical proteins. The dominance of α-helical structure manifests itself in the pronounced peak centered at $\alpha \approx -129°$ and $\tau \approx 90°$. The same distributions, for our $N = 21$ sequence, are plotted on the right side of Fig. 6.2. The results for the other two simulated sequences are very similar. This is in line with the behavior of the AB model, where the angle distributions are

[1] The PDB accession codes are: 3sdh, 1ctj, 1enh, 2erl, 2end, 1lis, 1hme, 1nfn, 1bcf, 1rhg, 1acp, 1rop, 1coo, 4icb, 1utg, 1fia, 2wrp, 2tct, 1fps, 1ecm, 1aep, 1axn, 1hiw, 2abk, 1eci, 2abd, 1c5a.

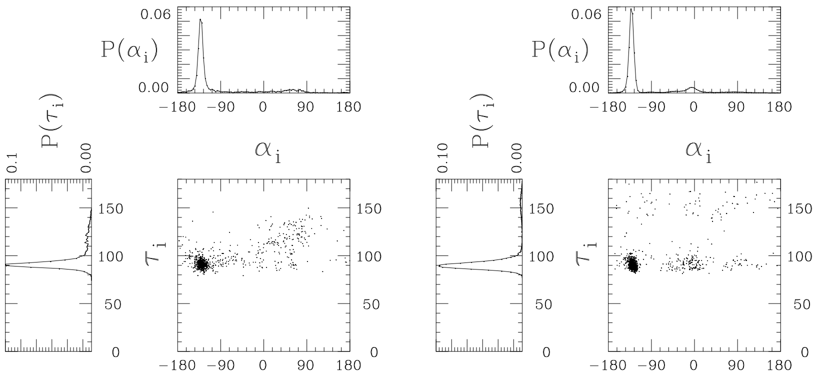

Figure 6.2: Bond (τ_i) and torsional (α_i) angle distributions. Left: For the 27 α-helical proteins taken from the PDB database. Right: Model results for the $N=21$ sequence at $T=0.15$. The other two sequences simulated, with $N=33$ and $N=45$, give very similar distributions. Both scatter plots contain 1000 data points.

fairly sequence-independent as well [11]. Not surprisingly, we find that there is indeed a peak in the distribution centered at the position of the ideal α-helix. This peak contains roughly the same amount of probability as for real α-helical proteins. Moreover, the width of this peak is very similar to that of the experimental structure, which justifies our ad hoc value for the width w of the exponential in Eq. 6.2. Another and much weaker peak in the angle distribution for our model is found at $(\alpha,\tau)=(0°,90°)$, which is an artifact of our model not present in real proteins.

Our second way to monitor local structure are the bond-bond correlations, which are studied using the function

$$C_b(d) = \left\langle \frac{\vec{b}_i \cdot \vec{b}_{i+d}}{|\vec{b}_i||\vec{b}_{i+d}|} \right\rangle \qquad C_b(d) \in [-1,1], \tag{6.4}$$

where $\langle\cdot\rangle$ denotes the average over all positions i and observed structures (with d fixed). $C_b(d)$ is a measure of the average alignment of bond vectors at a given topological distance d along the chain. It is normalized so that $C_b(0)=1$. In Fig. 6.3 (dotted line, left plot), we show the correlation function $C_b(d)$ for the 27 α-helical proteins. As can be seen from this figure, there are significant

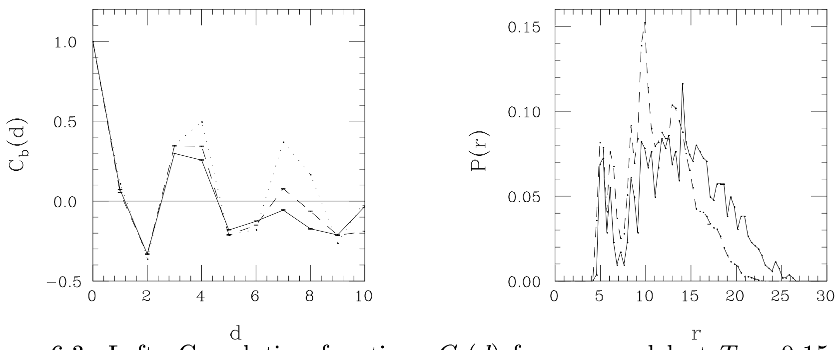

Figure 6.3: Left: Correlation functions $C_b(d)$ for our model at $T = 0.15$ for $N = 21$ (solid line) and $N = 33$ (dashed line). $C_b(d)$ for the 27 α-helical proteins listed in the text is given by the dotted line. Right: C_α-C_α distance distribution $P(r)$ for the (molecular dynamics refined) experimental structure of the DNA binding protein 1enh ($N = 54$, solid line), and our model ($N = 45$, dashed line) at $T = 0.15$. Nearest-neighbor pairs are not included.

correlations at least out to separations of about ten residues. The oscillations can be related to the presence of α-helical structure, which has a period of 3.6. In [11] it was shown that the AB model gives roughly the right correlation length, but fails to reproduce the periodicity of 3.6. $C_b(d)$ for our model is shown in Fig. 6.3, using $N = 21$ (solid line) and $N = 33$ (dashed line). Although the correlations decay somewhat faster in the model, the results are very similar to those for the real proteins.

Global Structure and Stability

It is an essential feature of any plausible model for protein folding that it allows for structural stability. In the following section, we examine our model with respect to this important aspect using two methods. First, we study structural stability by measuring mean square fluctuations, $\langle \delta^2 \rangle$. Second, the structure of the folded conformations will be characterized by using a measure for α-helix formation.

