Molecular Physics, Vol. 105, No. 4, 20 February 2007, 375–385

Taylor & Francis
Taylor & Francis Group

Simple off-lattice model to study the folding and aggregation of peptides

NICOLAS COMBE*†‡ and DAAN FRENKEL§

†Laboratoire de Physique des solides de Toulouse, UMR 5477, Université Paul Saltier, 118 route de Narbonne, 31062 Toulouse cedex 4, France
‡Centre d'Elaboration de matériaux et d'Etudes Structurales (CEMES), CNRS UPR 8011, 29 rue J. Marvig, BP 94347, 31055 Toulouse cedex 4, France
§FOM Institute for Atomic and Molecular Physics, Kruislaan 407, 1098 SJ Amsterdam, The Netherlands

(*Received 31 October 2006; in final form 14 December 2006*)

We present a numerical study of a new protein model. This off-lattice model takes into account both the hydrogen bonds and the amino-acid interactions. It reproduces the folding of a small protein (peptide): morphological analysis of the conformations at low temperature shows two well-known substructures α-helix and β-sheet depending on the chosen sequence. The folding pathway in the scope of this model is studied through a free-energy analysis. We then study the aggregation of proteins. Proteins in the aggregate are mainly bound via hydrogen bonds. Performing a free-energy analysis we show that the addition of a peptide to such an aggregate is not favourable. We qualitatively reproduce the abnormal aggregation of proteins in prion diseases.

Keywords: Peptides folding; Aggregation; Molecular dynamics simulation

1. Introduction

The collective behaviour of polymers has been studied extensively, but this is not the case for bio-macromolecules. However, the aggregation of macromolecules such as proteins is of great practical importance. It is generally believed to play a role in prion diseases [1, 2], Alzheimer's disease [3, 4] and the formation of cataracts [5]. All these conditions appear to be related to the abnormal aggregation of proteins. But a good understanding of protein aggregation is also important for processes in the pharmaceutical [6] and food industry [7]. Yet, in spite of its importance, the physics of bio-molecular aggregation is poorly understood.

Early numerical studies of protein aggregation were reported by Gupta *et al*. [8]. Since then, many other model studies of this phenomenon have been reported [9–11]. The problem with the simulation of bio-molecular aggregation is that it requires a model that is sufficiently detailed to account for the specific intermolecular interactions that drive the aggregation, yet sufficiently cheap to allow numerical simulations of the collective behaviour of many bio-molecules.

In a recent study, we considered a lattice model of a protein solution based on the Gō Model [12]. The simplicity of this lattice model allowed us to determine the complete phase diagram of the system [13]. But lattice models suffer from serious drawbacks: in particular, their representation of the conformations of bio-molecules is so oversimplified that they can hardly be considered as representative of real bio-macromolecules. Clearly, to make progress in the modeling of the aggregation of real biomolecules, one must use more realistic models. Ideally, one would use a model where all the atoms of the protein and of the solvent are represented explicitly [14–16]. Unfortunately, the computational cost of such a model is such that it cannot be used to simulate the collective behaviour of a solution containing many realistic proteins.

This problem is, of course, well known and hence several authors have proposed more simple off-lattice models to study the collective behaviour of systems containing many proteins: for instance, Voegler-Smith and Hall [17, 18] studied the competition

*Corresponding author. Email: combe@cemes.fr

Molecular Physics
ISSN 0026–8976 print/ISSN 1362–3028 online © 2007 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/00268970601175483

between refolding and aggregation using such a model. Their work suggests that, in order to mimic the behaviour of real proteins, a model needs to account both for the effect of hydrogen bonding between amino acids and for the interaction between different amino-acid residues. The model used in [17, 18] is based on discontinuous potentials and can be studied by Monte Carlo or event-driven molecular dynamics simulations. Whilst this choice is computationally cheap, it cannot account for long-ranged interactions. Moreover, the use of discontinuous forces may lead to unrealistic folding dynamics.

In this paper, we propose a simple, off-lattice protein model that takes into account both the hydrogen bonds that are essential for the creation of secondary structures such as α-helices and β-sheets, and the interactions between side chains. We performed Molecular Dynamics simulations to study the thermal properties of this model and Monte Carlo simulations to gather information on the free energy of folding and on the aggregation of model proteins.

In the first section, we describe the model used. The second section is devoted to the study of the behaviour of a single protein. In the final section, we look at the aggregation of a folded protein and an existing aggregate.

2. Protein model

Any protein model, however simplified, must reflect some features that are dictated by the chemical structure of polypeptides, i.e. chains of amino acids. For the construction of proteins, Nature makes use of 20 different types of amino acids that only differ in their side chains [19]. To form a protein from these amino acids, units are linked by peptide bonds: the carboxyl group C′=O is linked to the nitrogen group N–H by an amino bond. As the nitrogen lone-pair is partially conjugated with the π-bond of the C′=O group, the four atoms NH–C′O are fixed in the same plane [20].

Let us consider the interactions that drive the folding and the aggregation of proteins. The two most important interactions are side chain–side chain interactions and hydrogen bonds. Among the side chain–side chain interactions, the hydrophobic interactions are thought to be the main driver for the folding of proteins. Following the usual description [19, 20], there are three types of side chains: hydrophobic, charged and polar. In addition, there can also be disulfide bridges. Such disulfide bridges are usually not found in intracellular proteins, but they occur quite frequently in extracellular proteins [20]. In this study, we will not take into account the disulfide bridge interactions.

Proteins occur in an aqueous environment (at least in living systems) and water can play a role in the folding. Indeed, hydrophobic side chains tend to pack in the interior of the protein. Charged and polar side chains interact through both Coulomb and Van der Waals forces [19].

Finally, hydrogen bonds occur between the oxygen lone-pair of C′=O and the hydrogens of the N–H of two different amino acids spatially close together. These hydrogen bonds play an important role in the most prevalent secondary protein structures, namely the α-helix and the β-sheet [19, 20].

2.1. *Model*

Our aim is to describe a protein by a simple model that retains the essential features of the interactions in real proteins, yet is sufficiently simple to make it computationally cheap. We therefore retain in our model the interactions between side chains and the hydrogen bonds. A related approach has recently been proposed independently by Chen *et al*. [21].

In order to reduce the computational cost of our simulations, we wish to minimize the number of particles in the model. To this end, we make the following approximations.

