LU TP 98-13
June 18, 1998

# A Minimal Off-Lattice Model for $\alpha$-helical Proteins

Frank Potthast[1]

Complex Systems Group, Department of Theoretical Physics
University of Lund, Sölvegatan 14A, S-223 62 Lund, Sweden
http://www.thep.lu.se/tf2/complex

Submitted to *Journal of Computational Biology*

Abstract:

A minimal off-lattice model for $\alpha$-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 $\alpha$-helical structure. They model chirality and $\alpha$-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 $\alpha$-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

[1]frank@thep.lu.se

# 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, 7, 8, 9, 10, 11]. The main difference between our model and those studied previously is that it is deliberately constructed so as to model the $C_\alpha$ backbone geometry of $\alpha$-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 $\alpha$-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 $\alpha$-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.

# 2 Methods

## 2.1 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 $\alpha$-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 $\alpha$-helical structure found in real proteins an **ideal $\alpha$-helix**; this helix has 3.6 amino acids per turn and a translation of 5.4Å per turn. An ideal $\alpha$-helix is completely described by the geometry of three successive (virtual) $C_\alpha$-$C_\alpha$ bonds; that is by one torsional ($\alpha_i$) and two bend angles ($\tau_i, \tau_{i+1}$). Equivalently, one may describe an ideal $\alpha$-helix by the mutual distances $r^\alpha_{ij}$ of the four participating $C_\alpha$ atoms (plus the correct chirality).

In our model, each residue is represented by a single site corresponding to the $C_\alpha$ 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

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

The first term in this equation consists of Lennard-Jones (LJ) interactions. The LJ parameters $A_{ij}$ are chosen as $A_{ij} = 2^{-1/6} r^\alpha_{ij}$ for $|j-i| \leq 4$, where $r^\alpha_{ij}$ denotes the distance in an ideal $\alpha$-helix. This choice of $A_{ij}$ gives the corresponding terms in eq. 1 a minimum at $r^\alpha_{ij}$. 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 $\alpha$-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 a negative gaussian with unit depth and width $w$;

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

where $\hat{\alpha}$ and $\hat{\tau}$ are the torsional and bend angles in an ideal $\alpha$-helix. $E_l(i)$ is close to minus one if and only if the three participating bond vectors are close to $\alpha$-helical structure, which is illustrated in Fig. 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 3.1. 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 $\alpha$-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 3.1, it will be shown that, not surprisingly, the model indeed gives rise to $\alpha$-helical structure. This is done by comparing the local structure of our model to that of real $\alpha$-helical proteins. Quantitatively very similar results are found. Furthermore, we examine the overall topology of the simulated sequences. As will be shown in section 3.2, we find topologies corresponding to 2- and

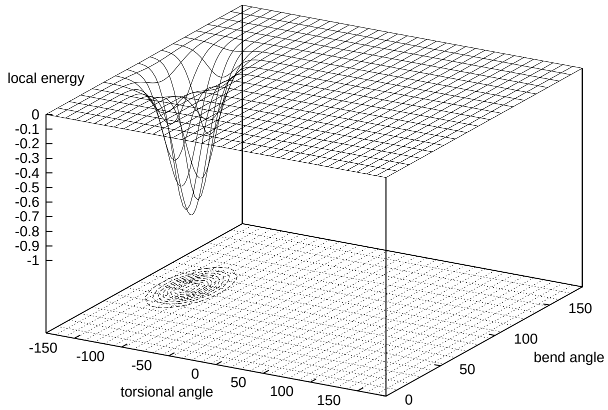


Figure 1: Local energy $E_l(i)$ (Eq. 2) displayed in one bond and one torsional angle. The second bend angle is set to $\hat{\tau}$, the bend angle of an ideal $\alpha$-helix. $E_l(i)$ has a single minimum of depth one around the $\alpha$-helical conformation and is close to zero elsewhere. Equipotential lines for $E_l(i) \in \{-0.1, -0.2, ..., -0.9\}$ are displayed with circles.

