Proceedings of the
International Conference on Complex Systems
Nashua, NH, 21–26 Sept. 1997.

Y. Bar-Yam (ed.), New England Complex Systems Institute (1997) .
Also on-line in the *InterJournal*.

Folding Properties in Off-Lattice Protein Models

Frank Potthast
Lund University, Theoretical Physics, Complex Systems Group.
Ola Sommelius
NORDITA, Copenhagen.

The thermodynamic behavior of an off-lattice model for protein folding is probed in two and three dimensions. The model has only two types of residues, hydrophobic and hydrophilic.

Our ability to map out the thermodynamics of the off-lattice models relies heavily upon the efficiency of the simulated tempering algorithm. In this approach, the temperature is a fluctuating variable, enabling the crossing of free energy barriers.

In absence of local interactions, native structure formation does not occur in 3D. By including sequence independent local interactions, which qualitatively reproduce local properties of real 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. 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 3D.

A statistical analysis of hydrophobicity patterns is performed. Deviations from randomness are found for real proteins. The same type of deviations from randomness are found for good-folding sequences in 2D. An interpretation of the deviations is given in terms of an Ising model.

1 Introduction

Proteins are biological macromolecules that perform a large variety of functions in living organisms. Their function is encoded in the well defined three dimensional structure which they attain under physiological conditions. It is known that this functional (native) structure is determined from the sequence of amino acids alone [1]. To understand how the native state is selected among the vast number of available conformations is a difficult problem, often denoted the protein folding problem, and one of the great challenges of molecular biology.

One way of unveiling central issues in the thermodynamics and kinetics of protein folding is to study simple, coarse grained models in which the amino acids constitute the basic entities. Models of this type that have gained increasing popularity are lattice models with contact interactions only. Despite their obvious shortcomings lattice models are able to catch several non-trivial features of the protein folding problem. This does however not imply that the appoximations involved are understood [8]. It is therefore important to pursue the study of alternative models.

Here we present results for an off-lattice model in two [5, 7] and three [9] dimensions. The model contains two types of amino acids, hydrophobic and hydrophilic. The key part of the interactions is a sequence dependent Lennard-Jones term that favors the formation of a hydrophobic core. In addition the model also contains sequence independent local interactions

A new Monte-Carlo scheme, simulated tempering [11], is crucial to our ability to map out the thermodynamics. In this approach, the temperature is a fluctuating variable, enabling the crossing of free energy barriers.

In 2D the folding properties of 300 sequences are studied. Out of these 10% can be classified as good folding sequences. Statistical methods and artificial neural networks are used to reveal functional dependencies between sequence patterns and the thermodynamic stability.

Hydrophobicity is known to play an important role in protein folding [10]. Therefore, a statistical analysis of the hydrophobicity patterns is performed. The results for good folding sequences are found to deviate from the results for random sequences in a way that strongly resembles the deviations found for real proteins.

The key finding in 3D is the necessity of local interactions for structural stability in the model. This property does not depend on the details of local interactions.

This paper is organized as follows: In §2 we describe the general form of the model, simulated tempering and the observables. Thermodynamic and kinetic properties of the 2D-model are discussed in §3. The analysis of hydrophobicity patterns is described in §4, followed by the results for the 3D-model in §5. We finish with a summary and outlook in §6.

2 Methods

2.1 The Model

Each residue is represented by a single site. The sites are linked by rigid bond vectors of unit length, $\hat{b}_i$, see Fig. 1. The shape of an N-mer is specified by the $N-1$ bond vectors $\hat{b}_i$.

The model contains two kinds of residues, A and B, which behave as hydrophobic (σ_i=+1) and hydrophilic ($\sigma_i = -1$) residues respectively. A binary string σ of length N therefore determines the sequence of an N-mer.

The energy function has the general form

$$E(\hat{b};\sigma) = E_l(\hat{b}) + E_g(\hat{b};\sigma), \quad (1)$$

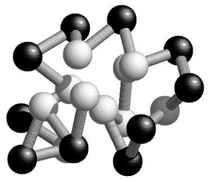

Figure 1: Illustration of hydrophobic core formation. The hydrophobic (A) and hydrophilic (B) monomers are represented by white and black spheres, respectively.

where the local interactions E_l are sequence-independent. The global interactions E_g are given by Lennard-Jones potentials. The exact forms will be given in §3.1 for 2D and §5.1 for 3D.

