Revised version
LU TP 02-36
November 24, 2002

A Minimalistic All-Atom Approach to Protein Folding

Anders Irbäck*

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

Submitted to *Journal of Physics: Condensed Matter*
(Proceedings of EPS Meeting on "Nano Physics in Life Systems",
Copenhagen, June 21-22, 2002)

Abstract:
Using simple sequence-based potentials, the folding properties of a designed three-helix-bundle protein, an α-helix and a β-hairpin are studied. The three-helix-bundle protein is modelled using 5–6 atoms per amino acid and is found to undergo a first-order-like folding transition at which chain collapse and helix formation cannot be separated, which is in accord with experimental data. The other two sequences are studied using a model that contains all atoms and are indeed found to make an α-helix and a β-hairpin, respectively, for exactly the same choice of parameters. Calculated melting curves are, moreover, in reasonable quantitative agreement with experimental data, for both peptides. The melting curves are found to be quite well described by a simple two-state model, although the energy distributions lack a clear bimodal shape.

PACS numbers: 87.15.Aa, 05.10.Ln

*E-mail: anders@thep.lu.se

1 Introduction

The folding of proteins to their functional states is a remarkable process [1]. In the cell, the folding process may require the assistance from helper molecules. However, as shown by refolding experiments, many proteins have the ability to fold spontaneously to their native states. This implies that the amino acid sequence contains all the information needed for the formation of the functional state [2]. The questions of how the folding process takes place and how the structure is encoded in the sequence are fascinating and in the focus of both experimental and theoretical research. In recent years there have been many advances in this area (for two recent reviews, see [3, 4]). However, to be able to simulate the folding process on the computer in atomic detail, is a goal that remains to be achieved.

The reason why simulating protein folding is a challenge is in part computational, but the computational difficulties are not necessarily insurmountable, as shown by recent all-atom studies [5, 6] of Gō-type [7] models with a bias towards the native structure. The most challenging part appears instead to be the search for suitable potentials. Extending the calculations of [5, 6] to entirely sequence-based potentials is indeed a task that remains to be accomplished.

There exist a number of semi-empirical (sequence-based) potentials that are being widely used to study various properties of proteins, usually through molecular dynamics simulations (for a review, see [8]). However, this is not yet a feasible method for studying the full folding process. To find a viable approach to this problem, it is of interest to study the behaviour of simpler and more transparent models, with fewer parameters to tune.

Here two simple sequence-based models are discussed, in which the folding process is driven by backbone hydrogen bonding and effective hydrophobicity forces (no explicit water). In the first model, the amino acid side chains are represented by large C_β atoms. This model has been applied to small helical proteins with about 50 amino acids [9–11]. The second model is an extension of the first one and contains all atoms. It has been tested on an α-helix with 21 amino acids and a β-hairpin with 16 amino acids [12].

The paper is organised as follows. Section 2 describes the two models. Sections 3 and 4 present results obtained using the large-C_β model and the all-atom model, respectively. A brief summary is given in section 5.

2 Models and Methods

2.1 Large-C_β model

In this model, each amino acid is represented by five or six atoms, three of which are the backbone atoms N, C_α and C'. Also included are the H and O atoms of the peptide units, which are used to define hydrogen bonds. The main simplification is that the side chain is represented by a single atom, a large C_β. The C_β atom can be either hydrophobic, polar or absent, which gives us three types of amino acids: H with hydrophobic C_β, P with polar C_β, and G (glycine) without C_β. All bond lengths, bond angles and peptide torsion angles (180°) are held fixed, which leaves us with two degrees of freedom per amino acid, the Ramachandran torsion angles ϕ and ψ.

The potential function

$$E = E_{\rm loc} + E_{\rm ev} + E_{\rm hb} + E_{\rm hp} \tag{1}$$

is composed of four terms. 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 excluded-volume term $E_{\rm ev}$ is given by a hard-sphere potential of the form

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

where the sum runs over all possible atom pairs except those consisting of two hydrophobic C_β. The parameter σ_{ij} is given by $\sigma_{ij} = \sigma_i + \sigma_j + \Delta\sigma_{ij}$, where $\Delta\sigma_{ij} = 0.625\,\text{Å}$ for $C_\beta C'$, $C_\beta N$ and $C_\beta O$ pairs that are connected by three covalent bonds, and $\Delta\sigma_{ij} = 0\,\text{Å}$ otherwise. The introduction of the parameter $\Delta\sigma_{ij}$ can be thought of as a change of the local potential.

