Revised Version
LU TP 96-28
December 4, 1997

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

Anders Irbäck[1], Carsten Peterson[2], Frank Potthast[3] and Ola Sommelius[4]

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

Published in
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.

[1]irback@thep.lu.se
[2]carsten@thep.lu.se
[3]frank@thep.lu.se
[4]ola@thep.lu.se

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. 2 the 3D model is defined and in Sec. 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. 4. The results and the summary are found in Sec. 5 and 6 respectively.

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. 1).

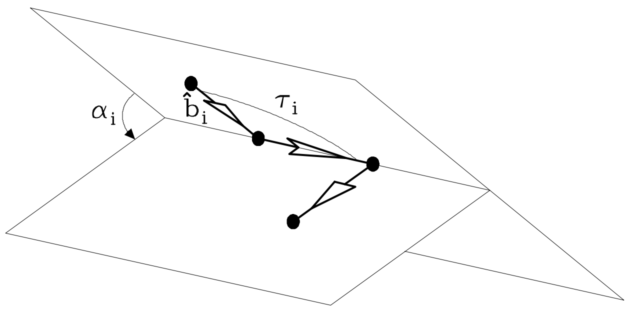

Figure 1: Definition of $\hat{b}_i$, τ_i and α_i.

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

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

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,

$$E_{\mathrm{LJ}}(r_{ij};\sigma_i,\sigma_j) = 4\epsilon(\sigma_i,\sigma_j)\left(\frac{1}{r_{ij}^{12}} - \frac{1}{r_{ij}^{6}}\right) . \tag{2}$$

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

$$\epsilon(\sigma_i, \sigma_j) = \begin{cases} 1 & \text{AA} \\ \\ \frac{1}{2} & \text{BB, AB} \end{cases} \tag{3}$$

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

$$Z(T; \sigma) = \int \left[\prod_{i=1}^{N-1} d\hat{b}_i \right] \exp(-E(\hat{b}; \sigma)/T) \ . \tag{4}$$

Let us stress that the interactions in Eq. 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 results to be presented in Sec. 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).

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.

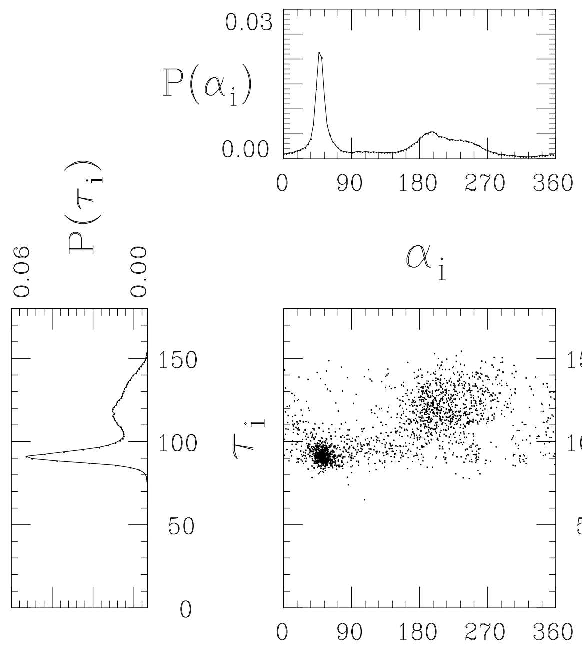

Figure 2: Bond (τ_i) and torsional (α_i) angle distributions for functional proteins.

In this section we describe the results one finds for functional proteins. These results will be used in Sec. 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. 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. 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 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

$$C_b(d) = \frac{1}{N-d-1} \sum_{i=1}^{N-d-1} \hat{b}_i \cdot \hat{b}_{i+d} \tag{5}$$

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. In Fig. 3a 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.

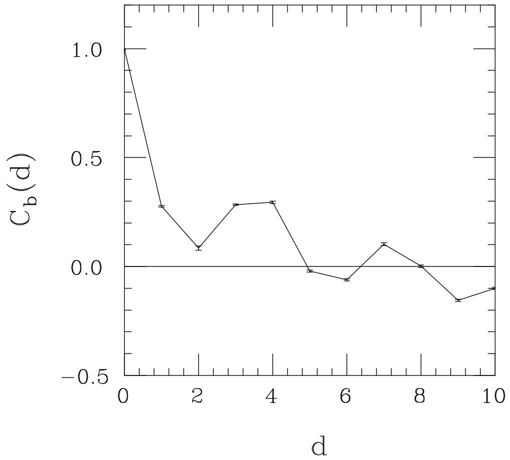

Figure 3: The correlation function $C_b(d)$ for functional proteins (see Eq. 5).