2.2 Monte Carlo: Simulated Tempering

Simulated tempering [11] is a method meant for systems with a rough free-energy landscape. In this approach, barriers of the free-energy landscape are circumvented by treating the temperature as a fluctuating variable.

To this end, the joint probability distribution

$$P(\hat{b},k) \propto \exp(-g_k - E(\hat{b},\sigma)/T_k) \quad (2)$$

is simulated, where T_k, $k = 1,\ldots,K$, are the allowed values of the temperature.

The g_k's are free parameters that determine the weights p_k of the different k values. In the simulations, we tried to adjust the g_k's for to get the p_k's uniformly distributed. Simulated tempering is free from systematic errors for any choice of the g_k's. Details may be found in [5]. When compared to Hybrid Monte Carlo [3], a speed up factor of almost 10^3 was observed [5] for $N = 10$ in 2D.

2.3 Observables

When talking about structural stability, one needs a measure for similarity between configurations. 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

$$\delta^2_{ab} = \min \frac{1}{N}\sum_{i=1}^{N} |\bar{x}_i^{(a)} - \bar{x}_i^{(b)}|^2. \quad (3)$$

The minimum is taken over translations, rotations and reflections. The corresponding distribution, $P(\delta^2)$, gives a good impression about the degree of structural stability[4].

The compactification of the chains is measured by the gyration radius $r^2_{\text{gyr}} = \frac{1}{N}\sum_{i=1}^{N}\left(\langle\bar{x}_i^2\rangle - \langle\bar{x}_i\rangle^2\right)$.

3 2D-simulations

3.1 The model in 2D

The general form of the energy function has been given in equation 1. Here, in the 2D-case, we use the form suggested by Frank Stillinger et al. [14]. The *sequence independent* local interactions $E_l(\hat{b})$ and the *sequence dependent* global interactions $E_g(\hat{b};\sigma)$ are given by

$$E_l(\hat{b}) = -\frac{1}{4}\sum_{i=1}^{N-2} \hat{b}_i \cdot \hat{b}_{i+1} \quad (4)$$

$$E_g(\hat{b};\sigma) = \sum_{i=1}^{N-2}\sum_{j=i+2}^{N} 4(r_{ij}^{-12} - C(\sigma_i,\sigma_j)r_{ij}^{-6}). \quad (5)$$

The sequence independent local interaction favors alignment of two successive bonds $\hat{b}_i$ and $\hat{b}_{i+1}$. The global interactions are sequence-dependent: $C(\sigma_i,\sigma_j)$, 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 [14] 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 role of the local interaction term $E_l(\hat{b})$ will be discussed below. In contrast to 3D, we will here find that $E_l(\hat{b})$ is not needed for to observe thermodynamically stable sequences.

3.2 Simulations: 300 2D-chains.

300 randomly selected $N = 20$ sequences containing 14 hydrophobic and 6 hydrophilic monomers have been simulated using simulated tempering. In these simulations, we used a set of $K = 13$ temperature values, ranging from $T = 0.15$ to $T = 0.6$. Each simulation took around 4 CPU hours on a DEC Alpha 2000.

In order to study the kinetic behavior of the 2D-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.

3.3 Thermodynamic and Kinetic Properties (2D)

At $T = 0.60$ the δ^2 distributions all have the same structureless broad appearance, the fluctuations in shape are large. At $T = 0.15$ $P(\delta^2)$ is, by contrast, strongly sequence dependent. Some chains exists in a state with well-defined shape, while others still undergo large structural fluctuations.

Although the chains have different structural stability, their extension is similar; the differences in radius of gyration are 10% or smaller at all the temperatures studied. The radius of gyration decreases gradually with decreasing temperature. No abrupt changes can be seen.

In order to check the importance of the local interactions we redid the simulations with $E_l = 0$ for 15 sequences.

We found strong correlation between the $\langle\delta^2\rangle$'s for these two types of potentials. Thus, $\langle\delta^2\rangle$ remains sequence dependent in 2D even if we leave out the local interaction term. This is in contrast to the behavior found in 3D, as will be discussed in §5.2.

