LU TP 00-22
May 26, 2000

Three-helix-bundle Protein in a Ramachandran Model

Anders Irbäck, Fredrik Sjunnesson and Stefan Wallin*

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

Submitted to *Proc. Natl. Acad. Sci. USA*

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

*E-mail: irback, fredriks, stefan@thep.lu.se

1 Introduction

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

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

In this model we in particular study a three-helix-bundle protein with 54 amino acids, which represents a truncated and simplified version of the four-helix-bundle protein *de novo* designed by Regan and DeGrado [11]. This example was chosen partly because there has been several earlier studies of three- and four-helix-bundle proteins using models at similar levels of resolution [12–16]. In addition to the three-helix sequence, to study size dependence, we also look at the behavior of the corresponding one- and two-helix segments. Using the method of simulated tempering [17–19], a careful study of the thermodynamic properties of these different chains is performed.

Not unexpectedly, it turns out that the behavior of the model is strongly dependent upon the relative strength of the hydrogen-bond and hydrophobicity terms. In fact, the situation is somewhat reminiscent of what has been found for homopolymers with stiffness [20–23], with hydrogen bonds playing the role of the stiffness term. Throughout this paper we focus on one specific, empirical choice of parameters.

For this choice of parameters, we in particular find that the three-helix-bundle protein has the following three properties. First of all, it does form a stable three-helix bundle (except for a twofold topological degeneracy). Second, its folding transition is abrupt,

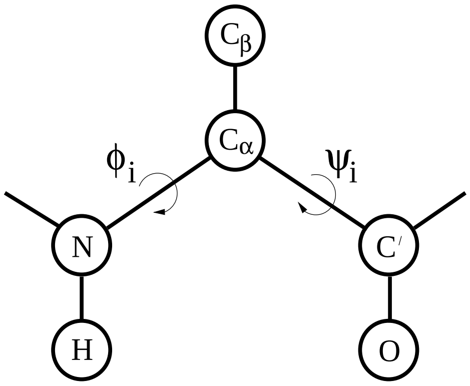

Figure 1: Schematic figure showing the representation of one amino acid.

from an expanded state to the native three-helix-bundle state. Third, compared to the one- and two-helix segments, it forms more stable secondary structure. It should be stressed that these properties are found without resorting to the popular Gō approximation [24], in which interactions that do not favor the desired structure are ignored.

2 The Model

The model we study is a reduced off-lattice model. The chain representation is illustrated in Fig. 1. As mentioned in the introduction, each amino acid is represented by five or six atoms. The three backbone atoms N, C_α and C′ are all included. Also included are the H and O atoms shown in Fig. 1, which we use to define hydrogen bonds. Finally, the side chain is represented by a single atom, C_β, which can be either hydrophobic or polar, or absent. This gives us the following three types of amino acids: A with hydrophobic C_β, B with polar C_β, and G (glycine) with no C_β.

The H, O and C_β atoms are all attached to the backbone in a rigid way. Furthermore, in the backbone, all bond lengths, bond angles and peptide torsional angles are held fixed.† This leaves us with two degrees of freedom per amino acid, the Ramachandran torsional angles ϕ_i and ψ_i (see Fig. 1).

Our energy function

$$E = E_{\mathrm{loc}} + E_{\mathrm{sa}} + E_{\mathrm{hb}} + E_{\mathrm{AA}} \tag{1}$$

†For a complete listing of geometry parameters, please send an e-mail request to the authors.

is composed of four terms. Roughly speaking, the first two terms, $E_{\rm loc}$ and $E_{\rm sa}$, enforce steric constraints, whereas the last two terms, $E_{\rm hb}$ and $E_{\rm AA}$, are the ones responsible for stability in this simplified force field.