4 Methods

4.1 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, 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

$$P(\hat{b}, k) \propto \exp(-g_k - E(\hat{b}, \sigma)/T_k) \tag{6}$$

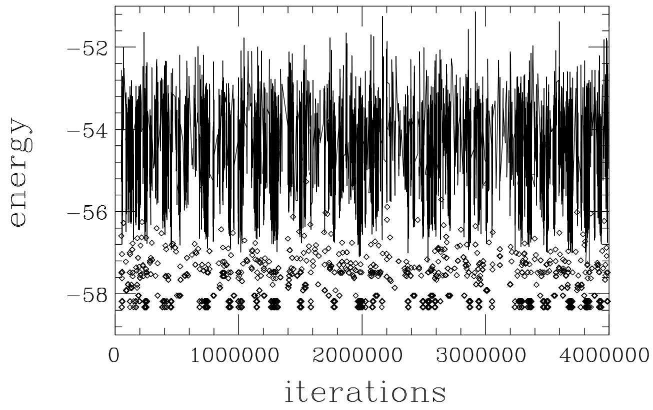

Figure 4: Evolution of the quenched (diamonds) and unquenched (line) energies in the simulation of sequence 1 (see Table 1). Measurements were taken every 10 iterations. Shown are the data corresponding to the lowest allowed temperature. The thermalization phase of 50000 sweeps is not shown.

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

In Fig. 4 we show the evolution of the quenched and unquenched energies in a typical simulation of a $N = 20$ chain.

4.2 Measurements

In Sec. 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

$$\delta^2_{ab} = \min \frac{1}{N} \sum_{i=1}^{N} |\bar{x}_i^{(a)} - \bar{x}_i^{(b)}|^2 \tag{7}$$

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

$$P(\delta^2) = \frac{1}{Z(T;\sigma)^2} \int d\hat{b}^{(a)} d\hat{b}^{(b)} \delta(\delta^2 - \delta^2_{ab}) \mathrm{e}^{-E(\hat{b}^{(a)};\sigma)/T} \mathrm{e}^{-E(\hat{b}^{(b)};\sigma)/T} \tag{8}$$

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

$$\langle \delta^2 \rangle = \int d\delta'^2 P(\delta'^2) \delta'^2 \tag{9}$$

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

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

$$C_V = \frac{1}{T^2} \left(\langle E^2 \rangle - \langle E \rangle^2 \right) \tag{10}$$

$$r^2_{\text{gyr}} = \frac{1}{N} \sum_{i=1}^{N} \left(\langle \bar{x}_i^2 \rangle - \langle \bar{x}_i \rangle^2 \right) \tag{11}$$

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. 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

$$p_0 = \int_0^{\Delta} P(\delta_0^2) d\delta_0^2 \tag{12}$$

with the parameter $\Delta = 0.04$; this choice of Δ is motivated by the δ_0^2 distributions (cf Figs. 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

$$Q = \frac{n^o}{n} \tag{13}$$

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

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 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.

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 $2 \leq |i - j| \leq 4$ 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.

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

$$(\kappa_1, \kappa_2) = \begin{cases} (0,0) \\ (-1,0) \\ (-1,0.5) \end{cases} \tag{14}$$

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 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.

5.1 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.

