Influence of the native topology on the folding barrier for small proteins

Lidia Prieto and Antonio Rey[a)]
Departamento de Química Física I, Facultad de Ciencias Químicas, Universidad Complutense, E-28040 Madrid, Spain

(Received 23 July 2007; accepted 14 August 2007; published online 2 November 2007)

The possibility of downhill instead of two-state folding for proteins has been a very controversial topic which arose from recent experimental studies. From the theoretical side, this question has also been accomplished in different ways. Given the experimental observation that a relationship exists between the native structure topology of a protein and the kinetic and thermodynamic properties of its folding process, Gō-type potentials are an appropriate way to approach this problem. In this work, we employ an interaction potential from this family to get a better insight on the topological characteristics of the native state that may somehow determine the presence of a thermodynamic barrier in the folding pathway. The results presented here show that, indeed, the native topology of a small protein has a great influence on its folding behavior, mostly depending on the proportion of local and long range contacts the protein has in its native structure. Furthermore, when all the interactions present contribute in a balanced way, the transition results to be cooperative. Otherwise, the tendency to a downhill folding behavior increases. © *2007 American Institute of Physics*. [DOI: 10.1063/1.2780154]

I. INTRODUCTION

Thermodynamically, the formation of tertiary structure in many single domain proteins has been shown to exhibit an all-or-none cooperative behavior. In these cases, two states are observed: the unfolded protein, which is a high entropy, high energy disordered state (taking into account only the conformational space available to the polypeptide chain), and the folded protein, which can be described as a low entropy, low energy phase.[1,2] According to the two-state mechanism, denaturation is an all-or-none transition in which the protein is in either of these two macrostates.[3] Therefore, two-stateness is signalized in a profile of free energy *versus* an unfolding coordinate by two minima, corresponding to the native and unfolded macrostates, whose relative depths depend on temperature. In order to determine if a protein folds in a two-state manner, the most widely used criterion is the calorimetric one.[2,4–8] However, this criterion has been revisited,[9] and it has been claimed that for very small proteins good agreement between the experimental values for the calorimetric enthalpy and the van't Hoff enthalpy is not solid proof of a two-state behavior, without a careful evaluation of the baselines involved.[10]

Deviations from two-state behavior can be interpreted in many ways, depending on the thermodynamic or kinetic character of the experiments involved. In the case of fast folding proteins, for example, an association has been established between fast kinetics and a potentially low or nonexistent (in thermodynamic terms) folding/unfolding barrier.[10–14] In these cases, we would be talking about a downhill scenario in which only a single minimum would appear in the free energy versus degree of unfolding profile, which would shift along the unfolding coordinate upon changing temperature. When such a transition takes place, it happens in a continuous manner, involving a single macrostate.[10]

The energy landscape theory predicts that under certain conditions, folding can proceed without crossing barriers.[1,15] Consequently, although many proteins have apparently been evolutionary selected to fold in a two-state manner, this is not a physico-chemical requirement for the folding process.[16] It has been proposed that some proteins may have evolved as fast folding proteins in order to avoid aggregation. However, it has also been argued that a low activation barrier might imply a larger tendency to aggregate.[17]

Despite the reasons that would have led to the existence or not of downhill protein folding, this type of thermodynamic transition is predicted by theory. The field of those processes that behave as first order or continuous transitions, depending on conditions, is accomplished by a well-known branch of thermodynamics: the theory of critical transitions. This theory has allowed Munoz and Sanchez-Ruiz to introduce a simple phenomenological model for the analysis of equilibrium protein folding experiments.[13] This model does not presume the free energy barrier height, nor other properties of the folding ensemble, but permits one to obtain them directly from the experimental data.

Some authors propose that, indeed, some fast folding proteins have a downhill folding transition. The absence of a thermodynamic barrier would explain the kinetic properties of the transition for these proteins.[18–22] More explicitly, it has been proposed that the thermal unfolding of the peripheral subunit binding domain from *Escherichia coli*'s 2-oxoglutarate dehydrogenase multienzyme complex [protein data bank (PDB) code: 1BBL] does not occur in a two-state manner, but follows a downhill type process.[12–14,18,23] There has been much controversy about this topic, with other au-

[a)]Author to whom correspondence should be addressed. Electronic mail: jsbach@quim.ucm.es

thors arguing against the one-state folding of 1BBL.[24–30]

In order to try to shed some light into this polemic topic, many authors have looked for a convenient way of predicting the existence or not of a thermodynamic folding barrier and its height. As previously said, Munoz and Sanchez-Ruiz have proposed for this aim the variable-barrier model,[13] which uses calorimetric experimental data. Other criteria are based on kinetic folding properties.[31–33] One of these is the folding "speed limit",[32] according to which no protein has been identified so far to fold without overcrossing a free energy barrier.[17] Some other works about downhill folding include numerical simulation studies to compare and explain experimental data.[25,33] Most of them are molecular dynamics simulations, specially centered on several kinetic properties of the folding transition.