A commmonly used measure of the similarity between two conformations a and b is the mean square distance δ^2_{ab}, which is defined as

$$\delta^2_{ab} = \min \frac{1}{N} \sum_{i=1}^{N} |\bar{x}_i^{(a)} - \bar{x}_i^{(b)}|^2 \tag{6.5}$$

where $|\bar{x}_i^{(a)} - \bar{x}_i^{(b)}|$ denotes the distance between the sites $\bar{x}_i^{(a)}$ and $\bar{x}_i^{(b)}$, and where the minimum is taken over translations and rotations. The probability distribution $P(\delta^2)$, measured on a thermodynamic ensemble with fixed

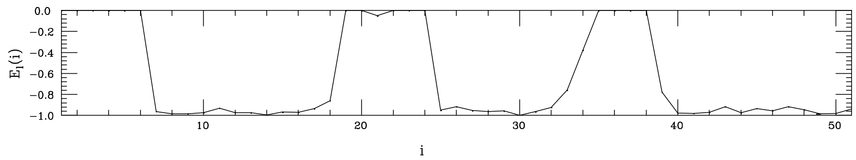

Figure 6.4: The function $E_l(i)$ of Eq. 6.2 for the experimental structure of DNA binding protein, PDB accession code 1enh ($N=54$). One easily identifies three α-helices from this plot; for α-helical structure $E_l(i)$ is close to minus one. The resulting α-helix assignments are in agreement with those in the PDB entry.

sequence and fixed temperature, gives valuable information about structural stability [6]. In particular, if there is a number of relatively well-defined structures present, the distribution will have a narrow peak close to $\delta^2 \approx 0$; in this case the ensemble contains many pairs of similar conformations.

Our second method for characterizing structure and stability is to monitor the formation of α-helical structure. A convenient way of doing this is to utilize the function $E_l(i)$ (Eq. 6.2); besides using it for modelling hydrogen-bonding in the α-helical structure, $E_l(i)$ can also be used to detect α-helices in a given structure. To illustrate this, we plot $E_l(i)$ for the experimental structure of the DNA binding protein 1enh in Fig. 6.4. From this plot, it is immediately clear that the structure contains three α-helices. The first α-helix is identified by the drop of $E_l(i)$ from zero to slightly above minus one in the region from position $i=7$ to $i=19$. It should be observed that 16 amino acids participate in this helix; the positions of four successive C_α-atoms determine one value of $E_l(i)$. Below, we will perform a similar analysis for our model proteins.

For the $N=21$ sequence at $T=0.15$, the distribution $P(\delta^2)$ has a mean value of $\langle\delta^2\rangle = 3.6 \pm 0.5\text{Å}^2$. This result may be compared to those of [11] for the AB model. There, six different $N=20$ sequences were studied, and $\langle\delta^2\rangle$ varied between 1Å^2 and 10Å^2. It should be pointed out, however, that the size of the monomers, as measured by the dimensionless ratio $A_{ij}/r_{i,i+1}$, is slightly larger in the present model (it is equal to one in the AB model).

One could be tempted to conclude from the $P(\delta^2)$ distribution, which is shown in Fig. 6.5 (right, dotted line), that the $N=21$ system has two similar but distinct states. A closer analysis reveals that there are actually four distinct states; the system spends almost all time, 95%, in the near vincinities of these. This can be shown by measuring the simultaneous δ^2 distances to all these four states. The mutual distances between these states range from 1.5Å^2 to 3.7Å^2. When it comes to helix content, these four states are very similar. The function $E_l(i)$, as displayed in Fig. 6.6, makes it evident that the overall structure is that of a two helix bundle. We also looked at the expectation values of the bend and torsional angles and their variances, which showed that the helices

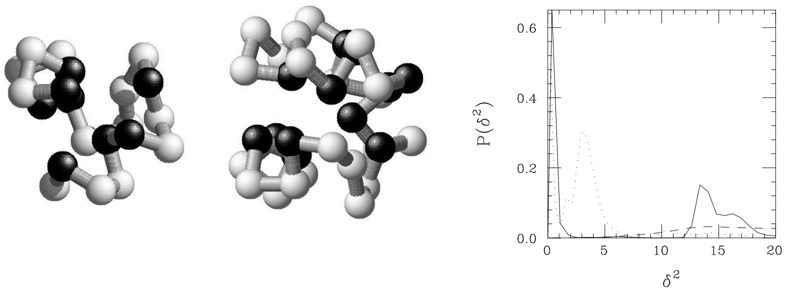

Figure 6.5: Representative structures for the most probable states for $N = 21$ (left) and $N = 33$ (middle). Hydrophobic and hydrophilic amino acids are represented by black and white spheres, respectively. On the right side, $P(\delta^2)$ is plotted for (1) $N = 21$, $T = 0.15$ (dotted line), (2) $N = 33$, $T = 0.15$ (solid line), and (3) $N = 33$, $T \approx 0.40$ (dashed line).

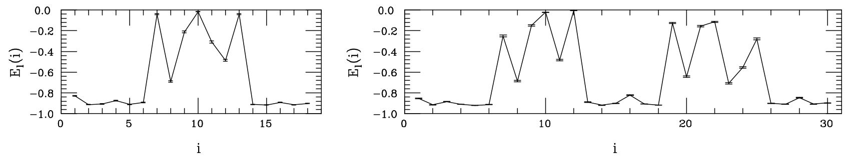

Figure 6.6: Thermodynamic averages of the local interactions $E_l(i)$ at $T = 0.15$ for $N = 21$ (left) and $N = 33$ (right). These curves clearly show that the dominating states for $N = 21$ and $N = 33$ contain 2 and 3 helices, respectively. This is also supported by the fact that fluctuations of bend and torsional angles are small in the helical regions.

are indeed very stable and common to the four different states. The structure of one of the four states is shown in Fig. 6.5 (the other three are very similar). From Fig. 6.5 it can also be seen that (1) we succeed to obtain a hydrophobic core and (2) the α-helices show the patterning intended in the construction of the sequences.