3-helix bundles. By looking at the $C_\alpha$-$C_\alpha$ distance distribution functions, it can be seen that even the linear dimensions of our model proteins are similar to those of real ones.

### 2.2 Sequences

We have studied the behavior of two sequences in this model. These were deliberately chosen so as to be consistent with $\alpha$-helical structure. Specifically, we start from a sequence segment of length 14, which is known to occur in amphiphilic $\alpha$-helices [15]. This segment has been used earlier for the design of (real) de novo $\alpha$-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 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 1: The two sequences studied. Basic building blocks of length 9 are connected by hydrophilic segments of length 3.

### 2.3 Monte Carlo Methods: Simulated Tempering

We calculate thermodynamic properties by using the method of simulated tempering [16, 17, 10]. 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

$$P(\vec{b},k) \propto \exp(-g_k - E(\vec{b},\sigma)/T_k) \,, \tag{3}$$

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.

# 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.

## 3.1 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 $\alpha$-helical proteins. The protein data were extracted using the $C_\alpha$ coordinates for 27 proteins containing mainly $\alpha$-helices and almost no $\beta$-sheets. All these 27 proteins are listed in the all-$\alpha$-helical class of the SCOP database [18]. The corresponding structures were taken from the Brookhaven Protein Data Bank (PDB) [19] [2]. One of these proteins, the $N = 54$ DNA-binding protein 1enh, which contains three $\alpha$-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.

[2]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.

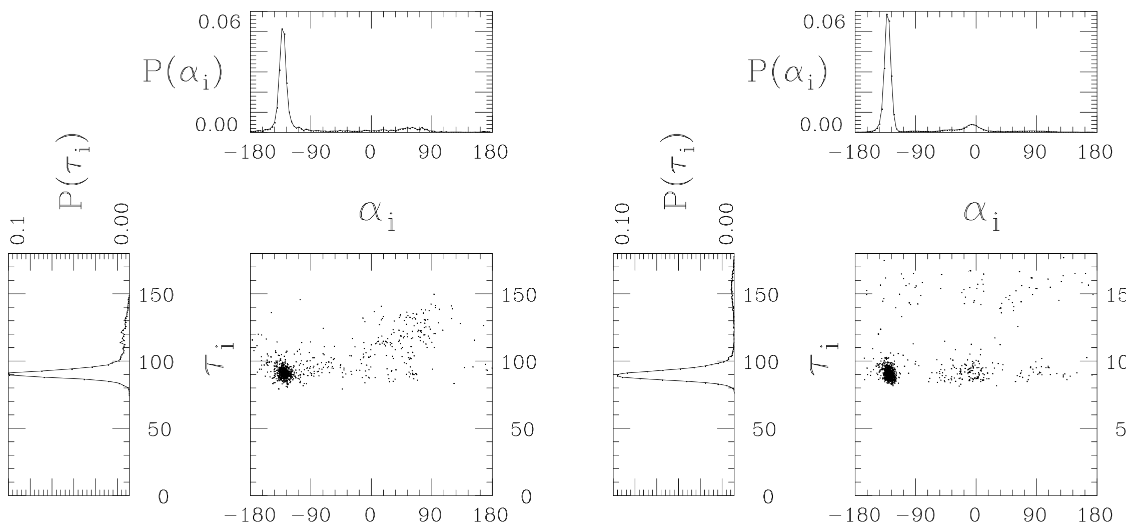


Figure 2: Bond ($\tau_i$) and torsional ($\alpha_i$) angle distributions. Left: For the 27 $\alpha$-helical proteins taken from the PDB database. Right: Model results for the $N=21$ sequence at $T=0.15$. The other two sequences simulated, with $N=33$ and $N=45$, give very similar distributions. Both scatter plots contain 1000 data points.