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. (3)). The simulations are started from random coils, and as a criterion of successful folding we use the condition $\delta_0^2 < 0.3$. For to get an impression about the dynamics, we measure the probability of successful folding within a given time, called the foldicity [12]. For each of the 300 $N = 20$ 2D-sequences we studied five different temperatures, $T = 0.15, 0.18, 0.24, 0.34$ and 0.60.

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 $0.15 \le T \le 0.60$. 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. It thus appears that the kinetic behavior has a relatively weak sequence dependence.

3.4 Folding Criteria (2D)

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 thermodynamic behavior at this temperature. If the sequence is found to exhibit a well-defined structure at $\tilde{T}$, then it is a good folder. In our determination of good folders we take the kinetic quantity $\tilde{T}$ to be the same for all sequences. This means that our classification is entirely determined by the thermodynamic behavior at a fixed temperature, which we take as $\tilde{T} = 0.15$. Our motivation for this simplifying approximation is that the kinetic behavior appears to have a weak sequence dependence, as discussed above in §3.3.

Defining a sequence to be a good folder if

$$\langle\delta^2\rangle < \tilde{\delta}^2 \qquad \text{or} \qquad P(\delta^2 < 0.1) > \tilde{P}, \tag{6}$$

with $\tilde{\delta}^2 = 0.3$ and $\tilde{P} = 0.35$, yields 37 good folders of 300 sequences in total.

4 Evidence for non-randomness

In this paragraph, we turn to the statistical analysis of hydrophobicity patterns.

The question of whether proteins originate from random sequences of amino acids is addressed. A statistical analysis is performed 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 real proteins. The results convincingly show that the amino acid sequences in real proteins differ from random sequences in a statistically significant way.

We have also analyzed results from the 2D model described above in §3. Performing the same statistical analysis on the sequences that fold well indicates similar deviations from randomness as for the real 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 [15] 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 might give rise to energy landscapes with poor folding properties and hence might not have survived the evolution.

In the following two sections, Block Fluctuations (§4.1) and Fourier Analysis (§4.2), we describe the toolbox used for to check the binary sequences for randomness. Following this, we give the results for real proteins (§4.3), the results found for the 2D-model (§4.4) and an interpretation in terms of an Ising model (§4.5).

The fundamental concept of the analysis is to rewrite the binary sequence σ to a *hydrophobicity walk representation*,

$$r_n = \sum_{i=1}^{n} \sigma_i, \tag{7}$$

which is done in order to capture long range correlation properties.

4.1 Block Fluctuations

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

$$\sigma_i^{(s)} = \sum_{j=1}^{s} \sigma_{(i-1)s+j} = r_{is} - r_{(i-1)s} \qquad i = 1, \ldots, N/s \tag{8}$$

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 protein sequences with a random distribution of hydrophobic residues. For this reason we average over all sequences with a fixed length and composition.

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

$$\psi_i^{(s)} = \frac{1}{K}\left(\sigma_i^{(s)} - \frac{s}{N}\sum_{j=1}^{N/s}\sigma_j^{(s)}\right)^2 \qquad i = 1,\ldots,N/s \tag{9}$$

The constant K is chosen such that $\langle\psi_i^{(s)}\rangle_{N,N_+} = s$ for all N and N_+.

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

$$\psi^{(s)} = \frac{1}{K}\frac{s}{N}\sum_{i=1}^{N/s}\left(\sigma_i^{(s)} - \frac{s}{N}\sum_{j=1}^{N/s}\sigma_j^{(s)}\right)^2 = \frac{s}{N}\sum_{i=1}^{N/s}\psi_i^{(s)} \tag{10}$$

Obviously, one has

$$\langle\psi^{(s)}\rangle_{N,N_+} = s \tag{11}$$

and that $\psi^{(1)} = 1$ is independent of configuration σ_i.

4.2 Fourier Analysis

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. Also, sequences that form β sheets tend to exhibit a periodicity in the hydrophobicity of about 2.3 residues. Here, the the full power spectrum for proteins is compared with that for random sequences.