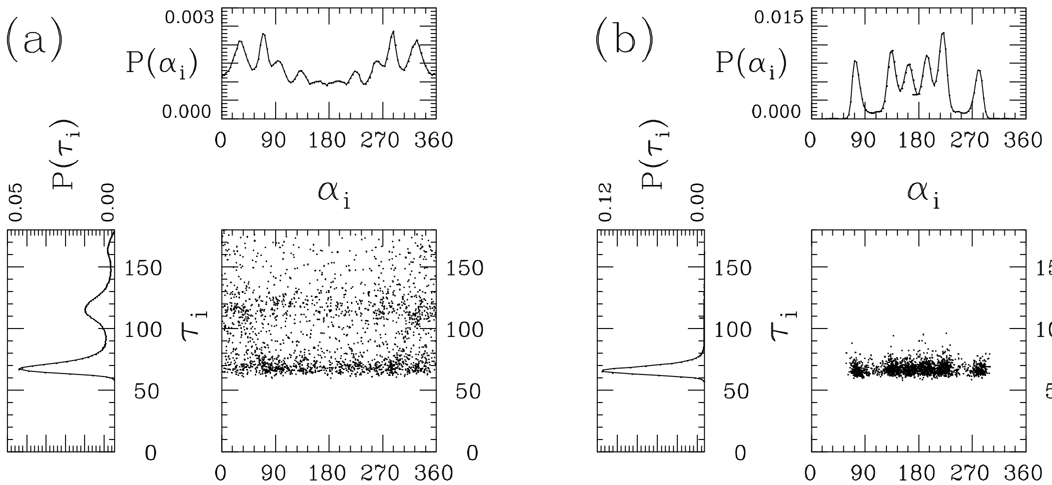

Figure 5: Bond (τ_i) and torsional (α_i) angle distributions at $T = 0.15$ for the six simulated sequences, **(a)** pure Lennard-Jones potential $(\kappa_1, \kappa_2) = (0, 0)$ and **(b)** $(\kappa_1, \kappa_2) = (-1, 0.5)$. The potential corresponding to $(\kappa_1, \kappa_2) = (-1, 0)$ yields a similar distribution as in (b).

In Figs. 5 and 6 the model counterparts of Figs. 2 and 3 are shown. The data in Fig. 5 were obtained at a fixed temperature, $T = 0.15$, while those in Fig. 6 were obtained using 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 and 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 and 6 alone in this respect, although one may argue that the range of the correlation $C_b(d)$ is still somewhat short for $(\kappa_1, \kappa_2) = (-1, 0)$.

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. 8, $P(\delta^2)$ is shown for sequence 4 in Table 1. In contrast to the local properties discussed above, $P(\delta^2)$ is strongly sequence dependent. Nevertheless, Fig. 8 illustrates some general trends seen in the data. First, the pure Lennard-Jones interaction

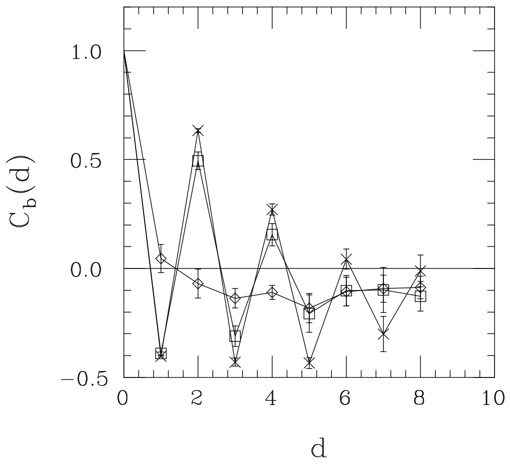

Figure 6: The correlation function $C_b(d)$ (Eq. 5) averraged over the six sequences in Table 1 using (κ_1, κ_2)=(0,0) (◇), (-1,0) (□) and (-1,0.5) (×).

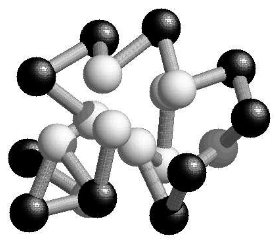

Figure 7: A representative native state (seq. 6) for $(\kappa_1, \kappa_2) = (-1, 0.5)$ illustrating the formation of a hydrophobic core. The hydrophobic (A) and hydrophilic (B) monomers are represented by white and black spheres respectively.

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 1.

In Fig. 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.

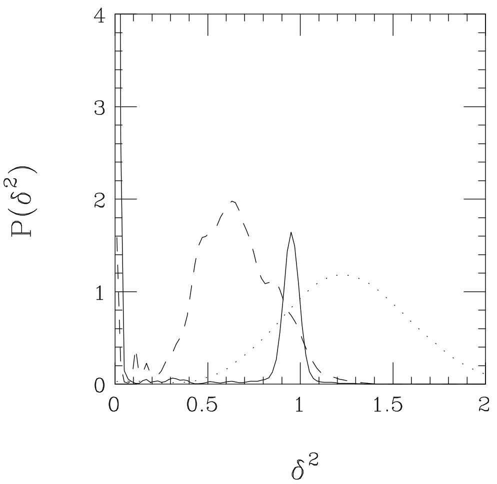

Figure 8: $P(\delta^2)$ at $T = 0.15$ for sequence 4 (cf Table 1); (κ_1,κ_2)=(0,0) (dots), (-1,0) (dashes), and (-1,0.5) (solid). The low-δ^2 peak for (κ_1,κ_2)=(-1,0.5) extends to 37.