In Fig. 2, left plot, bend and torsional angle distributions are displayed for the set of 27 $\alpha$-helical proteins. The dominance of $\alpha$-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. 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 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 $\alpha$-helix. This peak contains roughly the same amount of probability as for real $\alpha$-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. 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

$$C_b(d) = \left\langle \frac{\vec{b}_i \cdot \vec{b}_{i+d}}{|\,\vec{b}_i\,|\,|\,\vec{b}_{i+d}\,|} \right\rangle \qquad C_b(d) \in [-1,1], \tag{4}$$

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. 3 (dotted line, left plot), we show the correlation function

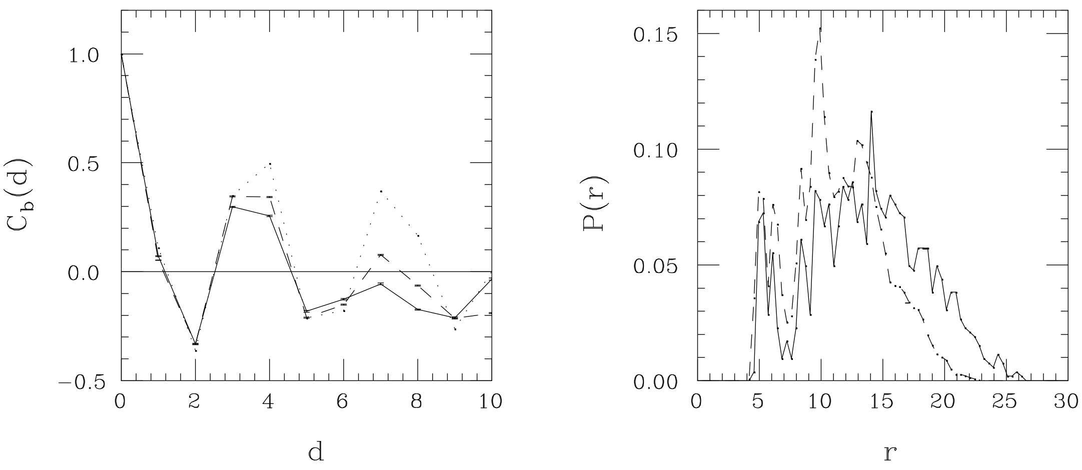


Figure 3: Left: Correlation functions $C_b(d)$ for our model at $T = 0.15$ for $N = 21$ (solid line) and $N = 33$ (dashed line). $C_b(d)$ for the 27 $\alpha$-helical proteins listed in the text is given by the dotted line. Right: $C_\alpha$-$C_\alpha$ distance distribution $P(r)$ for the (molecular dynamics 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.

$C_b(d)$ for the 27 $\alpha$-helical proteins. As can be seen from this figure, there are significant correlations at least out to separations of about ten residues. The oscillations can be related to the presence of $\alpha$-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. 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.

### 3.2 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 $\alpha$-helix formation.

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

$$\delta^2_{ab} = \min \frac{1}{N} \sum_{i=1}^{N} |\bar{x}_i^{(a)} - \bar{x}_i^{(b)}|^2 \tag{5}$$

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 sequence and fixed temperature, gives valuable information about

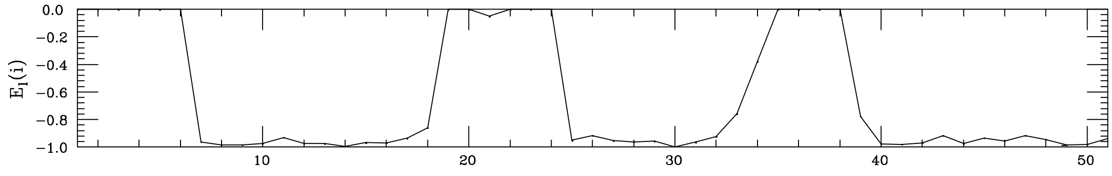


Figure 4: The function $E_l(i)$ of Eq. 2 for the experimental structure of DNA binding protein, PDB accession code 1enh ($N = 54$). One easily identifies three $\alpha$-helices from this plot; for $\alpha$-helical structure $E_l(i)$ is close to minus one. The resulting $\alpha$-helix assignments are in agreement with those in the PDB entry.