- We do not describe the solvent molecules. Solvent effects such as hydrophobic effects are taken into account through effective interactions between side chains.
- To account for hydrogen bonds, we do not explicitly simulate all the atoms NH–CO of each amino acid because they are in the same plane. Rather, we model this plane by a spin that can rotate perpendicularly to the $C_\alpha C_\alpha$ bond. Hydrogen bonds are taken into account through the interaction between spins.
- Side chains of proteins are represented by only one particle. They are thus modeled as spheres of different types, regardless of the size of the side chain and steric effects. We take into account only three different types of side chains: hydrophobic (H), polar positive (P) and polar negative (N). We did not introduce the 20 different types of amino acid because this would have increased the number of parameters of the model and made it more difficult to draw qualitative conclusions from our simulations. However, the model can easily be extended to account for the heterogeneity of amino acids.

Figure 1 shows a representation of our model. Since the NHCO group is modelled by a simple spin, the C_α carbon of an amino acid is not chiral. However, once

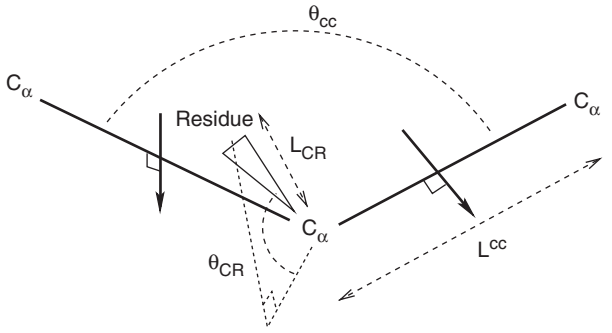

Figure 1. Model of an amino acid. The orientation of the CO–NH plane is described by a spin perpendicular to the $C_\alpha C_\alpha$ bond.

introduced into an amino-acid sequence, the C_α carbon becomes chiral most of the time and the polypeptide is chiral itself (except if the sequence of amino acids is symmetric).

Table 1 gives the values that we use for the different structural parameters. The values for the bond or pseudo-bond length and for the angles or pseudo-angles were calculated from known protein structural data: L^0_{CR} is fixed at the length of the usual sp^3 carbon–carbon bond, and θ^0_{CR} is fixed as in sp^3 carbon. L^0_{CC} was calculated from the atomic distances in amino acids [22]. Below, we briefly summarize the potential that determines the vibration and torsion of the peptide backbone.

(i) The lengths of the bonds and the angles between the bonds are constrained to be close to their equilibrium values L^0_{CC}, L^0_{CR} and θ^0_{CR} using harmonic potentials of strength $k_{\text{length–bond}}$ and k_{angle} respectively:

$$E_{L_{CC}} = 1/2k_{\text{length–bond}}(L_{CC} - L^0_{CC})^2, \quad (1)$$

$$E_{L_{CR}} = 1/2k_{\text{length–bond}}(L_{CR} - L^0_{CR})^2, \quad (2)$$

$$E_{\theta_{CR}} = 1/2k_{\text{angle}}(\theta_{CR} - \theta^0_{CR})^2. \quad (3)$$

In practice, the angle θ_{CC} depends on the orientation of the adjacent CO–NH planes and thus on the 'spins' in our model. Considering the different equilibrium distances between the atoms and the equilibrium angles between the bonds in a real amino acid [22], we calculated the angle θ_{CC} depending on the orientation of the CO–NH planes. The angle θ_{CC} varies between 1.4 and 2.4 rad. We do not explicitly take into account this dependence because the resulting potential would depend on the relative positions of three C_α atoms and on the orientation of two spins. Such a 'many-body' potential is computationally costly. Instead, we

Table 1. Structural data for our model.

L^0_{CC} (nm)	0.38
L^0_{CR} (nm)	0.15
θ^0_{CC} (rad)	1.910612
θ^0_{CR} (rad)	0.61549

allow θ_{CC} to vary freely between 1.4 and 2.4 rad using the following potential:

$$E_{\theta_{CC}} = 1/2k_{\text{angle}}(\theta_{CC} - 1.4)^2, \quad \text{if } \theta_{CC} < 1.4, \quad (4)$$

$$E_{\theta_{CC}} = 0, \quad \text{if } 1.4 < \theta_{CC} < 2.4, \quad (5)$$

$$E_{\theta_{CC}} = 1/2k_{\text{angle}}(\theta_{CC} - 2.4)^2, \quad \text{if } \theta_{CC} > 2.4. \quad (6)$$

(ii) The spins are located at the center of the $C_\alpha C_\alpha$ bond and are maintained perpendicular to the bond through a harmonic potential of strength $k_{\text{angle–spin}}$:

$$E_{\text{spin}_{CC}} = 1/2k_{\text{angle–spin}}(\theta_{\text{spin}_{CC}} - \pi/2)^2, \quad (7)$$

where $\theta_{\text{spin}_{CC}}$ is the angle between the spin and the $C_\alpha C_\alpha$ pseudo-bonds.

Finally, we need to define the potentials for the interactions between residues and for hydrogen bonds.

For the side chain–side chain interactions, we use a potential that can be either attractive or purely repulsive, depending on the type of side chain. The effective interactions between side chains account for solvent effects and for screened Coulomb interactions between charged particles. Two side chains of the same protein can interact only if they belong to amino acids separated by at least one amino acid in the chain. Table 2 lists the potentials used.

For the spin–spin interactions, we use the following potential:

$$V_{\text{spin}} = 4\epsilon_{\text{spin}}\left(\left[\frac{\sigma_{\text{spin}}}{r}\right]^{12} - \left(e^{-\lambda(\theta_I^2+\theta_J^2)} + e^{-\lambda((\Pi-\theta_I)^2+(\Pi-\theta_J)^2)}\right)\left[\frac{\sigma_{\text{spin}}}{r}\right]^8\right), \quad (8)$$

where r is the distance between the two spins, and θ_I and θ_J denote the angles between each spin and the line joining the two spins (see figure 2).

This potential is thus attractive if the two spins are parallel in the direction of the line joining the spins, and it becomes less and less attractive when the spins change their orientations. It can almost be purely repulsive if the angles are large compared with the value of λ. Two spins of the same protein can interact only if they belong to amino acids separated by at least two amino acids in the chain.

Table 2. Definition of the interactions between side chains. r is the distance between side chains. H indicates hydrophobic side chains, P polar positive side chains and N polar negative side chains. The values of ϵ_{HH}, ϵ_{HP}, ϵ_{PP}, σ_{HH}, σ_{HP} and σ_{PP} are defined in table 3.