The $P(\delta^2)$ distribution for $N = 33$, see Fig. 6.5 (right plot), also has a pronounced peak at $\delta^2 \approx 0$. However, a significant amount of probability is also found around $\delta^2 \approx 15\text{Å}^2$. The existence of this outlier in $P(\delta^2)$ shows that the overall geometry must be different for the different states present. A detailed analysis, similar to that for $N = 21$, shows that $P(\delta^2)$ is dominated by three states having mutual distances of around $13.8, 15.0$ and 16.3Å^2. Taken together, these three states contain 97% probability at $T = 0.15$, distributed as 68%, 17% and 12%. These three states share three common α-helices; the positions of the helices along the chain can be seen from Fig. 6.6. How do the three states differ? For the most probable state (68%), the three helices are coordinated in such a way that the overall topology has right-handed chirality , as can be seen from Fig. 6.5, middle plot. For this state, 26 monomers participate in the three helices. The other two states are described as follows: the first two helices and the connecting loop form a U-shaped entity (the helices are not really parallel). The third helix points back into the U, for one state on top and for the other below the U-plane. The right-handed state is slightly less probable (12%) than the left-handed one (17%). The helices of the most probable state are present in these two states as well; however, the helices of these two states are longer with 31 and 33 monomers participating in the sense that 22 and 24 values of $E_l(i)$ are close to minus one.

Knowing about this three-fold structural degeneracy could be of interest for protein design. If the design is done in terms of the hydrophobicity pattern, as for example in [13], the designer could easily knock out two unwanted states by introducing unfavorable residues.

As stated in section 6.2, we also tried to simulate a sequence with $N=45$, intended to design 4-helix topology. 1000 CPU hours were insufficient to obtain fully reliable results. However, we feel confident that we managed to adjust the weights g_k pretty well and obtained a selection of low-energy structures that we rate as representative when it comes to local structure and overall size of the chain. To give an idea about the size of our model chains, we display the C_α-C_α distance distribution function in Fig. 6.3 (right figure, dashed line). In this graph, we compare it with the same distribution for the DNA binding protein 1enh, which contains 54 amino acids. The general shape of the two curves is similar, taking into account the shorter length of the model protein. This shows that even the size of our model proteins resembles that of real ones.

6.4 Conclusions

A minimal off-lattice model for α-helical proteins has been presented. α-helical structure is obtained by setting local LJ parameters to values consistent with α-helical geometry and introducing local interactions modelling chirality and α-helical hydrogen bonding. The global LJ parameters are determined by the size of an average amino acid.

Thermodynamic results for two sequences of length 21 and 33 have been presented. The construction of these sequences was inspired by the method of Kamtekar [13] to design real de novo proteins. This turns out to be a successful strategy here as well, since we observe α-helices at the intended positions along the chains. Furthermore, the helices observed in our model show the intended hydrophobicity pattern with one side being hydrophobic and the other hydrophilic; the hydrophobic sides of the helices point inwards.

The local structure has been explored using angle distributions and bond-bond correlations. When we compare results from our model with those for real α-helical proteins, these observables are very similar.

By inspecting the mean square distribution, $P(\delta^2)$, it has convincingly be shown that the $N = 21$ sequence is structurally stable. The overall topology was found to be that of a 2-helix bundle. For $N=33$, we found three globally different states which dominate the low temperature phase. These three states have three α-helices in common and differ from each other by the orientation of these helices.

Simulated tempering is the key to our ability investigate the low-temperature thermodynamics. The simulations took 30 CPU hours for $N=21$ and 500 CPU hours for $N = 33$. 1000 CPU hours were not sufficient to obtain reliable results for $N=45$, but we feel confident that we managed to properly adjust the weights g_k. Therefore, we believe that simulating the $N = 45$ sequence in a reliable way will be possible in the near future. Also, we feel confident that the simulations can be speeded up by further algorithmic improvements, such as a better way of distributing the temperatures T_k (see e.g. [21]).

6.5 Appendix

The Ideal α-Helix: Model Parameters

Five parameters of our model are determined by the geometry of an ideal α-helix. A sixth parameter is determined by the average size of an amino acid. In this section, the calculation of these parameters is presented.

A discrete helix is uniquely described by its bend and torsional angles, or, equivalently, by its pitch, the number of entities per turn and its chirality. An ideal α-helix has a translation of 5.4Å and 3.6 amino acids per turn with right-handed chirality, as is illustrated in Fig. 6.7. This is easily reformulated in terms of bend and torsional angles which we denote by $\hat{\tau}$ and $\hat{\alpha}$, respectively.

In order to make everything well-defined, we have to give an exact definition of bend and torsional angles. Three succesive bond vectors $\vec{b}_i, \vec{b}_{i+1}, \vec{b}_{i+2}$ define the two bend angles τ_i,τ_{i+1} and the torsional angle α_i. The bend angle τ is defined to be 180° if the two involved bond vectors are parallel. The torsional angle α is taken to be the angle between the two vectors $\vec{b}_i \times \vec{b}_{i+1}$ and $\vec{b}_{i+1} \times \vec{b}_{i+2}$. Therefore, $\alpha=0$ if and only if the three involved bond vectors are situated in a plane. α ranges from $-180°$ to $180°$, and is chosen to be negative for a right-handed and positive for a left-handed system.

A single LJ term in Eq. 6.1,

$$4\epsilon\left(\left(\frac{A_{ij}}{r_{ij}}\right)^{12} - \left(\frac{A_{ij}}{r_{ij}}\right)^{6}\right), \tag{6.6}$$

Figure 6.7: Illustration of an ideal α-helix: The polypeptide backbone has 3.6 amino acids per turn with a pitch of 5.4Å, as indicated by the height of the cylinder. The cylinder has a diameter of 4.5..Å.

has its minimum of depth ϵ at $r_{ij} = A_{ij} \cdot 2^{1/6}$; the minimum can be moved by changing A_{ij}. We adjust A_{ij} for $|i-j| \leq 4$ such that the minimum is situated at the distance observed in an ideal α-helix, r^{α}_{ij}. For all other pairs, $|j-i| > 4$, we choose the LJ parameter A_{ij} by tuning the volume of the LJ sphere to that of an average amino acid, 161Å^3:

$$|i-j| > 4 \;\Rightarrow\; A_{ij} = 2^{-1/6} \times \sqrt[3]{\frac{3}{4\pi} 161\text{Å}^3} \approx 6.01\text{Å} \tag{6.7}$$

To summarize, one obtains the following values for $\hat{\tau}$, $\hat{\alpha}$, r^{α}_{ij} and A_{ij}:

$\hat{\tau} \approx 90.53°$ $\hat{\alpha} \approx -129.61°$

	$\lvert j-i\rvert = 1$	$\lvert j-i\rvert = 2$	$\lvert j-i\rvert = 3$	$\lvert j-i\rvert = 4$	$\lvert j-i\rvert > 4$
r^{α}_{ij} [Å]	3.8	5.39..	5.04..	6.19..	-
A_{ij} [Å]	-	4.80..	4.49..	5.52..	6.01..