It is convenient to introduce the *modified hydrophobicity walk* ρ_n:

$$\rho_n = r_n - n\frac{2N_+ - N}{N} \qquad n = 1,\ldots,N \tag{12}$$

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

$$f_k = \sum_{n=1}^{N-1}\rho_n \sin\frac{\pi k n}{N} \qquad k = 1,\ldots,N-1 \tag{13}$$

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

$$\langle f_k^2\rangle_{N,N_+} = \frac{2N_+(N-N_+)}{N-1}\frac{1}{(2\sin\frac{\pi k}{2N})^2} \tag{14}$$

In our calculations we have used the normalized squared amplitude

$$\tilde{f}_k^2 = \frac{f_k^2}{\langle f_k^2\rangle_{N,N_+}} \tag{15}$$

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 only study the relative strength of the different components.

4.3 Hydrophobicity Pattern Analysis: Real Proteins

The analysis described below was carried out on a set of 36765 protein sequences taken from the SWISS-PROT database [2].

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.

In Fig. 2 we show the block fluctuation $\psi_i^{(s=4)}$ 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.

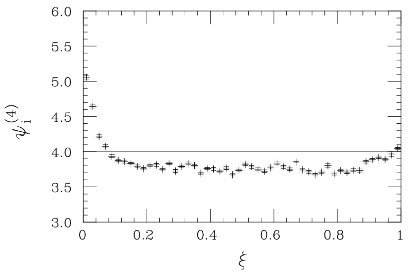

Figure 2: Average values of $\psi_i^{(s=4)}$ against the relative positions of the blocks along the sequence, ξ, for real proteins. The straight line is the result for random sequences.

The results for other values of s are similar. The straight 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 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.

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.

The Fraction of Hydrophobic Residues

Given a set of sequences, one may ask whether the total number of hydrophobic residues in a sequence follows a random pattern. This question can be addressed by studying the quantity

$$X = \frac{N_+ - Np}{\sqrt{Np(1-p)}} \tag{16}$$

where N_+ is the number of hydrophobic residues, N is the total number of residues, and $p \approx 0.291$ is the average of N_+/N over all sequences.

The distribution of X obtained has larger tails than the distribution for random sequences.

When studying correlations in hydrophobicity, we have divided the sequences into groups corresponding to different regions in X. It will be shown below that sequences with different X tend to display different types of correlations.

Hydrophobicity Pattern: Real Proteins

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. 3 results are shown corresponding to three different regions in X: $|X| < 0.5$, $|X| > 3$, and all X.

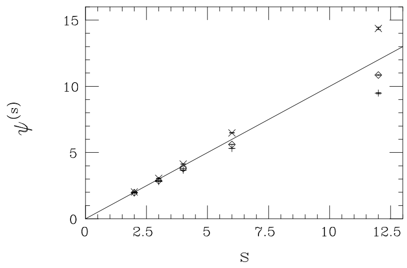

Figure 3: Mean-square fluctuation of the block variables, $\psi^{(s)}$, as a function of block size s for $|X| < 0.5$ (+ ;10154 qualifying proteins), $|X| > 3$ (×; 4928), and all X (⋄; 36765). The straight line is the result for random sequences.

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. These results suggest that there exist 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. The size dependence was found to be fairly weak.

Next we compare the behavior of the Fourier components for small and large $|X|$. 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 $|X|$. These components are suppressed for small $|X|$ and strong for large $|X|$.

The above analysis has also been carried out on a set of 486 non-homologous sequences [6]; the results were similar.

4.4 Hydrophobicity Pattern Analysis: 2D-model

Using the 37 good folding sequences found in the 2D-simulations (see §3.4), the analysis of the previous section has been repeated.

In Fig. 4 we show the mean-square fluctuation of the block variables, $\psi^{(s)}$.

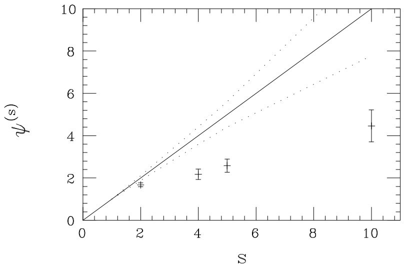

Figure 4: Mean-square fluctuation of the block variables, $\psi^{(s)}$, against block size s for good folding sequences in the 2D model. The straight line is the result for random sequences.

The squared Fourier amplitudes show clear deviations from the result for random sequences at large wave-length. 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 for real proteins with small $|X|$.