$V_{HH} = 4\epsilon_{HH}[(\sigma_{HH}/r)^{12} - (\sigma_{HH}/r)^{6}]$
$V_{HP} = 4\epsilon_{HP}(\sigma_{HP}/r)^{12}$
$V_{HN} = 4\epsilon_{HP}(\sigma_{HP}/r)^{12}$
$V_{PN} = 4\epsilon_{PP}[(\sigma_{PP}/r)^{12} - (\sigma_{PP}/r)^{6}]$
$V_{PP} = 4\epsilon_{PP}(\sigma_{PP}/r)^{12}$
$V_{NN} = 4\epsilon_{PP}(\sigma_{PP}/r)^{12}$

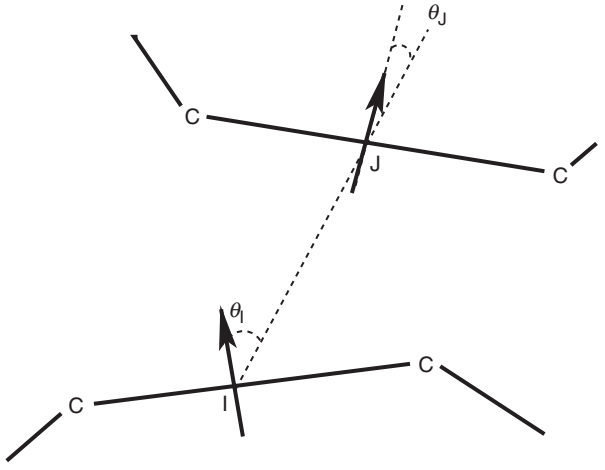

Figure 2. Definition of the angles θ_I and θ_J used in equation (6).

Table 3 gives the values of the different parameters for the side chain–side chain and the spin–spin potentials. $k_{length-bond}$ and k_{angle} are fixed to the values used for CC bonds in alkanes [23]. $k_{angle-spin}$ was chosen to give some flexibility to the hydrogen bonds. We did not attempt to optimize this parameter. ϵ_{spin} was chosen such that the depth of the spin–spin potential is 4.15 kT at 300 K, which is of the order of magnitude of known hydrogen-bond energies [24]. σ_{spin} is fixed to reproduce the hydrogen-bond length. ϵ_{HH}, ϵ_{HP} and ϵ_{PP} were adjusted such that the conformations of the proteins at low temperature depend on the sequence of amino acids: these three energies have been taken to be equal to reduce the number of parameters, although this is certainly not the case in reality. σ_{HH}, σ_{HP} and σ_{PP} are chosen to be reasonable estimates for the sizes of the groups they represent [18]. Finally, λ was fixed such that two spins experience an attraction if their directions

Table 3. Numerical values of the energy parameters of our model.

Bond potential parameters	
$k_{length-bond}$ (kcal mol^{-1} Å^{-2})	235.5
k_{angle} (kcal mol^{-1} rad^{-2})	60
$k_{angle-spin}$ (kcal mol^{-1} rad^{-2})	1
Side chain interaction parameters	
ϵ_{HH} (kcal mol^{-1})	1.1
ϵ_{HP} (kcal mol^{-1})	1.1
ϵ_{PP} (kcal mol^{-1})	1.1
σ_{HH} (nm)	0.36
σ_{HP} (nm)	0.36
σ_{PP} (nm)	0.36
Spin interaction parameters	
ϵ_{spin} (kcal mol^{-1})	4.36
λ (rad^{-2})	2
σ_{spin} (nm)	0.48

differ from the line joining the two spins by about 30°. Note that the spin–spin interactions are stronger than the side chain–side chain interactions. A similar trend has been observed in other models [17], where hydrogen bonds were six times stronger than side chain–side chain interactions.

In our model, the main chain does not interact with side chains, although it would be easy to add a repulsive interaction. However, side chain–side chain interactions prevent any overlap between the main chain and side chains: we have never experienced such a situation in our simulations.

This concludes the description of our model. In the next section we use this model to simulate the behaviour of an isolated protein depending on the sequence and on the temperature.

3. Properties of model isolated proteins

To investigate the properties of the protein model described above, we performed simulations to probe both the behaviour of isolated model proteins and of protein aggregates. We performed molecular dynamic simulations at constant temperature using a Nose–Hoover thermostat and a multiple-time-step integrator scheme [25]. As a demonstration, we used the present model to study oligopeptides consisting of 12 amino acid residues. Of course, such chains are short compared with most proteins, although it is worth stressing that several biologically active oligopeptides [26, 27] are known. In addition, oligopeptides can form amyloid fibers [14, 28]. We stress that there is no intrinsic limitation of the present model to short oligopeptides. We studied both the temperature

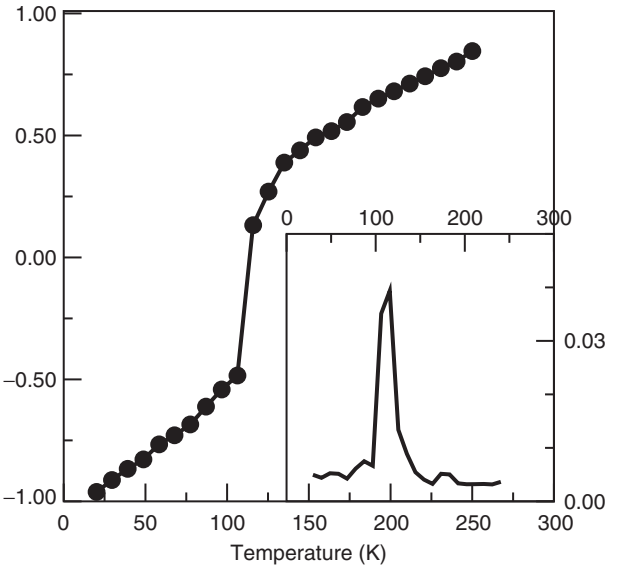

Figure 3. Temperature dependence of the average potential energy of the model protein shown in figure 4. The inset shows the heat capacity, i.e. the derivative of the potential energy with respect to temperature.

dependence of the internal energy of the model proteins and the conformational changes that the oligopeptide undergoes upon changing the temperature.

3.1. *Temperature dependence of the internal energy*

Figure 3 shows the average potential energy of a single model protein as a function of temperature. The steep part of the curve (corresponding to a peak in the heat capacity) is indicative of a transition between a coil state and a folded state. In the (high-temperature) coil state there are few hydrogen bonds or side chain–side chain interactions. Upon decreasing the temperature, the chain folds into a well-defined native state. The conformation of that state depends on the amino-acid sequence.

