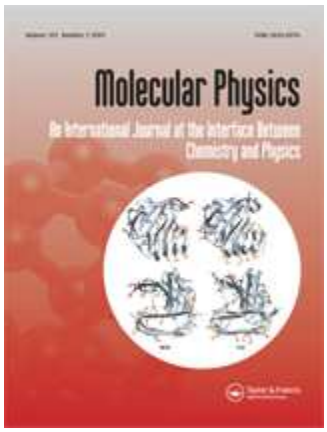
Simple off-lattice model to study folding and aggregation of peptides

Nicolas Combe, Daan Frenkel



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

Journal:	*Molecular Physics*
Manuscript ID:	TMPH-2006-0076.R1
Manuscript Type:	Full Paper
Date Submitted by the Author:	13-Dec-2006
Complete List of Authors:	Combe, Nicolas; Centre d'Elaboration de materiaux et d'Etudes Structurales, CNRS UPR 8011 Frenkel, Daan; FOM Institute for Atomic and Molecular Physics
Keywords:	protein, folding, aggregation

Molecular Physics, Vol. 00, No. 00, DD Month 200x, 1–9

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

Nicolas Combe,[a,b] and Daan Frenkel[c]

[a] Laboratoire de Physique des solides de Toulouse, UMR 5477,
Université Paul Saltier, 118 route de Narbonne, 31062 Toulouse cedex 4, France.
[b] Centre d'Elaboration de matériaux et d'Etudes Structurales (CEMES), CNRS UPR 8011,
29 rue J. Marvig, BP 94347, 31055 Toulouse cedex 4, France
E-mail:combe@cemes.fr

[c] FOM Institute for Atomic and Molecular Physics,
Kruislaan 407, 1098 SJ Amsterdam, The Netherlands.
E-mail: frenkel@amolf.nl

(*Received 00 Month 200x; In final form 00 Month 200x*)

We present a numerical study of a new protein model. This off-lattice model takes into account both the hydrogen bonds and the amino-acids interactions. It reproduces the folding of a small protein (peptide) : a morphological analysis of the conformations at low temperature exhibits the 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 trough a free energy analysis. We then study the aggregation of proteins. Proteins in the aggregate are mainly bounded trough hydrogen bonds. Performing a free energy analysis, we show that the addition of a peptide in such an aggregate is not favorable. We qualitatively reproduces the abnormal aggregation of proteins in prion diseases.

Keywords: protein,folding, aggregation

1 Introduction

The collective behavior of polymers has been extensively studied but this is not the case of bio-macromolecules. However, 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 (3; 4) and the formations of cataracts (5). All these conditions appear to be related to abnormal aggregation of proteins. But also for processes in the pharmaceutical (6) and food industry (7), a good understanding of protein aggregation is important. 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; 10; 11). The problem with the simulations of bio-molecular aggregation is that it requires a model that is detailed enough to account for the specific intermolecular interactions that drive the aggregation, yet sufficiently cheap to allow numerical simulations of the collective behavior 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 representation of real bio-macromolecules. Clearly, to make progress in the modeling of aggregation of real biomolecules one must use more realistic models. Ideally, one would use a model where all the atoms of a protein and of the solvent are represented explicitly (14; 15; 16). Unfortunately, the computational cost of such a model is such that it cannot be use to simulate the collective behavior 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 behavior of systems containing many proteins: for instance, Voegler-Smith and Hall (17; 18) have studied the competition between the refolding and aggregation using such a model. Their work suggests that, in order to mimic the behavior of real proteins, a model needs to account both for the effect of for hydrogen bonding between amino acids and for the interaction between different amino-acid residues. The model used in refs. (17; 18) is based on discontinuous potentials and can be studied by Monte Carlo or by event-driven molecular dynamics simulations. Whilst this choice is computationally cheap, it cannot account for long-ranged interaction. 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 thermal properties of this model and Monte Carlo simulations to gather information on the free-energy of folding and on the aggregation of the model proteins.