The local potential $E_{\rm loc}$ has a standard form with threefold symmetry,

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

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

$$E_{\rm sa} = \epsilon_{\rm sa}\sum_{i<j}{}' \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}, \tag{3}$$

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

$$E_{\rm hb} = \epsilon_{\rm hb}\sum_{ij} u(r_{ij})v(\alpha_{ij},\beta_{ij})\,, \tag{4}$$

where

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

$$v(\alpha_{ij},\beta_{ij}) = \begin{cases} \cos^2\alpha_{ij}\cos^2\beta_{ij} & \alpha_{ij},\beta_{ij} > 90^\circ \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

In Eq. 4 i and j represent H and O atoms, respectively, and r_{ij} denotes the HO distance, α_{ij} the NHO angle, and β_{ij} the HOC$'$ angle. Any HO pair can form a hydrogen bond. The last term in Eq. 1, the hydrophobicity term $E_{\rm AA}$, has the form

$$E_{\rm AA} = \epsilon_{\rm AA}\sum_{i<j}\left[\left(\frac{\sigma_{\rm AA}}{r_{ij}}\right)^{12} - 2\left(\frac{\sigma_{\rm AA}}{r_{ij}}\right)^{6}\right], \tag{7}$$

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

The parameters of our energy function were determined largely by trial and error. The final parameters can be found in Table 1. The parameters σ_{ij} of Eq. 3 are given by

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

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

					σ_i(Å)							
ϵ_ϕ	ϵ_ψ	$\epsilon_{\rm sa}$	$\epsilon_{\rm hb}$	$\epsilon_{\rm AA}$	N	C_α	C′	H	C_β	O	$\sigma_{\rm hb}$ (Å)	$\sigma_{\rm AA}$ (Å)
1	1	0.0034	2.8	2.2	1.65	1.85	1.85	1.0	2.5	1.65	2.0	5.0

Table 1: Parameters of the energy function. Energies are in dimensionless units, in which the folding transition occurs at $kT \approx 0.65$ for the three-helix-bundle protein (see below).

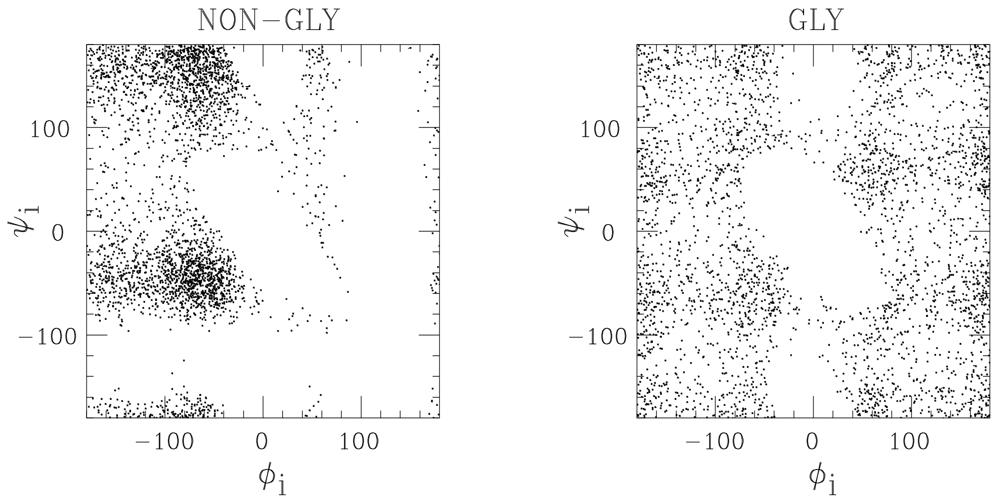

Figure 2: ϕ_i, ψ_i scatter plots for non-glycine and glycine, as obtained by simulations of the chains GXG for X=A/B and X=G, respectively, at $kT = 0.625$ (shown is ϕ_i, ψ_i for X).

where σ_i, σ_j can be found in Table 1, and $\Delta\sigma_{ij}$ is zero except for $C_\beta C'$, $C_\beta N$ and $C_\beta O$ pairs that are connected by three covalent bonds. In these three cases, we put $\Delta\sigma_{ij} = 0.625$Å. This could equivalently be described as a change of the local ϕ_i and ψ_i potentials. In Fig. 2 we show ϕ_i, ψ_i scatter plots for non-glycine (A and B) and glycine for our final parameters, which are in good qualitative agreement with the ϕ_i, ψ_i distributions of real proteins [8, 25].