The location of the peak in the heat capacity provides us with an estimate of the transition temperature: it is about $T_f = 115\,K$. In the following, all temperatures will be normalized by T_f. This temperature is low compared with the typical folding temperatures of real proteins (about 310 K) [29]. Of course, our results depend upon the choice of the energy parameters ϵ_{HH}, ϵ_{HP}, ϵ_{PP} and ϵ_{spin}. Within the constraints of the rather simple model that we use, we have chosen to take a realistic value for the hydrogen bonding. We have not attempted a systematic optimization of all force-field parameters in order to obtain, simultaneously, a realistic estimate for the energy of hydrogen bonding and a realistic folding temperature. The aim of the present work is primarily to illustrate that, even without much fine-tuning, our model exhibits protein-like behaviour. We expect (but have not tested) that more quantitative agreement with experiment can be achieved by force-field 'fitting'.

3.2. *Sequence dependence at low temperature*

As already mentioned, when the temperature decreases below the heat-capacity peak, the chain folds into a well-defined native state. We find that the low-temperature morphology depends on the sequence. Moreover, some of the observed conformations resemble the well-known α-helix and β-sheet. Depending on the sequence, our model can exhibit conformations that involve one or both substructures.

We stress that the folding in the α-helix conformation is not driven by a torsional potential [30]—rather, the protein spontaneously folds in that conformation. However, although the protein is chiral, our model does not distinguish between left-handed and right-handed helices: this drawback is a consequence of the non-chirality of the amino acids. Hence, the conformations that differ only in helicity are degenerate. This degeneracy between L and R helices can be broken by making the amino acids chiral. Figures 4 and 5 show two folded proteins, one in an α-helix conformation, the other in a β-sheet conformation. The only difference between the two conformations is the sequence of amino acids; all other parameters are the same.

In the α-helix conformation, hydrogen bonds are created between spins n and $n+3$ in such a way that they are roughly parallel to the axis of the α-helix. Because a spin in our model simulates the NH–CO group, this would correspond in real proteins to a hydrogen bond between C′=O of amino acid n and NH of amino acid $n+4$, as is indeed observed experimentally [20]. In other words, our model obtains the correct number of amino acids per α-helix turn. In β-sheets, our model generates hydrogen bonds perpendicular to the protein backbone, but within the plane of the β-sheet, as it should.

As a first conclusion, our model reproduces three important characteristics of real proteins.

- The protein can occur in two states depending on the temperature. At high temperatures, the protein is in a coil state, and at low temperatures it folds into a 'native' state.
- The conformation at low temperature is unique (except for handedness) and is sequence dependent.
- The conformations at low temperatures contain the same substructures (α-helix and β-sheet) as observed in real proteins.

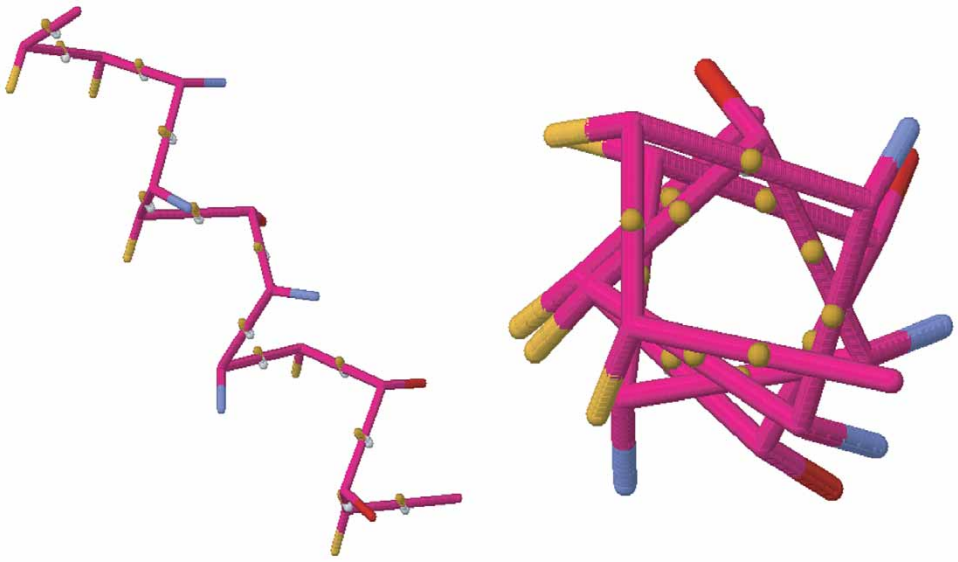

Figure 4. Folded conformation found at low temperature for the helix; the sequence used is $HP^2HN^2PHN^2H^2$. We show here the conformation from two different orientations. The colour code for the side chain is yellow for hydrophobic side chains, red for polar positive side chains, and blue for polar negative side chains. Spins are represented by a white and yellow stick to give their direction.

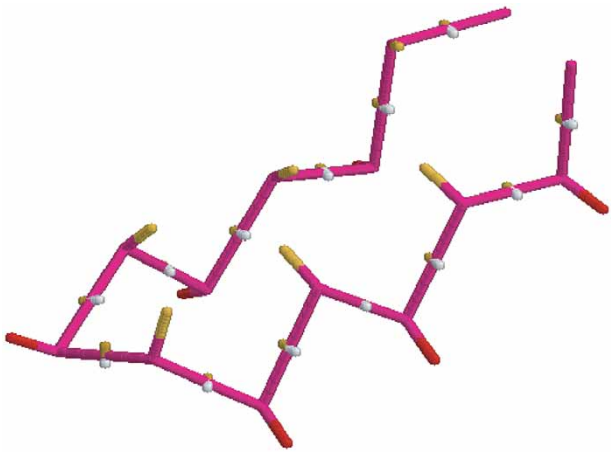

Figure 5. Folded conformation found at low temperature for the helix β-sheet; the sequence used is PHPHPHPHPHPH. The colour code for the side chain is as described in figure 4.