The AB model

The α-helical model presented in this paper is derived from the 3D AB model suggested recently by Irbäck et al. [11], which in turn is closely related to the 2D model proposed by Stillinger et. al [9,10]. To make comparisons easier, we here give the energy functions for both the AB model and the present model;

$$
\begin{aligned}
E(\vec{b};\sigma) &= \sum_{i=1}^{N-2}\sum_{j=i+2}^{N} 4\epsilon(\sigma_i,\sigma_j)\left(\left(\frac{A_{ij}}{r_{ij}}\right)^{12} - \left(\frac{A_{ij}}{r_{ij}}\right)^{6}\right) \\
&\quad -\kappa_1 \sum_{i=1}^{N-2} \frac{\vec{b}_i\cdot\vec{b}_{i+1}}{|\vec{b}_i||\vec{b}_{i+1}|} - \kappa_2 \sum_{i=1}^{N-3} \frac{\vec{b}_i\cdot\vec{b}_{i+2}}{|\vec{b}_i||\vec{b}_{i+2}|} \\
E(\vec{b};\sigma) &= \sum_{i=1}^{N-2}\sum_{j=i+2}^{N} 4\epsilon(\sigma_i,\sigma_j)\left(\left(\frac{A_{ij}}{r_{ij}}\right)^{12} - \left(\frac{A_{ij}}{r_{ij}}\right)^{6}\right) \\
&\quad - \sum_{i=1}^{N-3} \exp\left(-\frac{(\alpha_i-\hat{\alpha})^2+(\tau_i-\hat{\tau})^2+(\tau_{i+1}-\hat{\tau})^2}{2w^2}\right)
\end{aligned}
$$

They differ only in the choice of A_{ij} and in the local interactions; for the AB model, $A_{ij} = 3.8$Å independent of i and j. The dependence of the AB model on the strengths of the local interactions (κ_1, κ_2) was discussed to some extent in [11], where it was found that the local interactions are necessary in order to obtain regularities in the local structure and thermodynamic stable sequences. The final choice of (κ_1, κ_2) in [11] was $(\kappa_1, \kappa_2) = (-1, 0.5)$.

Acknowledgements

This project is supported by the “Swedish Foundation for Strategic Research”. Thanks to Anders Irbäck and Carsten Peterson for useful discussions. I appreciate helpful comments on the manuscript by Anders Irbäck, Robert van Leeuwen and Peter Sutton. This project would not have been started without the encouragement from Conrad Newton.

Bibliography

[1] See e.g. M. Karplus and A. Šali, "Theoretical Studies of Protein Folding and Unfolding", *Curr. Opin. Struct. Biol.* **5**, 58 (1995).

[2] K.F. Lau and K.A. Dill, "A Lattice Statistical Mechanics Model of the Conformational and Sequence Spaces of Proteins", *Macromolecules* **22**, 3986 (1989).

[3] C.J. Camacho and D. Thirumalai, "Minimum Energy Compact Structures of Random Sequences of Heteropolymers", *Phys. Rev. Lett.* **71**, 2505 (1993).

[4] K.A. Dill, S. Bromberg, K. Yue, K.M. Fiebig, D.P. Yee, P.D. Thomas and H.S. Chan, "Principles of Protein Folding — A Perspective from Simple Exact Models", *Protein Sci.* **4**, 561 (1995).

[5] A. Irbäck and E. Sandelin, "Local Interactions and Protein Folding: A Model Study on the Square and Triangular Lattices", *J. Chem. Phys.* **108**, 2245 (1998).

[6] G. Iori, E. Marinari and G. Parisi, "Random Self-Interacting Chains: A Mechanism for Protein Folding", *J. Phys.* **A 24**, 5349 (1991).

[7] M. Fukugita, D. Lancaster and M.G. Mitchard, "Kinematics and Thermodynammics of a Folding Heteropolymer", *Proc. Natl. Acad. Sci. USA* **90**, 6365 (1993).

[8] T. Veitshans, D.K. Klimov and D. Thirumalai, "Protein Folding Kinetics: Time Scales, Pathways, and Energy Landscapes in Terms of Sequence Dependent Properties", *Fold. Des.* **2**, 1 (1997).

[9] F.H. Stillinger, T. Head-Gordon and C.L. Hirschfeld, "Toy Model for Protein Folding", *Phys. Rev.* **E48**, 1469 (1993).

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

[11] A. Irbäck, C. Peterson, F. Potthast and O. Sommelius, "Local Interactions and Protein Folding: A 3D Off-Lattice Approach", *J. Chem. Phys.* **107**, 273 (1997).

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

[13] S. Kamtekar, J.M. Schiffer, H. Xiong, J.M. Babik and M.H. Hecht, "Protein Design by binary Patterning of polar and nonpolar Amino Acids", *Science* **262**, 1680 (1993).

[14] T. Creighton, *Proteins: Structures and Molecular Properties* (Freeman, New York, 1993).

[15] M.W. West and M.H. Hecht, "Binary Patterning of Polar and Nonpolar Amino Acids in the Sequences and Structures of Native Proteins", *Protein Sci.* **4**, 2032 (1995).

[16] A.P. Lyubartsev, A.A. Martsinovski, S.V. Shevkunov and P.N. Vorontsov-Velyaminov, "New Method to Monte Carlo Calculation of the Free Energy: Method of Expanded Ensembles", *J. Chem. Phys.* **93**, 1776 (1992).

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

[18] A.G. Murzin, S.E. Brenner, T. Hubbard and C. Chothia, "SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures", *J. Mol. Biol.* **247**, 536 (1995).

[19] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, "The Protein Data Bank: A Computer Based Archival File for Macromolecular Structures", *J. Mol. Biol.* **112**, 535 (1977).