Having determined local and self-avoidance parameters, we finally performed a set of trial runs to adjust the respective strengths $\epsilon_{\rm hb}$ and $\epsilon_{\rm AA}$ of the hydrogen-bond and hydrophobicity terms. These parameters were chosen based on the overall thermodynamic behavior of the chains. The values obtained this way do not seem unreasonable. At the folding temperature of the three-helix sequence (see below), we get $\epsilon_{\rm hb}/kT \approx 4.3$ and $\epsilon_{\rm AA}/kT \approx 3.4$.

In this model we study the three sequences shown in Table 2, which contain 16, 35

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

Table 2: The sequences studied.

and 54 amino acids, respectively. Following the strategy of Regan and DeGrado [11], the A and B amino acids are distributed along the sequence 1H in such a way that this segment can form a helix with all hydrophobic amino acids on the same side. The sequence 3H, consisting of three such stretches of As and Bs plus two GGG segments, is meant to form a three-helix bundle. This particular sequence was recently studied in a similar model with a different force field [16].

3 Results

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

An example of a simulated-tempering run is given in Fig. 3, which shows the Monte Carlo evolution of the energy E and radius of gyration R_g (calculated over all backbone atoms) in a simulation of the three-helix sequence. Also shown, bottom panel, is how the system jumps between the different temperatures. Two distinct types of behavior can be seen. In one case, E is high, fluctuations in size are large, and the temperatures visited are high. In the other case, E is low, the size is small and almost frozen, and the temperatures visited are low. Interesting to note is that there is one temperature, the next-lowest one, which is visited in both cases. Apparently, both types of behavior are possible at this temperature.

In Fig. 4a we show the specific heat as a function of temperature for the one-, two- and three-helix sequences. A pronounced peak can be seen, which gets stronger with increasing chain length. In fact, the increase in height is not inconsistent with a linear dependence on chain length, which is what one would have expected if it had been a conventional first-order phase transition with a latent heat.

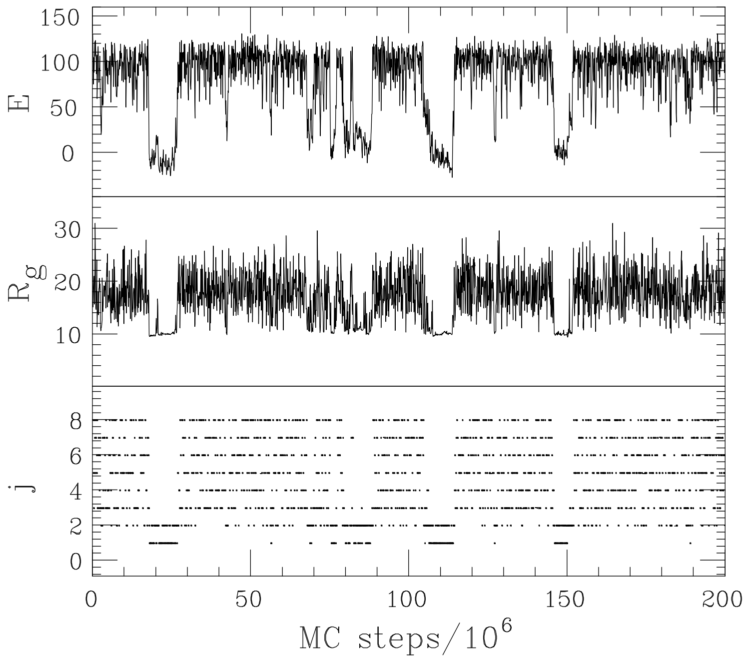

Figure 3: Monte Carlo evolution of the energy and radius of gyration in a typical simulation of the three-helix sequence. The bottom panel shows how the system jumps between the allowed temperatures T_j, which are given by $T_j = T_{\min}(T_{\max}/T_{\min})^{(j-1)/(J-1)}$ [26] with $kT_{\min} = 0.625$, $kT_{\max} = 0.9$ and $J = 8$.