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

2 Protein model

Any protein model, however much 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 twenty 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 the peptide bonds : the carboxyl group $C' = O$ is linked to the nitrogen group $N - H$ by a 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 hydrogens 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 side chains. In addition, there can also be disulfide bridges. Such disulfide bridges are usually not found in intracellular proteins, but they occur quite frequently among 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 the water can play a role in the folding. Indeed, hydrophobic side chains tend to pack in the interior of the proteins. Charged and polar side chains interact through both Coulomb and Van der Waals forces (19).
Finally, hydrogen bonds occur between the oxygen lone pair in $C' = O$ and the hydrogens of $N - H$ 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 in 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 the hydrophobic effects are taken into account trough effective interactions between side chains.
- To account for hydrogen bonds, we do not explicitly simulate all the atoms NH-CO of each amino-acids because they are in a 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 the size of the side chains and the steric effects. We take into account only tree different types of side chains : hydrophobic (H), polar positive (P) or polar negative(N). We have not introduced 20 different types of amino-acids because it would have increased the number of parameters of the model and makes 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.

Fig. 1 shows a representation of our model. Since the NHCO group is modelled by a simple spin, the carbon C_α of an amino-acid is not chiral. However, once introduced in a amino-acids sequence, carbon C_α becomes most of the time chiral and the polypeptide is chiral itself (except if the sequence of amino-acids is symmetric).
Table 1 gives the values that we used for the different structural parameters. The value of the bonds or pseudo-bonds lengths and of the angles or pseudo-angles has been calculated from the structural data known about proteins : L^0_{CR} is fixed as the length of the usual sp_3 carbon-carbon bond, θ^0_{CR} is fixed as in a sp_3 carbon. L^0_{CC} has been calculated from the atomic distances in an amino-acids (22). Below, we briefly summarize the potential that determines the vibration and torsion of the peptide backbone.

(i) The length of bonds and the angle between 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 $k_{length-bond}$ and k_{angle} respectively.

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

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

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

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 atoms and equilibrium angle between bond in a real amino-acid (22), we have calculated the angle θ_{CC} depending on the orientation $CO - NH$-planes. This angle θ_{CC} varies between 1.4 rad 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 allow θ_{CC} to vary freely between 1.4 and 2.4 rad using the following potential :

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

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

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

(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 $k_{angle-spin}$.

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

Where $\theta_{spin_{CC}}$ 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 the side chains. The effective interactions between side chains accounts for solvent effects and for screened Coulomb interactions between charged particles. Two side chains of the same proteins can interact only if they belong to amino-acids separated by at least one amino-acids in the chain. Table 2 lists the potentials used.
For the spin-spin interactions, we use the following potential :

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

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 Fig. 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 be almost purely repulsive if the angle are large compared to 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 3 gives the values of the different parameters for the side chain-side chain and the spin-spin potentials. $k_{\text{length bond}}$ and k_{angle} are fixed to the values used for CC bonds in alkanes (23). $k_{\text{angle-spin}}$ 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-bonds energy (24). σ_{spin} is fixed the reproduce the hydrogen bond length. ϵ_{HH}, ϵ_{HP} and ϵ_{PP} were adjusted such that conformations of proteins at low temperature depend on the sequence of amino-acids : these three energies have been taken equal to reduce the number of parameter though it is certainly not the case in reality. σ_{HH}, σ_{HP} and σ_{PP} are chosen to be reasonable estimates for the sizes of groups they represent (18). Finally, λ was fixed such that two spins feel an attraction if their directions differ from the line joining the two spins by about 30 degrees. Note that the spin-spin interactions are stronger than side chain-side chain interactions. A similar trend has been observed in other models (17) where hydrogen bonds are six times stronger than side chain-side chain interactions.
In our model , the main-chain does not interact with side-chains though it would be easy to add a repulsive interaction. However, side-chain side-chain interactions prevent any overlap between 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 behavior 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 behavior 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 oligo-peptides consisting of 12 amino acid residues. Of course, such chains are short compared to most proteins, although it is worth stressing that several biologically active oligo-peptides (26; 27) are known. In addition, oligo-peptides can form amyloid fibers (14; 28) . We stress that there is no intrinsic limitation of the present model to short oligo-peptides. We studied both the temperature dependence of the internal energy of the model proteins and the conformational changes that the oligo-peptide undergoes upon changing the temperature.