The hydrogen-bond term $E_{\rm hb}$ has the form

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

where the functions $u(r)$ and $v(\alpha,\beta)$ are given by

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

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

The sum in equation (4) runs over all possible HO pairs, and r_{ij} denotes the HO distance, α_{ij} the NHO angle, and β_{ij} the HOC′ angle. The last term of the potential, the hydrophobicity term $E_{\rm hp}$, is given by

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

where the sum runs over all possible pairs of hydrophobic C_β.

To speed up the calculations, a cutoff radius r_c is used, which is taken to be 4.5 Å for $E_{\rm ev}$ and $E_{\rm hb}$, and 8 Å for $E_{\rm hp}$. Numerical values of all energy and geometry parameters can be found in [9].

2.2 All-atom model

This model contains all atoms, heavy ones as well as hydrogens. All bond lengths, bond angles and peptide torsion angles (180°) are, as in the previous model, held constant. Hence, each amino acid has the Ramachandran torsion angles and a number of side-chain torsion angles as its degrees of freedom (for Pro, ϕ is held fixed at $-65°$).

The potential function has three instead of four terms in this model. It has the form

$$E = E_{\rm ev} + E_{\rm hb} + E_{\rm hp}\,, \tag{8}$$

where the three terms represent excluded-volume effects, hydrogen bonds and effective hydrophobicity forces, respectively. The local potential of the previous model is missing. Having included all atoms, it became possible to eliminate this term.

The excluded-volume term $E_{\rm ev}$ has the same functional form as in the previous model. It is given by

$$E_{\rm ev} = \epsilon_{\rm ev} \sum_{i<j} \left[\frac{\lambda_{ij}(\sigma_i + \sigma_j)}{r_{ij}} \right]^{12}, \tag{9}$$

where $\lambda_{ij} = 1$ for all pairs connected by three covalent bonds and for HH and OO pairs from adjacent peptide units, and $\lambda_{ij} = 0.75$ otherwise. The pairs for which $\lambda_{ij} = 1$ strongly influence the shapes of Ramachandran maps and rotamer potentials. The reason for using $\lambda_{ij} < 1$ for the large majority of all pairs is both computational efficiency and the restricted flexibility of chains with only torsional degrees of freedom.

The hydrogen-bond energy $E_{\rm hb}$ has the form

$$E_{\rm hb} = \epsilon_{\rm hb}^{(1)} \sum_{\substack{j<i-2 \\ \text{or } j>i+1}} u(r_{ij}) v(\alpha_{ij}, \beta_{ij}) + \epsilon_{\rm hb}^{(2)} \sum u(r_{ij}) v(\alpha_{ij}, \beta_{ij}) \,, \tag{10}$$

where the first sum represents backbone-backbone hydrogen bonds, and the second sum represents interactions between charged side chains and the backbone as well as interactions between oppositely charged side chains. These different interactions are, for convenience, taken to have the same form. The interactions of the second sum have a relatively weak influence on the thermodynamic behaviours of the systems studied ($\epsilon_{\rm hb}^{(2)}$ is smaller than $\epsilon_{\rm hb}^{(1)}$). The function $u(r)$ in equation (10) is the same as before, see equation (5). The function $v(\alpha, \beta)$ is given by

$$v(\alpha, \beta) \;\; = \;\; \begin{cases} (\cos\alpha \cos\beta)^{1/2} & \text{if } \alpha, \beta > 90^\circ \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

and differs from that in equation (6) in that the exponent of the cosines is 1/2 instead of 2. This change was necessary because when sticking to the exponent 2, the structures became too regular in our β-hairpin study. The exponent 1/2 gives a more permissive angular dependence.

The hydrophobicity potential $E_{\rm hp}$ assigns to each amino acid pair an energy that depends on the amino acid types and the degree of contact between the side chains. It can be written as

$$E_{\rm hp} = \epsilon_{\rm hp} \sum M_{IJ} C_{IJ} \,, \tag{12}$$

where the sum runs over all possible amino acid pairs IJ except nearest neighbours along the chain. The M_{IJ}'s (≤ 0) are taken as the contact energies of Miyazawa and Jernigan [13] shifted to zero mean, provided that amino acids I and J both are hydrophobic and that the shifted contact energy is negative; otherwise, $M_{IJ} = 0$. Eight of the amino acids are classified as hydrophobic, namely Ala, Val, Leu, Ile, Phe, Tyr, Trp and Met. The geometry factor C_{IJ} in equation (12) is a measure of the degree of contact between amino acids I and J. C_{IJ} is calculated using a predetermined set of N_I atoms, denoted by A_I, for each amino acid I. For the aromatic amino acids Phe, Tyr and Trp, A_I consists of the C atoms of the hexagonal ring. The other five hydrophobic amino acids each have an A_I containing all its non-hydrogen side-chain atoms. With these definitions, C_{IJ} can be written as