4.5 Interpretation of the Deviations from Randomness

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

$$P \propto \exp\Big(K \sum_{i=1}^{N-1} \sigma_i \sigma_{i+1}\Big) \qquad (17)$$

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 obtain a random system.

As in our 2D-simulations, we have taken $N = 20$ and $N_+ = 14$. At $K = 0.25$ the block fluctuations are large and the 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 sequences in the 2D-model and for real protein sequences with small $|X|$. On the other hand, the results for the ferromagnetic system resemble those for real proteins with large $|X|$.

4.6 Sequence Characteristics of Good Folders (2D)

In this section we analyze how $\langle \delta^2 \rangle$ depends upon the binary patterns of the 2D 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.

To this end, it turns out to be enlightening to transform the original binary sequences into more global variables prior to performing the statistical and ANN analysis. The following variables are formed:

- Hydrophobicity walk representations r_n (ref.eq.(7)).
- Block fluctuations – $\psi_i^{(s)}$ and $\psi^{(s)}$.
- The number N_{IA} of σ=+1 surrounded on both sides by σ=+1.
- N_C, the number of clumps of $\sigma = \pm 1$.

Correlations (2D)

$\langle\delta^2\rangle$ is found to correlate substantially with r_n, $\psi_i^{(4)}$, $\psi^{(4)}$, N_{IA} and N_C.

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 (hydrophobic) 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 above for real proteins.
- The strongest positive correlation is between $\langle\delta^2\rangle$ and the number of internal A's (N_{IA}). This may have a simple explanation: In the homopolymer limit, the energy landscape is degenerate, i.e. we have no structural stability. Therefore, one might intuitively expect long stretches of A's to be rare in good folders, and that N_{IA} tends to be low for such sequences. This is confirmed by this correlation.
- There is a strong anticorrelation between $\langle\delta^2\rangle$ and the number of clumps N_C; this is related to the positive correlation of $\langle\delta^2\rangle$ and N_{IA} above.

4.7 Predicting $\langle\delta^2\rangle$ given σ

Given the substantial correlations between $\langle\delta^2\rangle$ and the various quantities formed out of the sequence patterns σ, 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 , 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 approaches. In our case the feedforward ANN consists of an input layer representing the variables defined in §4.6, an output unit for $\langle\delta^2\rangle$ and a set of hidden units in order to model nonlinearities.

In Fig. 5 the resulting prediction for $\langle\delta^2\rangle$ is compared with the true values.

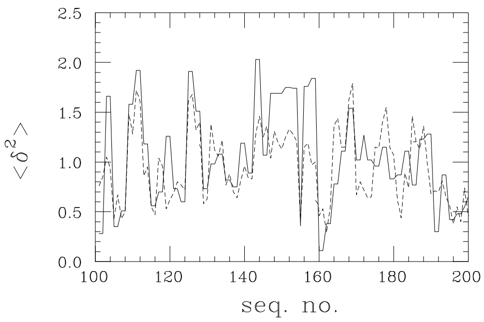

Figure 5: Comparison of the predicted values for $\langle\delta^2\rangle$ with the true ones. The full line represents the true $\langle\delta^2\rangle$'s and the dashed line the predicted ones.

5 3D-simulations

5.1 The model in 3D

The *species-dependent* global interactions are given by the Lennard-Jones potential,

$$E_g(\hat{b};\sigma) = \sum_{i=1}^{N-2} \sum_{j=i+2}^{N} 4\epsilon(\sigma_i,\sigma_j)\Big(\frac{1}{r_{ij}^{12}} - \frac{1}{r_{ij}^{6}}\Big) \ . \qquad (18)$$

The depth of the minimum is chosen to favor the formation of a core of hydrophobic A residues, $\epsilon(A,A) = 1$ and $\epsilon(A,B) = \epsilon(B,B) = 1/2$. The *species-independent* local interactions are defined to be

$$E_l(\hat{b}) = -\kappa_1 \sum_{i=1}^{N-2} \hat{b}_i \cdot \hat{b}_{i+1} - \kappa_2 \sum_{i=1}^{N-3} \hat{b}_i \cdot \hat{b}_{i+2} \qquad (19)$$