5.2 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 $(\kappa_1, \kappa_2)=(\frac{1}{4}, 0)$ (cf Eq. 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 $(\kappa_1, \kappa_2)=(0, 0)$. 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 $(\kappa_1, \kappa_2)=(0, 0)$ potential. This is in contrast to our findings in 3D, where all the sequences studied have a large $\langle\delta^2\rangle$ for $(\kappa_1, \kappa_2)=(0, 0)$.

5.3 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. 4.2 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.

5.3.1 Thermodynamic Behavior

In Fig. 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. 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. 9 we show the relative magnitude $\langle \delta^2 \rangle / r_{\rm gyr}^2$. 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

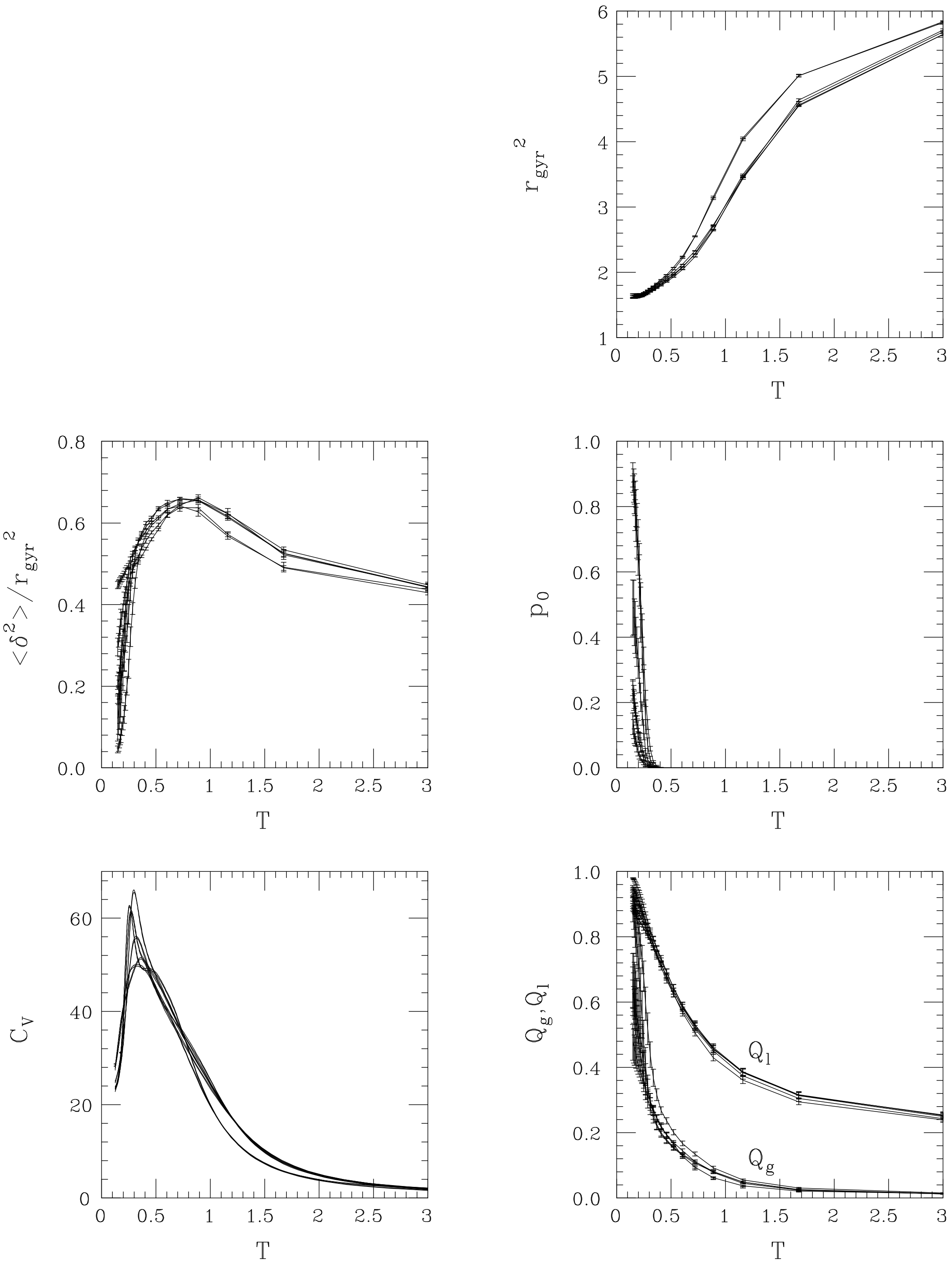

Figure 9: Thermodynamic properties for the six chains in Table 1 using (κ_1, κ_2)=(-1,0.5).