$$C_{IJ} = \frac{1}{N_I + N_J} \Big[\sum_{i \in A_I} f(\min_{j \in A_J} r_{ij}^2) + \sum_{j \in A_J} f(\min_{i \in A_I} r_{ij}^2) \Big] \,, \tag{13}$$

where the function $f(x) = 1$ if $x < A$, $f(x) = 0$ if $x > B$, and $f(x) = (B-x)/(B-A)$ if $A < x < B$ [$A = (3.5\,\text{Å})^2$, $B = (4.5\,\text{Å})^2$]. Roughly speaking, C_{IJ} is a measure

of the fraction of atoms in A_I or A_J that are in contact with some atom from the opposite side chain.

The potential function is evaluated using a cutoff of 4.5 Å for E_{hb} and a pair dependent cutoff of $4.3\lambda_{ij}$ Å for E_{ev}. Numerical values of all the parameters of the model can be found in [12].

2.3 Numerical methods

The thermodynamic behaviours of these models were studied by using simulated tempering [14–16], in which the temperature is a dynamical variable. Both the temperature update and all side-chain updates were standard Metropolis steps. For the backbone degrees of freedom, three different elementary moves were used: first, the simple non-local pivot move in which a single torsion angle is turned; second, a semi-local method [17] in which seven or eight adjacent torsion angles are turned in a coordinated way; and third, a non-local symmetry-based update of three randomly chosen backbone torsion angles. The third move was only used in the study of the all-atom model. To see how this move works, consider the three bonds corresponding to the randomly chosen torsion angles. The idea is then to reflect the mid bond in the plane defined by the two others, keeping the directions of these two other bonds fixed. This can be achieved by turning the three torsion angles considered.

The main reason for choosing Monte Carlo methods rather than molecular dynamics for these studies was computational efficiency. In fact, if one decides to use a generalised-ensemble method (for a review, see [18]) such as simulated tempering in order to speed up the simulations, there is no longer any obvious advantage in using molecular dynamics for the conformational search. Instead, one should then try to exploit the possibility of using large "unphysical" Monte Carlo moves. With the methods described above, our high-statistics simulations of these different systems required from a few days up to roughly one week on a standard desktop computer.

For the α-helix and the β-hairpin, Monte Carlo-based kinetic simulations were also performed. These simulations are only meant to mimic the time evolution of the system in a qualitative sense. The methods used for our kinetic simulations differ from those for the thermodynamic runs in two ways: first, the temperature was held constant; and second, to avoid large deformations of the chain, the non-local backbone updates were not used but only the semi-local method [17].

3 Designed three-helix bundle protein

Using the large-C_β model described in section 2.1, a designed three-helix-bundle protein with 54 amino acids was studied [9]. This sequence is a truncated three-letter version [19, 20] of a four-helix-bundle protein *de novo* designed by Regan and DeGrado [21]. It consists of three identical stretches of H and P amino acids, connected by two GGG segments. The HP segment is such that it can make an α-helix with all hydrophobic amino acids on the same side.

The thermodynamic behaviour of the model depends strongly on the parameters $\epsilon_{\rm hb}$ and $\epsilon_{\rm hp}$ that sets the strengths of the hydrogen bonds and hydrophobicity forces, respectively. For a suitable choice of these parameters, the designed sequence was found to have the following properties [9]:

- It does form a stable three-helix bundle, except for a twofold topological degeneracy.
- It makes more stable secondary structure than the corresponding one- and two-helix segments, which is in accord with experimental results.
- It undergoes a first-order-like folding transition, directly from an expanded state to the three-helix-bundle state. Hence, the folding process has two-state character, which is a hallmark of many small single-domain proteins [22].

Figure 1 is a schematic illustration of representative structures for the two topologies, as obtained by energy minimisation. The difference between the two possible topologies is that if one lets the first two helices form a U, then the third helix can be either in front of (FU) or behind (BU) that U. The helices in these states are all right-handed, as they should, so the model is able to discriminate between right-handed and left-handed helices.

The model is, however, unable to distinguish between the two ways of arranging the helices. This is not surprising, given that the hydrophobicity potential of the model is pairwise additive and that the two topologies have similar C_β-C_β contact patterns [11, 24]. The C_β-C_β contact patterns of the two topologies are indeed very similar when taking thermal fluctuations into account, although the "ideal" structures FU and BU have far from identical sets of C_β-C_β contacts [24]. In order for the model to be able to discriminate between the two topologies, it is probably necessary to include multi-body terms and/or full side chains.