3.1 *Temperature dependence of internal energy*

Fig. 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-acids sequence.
The location of the peak in the heat capacity provides us with an estimate of the transition temperature: it is about $T_f = 115K$. In the following, all temperature will be normalized by T_f. This temperature is low compared to typical folding temperatures of real proteins (about 310 K) (29). Of course, our results depend on 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 paper is primarily to illustrate that, even without much fine tuning, our model exhibits protein-like behavior. 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 temperature decreases below the heat-capacity peak, the chain folds in 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, though the protein is chiral, our model does not distinguish between left-handed and right-handed

helices : this drawback is a consequence 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. Figs. 4 and 5 show two folded proteins: one in a α-helix conformation, the other in a β-sheet conformation. The only difference between these 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 NH-CO group, this would correspond in real proteins, to a hydrogen bond between the $C' = O$ of amino-acid n and the 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 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 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.

To our knowledge, the present coarse-grained model is the simplest that reproduces both α-helix and β-sheet structures using only three amino-acids 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 show 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 favor a particular target state (the "native" state) because it assumes that only those side chains that are nearest neighbors in the native state can attract each other. The amino-acid alphabet for the Gō model is therefore unbounded, as it grows with the number of nearest-neighbor contacts in the native state. By comparison, our model has the advantage that it does not have 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 the strength of both interactions.

3.3 *Folding pathway*

In our model, protein folding is the result of a 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 favorable 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 Eq. 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 favor a small number of amino-acids per helix turn. However, this is frustrated by the energetic cost to decrease 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-carbon in the range between 1.4 rad and 2.4 rad. If we would have constrained this angle to take a value 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 Fig. 5 corresponds to a metastable state. However, in Molecular Dynamics simulations, the formation of these structures is often kinetically favored. This is so because one can choose a sequence of amino-acids that favors the β-sheet structure: as the attraction between these side chains is relatively long ranged, one finds that kinetics of the folding process can favor 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 in 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 an order parameter q that characterizes the degree of folding of the protein. We chose the following order parameter:

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

The sum is performed over all allowed couples [1] of spins in the chain and V_{spin} is given by Eq. 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: $W = 1/2\ k\ (q - q_0)^2$ where k and q_0 are parameters that can be varied at will. From these simulations, we get the free energy curve around q_0, up to a constant. To get 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 8 that fits the curves. Fig 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 a high value 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 bond. 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 protein. Fig 6 also allows to estimate the transition temperature: here between $t/T_f = 0.95$ and $T/T_f = 1.13$ in agreement with the data of Fig 3.
In summary: our model reproduces the two-state behavior of real short proteins and the resulting folded conformation contain secondary structures that resemble those of real proteins. Below, we consider the possible aggregation of proteins.

[1] Two spins can interact if they are separated by at least 2 amino-acids.

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 the 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 stabilized by inter-protein hydrogen bonds. These hydrogen bonds are perpendicular to the *beta*-sheet plane of the individual proteins.
This phenomenon is illustrated in Fig. 7 where we show that two proteins that have been designed to form an α helix, when isolated (see Fig. 4), form a stable, β-sheet-like dimer. We have verified that this aggregate stays stable over the range of temperatures where the native state of the isolated protein is stable. We emphasize that the stability of aggregate of Fig. 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 3.1. Fig. 8 shows the variation of the average potential energy as a function of temperature.
An 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 Fig. 7 to the one shown in Fig. 9.
This morphology is more stable than the one proposed in Fig 7. Here, the two proteins are linked by side chain-side chain interactions and especially by hydrogens 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 has to note, that some hydrophobic interactions are created between the two proteins: the side chain-side chain interactions drive therefore both the folding and the aggregation of proteins.
Also in larger aggregates we observe proteins arrangements similar to the one in Fig. 9. Comparing our results with the work of Petkova et al. (34), our simulation shows the spontaneous formation of aggregate rather similar to β-amyloid fibrils occurring in prion diseases. Indeed, Petkova et al. (34) have recently provided a structural model for the Alzheimer's β-amyloid fibrils based on experimental constraints. They showed that these fibrils may have a structure analogous to the one presented in Fig. 9 (with a large number of proteins). Moreover, it is suggested in ref. (34) that fibrils formation is driven by hydrophobic interactions. The picture of aggregation we present here is thus comparable to the one 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 interaction between β-sheet like proteins interactions is 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 conclusion.