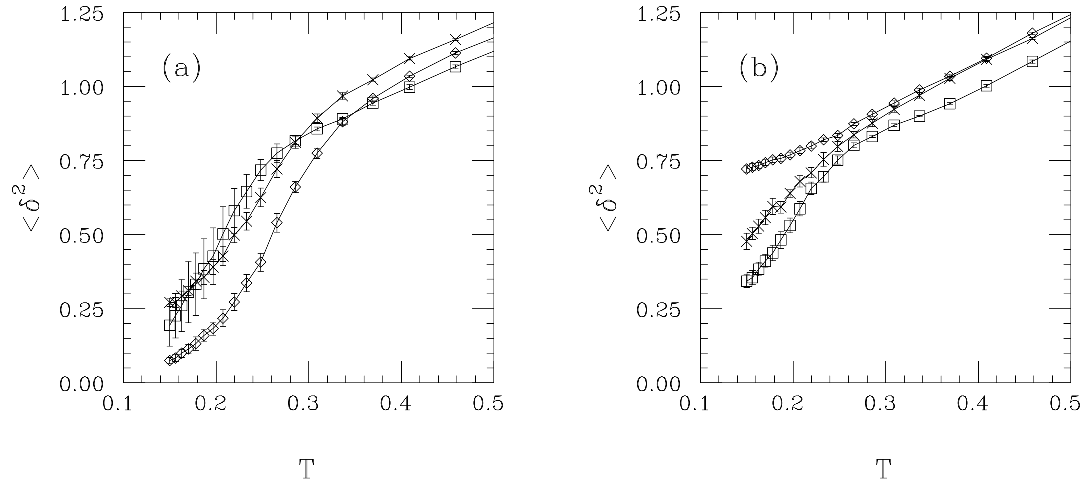

Figure 10: $\langle\delta^2\rangle$ as a function of T for the sequences of Table 1 using (κ_1,κ_2)=(-1,0.5) : **(a)** 4 ($\Box$), 2 ($\times$) and 3 ($\diamond$). **(b)** 6 ($\Box$), 1 ($\times$) and 5 ($\diamond$).

$\langle\delta^2\rangle$ and $\langle\delta^2\rangle/r^2_{\text{gyr}}$ are strongly sequence dependent. In Fig. 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. 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 1. For three of the six sequences T_f is smaller than 0.15, as can be seen from Fig. 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.

From Figs. 9 and 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 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. 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.

5.3.2 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,

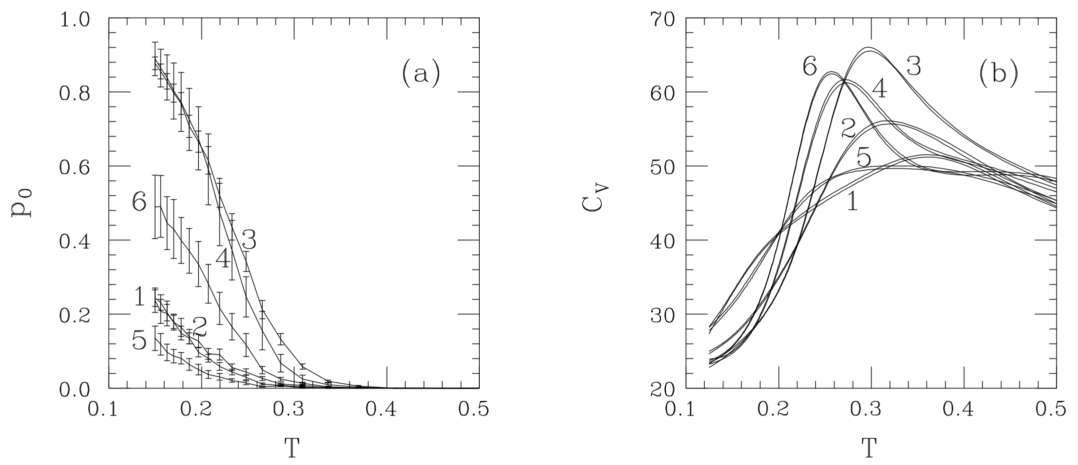

Figure 11: Enlargements of **(a)** p_0 and **(b)** C_V in Fig. 9, (κ_1,κ_2)=(-1,0.5). The bands in (b) correspond to $C_V \pm \sigma$, where σ is the standard deviation.

whereas sequence 6 represents a more typical T_f.