To our knowledge, the present coarse-grained model is one of the simplest that reproduces both the α-helix and β-sheet structures using only three amino-acid types. Of course, there exist other coarse-grained models that reproduce the α and β structures. However, this is either achieved by imposing a dihedral potential that facilitates helix formation [31] or by using a more complex (20 amino acid) 'alphabet' [32] where the strength of the interactions between side chains is estimated on the basis of the observed frequency of contacts between a specific side chain pair. Recently, a model has shown that the native-state folds of proteins can emerge on the basis of considerations of geometry and symmetry [33]. Finally, some aspects of protein folding can be reproduced with models based on a Gō model [12]. However, the Gō model is designed to favour a particular target state (the 'native' state) because it assumes that only those side chains that are nearest neighbours in the native state can attract each other. The amino-acid alphabet for the Gō model is therefore unbounded, as it increases with the number of nearest-neighbour contacts in the native state. By comparison, our model has the advantage that it does not have the properties of the native state built in and, moreover, it is very simple, as we introduce only two kinds of interactions and three types of amino acids. Below, we discuss the role and strength of both interactions.

3.3. *Folding pathway*

In our model, protein folding is the result of competition between the formation of hydrogen bonds and side chain–side chain interactions. Looking at the numerical values for the interaction strengths in table 3, it is clear that a hydrogen bond is more favourable energetically than a side chain–side chain bond, but the attraction between 'spins' is of shorter range than that between side chains (see table 2 and equation (8)). Due to the strong binding energy between spins, the lower-energy state of our model is always the α-helix conformation regardless

of the sequence of amino acids: the α-helix conformation maximizes the number of hydrogen bonds. Hydrogen bonds would favour a small number of amino acids per helix turn. However, this is frustrated by the energetic cost to reduce the angle between α-carbon atoms. The lowest-energy state corresponds to 3.5 amino acids per turn. As explained in section 2.1, our definition of the potential E_{CC} allows free variation of the angle between consecutive α-carbons in the range between 1.4 and 2.4 rad. If we would have constrained this angle to take a value of around 1.9 rad, the lowest-energy structure would be one where hydrogen bonds form between spins n and $n+4$, something that is not observed in real proteins.

As a consequence of the strong binding energy between spins, an isolated β-sheet, as shown in figure 5, corresponds to a metastable state. However, in Molecular Dynamics simulations, the formation of these structures is often kinetically favoured. This is so because one can choose a sequence of amino acids that favours the β-sheet structure: as the attraction between these side chains is relatively long ranged, one finds that the kinetics of the folding process can favour β-sheet formation, even though, for an isolated protein, this is not necessarily the most stable state. As a result, the conformation of the folded protein is sequence dependent.

3.4. *The coil–native transition*

To gain insight into the relative stability of different protein conformations, we performed Monte Carlo simulations using local moves. Using Umbrella Sampling (see, e.g., [25]), we computed the free energy of the system as a function of order parameter q that characterizes the degree of folding of the protein. We chose the following order parameter:

$$q = -\sum \frac{V_{\text{spin}}}{\epsilon_{\text{spin}}}. \tag{9}$$

The sum is performed over all allowed couples (two spins can interact if they are separated by at least two amino acids) of spins in the chain and V_{spin} is given by equation (8). This parameter thus approximately corresponds to the number of hydrogen bonds in the system.

In our Umbrella Sampling simulations, we bias the Hamiltonian by a harmonic potential of the form $W = 1/2k(q - q_0)^2$, where k and q_0 are parameters that can be varied at will. From these simulations, we obtain the free-energy curve around q_0, up to a constant. To obtain the full curve, we use the continuity of the free energy as a function of q. Actually, we look for the best polynomial of order eight that fits the curves.

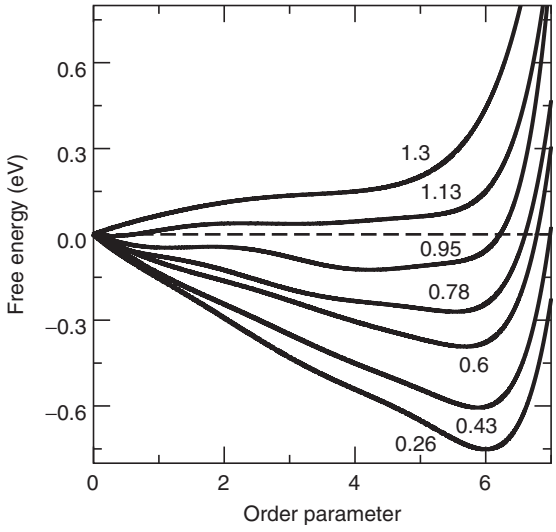

Figure 6. Free energy of the protein shown in figure 4 as a function of the order parameter for different temperatures T/T_f: 0.26, 0.43, 0.6, 0.78, 0.95, 1.13 and 1.3 (from bottom to top).

Figure 6 shows the free-energy curve as a function of the order parameter for different temperatures.

At temperatures well below the coil–native transition temperature ($T/T_f = 0.26$, for instance), the free energy is lowest for high values of the order parameter (folded chain). At high temperatures ($T/T_f = 1.3$, for instance) the stablest state has a low value of the order parameter (coil state) with almost no hydrogen bonding. At the transition temperature, a small free-energy barrier separates the two states, suggesting that the transition may become first order for a sufficiently long chain. However, we have not computed this free-energy barrier as a function of the size of the protein. Figure 6 also allows us to estimate the transition temperature: here between $T/T_f = 0.95$ and $T/T_f = 1.13$, in agreement with the data of figure 3.

In summary, our model reproduces the two-state behaviour of real short proteins and the resulting folded conformation contains secondary structures that resemble those of real proteins. Below, we consider the possible aggregation of proteins.

4. Aggregation of proteins

In spite of the fact that our model proteins are computationally much cheaper than full-atom models, it is still prohibitively expensive to compute a full phase diagram, using systems containing many hundreds of proteins. Instead, we have studied the aggregation of a small number of proteins.

4.1. *Stability of aggregates*

In section 3, we showed that folding is initially driven by long-ranged side chain–side chain interactions and that the short-ranged hydrogen bonds stabilize the resulting structure. For isolated proteins, α-helices are more stable than β-sheets, because the latter have fewer hydrogen bonds. However, this energetic disadvantage of β-sheets does not apply if the remaining hydrogen bonds are involved in inter-protein interactions. This suggests that β-sheets could be stabilized by the formation of protein aggregates involving inter-protein hydrogen bonds. These hydrogen bonds are perpendicular to the β-sheet plane of the individual proteins.

This phenomenon is illustrated in figure 7 where we show that two proteins that have been designed to form an α-helix, when isolated (see figure 4), form a stable β-sheet-like dimer. We emphasize that the stability of the aggregate shown in figure 7 is due to inter-molecular hydrogen bonds.

To investigate the stability of such an aggregate, we performed a Molecular Dynamics simulation of the aggregate, varying the temperature from $T/T_f = 0.2$ to $T/T_f = 1.7$, where T_f is the folding temperature introduced in section 3.1. Figure 8 shows the variation of the average potential energy as a function of temperature.