4.2 *Growth of the aggregate*

An aggregate such as the one in fig. 9 can grow by addition of another protein. But, since the temperature is lower than the folding temperature, an added protein will have first to unfold to be able to make some 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 peptide in the aggregate. Thus, by adding a folded protein in the aggregate, the number of unsatisfied hydrogen bonds should be reduced by the number of unsatisfied hydrogens bonds in the initial added chain: we can thus expect such an operation to be energetically favorable. The number of unsatisfied bonds gained is $2S$.
This analysis is incorrect for the formation of the first aggregate of 2 proteins since in that case, the initial system in composed of 2 folded proteins in a helix conformation and no aggregate exists yet. Thus the price in hydrogen bonds is $L - 4S$ and for long proteins, L is higher than $2S$.
Following this last analysis based on energy arguments, we can draw a schematic representations of the expected energy landscape presented in Fig. 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).
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:

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

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 an helix conformation. The initial configuration is the one shown in Fig. 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 8 that fits the simulated free-energy data. Note that the free energy is only determined up to an additive constant. Fig. 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 $q \approx 9 - 10$.. The free energy (Fig. 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 the two minima ($q \approx 10$ and $q \approx 14$). One minimum corresponds to the original aggregate $q \approx 10$, 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 an 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 Fig 10, we see that the relative propensity for aggregation and folding depends on 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 an protein incorporated in an aggregate in a helix conformation **Especially, our calculation only partly takes into account the entropy associated to the volume of simulations cell : before aggregate to a protein, the protein has first 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 Fig. 11 is of the order of 1eV. This value should be compared to the energy of a single hydrogen bond energy: 103meV. 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 to a fibril.

5 Conclusion

In this paper we analyzed the properties of an off-lattice, coarse-grained protein model. Even though it is very simple, it does qualitatively reproduce several key properties of real proteins. In particular, the model proteins can fold in a β-sheet or α-helix structure, depending on the sequence of amino-acids. Moreover, we have shown that small aggregates spontaneously organize in 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 hydrogen bond is realistic ($4kT$ for our model) whilst the energy for side chain-side chain interactions is rather small. The value of these energies have been chosen so that proteins can fold in 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. Especially, the folding and the structure of the aggregate (Fig. 3 and Fig. 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 figs. 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 with 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 it illustrates that the present simple model is likely to be useful for qualitative studies of the competition between aggregation and folding, .

6 Acknowledgments

We gratefully acknowledge discussions with P.R. ten Wolde. This research has been 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 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, *Mol. Hum. Reprod.*, 1997, **3**, 529, (1997)
[2] A. Slepoy, R. R. P. Singh, F. Pázmándi, R. V. Kulkarni, and D. L. Cox, *Phys. Rev. Lett.*, **87**(5), 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, *Mol. Pharmacol.*, **45**, 373 (1994).
[4] D. J. Selkoe, *J. NIH Res.*, **7**, 57 (1995).
[5] J. Clark and J. Steele, *Proc. Natl. Acad. Sci. USA*, **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, *Proteins Science*, **7**, 2642 (1998).
[9] P. M. Harrison, H. S. Chan, S. B. Prusiner, and F. E. Cohen, *Proteins Science*, **10**, 819 (2001).
[10] G. Giugliarelli, C. Micheletti,J. R. Banavar and A. Maritan, *J. Chem. Phys.*, **113**(12), 5072 (2000)
[11] P. Gupta, C. K. Hall, and A. Voegler, *Fluid Phase Equilibria*, **158-160**, 87 (1999).
[12] N. Go, *J. Stat. Phys.*, **30**(2), 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. Mol. Biol.*, **312**, 187 (2001)
[18] S. W. Voegler-Smith and C. K. Hall, *Proteins : Struc. Func. Gen.*, **44**, 344 (2001).
[19] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the cell*, third ed., Garland, New York, (1994).
[20] C. Branden and J.Tooze, *Introduction to protein structure*, seconde 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. Natl. Acad. Sci.*, **37**, 235 (1951).
[23] C. Das and D. Frenkel, *J. Chem. Phys.*, **118**(20), 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*, , second 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(4), e52 (2006).
[27] J. C. McKnight, D. S. Doering, P. T. Matsudaira and P. S. Kim *J. Mol. 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. Mol. Biol.*, **295**, 1055 (2000).
[29] O. Collet, *Europhys. Lett.*, **53**(1), 93 (2001).
[30] D. K. Klimov and D. Thirumalai, *Proc. Natl. Acad. Sci. USA*, **97**, 2544 (2000).
[31] T. Veitshans, D. Klimov and D. Thirumalai, *Folding and Design*, **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. Natl. 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. Natl. Acad. Sci. USA*, **99**(26), 16742 (2002).
[35] R. Nelson, M. R. Sawaya, M. Balbirnie, A. O. Madsen, C. Riekel, R. Grothe and D. Eisenberg, *Nature*, **435**(9), 773 (2005).
[36] J. I. Lauritzen and J. D. Hoffman, *J. Res. Nat. Bur. Stds.*, **64**, 73 (1960).
[37] D. Frenkel and T. Schilling, *Phys. Rev. E*, **66**, 041606 (2002).