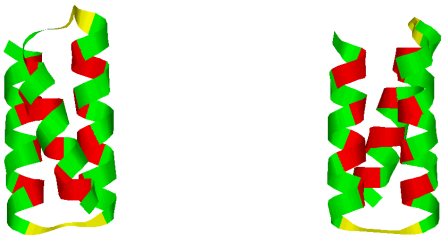

Figure 1: Representative structures for the two topologies, FU and BU. Drawn with RasMol [23].

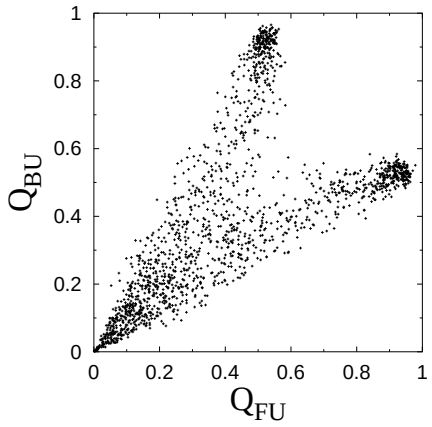

Figure 2: Distribution of the similarity parameters Q_{FU} and Q_{BU} at the collapse temperature. Q_{FU} is defined as $Q_{\mathrm{FU}} = \exp(-\delta_{\mathrm{FU}}^2/100\text{Å}^2)$, where δ_{FU} denotes the root-mean-square deviation (rmsd) from the state FU, calculated over all backbone atoms. The corresponding parameter for BU, Q_{BU}, is similarly defined.

Figure 2 illustrates the first-order character of the folding transition. It shows the joint distribution of two parameters Q_{FU} and Q_{BU} that are measures of similarity with the states FU and BU (see figure 1), respectively, at the collapse temperature. At this temperature, it can be seen that the folded state, where either Q_{FU} or Q_{BU} is close to 1, coexists with a state where both Q_{FU} and Q_{BU} are small. This shows that the folding transition takes place at the collapse temperature, and is first-order-like.

A fundamental issue in the characterisation of the folding process is whether chain collapse occurs before or after secondary-structure formation. In our model, these two processes cannot be separated [9,10]. This behaviour is in agreement with recent experiments on small helical proteins [25], and inconsistent with theories stipulating that either of these two processes must occur before the other.

One may ask how robust this conclusion is, because how fast the collapse is relative to helix formation depends strongly on the relative strength of the parameters $\epsilon_{\rm hb}$ and $\epsilon_{\rm hp}$. It is therefore important to stress that if the model is to have a compact native state and show two-state folding, there is not much freedom left in the choice of the ratio $\epsilon_{\rm hb}/\epsilon_{\rm hp}$ [10]. If $\epsilon_{\rm hb}/\epsilon_{\rm hp}$ is too large, the chain will not fold to a compact helical bundle. If, on the other hand, $\epsilon_{\rm hb}/\epsilon_{\rm hp}$ is too small, the folding transition gets weak; for example, it turns out that a relatively small decrease of $\epsilon_{\rm hb}/\epsilon_{\rm hp}$ is sufficient to make the peak in the specific heat (data not shown) much weaker [10].

It is also interesting to note that the phase behaviour of this model is somewhat reminiscent of that of a recently studied homopolymer model with stiffness [26–29]. The hydrogen bonds are then thought of as a stiffness term. For small but nonzero stiffness, the homopolymer model exhibits first a collapse and then a freezing transition with decreasing temperature. For large stiffness, these two transitions coincide, so that freezing occurs directly, without the formation of an intermediate globular state.

After adding two more amino acid types, the large-C_β model was also applied [11] to a real three-helix-bundle protein, the 10–55-amino acid fragment from the B domain of staphylococcal protein A (PDB code 1bdd), with quite good results. For example, energy minimisation restricted to the thermodynamically favoured three-helix-bundle topology gave a structure with an rmsd of 1.8 Å from the native structure, as determined by NMR [30].

4 α-helix and β-hairpin

Let us now turn to the two peptides. The β-hairpin studied is the second β-hairpin from the protein G B1 domain (amino acids 41–56), which has been the subject of seminal experimental studies. First, Blanco *et al.* [31] analysed this peptide in solution by NMR, and were able to show that the excised fragment adopts a structure similar to that in the full protein. Muñoz *et al.* [32] then showed, by tryptophan (Trp43) fluorescence experiments, that this β-hairpin shows two-state folding, like many small proteins. These experiments have stimulated a number of theoretical studies of this peptide, including simulations of atomic models with semi-empirical potentials [33–38]. Reproducing the melting behaviour of the β-hairpin has, however, proven non-trivial, as was recently pointed out by Zhou *et al.* [38]. The presence of a hydrophobic cluster (Trp43, Tyr45, Phe52, Val54) and sequence-specific hydrogen bonds in the turn region are two factors believed to be crucial for the stability of the

isolated β-hairpin [39].