[20] C.R. Kissinger, B. Liu, E. Martin-Blanco, T.B. Kornberg and C.O. Pabo, "Crystal Structure of an engrailed Homeodomain-DNA Complex at 2.8 Resolution: A Framework for understanding Homeodomain-DNA Interactions", *Cell* **63**, 579 (1990).

[21] U.H.E. Hansmann, "Parallel Tempering Algorithm for Conformational Studies of Biological Molecules", *Chem. Phys. Lett.* **281**, 140 (1997).

Paper VII

A Novel Monte Carlo Procedure for Protein Design

Anders Irbäck, Carsten Peterson,
Frank Potthast and Erik Sandelin
Complex Systems Group, Department of Theoretical Physics,
University of Lund, Sölvegatan 14A,
S-223 62 Lund, Sweden

LU TP 97-33
Submitted to *Physical Review* **E**

Abstract:

A new method for sequence optimization in protein models is presented. The approach, which has inherited its basic philosophy from recent work by Deutsch and Kurosky [*Phys. Rev. Lett.* **76**, 323 (1996)] by maximizing conditional probabilities rather than minimizing energy functions, is based upon a novel and very efficient multisequence Monte Carlo scheme. By construction, the method ensures that the designed sequences represent good folders thermodynamically. The algorithm, which is successfully explored on the two-dimensional HP model with chain lengths $N = 16$, 18 and 32, can easily be generalized to off-lattice models. Also, a bootstrap procedure for the sequence space search is devised, which makes very large chains feasible.
PACS numbers: 87.15.By, 87.10.+e

7.1 Introduction

The "inverse" of protein folding, sequence optimization, is of utmost relevance in the context of drug design. This problem, which amounts to finding optimal amino acid sequences given a target structure, has also been investigated in the context of understanding folding properties of coarse-grained models for protein folding. Such models are described by energy functions $E(r, \sigma)$, where $r = \{\mathbf{r}_1, \mathbf{r}_2, .., \mathbf{r}_N\}$ denotes the amino acid coordinates and $\sigma = \{\sigma_1, \sigma_2, .., \sigma_N\}$ the amino acid sequence.

Good folding sequences fold fast and in a stable way into the desired target structure. A brute force search for sequences meeting these criteria is prohibitively time-consuming even in minimalist models for protein folding. Although it has been possible to apply this type of criteria to a simple helix-coil model [1], it is essential to find more efficient strategies. A fairly drastic simplification was proposed in [2], where the problem is approached by minimizing E with respect to σ with r clamped to the target structure, r_0. This method is very fast since no exploration of the conformational space is involved, but, unfortunately, it fails for a number of examples (see e.g. [1, 3, 4]). Recently, a more generic scheme was suggested [3], which aims at optimizing the conditional probability $P(r_0|\sigma)$, i.e. the Boltzmann weight, rather than $E(r_0, \sigma)$. Using statistics language this corresponds to Bayesian regression rather than fitting a model to data points (desired structure), with the width of the probability distribution given by the temperature T. This approach has the advantage that entropy effects are taken into account, but its usefulness is not obvious since maximizing $P(r_0|\sigma)$ is a non-trivial task. In fact, the calculations in [3] involved simplifying assumptions about both the form of $P(r_0|\sigma)$ and the conformational space. In this letter we present a practical Monte Carlo (MC) procedure for performing the maximization of $P(r_0|\sigma)$.

Thermodynamical characteristics for good folders are that the ground state minima are well separated from other states — at finite T the system spends a long time in the ground state well. In lattice models, where for relatively small chains the states are enumerable, this is often taken as non-degeneracy of ground state and that the latter is well separated from higher energy states. One expects that working with finite T distributions in the matching process singles out those optimal sequences that have good folding properties in terms of non-degeneracy. Indeed, in [3] when exploring the technique on lattice models, superior results are obtained when comparing with what was obtained in [2], where $E(r_0, \sigma)$ was minimized.

Computationally, straightforward MC approaches for maximizing $P(r_0|\sigma)$ are extremely tedious. Our novel MC methodology is based on the *multisequence method* [5], where both sequence and coordinate degrees of freedom are subject to simultaneous moves. The basic idea is to perform a single simulation of a joint probability distribution $P(r, \sigma)$ rather than repeated simulations of

$P(r|\sigma)$ for different fixed σ. Hence, our approach is fundamentally different from that of [4].

Our method for maximizing $P(r|\sigma)$ is explored on the two-dimensional HP lattice model [6] with chain lengths $N = 16$, 18 and 32. For N=16 we study an example used in [2–4]. The results for both N=16 and 18 are checked against exact enumerations, whereas for $N = 32$ we use a target structure constructed "by hand". Our method reproduces the exact results extremely rapidly whenever comparisons are feasible. Furthermore, the method has quite some potential to deal with very large chains.

7.2 The Basic Idea

The problem of finding thermodynamically optimal sequences given a target structure r_0 is simple to formulate mathematically — maximize with respect to σ the conditional probability

$$P(r_0|\sigma) = \frac{1}{Z(\sigma)} \exp(-E(r_0,\sigma)/T) \tag{7.1}$$

$$Z(\sigma) = \sum_r \exp(-E(r,\sigma)/T) \tag{7.2}$$

where Eq. (7.1) can be rewritten in terms of the free energy $F(\sigma) = -T \ln Z(\sigma)$ as

$$P(r_0|\sigma) = \exp[-(E(r_0,\sigma) - F(\sigma))/T] \tag{7.3}$$

Hence, for each σ one needs to estimate $P(r_0|\sigma)$ which in turn involves a sum over all possible r. The situation is shown in Fig. 7.1, where the horizontal line represents the region probed in protein folding. In the simplified approach to the inverse problem [2], minimizing $E(r_0,\sigma)$, one works along the vertical line. Maximizing $P(r_0|\sigma)$ is a real challenge since it requires sampling of the entire (r,σ)-plane. Refs. [3,4] approached this problem by using simulated annealing

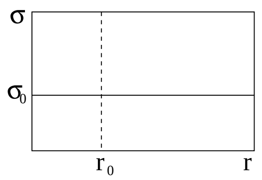

Figure 7.1: The (r,σ)-plane. The horizontal and vertical lines represent protein folding and "naive" sequence optimization respectively (see text).

in sequence space. The key difficulty then is to estimate the partition function $Z(\sigma)$ (Eq. (7.2)). In [3] this was done using the lowest-order term in the cumulant expansion of $F(\sigma)$. This approximation is valid at high temperature but it is unclear how good it is in the temperature regime of interest here. Ref. [4], on the other hand, used a chain growth MC method to estimate $Z(\sigma)$. In this case one has a nested MC, where the inner part by itself is far from trivial.