It is obvious that the global interactions play a key role in the model. They are responsible for the compaction of the chain and the formation of a hydrophobic core. The importance of local interactions is less clear. Therefore initially the behaviour of the model is examined for different values of the two parameters (κ_1, κ_2) govering the strength of the local interactions. The final choice will be (κ_1,κ_2)=$(-1, 0.5)$. The numerical results are based on simulated tempering simulations of six different sequences, of length 20, chosen to represent a variety of behavior. The joint distribution $P(\hat{b}, k)$ (see eq. (2)) is simulated using separate Metropolis steps in k and $\hat{b}$. For $\hat{b}$ we use two types of elementary moves: Rotations of single bonds and moves of pivot type. A set of $K = 25$ temperatures is used ranging from 0.15 to 3.0. Details may be found in [9].

5.2 Local Interactions

In a 3D off-lattice model it is possible to directly compare the local properties of the model against real proteins. This we have done for two properties: first the distribution of bond and torsional angles and second the local correlations along the chain.

The bond (τ_i) and torisonal (α_i) angles for real proteins, defined from the backbone of (virtual) bonds connecting the C_α atoms, are displayed in Fig. 6(a) and show the strong regularities present in the local structure. The

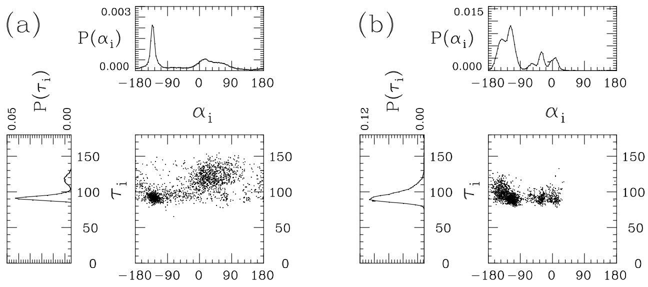

Figure 6: Bond (τ_i) and torsional (α_i) angles. **(a)** for real proteins. **(b)** for our 3D-model at $T = 0.15$, using local interactions of type $(\hat{b}_i \times \hat{b}_{i+1}) \cdot \hat{b}_{i+2}$ and an increased LJ-radius.

region $\tau_i \in$[85°,100°] and $\alpha_i \in$ [35°,70°] corresponds to right-handed α-helix and the other, $\tau_i \in$ [105°,145°] and $\alpha_i \in$ [170°,250°], corresponds to β-sheet. A simple model like the one here studied can not be expected to reproduce these distributions, but the results for different values of (κ_1,κ_2) strongly indicate the need for local interactions. The pure Lennard-Jones chains (κ_1,κ_2)=(0,0) to a large extent lack the regularities present in real proteins: the torsional angle distributions are much flatter (see Fig. 7(a)) and the bond-bond correlation function, $\langle \hat{b}_i \cdot \hat{b}_{i+d} \rangle$, has a very short range. The inclusion of local interactions leads to a significant improvement. The torsional angle distributions becomes concentrated to a few regions (see Fig. 7(b)) and the bond-bond correlation function has a range comparable to that found in real proteins. The local properties

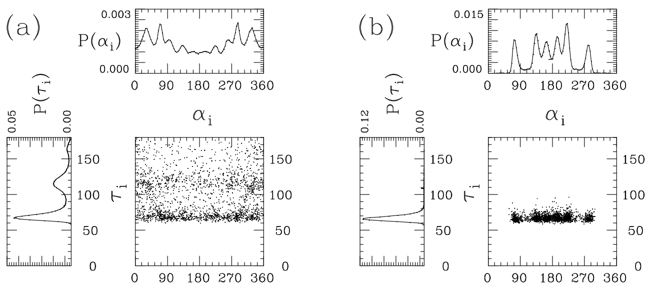

Figure 7: Bond (τ_i) and torsional (α_i) angle distributions at $T = 0.15$. **(a)** pure Lennard-Jones potential $(\kappa_1, \kappa_2) = (0, 0)$ and **(b)** $(\kappa_1, \kappa_2) = (-1, 0.5)$. The potential corresponding to $(\kappa_1, \kappa_2) = (-1, 0)$ yields a similar distribution as in (b).

discussed above were found to be rather sequence independent.