Our results for the radius of gyration (not displayed) show that the specific heat maximum can be viewed as the collapse temperature. The specific heat maximum is also where hydrogen-bond formation occurs, as can be seen from Fig. 4b. Important to note in this figure is that the decrease in hydrogen-bond energy *per amino acid* with decreasing temperature is most rapid for the three-helix sequence, which implies that, compared to the shorter ones, this sequence forms more stable secondary structure. The results for the chain entropy shown in Fig. 4c provide further support for this; the entropy loss per amino acid with decreasing temperature is largest for the three-helix sequence.

We now turn to the three-dimensional structure of the three-helix sequence in the collapsed phase. It turns out that it does form a three-helix bundle. This bundle can have two distinct topologies; if we let the first two helices form a U, then the third helix can be either in front of or behind that U. The model is, not unexpectedly, unable to discriminate between these two possibilities. To characterize low-temperature conformations, we therefore determined two representative structures, one for each

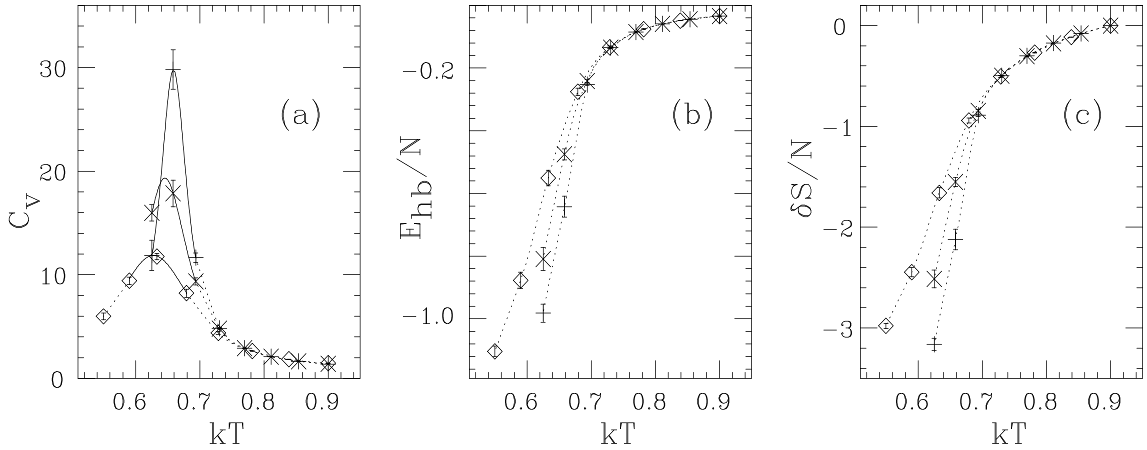

Figure 4: Thermodynamic functions against temperature for the sequences 1H ($\diamond$), 2H ($\times$) and 3H ($+$) in Table 2. (a) Specific heat $C_{\rm v} = (\langle E^2\rangle - \langle E\rangle^2)/Nk^2T^2$, N being the number of amino acids. (b) Hydrogen-bond energy per amino acid, $E_{\rm hb}/N$. (c) Chain entropy per amino acid, $\delta S/N = [S - S(kT = 0.9)]/N$. The full lines in (a) represent single-histogram extrapolations [27]. Dotted lines are drawn to guide the eye.