In [3,4] these methods were successfully tested on examples where a simple minimization of $E(r_0, \sigma)$ along the vertical line in Fig. 7.1 fails. The chains were short enough for the results to be tested against exact enumerations. Another difference between optimizing $P(r_0|\sigma)$ versus $E(r_0, \sigma)$ is that the latter requires an optimization constrained to a preset net hydrophobicity.

In this letter we take a quite different path capitalizing on the multisequence method [5]. Here the basic strategy is to create an enlarged configuration space; the sequence σ becomes a dynamical variable. Hence, r and σ are put on a more equal footing, which, in particular, enables us to avoid a nested MC.

Our starting point is the joint probability distribution

$$P(r, \sigma) = \frac{1}{Z} \exp(-g(\sigma) - E(r, \sigma)/T) \tag{7.4}$$

$$Z = \sum_{\sigma} \exp(-g(\sigma)) Z(\sigma) \tag{7.5}$$

where $\{g(\sigma)\}$ is a set of tunable parameters, which govern the marginal distribution

$$P(\sigma) = \sum_r P(r, \sigma) = \frac{1}{Z} \exp(-g(\sigma)) Z(\sigma) \tag{7.6}$$

From the Bayes relation $P(r, \sigma) = P(r|\sigma)P(\sigma)$ one obtains the desired conditional probabilities, Eq. (7.1), which of course are independent of the choice of $g(\sigma)$.

The choice of $g(\sigma)$ is crucial for the efficiency of the method. At first sight, it may seem that one would need to estimate $Z(\sigma)$ in order to obtain reasonable $g(\sigma)$. However, a convenient choice is

$$g(\sigma) = -E(r_0, \sigma)/T \tag{7.7}$$

Maximizing $P(r_0|\sigma)$ in Eq. (7.1) corresponds to minimizing the quantity

$$\Delta F_0 = -T \ln P(r_0|\sigma) = E(r_0, \sigma) - F(\sigma) \tag{7.8}$$

which for the choice in Eq. (7.7) can be rewritten as

$$\Delta F_0 = T \ln P(\sigma) + T \ln Z \tag{7.9}$$

Hence, neglecting an unimportant constant, ΔF_0 can be obtained directly from the marginal distribution $P(\sigma)$.

The joint distribution $P(r,\sigma)$ can be simulated by using separate ordinary r and σ updates. This single simulation of $P(r,\sigma)$ replaces simulations of $P(r|\sigma)$ for a number of different fixed σ. This is quite convenient. However, it should be stressed that the major motivation for using this scheme is its efficiency. In fact, it was demonstrated in [5] that the simulation of $P(r,\sigma)$ can be much faster than the simulation of $P(r|\sigma)$ even for *a single* σ; the exploration of the conformational space becomes more efficient when the sequence is allowed to fluctuate.

The number of sequences that can be studied in a multisequence simulation is of course limited. It is therefore desirable to incorporate a step in which "bad" sequences are removed. This elimination step can be formulated in different ways. In our calculations a sequence σ is removed as soon as some structure $r \neq r_0$ is encountered for which $E(r,\sigma) \leq E(r_0,\sigma)$. This process is continued until the remaining sequences can be studied through a final multisequence run. If the elimination proceeds for a sufficiently long time, then the surviving sequences are, by construction, those that have the desired structure as their non-degenerate ground state. It should be pointed out that the elimination process serves two purposes. In addition to bringing down the number of sequences to a manageable level, it also tends to make the distribution $P(\sigma)$ more uniform, which is instrumental for the efficiency of the final multisequence run. Note that for a set of sequences all having r_0 as their unique ground state, one has $\exp(-g(\sigma))Z(\sigma) \approx 1$ to leading order at low T, independent of σ, which implies that $P(\sigma)$ becomes uniform in the zero temperature limit.

7.3 Numerical Implementation and Results

We have performed extensive numerical explorations on the HP model [6] for a variety of sizes and target structures and find that the approach consistently identifies the appropriate sequences in an efficient way. In this letter we report on results for $N = 16$ and 18, which have been previously used for evaluating design algorithms [3,4]. Also results for $N = 32$ are presented in some detail.

The HP model [6] is defined by

$$E(r,\sigma) = \sum_{i<j} \epsilon_{\sigma_i \sigma_j} \Delta(r_i - r_j) \tag{7.10}$$

where $\Delta(r_i - r_j) = 1$ if r_i and r_j are nearest neighbor monomers but non-adjacent along the chain and zero otherwise. Dependent upon whether σ_i is hydrophobic (H) or polar (P), σ_i=1 and 0, respectively, one has $\epsilon_{\mathrm{HH}} = -1$ and $\epsilon_{\mathrm{PP}} = \epsilon_{\mathrm{HP}} = 0$. We work on the square lattice, for which it is known that sequences with non-degenerate ground states are not too rare [7]. All simulations are performed using standard Metropolis steps [8] in σ. In r we use three types of elementary moves: one-bead, two-bead and pivot [9, 10]. A sweep refers to a combination of these three move types followed by one attempt to update σ.

We first test our method for the $N = 16$ target structure studied in [4] (see Fig. 1 in [4]). There is one sequence which has this structure as its non-degenerate ground state, as can be shown by exact enumeration. It turns out that the design procedure in [4] is able to find this sequence, while the methods in [2, 3] fail to do so. Our calculation is carried out starting from the set of all 2^{16} possible sequences. After 8000 MC sweeps, corresponding to a few CPU seconds on a DEC Alpha 200, all sequences except the correct one have been removed.