The α-helix considered, the designed so-called F_S peptide, has also been extensively studied both experimentally [40–43] and theoretically [44]. The amino acid sequence of the F_S peptide is $AAAAA(AAARA)_3A$, where A is Ala and R is Arg.

Using the all-atom model described in section 2.2, these two sequences were found to have the following properties [12]:

- The two sequences do make a β-hairpin with the native topology and an α-helix, respectively. For a β-hairpin, there are two topologically distinct states with similar backbone folds but oppositely oriented side chains. The reason why the model prefers the native topology over the non-native one is that the formation of the hydrophobic cluster is sterically difficult to accomplish in the non-native topology.
- Melting curves for both peptides can to a good approximation be described by a simple (first-order) two-state model, with parameters that are in reasonable agreement with experimental data.
- Despite the apparent two-state character of the melting curves, the energy distributions lack a clear bimodal shape, which shows that a simple two-state description of the transition is an oversimplification.

Figure 3 is a free-energy plot for the β-hairpin at the temperature $T = 273\,\mathrm{K}$; it shows the free energy $F(\Delta, E)$ as a function of rmsd from the native β-hairpin, Δ, and energy, E. The global minimum of $F(\Delta, E)$ is found at 2–4 Å in Δ and corresponds to a β-hairpin with the native topology and the native set of hydrogen bonds between the two strands; the main difference between structures within this minimum lies in the shape of the turn. There are also two local minima, corresponding to a β-hairpin with the non-native topology ($\Delta \approx 5\,\text{Å}$) and an α-helix ($\Delta \approx 10\,\text{Å}$), respectively.

Figures 4a and 5a show melting curves for the β-hairpin and the α-helix, respectively. For the β-hairpin, the hydrophobicity energy E_{hp} is shown, which should be strongly correlated with Trp43 fluorescence. For the α-hairpin, the hydrogen-bond energy E_{hb} is studied. Also shown in these figures are fits of the data to a first-order two-state model. The fits are not perfect, but this can be detected only because the statistical errors are very small at the highest temperatures ($\sim 0.1\,\%$). For most practical purposes, the two-state description is accurate enough, as is evident from the figures. To further illustrate this point, a second fit to the same model was performed, this

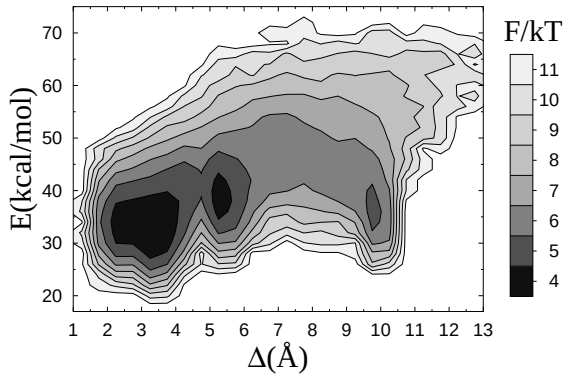

Figure 3: Free energy $F(\Delta, E) = -kT \ln P(\Delta, E)$ for the β-hairpin at $T = 273\,\mathrm{K}$. Δ denotes rmsd from the native structure, calculated over all non-hydrogen atoms (a backbone rmsd would be unable to distinguish the two possible β-hairpin topologies). $P(\Delta, E)$ is the joint distribution of Δ and energy, E. In the absence of a complete structure for the isolated β-hairpin, the native structure was taken from data for the full protein, as obtained by NMR [45] (PDB code 1gb1).

time assigning each data point an artifical uncertainty of 1%. This gave fits with $\chi^2/\mathrm{dof} \sim 1$ for both peptides.

In these calculations, the energy scale was set by taking the specific heat maximum for the β-hairpin (data not shown) to be the midpoint temperature $T_\mathrm{m} = 297\,\mathrm{K}$, as determined by Muñoz *et al.* [32]. Having calibrated the model this way, the specific heat maximum was found to occur at $T_\mathrm{m} = 310\,\mathrm{K}$ for the α-helix. Analyses of circular dichroism and infrared (IR) spectroscopy data for this peptide have given $T_\mathrm{m} = 303$, $308\,\mathrm{K}$ [41, 43] and $T_\mathrm{m} = 334\,\mathrm{K}$ [42], respectively.