In Figs. 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. 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. 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. 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.

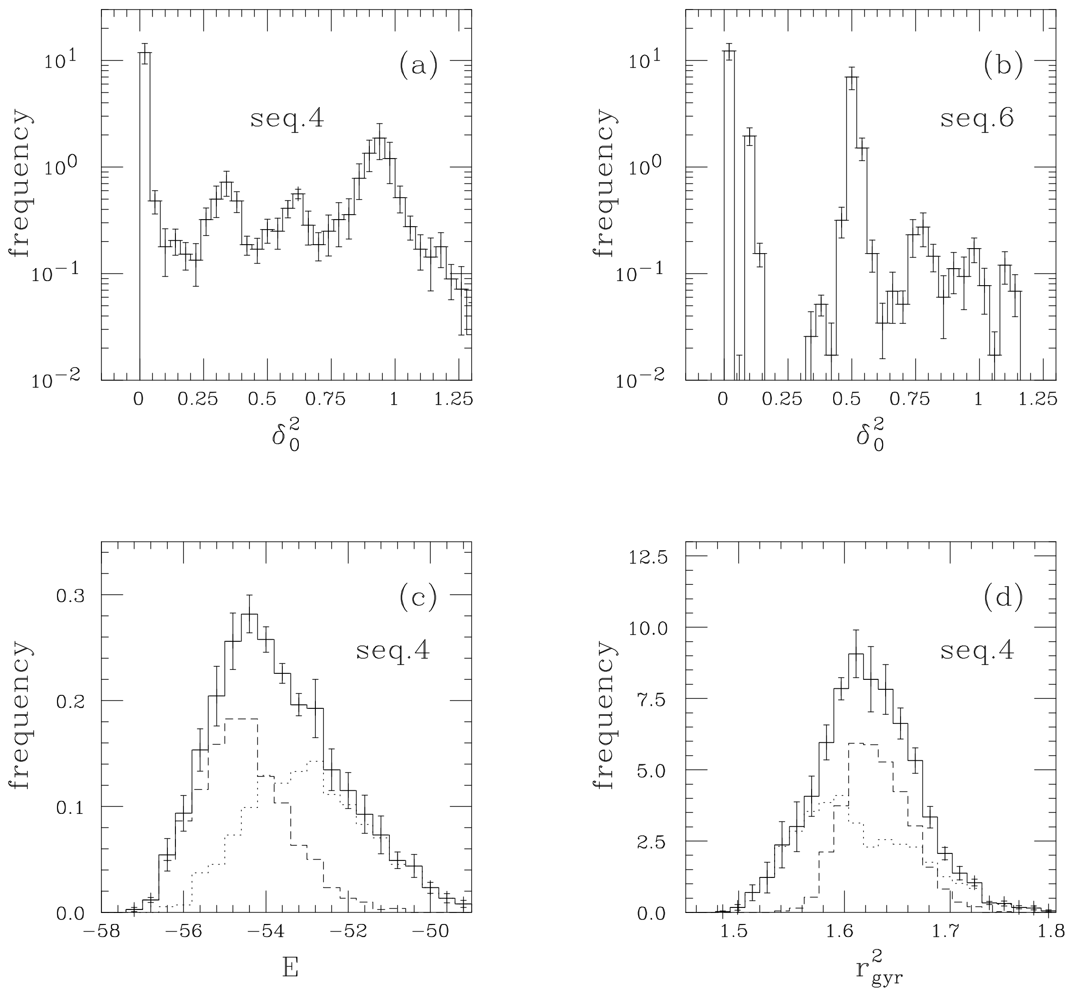

Figure 12: Histograms at $T \approx T_f$ for sequence 4 ($T = 0.22$) and 6 ($T = 0.15$), $(\kappa_1, \kappa_2) = (-1, 0.5)$. **(a)** δ_0^2 for seq. 4. **(b)** δ_0^2 for seq. 6. **(c)** E for seq. 4. **(d)** $r^2_{\rm gyr}$ for seq. 4. The dashed and dotted lines in (c) and (d) represent the contributions corresponding to $\delta_0^2 < 0.04$ and $\delta_0^2 > 0.04$, respectively.

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 $r^2_{\rm gyr}$ was used as a measure of compactness. In contrast to the results shown in Figs. 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.

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.

References

[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).