Next we turn to how the key folding property – the distribution $P(\delta^2)$ – is affected by the local interactions. In Fig. 8 $P(\delta^2)$ is displayed for one sequence at low temperature. Despite the fact that $P(\delta^2)$ is strongly sequence dependent Fig. 8 ilustrates some general trends seen in the data. The pure Lennard-Jones interaction yields a very broad distribution indicating that folding properties are poor. This was the case for all the sequences studied. This observatoin is

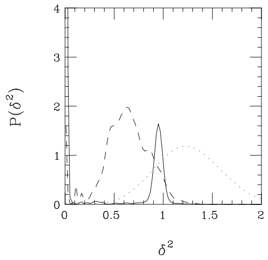

Figure 8: The distribution $P(\delta^2)$ for one sequence with different types of local interaction: $(\kappa_1, \kappa_2) = (0, 0)$ (dots), (-1,0) (dashes) and (-1,0.5)(solid). $T = 0.15$. The low peak for (-1,0.5) extends to 37.

reminiscent of existing lattice model studies [13, 16] and could be taken as an indication that it is necessary to have more than two types of amino acids. Here we have explored another possibility, good folding sequences can be obtained by including local interactions.

Several other settings of (κ_1, κ_2) have been tested, all giving rise to higher structural stability when compared to the pure LJ-potential. We also tried local interactions of the form $(\hat{b}_i \times \hat{b}_{i+1}) \cdot \hat{b}_{i+2}$ and an increased LJ-radius, see Fig. 6(b). The local interactions give rise to higher structural stability in this case as well.

Thus the inclusion of local interactions is essential in this model in order to get local structure and stuctural stability.

5.3 Folding properties

When studying the folding properties of the chains we focus on the thermodynamics. Furthermore we restrict ourselves to local interactions of the type $(\kappa_1, \kappa_2) = (-1, 0.5)$.

In Fig. 9 we see that the overall size of the chain, as measured by r^2_{gyr}, decreases substantially as the temperature is lowered. The upper curves corresponds to sequences with composition 10A+10B and the lower curves to composition 14A+6B. In the same figure p_0, the relative population of

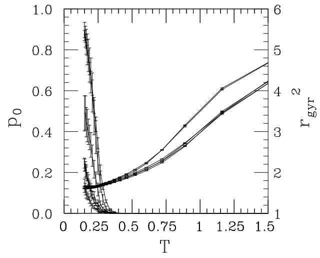

Figure 9: The radius of gyration (squared) r^2_{gyr} and the relative population of the native state, p_0, as a function of temperature.

the native state, is displayed. As can be seen a sequence dependent folding transition occurs in the compact state. The native contacts at short topological distance along the chain are formed in a gradual process prior to this transition. This is in contrast to the global i.e. long range, native contacts which are formed during the folding transition.

To study the beavior at the folding temperature T_f, where $p_0 = 1/2$, in some more detail we examine in Fig. 10

the histograms of δ_0^2, the mean-square distance to the native state, for two sequences. Both histograms exhibit a multipeak structure but, as can be seen from Fig. 10 the sequence with the lower folding temperature has a more

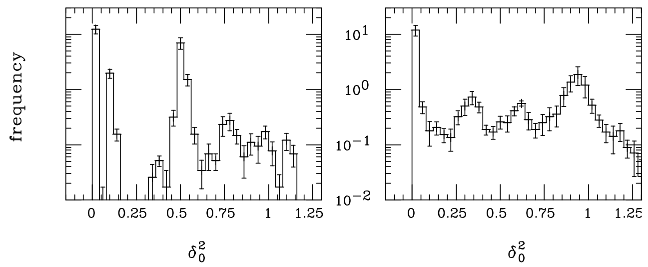

Figure 10: δ_0^2-histograms for two sequences at $T \approx T_f$, (κ_1, κ_2)=$(-1.0, 0.5)$. Left: $T_f \approx 0.15$. Right: Good folder, $T_f \approx 0.22$.

rugged distribution with strongly surpressed regions. This is important from the viewpoint of kinetics and suggests that, at $T = T_f$, folding is fastest for the sequence with the highest T_f.

6 Summary and Outlook

An off-lattice model for protein folding has been probed in two and three dimensions. The model has only two types of residues, hydrophobic and hydrophilic.