The energy change ΔE obtained from the two-state fits (see figures 4 and 5) can also be compared with experimental results. Our fitted ΔE for the β-hairpin is $\Delta E = 9.3 \pm 0.3\,\mathrm{kcal/mol}$, which is $\sim 20\%$ smaller than the value $\Delta E = 11.6\,\mathrm{kcal/mol}$ obtained in [32], by a similar fit of tryptophan fluorescence data. This shows that the shape of the melting transition in our model is comparable to experimental data [32]. Atomic models studied previously have, by contrast, given a much weaker temperature dependence [38]. Our fitted ΔE for the α-helix is $\Delta E = 16.1 \pm 0.9\,\mathrm{kcal/mol}$, which may be compared to the value $\Delta E = 12 \pm 2\,\mathrm{kcal/mol}$ obtained by a two-state fit of IR data [42].

The first-order two-state fits in figures 4a and 5a look good and can be readily ex-

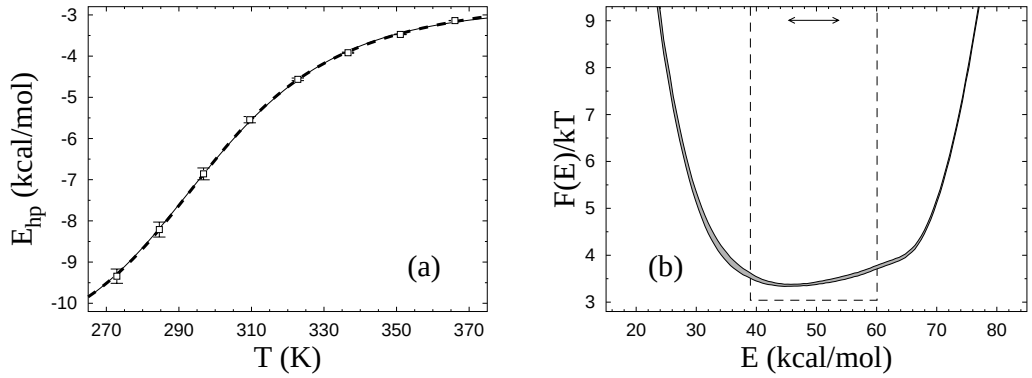

Figure 4: Unfolding of the β-hairpin. (a) Temperature dependence of the hydrophobicity energy E_{hp}. Squares represent the results from our simulations, with statistical 1σ errors. The solid and dashed curves (essentially coinciding) are fits of the data to the two-state expression $E_{\mathrm{hp}} = (E_{\mathrm{hp}}^{\mathrm{u}} + K E_{\mathrm{hp}}^{\mathrm{f}})/(1+K)$ and the square-well model (see text), respectively. The effective equilibrium constant K of the two-state fit has the first-order form $K = \exp[(1/kT - 1/kT_{\mathrm{m}})\Delta E]$. Both fits have three free parameters, whereas $T_{\mathrm{m}} = 297\,\mathrm{K}$ is held fixed. (b) Free-energy profile $F(E) = -kT \ln P(E)$ at $T = T_{\mathrm{m}}$. The shaded band is centred around the expected value and shows statistical 1σ errors. The double-headed arrow indicates ΔE of the two-state fit. The dashed line shows the free-energy profile corresponding to the square-well fit.

tended to higher orders, which may give the impression that the behaviours of these systems can be fully understood in terms of a two-state model. However, the two-state picture is far from perfect. This can be easily seen from the free-energy profiles $F(E)$ shown in figures 4b and 5b, which lack a clear bimodal shape. In particular, this shows that the parameters of the two-state fit must be interpreted with care, even if the fit is good. Given the calculated shapes of $F(E)$, it is instructive to perform an alternative fit, based on the assumptions that 1) $F(E)$ has the form of a square well of width ΔE_{sw} at $T = T_{\mathrm{m}}$, and that 2) the observable analysed varies linearly with E.[†] These square-well fits are shown in figures 4a and 5a, and the corresponding free-energy profiles $F(E)$ (at $T = T_{\mathrm{m}}$) are indicated in figures 4b and 5b. As expected, the square-well fits are somewhat better than the two-state fits. However, the difference is strikingly small, given the large difference between the underlying energy distributions.

[†]With these two assumptions, one finds that the average value of an arbitrary observable O at temperature T is given by $O(T) = \int_0^1 (O^{\mathrm{u}}(1-t) + O^{\mathrm{f}} t)\lambda^t dt \Big/ \int_0^1 \lambda^t dt = O^{\mathrm{u}} + (O^{\mathrm{f}} - O^{\mathrm{u}})(\frac{\lambda}{\lambda - 1} - \frac{1}{\ln \lambda})$, where $\lambda = \exp[(1/kT - 1/kT_{\mathrm{m}})\Delta E_{\mathrm{sw}}]$ and O^{u} and O^{f} are the values of O at the respective edges of the square well.