topology, which, following [16], are referred to as FU and BU, respectively. These structures are shown in Fig. 5. They were generated by quenching a large number of low-T structures to zero temperature, and we feel convinced that they provide good approximations of the energy minima for the respective topologies. Given an arbitrary conformation, we then measure the root-mean-square distances δ_i ($i = \mathrm{FU}, \mathrm{BU}$) to these two structures (calculated over all backbone atoms). These distances are converted into similarity parameters Q_i by using

$$Q_i = \exp(-\delta_i^2/100\text{Å}^2)\,. \tag{8}$$

At temperatures above the specific heat maximum, both Q_i tend to be small. At temperatures below this point, the system is found to spend most of its time close to one or the other of the representative structures; either $Q_{\rm FU}$ or $Q_{\rm BU}$ is close to 1. Finally, at the peak, all three of these regions in the $Q_{\rm FU}, Q_{\rm BU}$ plane are populated, as can be seen from Fig. 6a. In particular, this implies that the folding transition coincides with the specific heat maximum.

The folding transition can be described in terms of a single "order parameter" by taking $Q = \max(Q_{\rm FU}, Q_{\rm BU})$ as a measure of nativeness. Correspondingly, we put $\delta = \min(\delta_{\rm FU}, \delta_{\rm BU})$. In Fig. 6b we show the probability distribution of Q at the folding temperature. The distribution has a relatively sharp peak at $Q \approx 0.9$, corre-

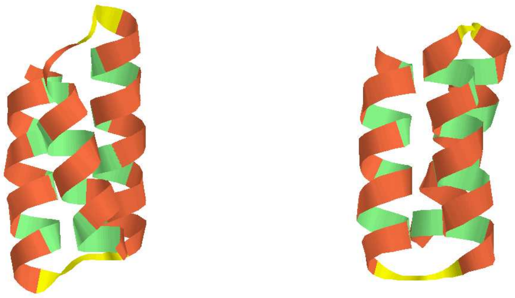

Figure 5: Representative low-temperature structures, FU and BU, respectively. Drawn with RasMol [28].

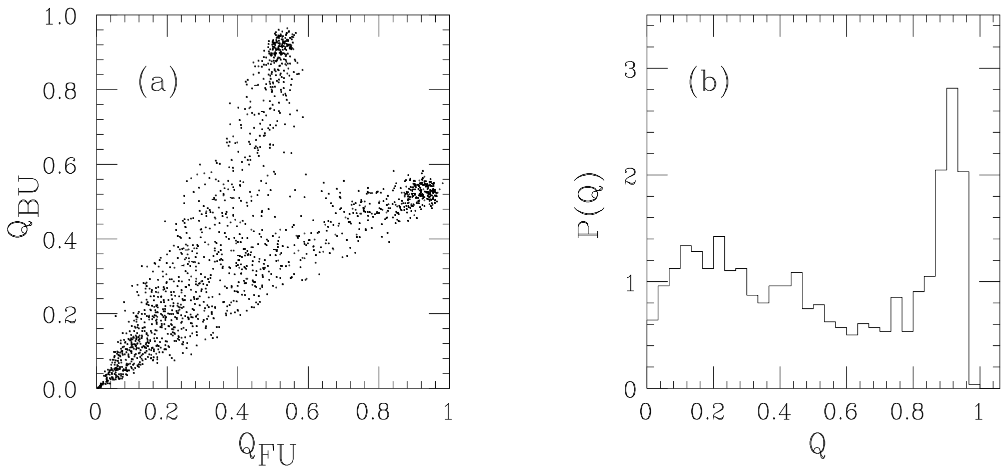

Figure 6: (a) $Q_{\rm FU}, Q_{\rm BU}$ (see Eq. 8) scatter plot at the specific heat maximum ($kT = 0.658$). (b) Histogram of Q at the same temperature.

sponding to $\delta \approx 3$Å. This is followed by a weak suppression of Q values around 0.7, corresponding to $\delta \approx 6$Å. Finally, there is a broad peak at small Q, where $Q = 0.2$ corresponds to $\delta \approx 13$Å.