In 2D a large number of sequences are examined. Correlations between thermodynamic stability and various functions of hydrophobicity patterns are obtained. The good folding sequences are identified and their hydrophobicity pattern is found to deviate from randomness in a way similar to what is found for real proteins.

In 3D the emphasis is put on local interactions. Without local interactions the sequence dependence is low and all sequences have poor folding proterties. By including local interactions the chains become more protein-like on short lengthscales and furthermore the structural stability is enhanced.

Our ability to map out the thermodynamics of the off-lattice models relies heavily upon the efficiency of the simulated tempering algorithm [11, 5]. As it stands, the results for each $N = 20$ 3D-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. Simulations for $N \approx 40$ seem to be possible.

7 Acknowledgements

We thank our co-workers Anders Irbäck and Carsten Peterson for useful discussions and kind support.

References

[1] ANFINSEN, Christian B. "Principles that Govern the Folding of Protein Chains", *Science* **181** (1973), 223–230.

[2] BAIROCH, Amos, and Brigitte BOECKMANN "The SWISS-PROT Protein Sequence Data Bank: Current Status", *Nucleic Acids Res.* **22** (1994), 3578–3580.

[3] DUANE, Simon , Anthony D. KENNEDY, Brian J. PENDLETON and Duncan ROWETH "Hybrid Monte Carlo", *Phys. Lett. B* **195** (1987), 216–222.

[4] IORI, Giulia, Enzo MARINARI, and Giorgio PARISI "Random Self-Interacting Chains: A Mechanism for Protein Folding", *J. Phys. A.* **24** (1991), 5349–5362.

[5] IRBÄCK, Anders, and Frank POTTHAST "Studies of an Off-Lattice Model for Protein Folding: Sequence Dependence and Improved Sampling at Finite Temperature", *J. Chem. Phys.* **103 (23)** (1995), 10298–10305.

[6] IRBÄCK, Anders, Carsten PETERSON, and Frank POTTHAST "Evidence for non-random hydrophobicity structures in protein chains", *Proc. Natl. Acad. Sci. USA* **93** (1996), 9533–9538.

[7] IRBÄCK, Anders, Carsten PETERSON, and Frank POTTHAST "Identification of Amino Acid Sequences with Good Folding Properties in an Off-Lattice Model", *Phys. Rev. E* **55** (1997), 860–867.

[8] IRBÄCK, Anders, and Erik SANDELIN "Local Interactions and Protein Folding: A Model Study on the Square and Triangular Lattices", *LU TP 97-17* submitted to *J. Chem. Phys.* (1997).

[9] IRBÄCK, Anders, Carsten PETERSON, Frank POTTHAST, and Ola SOMMELIUS "Local Interactions and Protein Folding : A three-dimensional off-lattice approach", *J. Chem. Phys.* **107 (1)** (1997), 273–282.

[10] KAUZMANN, Walter J. "Some Factors in the Interpretation of Protein Denaturation", *Adv. Protein Chem.* **14** (1959), 1–63.

[11] MARINARI, Enzo, and Giorgio PARISI "Simulated Tempering: A New Monte Carlo Scheme", *Europhys. Lett.* **19** (1992), 451–458.

[12] ŠALI, Andrej, Eugene SHAKHNOVICH, and Martin KARPLUS "Kinetics of protein folding. A lattice model study of the requirements for folding to the native state", *J. Mol. Biol.* **235** (1994), 1614–1636.

[13] SHAKHNOVICH, Eugene I. "Proteins with selected sequences fold into unique native conformation", *Phys. Rev. Let.* **72** (1994), 3907–3910.

[14] STILLINGER, Frank H., Theresa HEAD-GORDON, and Catherine L. HIRSHFELD "Toy Model for Protein Folding", *Phys. Rev. E* **48** (1993), 1469–1477.

[15] WHITE, Stephen H., and Russell E. JACOBS "Statistical distribution of hydrophobic residues along the length of protein chains. Implications for protein folding and evolution", *Biophys. J.* **57** (1990), 911–921.

[16] YUE, Kaizhi, Klaus M. FIEBIG, Paul D. THOMAS, Hue Sun CHAN, Eugene I. SHAKHNOVICH, and Ken A. DILL "A test of lattice protein folding algorithms", *Proc. Natl. Acad. Sci. USA* **92** (1995), 325–329.