Structural data	
$L^0_{CC}(nm)$	0.38
$L^0_{CR}(nm)$	0.15
$\theta^0_{CC}(rad)$	1.910612
$\theta^0_{CR}(rad)$	0.61549

Table 1. Structural data of our model

$$
\begin{aligned}
V_{HH} &= 4\epsilon_{HH}\left[\left(\tfrac{\sigma_{HH}}{r}\right)^{12} - \left(\tfrac{\sigma_{HH}}{r}\right)^{6}\right] \\
V_{HP} &= 4\epsilon_{HP}\left(\tfrac{\sigma_{HP}}{r}\right)^{12} \\
V_{HN} &= 4\epsilon_{HP}\left(\tfrac{\sigma_{HP}}{r}\right)^{12} \\
V_{PN} &= 4\epsilon_{PP}\left[\left(\tfrac{\sigma_{PP}}{r}\right)^{12} - \left(\tfrac{\sigma_{PP}}{r}\right)^{6}\right] \\
V_{PP} &= 4\epsilon_{PP}\left(\tfrac{\sigma_{PP}}{r}\right)^{12} \\
V_{NN} &= 4\epsilon_{PP}\left(\tfrac{\sigma_{PP}}{r}\right)^{12}
\end{aligned}
$$

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

Bonds potential parameters	
$k_{\text{length bond}}(kcal.mol^{-1}.A^{-2})$	235.5
$k_{\text{angle}}(kcal.mol^{-1}.rad^{-2})$	60
$k_{\text{angle-spin}}(kcal.mol^{-1}.rad^{-2})$	1
Side Chains interactions parameters	
$\epsilon_{HH}(kcal.mol^{-1})$	1.1
$\epsilon_{HP}(kcal.mol^{-1})$	1.1
$\epsilon_{PP}(kcal.mol^{-1})$	1.1
$\sigma_{HH}(nm)$	0.36
$\sigma_{HP}(nm)$	0.36
$\sigma_{PP}(nm)$	0.36
Spin interactions parameters	
$\epsilon_{spin}(kcal.mol^{-1})$	4.36
$\lambda(rad^{-2})$	2
$\sigma_{spin}(nm)$	0.48

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

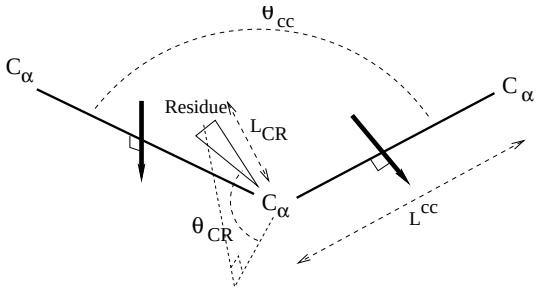


Figure 1. Model of amino-acid. The orientation of the $CO-NH$ plane is described by a spin perpendicular to the $C_\alpha C_\alpha$ bond.

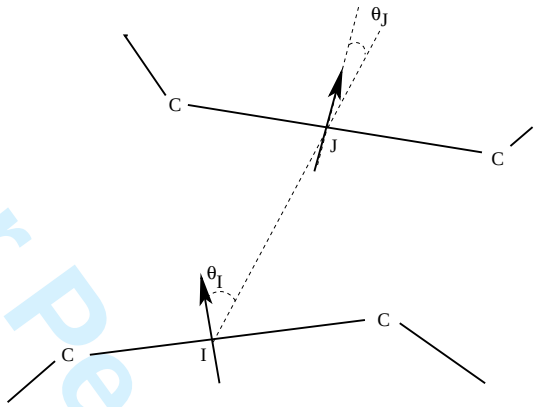


Figure 2. Definition of the angle θ_I and θ_J used in Eq. 8.