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 $\alpha$-helical structure. A convenient way of doing this is to utilize the function $E_l(i)$ (Eq. 2); besides using it for modelling hydrogen-bonding in the $\alpha$-helical structure, $E_l(i)$ can also be used to detect $\alpha$-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. 4. From this plot, it is immediately clear that the structure contains three $\alpha$-helices. The first $\alpha$-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 $C_\alpha$-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 $1\text{Å}^2$ and $10\text{Å}^2$. 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. 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 $\delta^2$ distances to all these four states. The mutual distances between these states range from $1.5\text{Å}^2$ to $3.7\text{Å}^2$. When it comes to helix content, these four states are very similar. The function $E_l(i)$, as displayed in Fig. 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 are indeed very stable and common to the four different states. The structure of one of the four states is shown in Fig. 5 (the other three are very similar). From Fig. 5 it can also be seen that (1) we succeed to obtain a hydrophobic core and (2) the $\alpha$-helices show the patterning intended in the construction of the sequences.

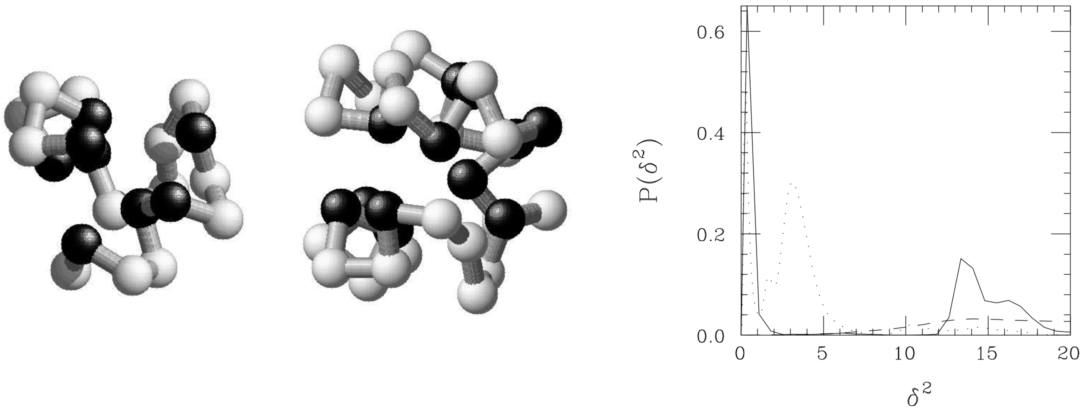


Figure 5: Representative structures for the most probable states for $N = 21$ (left) and $N = 33$ (middle). Hydrophobic and hydrophilic amino acids are represented by black and white spheres, respectively. On the right side, $P(\delta^2)$ is plotted for (1) $N=21$, $T=0.15$ (dotted line), (2) $N=33$, $T=0.15$ (solid line), and (3) $N=33$, $T\approx0.40$ (dashed line).

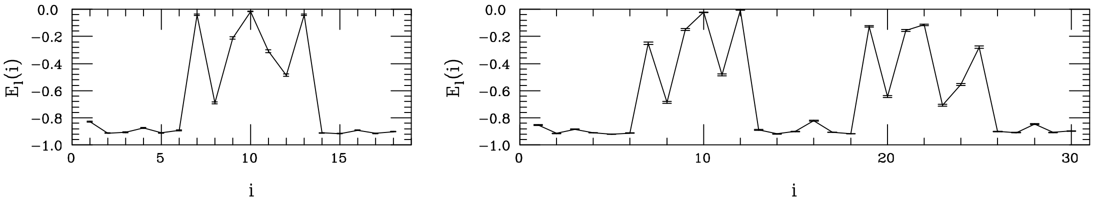


Figure 6: Thermodynamic averages of the local interactions $E_l(i)$ at $T = 0.15$ for $N = 21$ (left) and $N=33$ (right). These curves clearly show that the dominating states for $N=21$ and $N=33$ contain 2 and 3 helices, respectively. This is also supported by the fact that fluctuations of bend and torsional angles are small in the helical regions.