Next, we turn to the $N = 18$ target structure shown in Fig. 7.2a. The seven sequences listed in Table 7.1 were found by the design procedure. Non-degeneracy is tested and confirmed as in the $N = 16$ case above. The multisequence simulation is then continued with only the seven sequences left in order to estimate $P(\sigma)$ and $P(r_0|\sigma)$. As can be seen from Table 7.1, the results agree very well with the exact results. Since all seven sequences have r_0 as their unique ground state, $P(\sigma)$ is constant for $T = 0$. At $T = 1/3$, the temperature studied here, the $P(\sigma)$'s are not perfectly equal, but similar enough to allow for good mobility in sequence space. In summary, our method has removed those sequences that do not have r_0 as their unique ground state, and it also provides $P(r_0|\sigma)$ for all the surviving sequences. Note that the remaining sequences all have the same monomer type at 12 of the 18 positions. These are marked by filled (H) and open (P) circles in Fig. 7.2a.

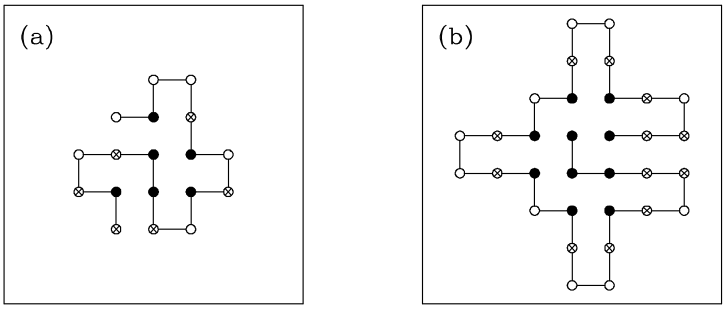

Figure 7.2: Target structures for (a) $N = 18$ and (b) $N = 32$. Symbols are explained in the text. The best sequence found for (b) is : HHPPHHPPPPHPH-PPPPHPHPPPPHHPPHHHH.

Sequence	$P(\sigma)$ MC	$P(\sigma)$ exact	$P(r_0\|\sigma)$ MC	$P(r_0\|\sigma)$ exact
PHPPPHPPHPPHHPPPHP	0.2101 ± 0.0008	0.211171	0.0175 ± 0.0005	0.017503
PHPPHHPPHPPHHPPPHP	0.0625 ± 0.0009	0.061675	0.0599 ± 0.0017	0.059929
PHPPPHPPHPHHHPPPHP	0.3104 ± 0.0018	0.311250	0.0120 ± 0.0003	0.011875
PHPPPHPPHPPHHHPPHP	0.0495 ± 0.0003	0.049458	0.0748 ± 0.0020	0.074733
PHPPPHPHHPPHHHPPHP	0.1765 ± 0.0019	0.175692	0.0211 ± 0.0006	0.021038
PHPPPHPPHPPHHHPHHP	0.1102 ± 0.0007	0.111024	0.0333 ± 0.0010	0.033291
PHPPPHPPHPHHHHPPHH	0.0807 ± 0.0019	0.079730	0.0473 ± 0.0017	0.046358

Table 7.1: The relative probabilities $P(\sigma)$ and the ground state occupancies $P(r_0|\sigma)$ for those seven $N = 18$ sequences that design the structure shown in Fig. 7.2a ($T = 1/3$). Listed are both the results from our multisequence simulation (MC) and the exact results, obtained by enumeration.

7.4 Results

In the second part of our simulations, where $P(r_0|\sigma)$ is estimated, it is clear that the stochastic sequence moves are essential. How useful these moves are in the first part, the elimination process, is less clear. To investigate this, we performed calculations both with and without these moves. In the latter case the simulated sequence is replaced only if it is to be removed from the simulation, and is then replaced by a randomly chosen sequence among the remaining ones. In Fig. 7.3 we show the number of MC sweeps needed to remove all sequences except those in Table 7.1, as a function of $1/T$. The results show that the required number of sweeps can be reduced by more than a factor of 10 by adding the stochastic sequence moves. Furthermore, the efficiency is less T dependent. The cost of the sequence moves is negligible.

We next turn to the $N = 32$ target structure shown in Fig. 7.2b, which is designed by hand since exhaustive enumeration is impracticable for this problem size. It is readily verified that this structure represents the minimum energy for the sequence with H at the filled circles and P at all the other positions along the chain (see Fig. 7.2b).

As with any other method, it is for large chains not feasible to explore the entire sequence space with our multisequence method. However, as in the $N = 18$ example above, a given structure typically exhibits several positions where σ_i is effectively frozen to H or P (see e.g. [11,12]). It turns out that such positions can be easily detected by means of a trial run. This leads us to a two-step bootstrap procedure.

The first step amounts to picking sets of random sequences and gauging σ_i for the surviving sequences. The σ_i profile obtained this way for the target structure in Fig. 7.2b is shown in Fig. 7.4, from which it is clear that indeed many σ_i exhibit a clear preference for either P or H. Based on this, we divide the positions along the chain into three groups corresponding to $\sigma_i > \sigma^{(1)}$ (filled

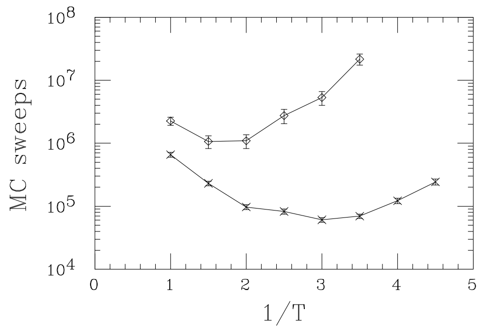

Figure 7.3: The number of MC sweeps needed to single out the seven sequences in Table 7.1. Each data point represents an average over 50 experiments, each started from the full set of all 2^{18} sequences but with different random number generator seeds. Shown are both results obtained with ($\times$) and without ($\diamond$) stochastic sequence moves.

circles in Fig. 7.2b), $\sigma_i < \sigma^{(2)}$ (open circles) and $\sigma^{(2)} < \sigma_i < \sigma^{(1)}$ (crosses), as indicated in Fig. 7.4. We then rerun the algorithm with those σ_i in the first two groups clamped to H ($\sigma_i = 1$) and P ($\sigma_i = 0$), respectively, and those in the third group left open, which corresponds to a set of 2^{12} sequences. In 20 CPU minutes ($5 \cdot 10^6$ MC sweeps) this set was reduced to 200 sequences. These can be readily studied in a final multisequence run, and the best sequence found is given in Fig. 7.2. Stability was confirmed by repeating the procedure for different random seeds. We also performed runs where the elimination process was continued for much longer, which finally contained 167 surviving sequences. While it could be that this is still not the proper asymptotic value, we feel confident that the best sequence is correctly identified by our method.