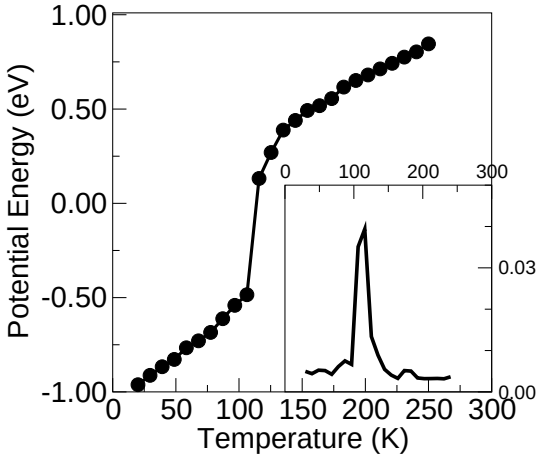


Figure 3. Temperature dependence of the average potential energy of the model protein shown in Fig. 4.The inset, shows the heat capacity, i.e. the derivative of the potential energy with respect to temperature.

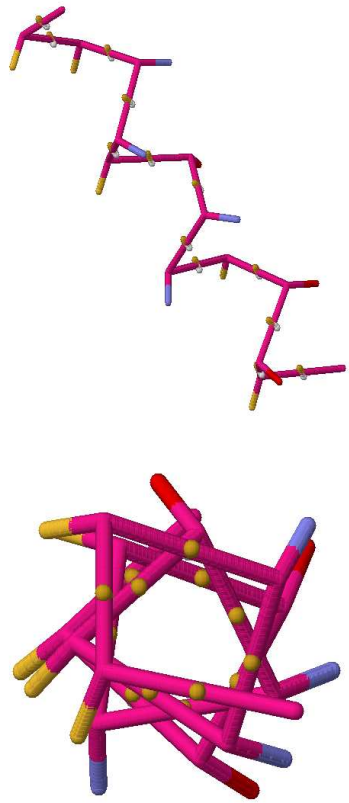


Figure 4. (Color online) Folded conformation found at low temperature for the helix, the sequence used is $HP^2HN^2PHN^2H^2$. We show here the conformation from two different orientations. The color code for the side chain is: yellow for hydrophobic side chains, red for polar positive, and blue for polar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

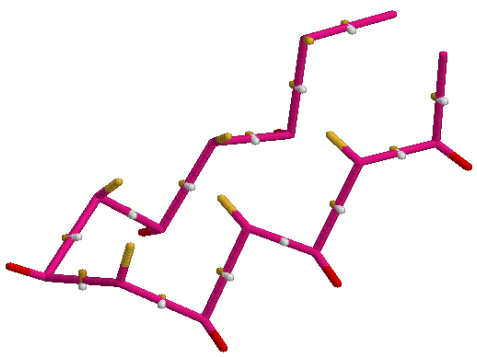


Figure 5. (Color online) Folded conformation found at low temperature for the helix β-sheet, the sequence used is $PHPHPHPHPHPH$. The color code for the side chain is as the one describe in Fig. 4

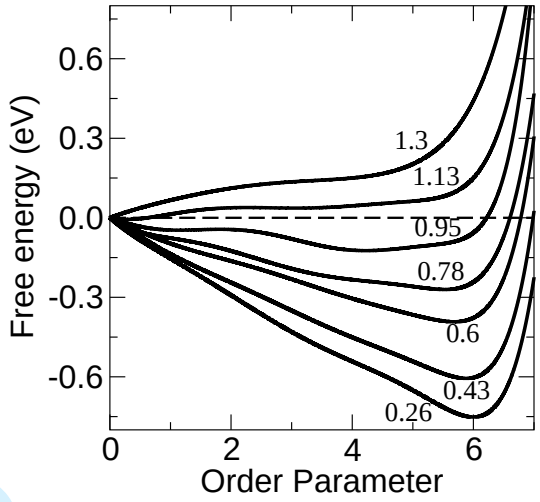


Figure 6. Free Energy of the protein shown in Fig. 4 as a function of the order parameter for different temperatures T/T_f : 0.26, 0.43, 0.6, 0.78 , 0.95 , 1.13 and 1.3 (from bottom to top).