The $P(\delta^2)$ distribution for $N=33$, see Fig. 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 $16.3\text{Å}^2$. Taken together, these three states contain 97% probability at $T=0.15$, distributed as 68%, 17% and 12%. These three states share three common $\alpha$-helices; the positions of the helices along the chain can be seen from Fig. 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. 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 2.3, 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 $C_\alpha$-$C_\alpha$ distance distribution function in Fig. 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.

# 4 Conclusions

A minimal off-lattice model for $\alpha$-helical proteins has been presented. $\alpha$-helical structure is obtained by setting local LJ parameters to values consistent with $\alpha$-helical geometry and introducing local interactions modelling chirality and $\alpha$-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 $\alpha$-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 $\alpha$-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 $\alpha$-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

Figure 7: Illustration of an ideal $\alpha$-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.

e.g. [21]).

# 5 Appendix

## 5.1 The Ideal $\alpha$-Helix: Model Parameters

Five parameters of our model are determined by the geometry of an ideal $\alpha$-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 $\alpha$-helix has a translation of 5.4Å and 3.6 amino acids per turn with right-handed chirality, as is illustrated in Fig. 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 $\tau_i$,$\tau_{i+1}$ and the torsional angle $\alpha_i$. The bend angle $\tau$ is defined to be 180° if the two involved bond vectors are parallel. The torsional angle $\alpha$ 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. $\alpha$ 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. 1,

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

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

$$|i-j|>4 \Rightarrow A_{ij} = 2^{-1/6} \times \sqrt[3]{\frac{3}{4\pi}161\text{Å}^3} \approx 6.01\text{Å} \tag{7}$$

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

| | | $\hat{\tau} \approx 90.53°$ | $\hat{\alpha} \approx -129.61°$ | | |
|---|---|---|---|---|---|
| | $\|j-i\|=1$ | $\|j-i\|=2$ | $\|j-i\|=3$ | $\|j-i\|=4$ | $\|j-i\|>4$ |
| $r_{ij}^{\alpha}$ [Å] | 3.8 | 5.39.. | 5.04.. | 6.19.. | - |
| $A_{ij}$ [Å] | - | 4.80.. | 4.49.. | 5.52.. | 6.01.. |

## 5.2 The AB model

The $\alpha$-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;

$$E(\vec{b};\sigma) = \sum_{i=1}^{N-2}\sum_{j=i+2}^{N} 4\epsilon(\sigma_i,\sigma_j)\left(\left(\frac{A_{ij}}{r_{ij}}\right)^{12} - \left(\frac{A_{ij}}{r_{ij}}\right)^{6}\right) - \kappa_1 \sum_{i=1}^{N-2}\frac{\vec{b}_i\cdot\vec{b}_{i+1}}{|\vec{b}_i||\vec{b}_{i+1}|} - \kappa_2 \sum_{i=1}^{N-3}\frac{\vec{b}_i\cdot\vec{b}_{i+2}}{|\vec{b}_i||\vec{b}_{i+2}|}$$

$$E(\vec{b};\sigma) = \sum_{i=1}^{N-2}\sum_{j=i+2}^{N} 4\epsilon(\sigma_i,\sigma_j)\left(\left(\frac{A_{ij}}{r_{ij}}\right)^{12} - \left(\frac{A_{ij}}{r_{ij}}\right)^{6}\right) - \sum_{i=1}^{N-3}\exp\left(-\frac{(\alpha_i-\hat{\alpha})^2+(\tau_i-\hat{\tau})^2+(\tau_{i+1}-\hat{\tau})^2}{2w^2}\right)$$

They differ only in the choice of $A_{ij}$ and in the local interactions; for the AB model, $A_{ij} = 3.8\text{Å}$ independent of $i$ and $j$. The dependence of the AB model on the strengths of the local interactions $(\kappa_1, \kappa_2)$ 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 $(\kappa_1, \kappa_2)$ in [11] was $(\kappa_1, \kappa_2) = (-1, 0.5)$.

# 6 Acknowledgements

This project is supported by the "Swedish Foundation for Strategic Research". Thanks to Anders Irbäck and Carsten Peterson for useful discussions. I appretiate 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.

## References


[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.S. 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).