What does the non-native population at the folding temperature correspond to in terms of $R_{\rm g}$ and $E_{\rm hb}$? This can be seen from the $Q, R_{\rm g}$ and $Q, E_{\rm hb}$ scatter plots in Fig. 7. These plots show that the low-Q part of the Q distribution, $Q = 0.2$ say, correspond to expanded structures with a varying but not high secondary-structure content. Although a detailed kinetic study is beyond the scope of this paper, we furthermore note that the free-energy landscapes corresponding to the distributions in Fig. 7 are relatively smooth. Consistent with that, we found that standard fixed-temperature Monte Carlo simulations were able to reach the native state, starting

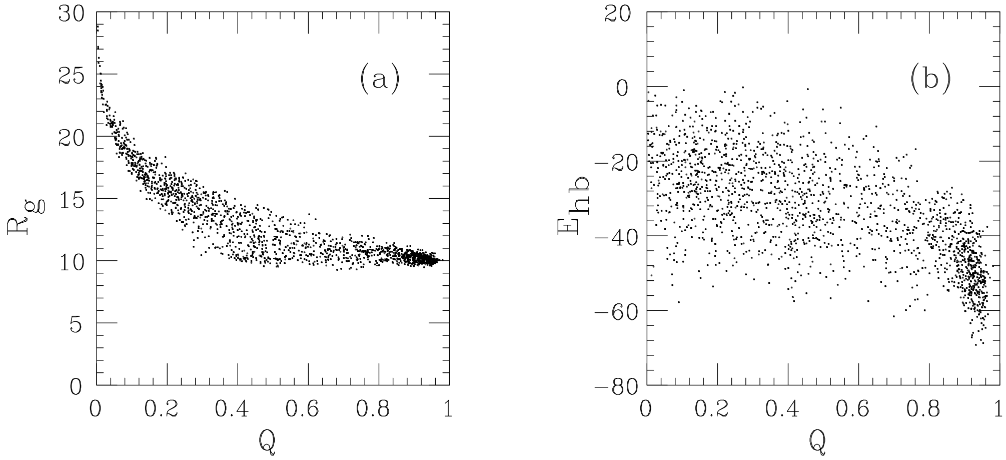

Figure 7: (a) Q, R_{g} and (b) Q, E_{hb} scatter plots at the folding temperature ($kT = 0.658$).

from random coils.

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

4 Summary and Outlook

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

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

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

A folding transition that takes the system directly from the unfolded state to the native one is what one expects for small fast-folding proteins. In order for the model to show this behavior, careful tuning of the relative strength of the hydrogen-bond and hydrophobicity terms, $\epsilon_{\rm hb}/\epsilon_{\rm AA}$, is required. This $\epsilon_{\rm hb}/\epsilon_{\rm AA}$ dependence may at first glance seem unwanted but is not physically unreasonable; $\epsilon_{\rm hb}$ can partly be thought of as a stiffness parameter, and chain stiffness has important implications for the phase structure, as shown by recent work on homopolymers [20–23]. Note also that incorporation of full side chains makes the chains intrinsically stiffer, which might lead to a weaker $\epsilon_{\rm hb}/\epsilon_{\rm AA}$ dependence.

Successful statistical-mechanical studies of proteins with similar lengths have recently been reported for C_α models [5, 14, 15] and also for a model at atomic resolution [29]. A fundamental difference between these studies and ours is that we do not use the Gō approximation; in fact, we use a simple three-letter alphabet.

Acknowledgements

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[29] Shimada, J., Kussell, E.L. & Shakhnovich, E.I. (2000) "The Folding Thermodynamics and Kinetics of Crambin Using a Monte Carlo Simulation at Atomic Resolution", submitted to *Proc. Natl. Acad. Sci. USA*.