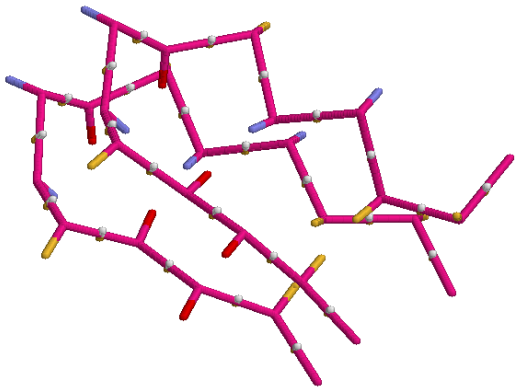


Figure 7. (Color online) Aggregate formed with 2 proteins identical to the one of Fig. 4.

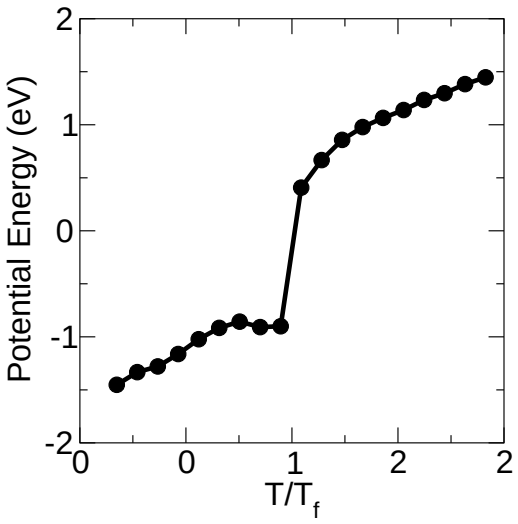


Figure 8. Potential energy of the aggregate of Fig. 7 as a function of the reduced temperature T/T_f where T_f is the folding temperature of the protein.

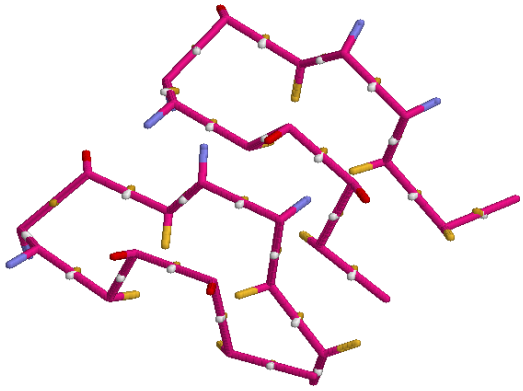


Figure 9. (Color online) Morphology of the aggregate at $T/T_f = 0.94$

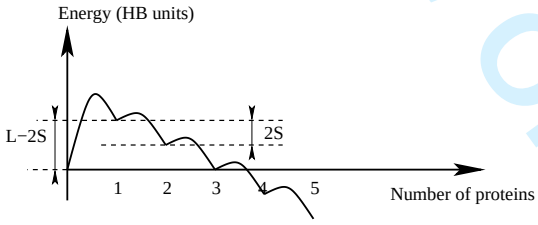


Figure 10. Sketch of the expected energy landscape in hydrogen bonds units as a function of the number of proteins in the aggregate. $L/2$ denotes the number of hydrogens bonds that connect the two peptides of Fig. 9 and S denotes the number of unsatisfied hydrogens bonds on one end of an helix like the one in Fig /refhelix.

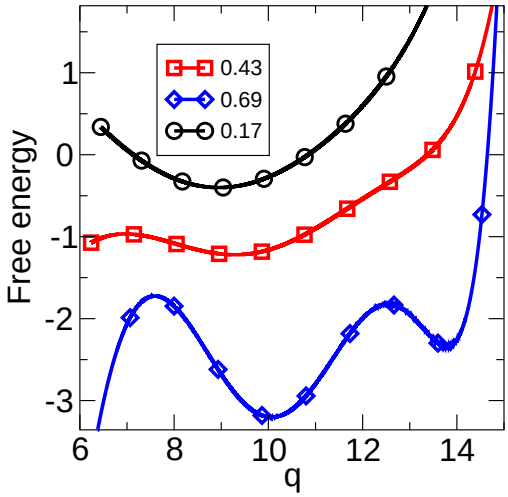


Figure 11. (Color online) Free-energy change (in eV) associated with the detachment of a protein from an existing aggregate. The free energy is plotted as a function of the order parameter q defined in the text. The legend mention the temperature in unit of the folding temperature T_f.