A clear change occurs around $T/T_f = 1.1$. Direct inspection of the structures generated in the simulation shows that this transition corresponds to the break up of the aggregate. This major transition is preceded by a smaller one around $T/T_f = 0.8$. This transition corresponds to a reorganization of the aggregate from the structure shown in figure 7 to that shown in figure 9.

This morphology is more stable than that proposed in figure 7. Here, the two proteins are linked by side chain–side chain interactions and, especially, by hydrogen bonds. We emphasize that here the proteins in the aggregate are identical and that their sequence is designed so that an isolated protein folds in a helix substructure. Also, one should note that some hydrophobic interactions occur between the two proteins: the side chain–side chain interactions therefore drive both the folding and the aggregation of proteins.

Also in larger aggregates we observe protein arrangements similar to that shown in figure 9. Comparing our results with the work of Petkova *et al.* [34], our simulation shows the spontaneous formation of an aggregate rather similar to the β-amyloid fibrils occurring in prion diseases. Indeed, Petkova *et al.* [34] have recently provided a structural model for Alzheimer's β-amyloid fibrils based on experimental constraints. They showed that these fibrils may have a structure

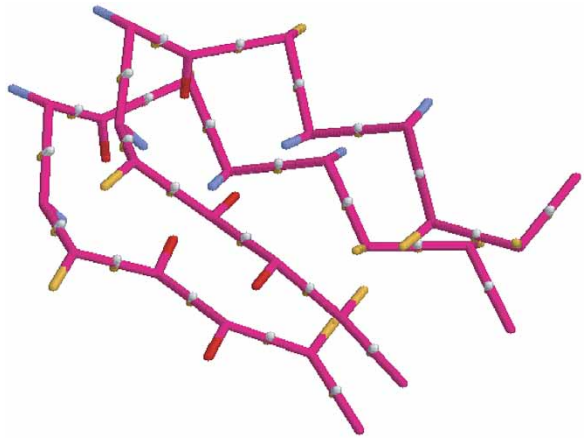

Figure 7. Aggregate formed with two proteins identical to that shown in figure 4.

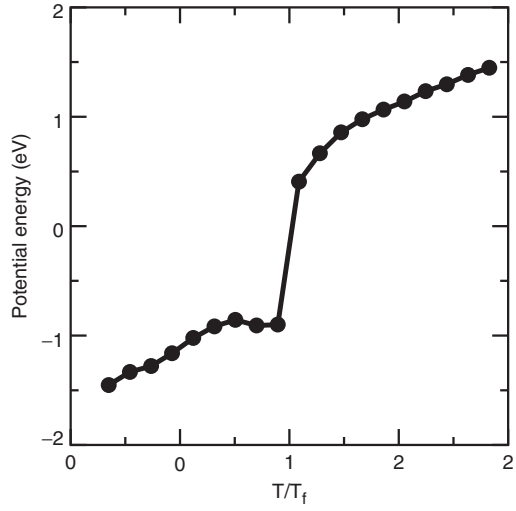

Figure 8. Potential energy of the aggregate shown in figure 7 as a function of the reduced temperature T/T_f where T_f is the folding temperature of the protein.

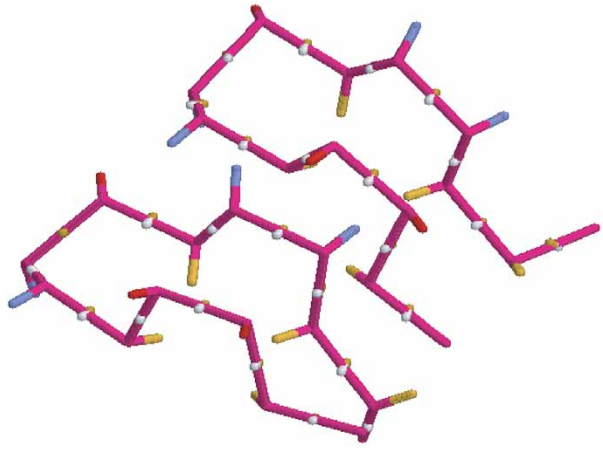

Figure 9. Morphology of the aggregate at $T/T_f = 0.94$.

analogous to that presented in figure 9 (with a large number of proteins). Moreover, it has been suggested [34] that fibril formation is driven by hydrophobic interactions. The picture of aggregation we present here is thus comparable to that of Petkova *et al*. except that, in our case, the final state of the aggregate is mainly stabilized by the hydrogen bonds, whereas one would reasonably expect a stronger stabilization from hydrophobic effects as suggested by Petkova. Recently, Nelson *et al*. [35] have even suggested that "opposing side chains do not form hydrogen bonds" and that interactions between β-sheet-like proteins are due to van der Waals interactions. Clearly, our model effectively attributes this hydrophobic effect to hydrogen bonds and this may explain why we over-estimate the strength of hydrogen bonds. We return to this point in the conclusions.

4.2. *Growth of the aggregate*

An aggregate such as that shown in figure 9 can grow by addition of another protein. But, since the temperature is lower than the folding temperature, the added protein will first have to unfold to be able to form hydrogen bonds with the existing aggregate. Since hydrogen bonds involve the strongest interaction of the model in terms of bond energy, we can do a very simple estimation of the energy balance of such an operation. In fact, we can just compare the number of unsatisfied hydrogen bonds in the case of an aggregate of n macromolecules and a free helix protein, and an aggregate of $n+1$ molecules. Basically, in the helix, only the top and bottom surfaces of the cylinder show unsatisfied hydrogen bonds: we denote this number by $2S$. On the aggregate, only the lateral surfaces have unsatisfied hydrogen bonds: this number we denote by $2 * L/2$ where L is a measure of the length of the peptide. A crucial point is that the number of unsatisfied hydrogen bonds in the aggregate is independent of the number of peptides in the aggregate. Thus, by adding a folded protein to the aggregate, the number of unsatisfied hydrogen bonds should be reduced by the number of unsatisfied hydrogen bonds in the initial added chain: we can thus expect such an operation to be energetically favourable. The number of unsatisfied bonds gained is $2S$.

This analysis is incorrect for the formation of the first aggregate of two proteins since, in that case, the initial system in composed of two folded proteins in a helix conformation and no aggregate exists. Thus the price in hydrogen bonds is $L-4S$ and, for long proteins, L is greater than $2S$.