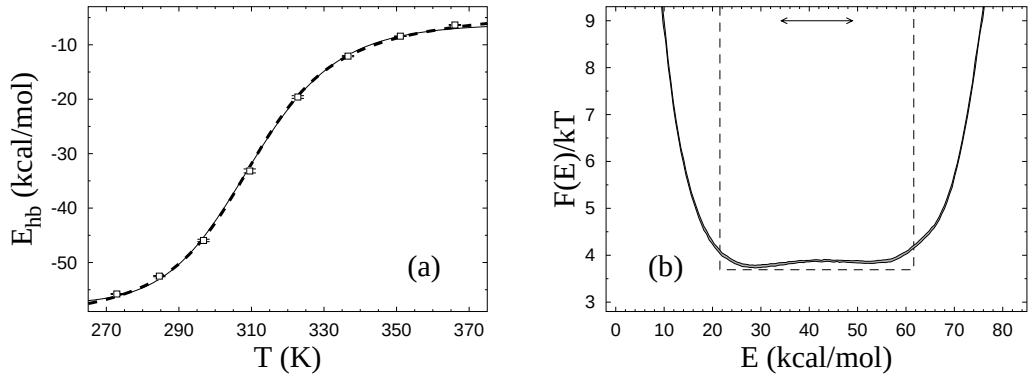

Figure 5: Unfolding of the α-helix. (a) Temperature dependence of the hydrogen-bond energy $E_{\rm hb}$, with the same two types of fits as in figure 4a (same symbols). (b) Free-energy profile $F(E) = -kT \ln P(E)$ at $T = T_{\rm m}$. Symbols as in figure 4b.

As mentioned in section 2.3, Monte Carlo-based kinetic simulations were also carried out for these peptides. Starting from equilibrium conformations at $T = 366\,\mathrm{K}$, the relaxation of ensemble averages was investigated at the respective melting temperatures, $T_{\rm m}$. The ensembles consisted of 1500 independent runs for each peptide. Except for a brief initial period of rapid change, the data were fully consistent with single-exponential relaxation for both peptides. This means that the clear deviations from perfect two-state behaviour seen in figures 4b and 5b are difficult to detect not only in the melting curves, but also in the relaxation data. The fitted relaxation time was larger for the β-hairpin than for the α-helix, by approximately a factor of 5. The corresponding factor is around 30 for experimental data [32, 42, 43].

5 Summary

Two simple protein models have been discussed, in which the process of folding is driven by backbone hydrogen bonding and effective hydrophobic attraction.

Using the large-C_β model, a designed three-helix-bundle protein was studied. The results show that this chain folds in a two-state manner, provided that there is a proper balance between hydrogen bonds and hydrophobicity forces. For this choice of parameters, it was further found that helix formation and chain collapse proceed in parallel; neither of the two processes can be said to occur before the other. It is worth noting that these calculations were carried out without resorting to the widely used

Gō prescription which, unless very carefully implemented, can make helix formation artifically fast relative to chain collapse [46].

Using the all-atom model, a β-hairpin and an α-helix were studied. The melting curves obtained for these sequences were in reasonable quantitative agreement with experimental data. The temperature dependence could, furthermore, to a good approximation be described by a simple two-state model, for both peptides. However, the energy distributions did not show a clear bimodal shape. This absence of bimodality is, in itself, maybe not surprising, because these systems are small and fluctuations therefore relatively large. What is striking is how difficult it is to detect deviations from the simple two-state picture when analysing the melting curves. These examples clearly demonstrate that drawing conclusions about the precise character of the folding transition can be a delicate task.

The interaction potentials of these models were deliberately kept simple. Extending the calculations to more general amino acid sequences will impose new conditions on the potential, and thereby make it possible and necessary to refine it.

Acknowledgements

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

References

[1] Branden C and Tooze J 1991 *Introduction to Protein Structure* (New York, Garland Publishing)

[2] Anfinsen C B 1973 *Science* **181** 223

[3] Dinner A R, Săli A, Smith L J, Dobson C M and Karplus M 2000 *Trends Biochem. Sci.* **25** 331

[4] Chan H S, Kaya H and Shimizu S 2002 *Current topics in Computational Biology* (eds. Jiang T, Xu Y and Zhang M Q; Cambridge, Massachusetts, MIT Press) pp. 403–447

[5] Kussell E, Shimada J and Shakhnovich E I 2002 *Proc. Natl. Acad. Sci. USA* **99** 5343

[6] Clementi C, García A E and Onuchic J N 2002 "Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: all-atom representation study of Protein L", preprint, submitted to *J. Mol. Biol.*