It is clear that this method can be generalized to a corresponding multi-step procedure for very large chains.

7.5 Summary

In summary, we have developed a new efficient MC method for protein design by maximizing conditional probabilities using the multisequence method. The method circumvents calculations of partition functions by a judicious choice of the multisequence sample weights. Large chains are feasible with the approach by means of a bootstrap procedure that limits the search in sequence space. The method, which is successfully explored on two-dimensional lattice models, easily lends itself to be used for off-lattice models [13]. Indeed, recent results for such

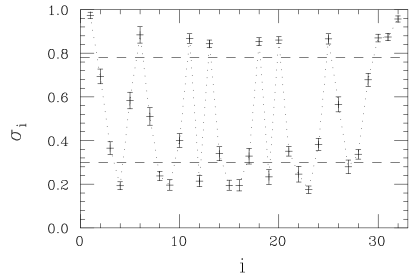

Figure 7.4: Average of σ_i against i for the surviving sequences from 10 runs, each started with a set of 10^5 random sequences ($N = 32$). The upper and lower lines represent $\sigma^{(1)}$ and $\sigma^{(2)}$, respectively (see text). The length of each run is 20000 MC sweeps (45 CPU seconds) and the number of surviving sequences varies between 15 and 57.

systems look very promising [14]. It should be mentioned that an alternative to pruning σ-space by removing sequences satisfying $E(r, \sigma) \leq E(r_0, \sigma)$ is to discard sequences with high $P(\sigma)$ (see Eq. (7.9)) [14].

Bibliography

[1] M. Ebeling and W. Nadler, *Proc. Natl. Acad. Sci. USA* **92**, 8798 (1995).

[2] E.I. Shakhnovich and A.M. Gutin, *Proc. Natl. Acad. Sci. USA* **90**, 7195 (1993).

[3] T. Kurosky and J.M. Deutsch, *J. Phys.* **A 27**, L387 (1995); J.M. Deutsch and T. Kurosky, *Phys. Rev. Lett.* **76**, 323 (1996)

[4] F. Seno, M. Vendruscolo, A. Maritan, and J.R. Banavar, *Phys. Rev. Lett.* **77**, 1901 (1996).

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

[6] K.F. Lau and K.A. Dill, *Macromolecules* **22**, 3986 (1989).

[7] K.A. Dill, S. Bromberg, K. Yue, K.M. Fiebig, P.D. Thomas and H.S. Chan, *Protein Sci.* **4**, 561 (1995).

[8] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth and A.H. Teller, *J. Chem. Phys.* **21**, 1087 (1953).

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

[10] A.D. Sokal, in *Monte Carlo and Molecular Dynamics Simulations in Polymer Science*, ed. K. Binder (Oxford University Press, New York, 1995).

[11] K. Yue and K.A. Dill, *Proc. Natl. Acad. Sci. USA* **92**, 146 (1995).

[12] H. Li, R. Helling, C. Tang and N. Wingreen, *Science* **273**, 666 (1996).

[13] A. Irbäck, C. Peterson, F. Potthast and O. Sommelius, *J. Chem. Phys.* **107**, 273 (1997).

[14] A. Irbäck, C. Peterson, F. Potthast, E. Sandelin and O. Sommelius, in preparation.

Leftovers

Although the results are not unexpected, the method is original and may have further applications.

Anonymous referee from *Protein Engineering*, on paper I.

The studied protein families are limited to clusters of close homologous sequences aligned with no gaps. At such low evolutionary distances, one can expect another coding bias due to the relative amino acid codon similarities.

Anonymous referee from *Protein Engineering*, on paper I.

The paper has been examined by two members of our editorial board with expertise in the relevant field. In their opinion, the conclusion reached, that proteins have non-random sequences, is well-known by biologists and structural biologists. Your paper adds little what is new and is therefore unsuitable for publication in JMB.

Editorial Office of *J. Mol. Biol.* , on Paper II.

This paper presents a solid statistical evidence that hydrophobic amino acids are anticorrelated along protein sequences.

Eugene I Shakhnovich in "Theoretical studies of protein-folding thermodynamics and kinetics", *Current Opinion in Structural Biology* 1997, **7**, on paper II.

This is a well done and important paper that addresses the question of whether proteins originate from random sequences. Previous investigations have indicated that this is the case.

Anonymous referee from *Proc. Natl. Acad. Sci. USA* , on Paper II.

In real proteins the local correlations in a denatured state and in the folded state are very similar. Paradoxically, the oversimplified lattice models of Dill, Shakhnovich and Karplus (and others) in some way account for this physical fact. This is not the case for the model proposed in the manuscript.

Anonymous referee from *The Journal of Chemical Physics*, on paper V.

To be brutally honest, my impression is that what is good about this manuscript is not new and what is new is not too good

Anonymous referee from *Physical Review E*, on paper IV.

....Thus the work is relevant and provides interesting new contributions to this field,....

Anonymous referee from *Physical Review E*, on paper IV.

In my opinion, the authors have quite adequately addressed all the concerns raised by previous referees. Any intelligent reader would realize that this is a study on a model, not on real proteins.

Anonymous referee from *Physical Review E*, on paper IV.

Va fan e Rimini mot Ribban o' en kasse bier

Jaques Werup.

ISBN 91-628-3071-6