Following this last analysis based on energy arguments, we can draw a schematic representation of the expected energy landscape presented in figure 10 as a function of the number of proteins in the aggregate. Similar free-energy landscapes play a role in the kinetics of formation of lamellar polymer crystals [36, 37].

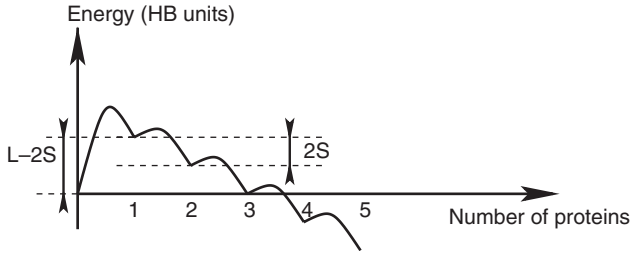

Figure 10. Schematic representation of the expected energy landscape in hydrogen bond units as a function of the number of proteins in the aggregate. $L/2$ denotes the number of hydrogen bonds that connect the two peptides shown in figure 9 and S denotes the number of unsatisfied hydrogen bonds on one end of a helix like that shown in figure 4.

To check this scenario, we performed a Monte Carlo simulation using Umbrella Sampling to study the free-energy landscape for aggregation.

To measure the progress of the aggregation, we define an order parameter q:

$$q = \sum_i \left[\overrightarrow{C_iC_{i+1}} \wedge \overrightarrow{C_{i+1}C_{i+2}}\right].\left[\overrightarrow{C_{i+1}C_{i+2}} \wedge \overrightarrow{C_{i+2}C_{i+3}}\right], \tag{10}$$

where C_i denotes the ith alpha carbon of the chain. The definition of q has been chosen such that q is large when the chain is folded in a helix conformation. The initial configuration is that shown in figure 9; q is calculated only for one chain and the other chain is fixed during the simulation. We performed a free-energy calculation using Umbrella Sampling [25] over a range of temperatures using a bias potential of the form $W = 1/2k(q-q_0)^2$. Twenty values of q_0 were explored, ranging from $q_0 = 7$ to $q_0 = 18$. In every 'window' we performed twenty million Monte Carlo cycles. The biasing allows us to explore the regions of configuration space where the protein detaches from the aggregate whilst folding into a helix conformation. The free-energy curves are estimated from these simulations by determining a polynomial of order eight that fits the simulated free-energy data. Note that the free energy is only determined up to an additive constant. Figure 11 shows how the free energy of the system varies with the order parameter q. In a helix conformation, the value of q is high, roughly about 13 to 14, whereas for a protein (β-sheet-like conformation) in an aggregate, this value is smaller, $q \approx 9$–10.

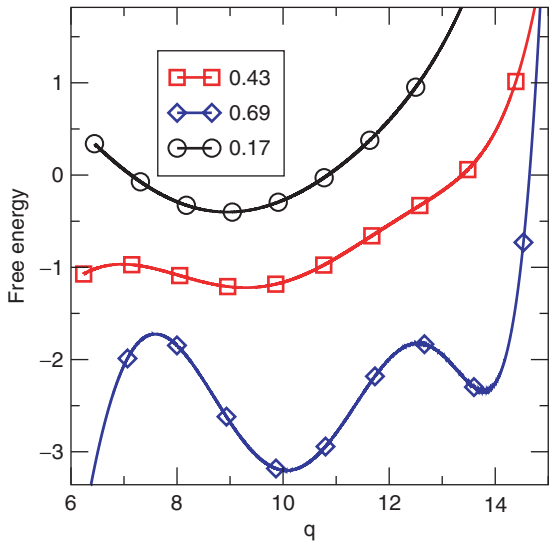

Figure 11. Free-energy change (in eV) associated with the detachment of a protein from an existing aggregate. The free energy is plotted as a function of the order parameter q defined in the text. The temperature is in units of the folding temperature T_f.

The free energy (figure 11) shows that the aggregate is very stable at low temperatures ($T/T_f = 0.17$, $T/T_f = 0.43$) since only one minimum appears, at $q \approx 10$, and no metastable states exist. At higher temperature ($T/T_f = 0.69$), the free-energy landscape exhibits two minima ($q \approx 10$ and $q \approx 14$). One minimum corresponds to the original aggregate, $q \approx 10$, and the other to an aggregate with one α-helix almost detached, $q \approx 14$. The latter minimum, present at very small values, $q < 6$, is an artefact of our simulations. The metastable configuration is one where the helix is still close to the aggregate: the two structures are still connected through the last hydrogen bonds on the top of the helix. Thus, coming back to the scenario sketched in figure 10, we see that the relative propensity for aggregation and folding depends on the temperature, emphasizing the importance of entropy in the aggregation process.

Our simulations suggest that, at low temperatures ($T/T_f \lessapprox 0.7$), proteins should spontaneously aggregate to existing fibrils. This tentative conclusion should be treated with caution as our calculation does not model the addition of a completely free helix protein to a fibril, but the folding of a protein incorporated in an aggregate in a helix conformation. In particular, our calculation only partly takes into account the entropy associated with the volume of the simulation cell: before aggregating to a protein, the protein first has to find the aggregate in the cell, which is not described in our case. Computation of the free energy needed to add a free helix protein to an existing fibril is feasible, but expensive.

Interestingly, the free-energy barrier for aggregation at the higher temperature in figure 11 is of the order of 1 eV. This value should be compared with the energy of a single hydrogen bond: 103 meV. This means that the order of magnitude of the free-energy barrier for aggregation is 10 hydrogen bonds. This barrier corresponds to the free energy needed to unfold a protein in the helix state and aggregate it into a fibril.

5. Conclusion

In this paper we have analysed the properties of an off-lattice, coarse-grained protein model. Even though it is very simple, it qualitatively reproduces several key properties of real proteins. In particular, the model proteins can fold into a β-sheet or an α-helix structure, depending on the amino-acid sequence. Moreover, we have shown that small aggregates spontaneously organize into fibrils. Considering the simplicity of the model, it is encouraging that it can account for these important properties of real proteins. However, the model also has some serious drawbacks. Most important among these is the role attributed to hydrogen bonds. The energy for the hydrogen bond is realistic ($4kT$ for our model), whilst the energy for side chain–side chain interactions is rather small. The values of these energies were chosen so that proteins can fold into native structures that depend on the amino-acid sequence. The energy parameters of our model are likely to depend strongly on the choice for the functional form of the effective interaction potentials: we would not expect real (short-ranged) hydrophobic interactions to be modelled adequately by a simple Lennard–Jones potential. Of course, our model could be improved by including more cooperativity and by using a more realistic description of the side chain–side chain interactions. However, such improvements would come at a considerable computational cost.