[7] Gō N and Abe H 1981 *Biopolymers* **20** 991

[8] Karplus M and McCammon J A 2002 *Nat. Struct. Biol.* **9** 646

[9] Irbäck A, Sjunnesson F and Wallin S 2000 *Proc. Natl. Acad. Sci. USA* **97** 13614

[10] Irbäck A, Sjunnesson F and Wallin S 2001 *J. Biol. Phys.* **27** 169

[11] Favrin G, Irbäck A and Wallin S 2002 *Proteins Struct. Funct. Genet.* **47** 99

[12] Irbäck A, Samuelsson B, Sjunnesson F and Wallin S 2002 preprint LU TP 02-28, submitted to *Proc. Natl. Acad. Sci. USA*
available at `www.thep.lu.se/complex/publications.html`

[13] Miyazawa S and Jernigan R L 1996 *J. Mol. Biol.* **256** 623

[14] Lyubartsev A P, Martsinovski A A, Shevkunov S V and Vorontsov-Velyaminov P N 1992 *J. Chem. Phys.* **96** 1776

[15] Marinari E and Parisi G 1992 *Europhys. Lett.* **19** 451

[16] Irbäck A and Potthast F 1995 *J. Chem. Phys.* **103** 10298

[17] Favrin G, Irbäck A and Sjunnesson F 2001 *J. Chem. Phys.* **114** 8154

[18] Hansmann U H E and Okamoto Y 1999 *Curr. Opin. Struct. Biol.* **9** 177

[19] Guo Z and Thirumalai D 1996 *J. Mol. Biol.* **263** 323

[20] Takada S, Luthey-Schulten Z and Wolynes P G 1999 *J. Chem. Phys.* **110** 11616

[21] Regan L and DeGrado W F 1988 *Science* **241** 976

[22] Jackson S E 1998 *Fold. Des.* **3** R81

[23] Sayle R and Milner-White E J 1995 *Trends Biochem. Sci.* **20** 374

[24] Wallin S, Farwer J and Bastolla U 2003 *Proteins Struct. Funct. Genet.* **50** 144

[25] Krantz B A, Srivastava A K, Nauli S, Baker D, Sauer R T and Sosnick T R 2002 *Nat. Struct. Biol.* **9** 458

[26] Kolinski A, Skolnick J and Yaris R 1986 *Proc. Natl. Acad. Sci. USA* **83** 7267

[27] Doniach S, Garel T and Orland H 1996 *J. Chem. Phys.* **105** 1601

[28] Bastolla U and Grassberger P 1997 *J. Stat. Phys.* **89** 1061

[29] Doye J P K, Sear R P and Frenkel D 1998 *J. Chem. Phys.* **108** 2134

[30] Gouda H, Torigoe H, Saito A, Sato M, Arata Y and Shimada I 1992 *Biochemistry* **31** 9665

[31] Blanco F J, Rivas G and Serrano L 1994 *Nat. Struct. Biol.* **1** 584

[32] Muñoz V, Thompson P A, Hofrichter J and Eaton W A 1997 *Nature* **390** 196

[33] Roccatano D, Amadei A, Di Nola A and Berendsen, H J C 1999 *Protein Sci.* **8** 2130

[34] Pande V S and Rokhsar D S 1999 *Proc. Natl. Acad. Sci. USA* **96** 9062

[35] Dinner A R, Lazaridis T and Karplus M 1999 *Proc. Natl. Acad. Sci. USA* **96** 9068

[36] García A E and Sanbonmatsu K Y 2001 *Proteins Struct. Funct. Genet.* **42** 345

[37] Zagrovic B, Sorin E J and Pande V 2001 *J. Mol. Biol.* **313** 151

[38] Zhou R, Berne B J and Germain R 2001 *Proc. Natl. Acad. Sci. USA* **98** 14931

[39] Kobayashi N, Honda S, Yoshii H and Munekata E 2000 *Biochemistry* **39** 6564

[40] Lockhart D J and Kim P S 1992 *Science* **257** 947

[41] Lockhart D J and Kim P S 1993 *Science* **260** 198

[42] Williams S, Causgrove T P, Gilmanshin R, Fang K S, Callender R H, Woodruff W H and Dyer R B 1996 *Biochemistry* **35** 691

[43] Thompson P A, Eaton W A and Hofrichter J 1997 *Biochemistry* **36** 9200

[44] García A E and Sanbonmatsu K Y 2002 *Proc. Natl. Acad. Sci. USA* **99** 2782

[45] Gronenborn A M, Filpula D R, Essig N Z, Achari A, Whitlow M, Wingfield P T and Clore G M 1991 *Science* **253** 657

[46] Shimada J, Kussell E L and Shakhnovich E I 2001 *J. Mol. Biol.* **308** 79