The advantage of the present model should become pronounced when studying longer proteins. In particular, the folding and structure of the aggregate (figures 3 and 8) could be obtained for much longer proteins (50 or 100 amino acids). The key issue of such simulations would be to ensure that we actually find the lowest-energy states: this would require numerous simulations starting from different initial conditions.

The computational cost of figures 6 and 11 would be much higher, as these figures result from the averaging of tens of millions of configurations for different values of the relevant order parameter. Such curves are computationally expensive and are, at present, hard to obtain for longer chains. As one of the main objectives

of the present work was to illustrate the calculation of free-energy curves, we focused our study on relatively short poly-peptides, as this illustrates that the present simple model is likely to be useful for qualitative studies of the competition between aggregation and folding.

Acknowledgements

We gratefully acknowledge discussions with P.R. ten Wolde. This research was supported by a Marie Curie Fellowship of the European Community program 'Improving Human Research Potential and the Socio-economic Knowledge Base' under contract number HPMF-CT-2001-01212. Disclaimer: the authors are solely responsible for the information communicated and the European Commission is not responsible for any views or result expressed. The work of the FOM Institute is part of the research program of FOM and is made possible by financial support from the Netherlands Organization for Scientific Research (NWO).

References

[1] R. C. Moore and D. W. Melton, Molec. Hum. Reprod. **3,** 529 (1997).
[2] A. Slepoy, R. R. P. Singh, F. Pázmándi, R. V. Kulkarni, and D. L. Cox, Phys. Rev. Lett. **87,** 058101 (2001).
[3] L. K. Simmons, P. C. May, K. J. Tomaselli, R. E. Rydel, K. S. Fuson, E. F. Brigham, S. Wright, I. Lieberburg, G. W. Becker, and D. N. Brems, Molec. Pharmacol. **45,** 373 (1994).
[4] D. J. Selkoe, J. NIH Res. **7,** 57 (1995).
[5] J. Clark and J. Steele, Proc. Natn. Acad. Sci. U.S.A. **89,** 1720 (1992).
[6] H. R. Costantino, R. Langer, and A. M. Klibanov, Biotechnology **13,** 493 (1995).
[7] K. M. Personn and V. Gekas, Process. Biochem. **29,** 89 (1994).
[8] P. Gupta, C. K. Hall, and A. C. Voegler, Protein Sci. **7,** 2642 (1998).
[9] P. M. Harrison, H. S. Chan, S. B. Prusiner, and F. E. Cohen, Protein Sci. **10,** 819 (2001).
[10] G. Giugliarelli, C. Micheletti, J. R. Banavar, and A. Maritan, J. Chem. Phys. **113,** 5072 (2000).
[11] P. Gupta, C. K. Hall, and A. Voegler, Fluid Phase Equilib. **158–160,** 87 (1999).
[12] N. Go, J. Statist. Phys. **30,** 413 (1983).
[13] N. Combe and D. Frenkel, J. Chem. Phys. **118,** 9015 (2003).
[14] D. Zanuy, B. Ma, and R. Nussinov, Biophys. J. **84,** 1884 (2003).
[15] H. H. Tsai, D. Zanuy, N. Haspel, K. Gunasekaran, B. Ma, C. J. Tsai, and R. Nussinov, Biophys. J. **87,** 146 (2004).
[16] G. Hummer, A. E. Garcia, and S. Garde, Phys. Rev. Lett. **85,** 2637 (2000).
[17] A. V. Smith and C. K. Hall, J. Molec. Biol. **312,** 187 (2001).
[18] S. W. Voegler-Smith and C. K. Hall, Proteins: Struct. Funct. Gen. **44,** 344 (2001).
[19] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*, 3rd ed. (Garland, New York, 1994).
[20] C. Branden and J.Tooze, *Introduction to Protein Structure*, 2nd ed. (Garland, New York, 1998).
[21] N. Y Chen, Z. Y. Su, and C. Y. Mou, Phys. Rev. Lett. **96,** 078103 (2006).
[22] L. Pauling and R. B. Corey, Proc. Natn. Acad. Sci. **37,** 235 (1951).
[23] C. Das and D. Frenkel, J. Chem. Phys. **118,** 9433 (2003).
[24] C. Pace, B. Shirley, M. McNutt, and K. Gajiwala, FASEB J. **10,** 75 (1996).
[25] D. Frenkel and B. Smit, *Understanding Molecular Simulation*, 2nd ed. (Academic Press, London, 2002).
[26] M. C. Frith, A. R. Forrest, E. Nourbakhsh, K. C. Pang, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, T. L. Bailey, and S. M. Grimmond, PLoS Genet. **2**, e52 (2006).
[27] J. C. McKnight, D. S. Doering, P. T. Matsudaira, and P. S. Kim, J. Molec. Biol. **260**, 126 (1996).
[28] K. Tenidis, M. Waldner, J. Bernhagen, W. Fischle, M. Bergmann, M. Weber, M. L. Merkle, W. Voelter, H. Brunner, and A. Kapurniotu, J. Molec. Biol. **295,** 1055 (2000).
[29] O. Collet, Europhys. Lett. **53,** 93 (2001).
[30] D. K. Klimov and D. Thirumalai, Proc. natn. Acad. Sci. U.S.A. **97,** 2544 (2000).
[31] T. Veitshans, D. Klimov, and D. Thirumalai, Fold. Des. **2,** 1 (1996).
[32] A. Kolinski, A. Godzik, and J. Skolnick, J. Chem. Phys. **98,** 7420 (1993).
[33] T. X. Hoang, A. Trovato, F. Seno, J. R. Banavar, and A. Maritan, Proc. Natn. Acad. Sci. **101,** 7960 (2004).
[34] A. T. Petkova, U. Ishii, J. J. Balbach, O. N. Antzutkin, R. D. Leapman, F. Delaglio, and R. Tycko, Proc. Natn. Acad. Sci. U.S.A. **99,** 16742 (2002).
[35] R. Nelson, M. R. Sawaya, M. Balbirnie, A. O. Madsen, C. Riekel, R. Grothe, and D. Eisenberg, Nature **435,** 773 (2005).
[36] J. I. Lauritzen and J. D. Hoffman, J. Res. Natn. Bur. Stand. **64,** 73 (1960).
[37] D. Frenkel and T. Schilling, Phys. Rev. E **66,** 041606 (2002).