While some authors have approached this problem using semiempirical potential energy functions applied to all-atom simulations,[25] others have studied the possible folding in absence of free energy barriers by simulating coarse grained models with Monte Carlo algorithms.[34] A common way of computationally studying the possible presence of thermodynamic barriers in protein folding is by using a Gō-type potential.[35,36] These name represents nowadays a broad family of interaction potentials based on the native structure of a protein, which have been mainly applied to kinetic studies of the folding transition.[33] It has been shown that protein folding rates correlate with protein size[19] and relative contact order.[37] As free energy barriers can be estimated from kinetic data, it can be said that native topology has a direct influence on the presence or not and on the height of the free energy barrier along the protein folding/unfolding process.

In this study, we use a Gō-type potential to study the influence of the native topology on the thermodynamic properties of the simulated folding process of a coarse grained protein model. A similar study has been very recently published by other authors,[38] a work that we have been aware of during the calculations involved in this project. However, Zuo *et al.* use Langevin Dynamics to approach the topic, as well as a different potential function and contact classification. Here, on the other hand, we use a parallel tempering Monte Carlo algorithm,[39] which, according to our experience, represents a more suitable methodology to study the thermodynamic behavior of a set of proteins represented by a simplified model. The aim of this work, as already said, is to try to understand which characteristics, if any, of the native topology may determine the properties of the transition studied. It has been previously pointed out how the definition of the Gō-type interaction itself (as a function of a proper description of the contact map of the native state) may be determinant for the height of the free energy barrier of folding.[39,40] Furthermore, the relative strength of the different types of interactions, i.e., local and long range interactions, also has a direct effect on the thermodynamic properties of the transition.[41,42]

In this work, using the parameters which provided reasonable results in our previous analysis of the simulation model,[39,41] we have used our carefully designed Gō-type potential to gain a better insight into the topological determinant factors for the type of folding transition.

It has to be clearly stated that Gō models, as the one we are using here, represent a drastic approximation for the folding process, which cannot have any predictive role about the final structure, nor on folding mechanisms which may involve non-native contacts. For these situations, more detailed models which also consider the protein sequence and correspondingly include a more detailed interaction scheme are mandatory, even at the level of coarse-grained studies.[43,44] However, Gō-type models are perfectly adequate when the sequence is voluntarily ignored, as it is here, to check the effects of the native state topology alone.

II. THE MODEL AND SIMULATION METHOD

As in previous works,[39,41] we use a coarse grained representation of the protein and its interactions. It is an off-lattice α-carbon representation of the protein; i.e., every residue is represented by a hard sphere centered at its α carbon. Two consecutive units in the model are kept at a fixed distance of 3.8 Å, corresponding to a *trans* peptide bond.

The interaction between pairs of amino acids includes a hard sphere repulsion and a Gō-type potential defining the attractive interactions. The form of the Gō-type potential we are using is a harmonic well. This well is centered at the native distance d_{ij}^{nat} (where i and j correspond to different amino acids of the chain). The mathematical definition of this potential for two residues at a distance r_{ij} is

$$u_{ij}(r_{ij}) = \begin{cases} w_{ij}[(r_{ij}-d_{ij}^{\mathrm{nat}})^2 - a^2] & \text{if } d_{ij}^{\mathrm{nat}} - a < r_{ij} < d_{ij}^{\mathrm{nat}} + a \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The potential is truncated at a distance $a=0.7$ Å, a value which provided very reasonable results (specially rather narrow transitions for a two-state protein) in a previous study.[39]

In this potential, three types of interactions are included:

- Virtual bond angle interactions: They happen between residues i and $i+2$. In these cases, $w_{ij} \neq 0$ for every pair of residues.
- Virtual torsion angle interactions: Those that take place between residues i and $i+3$. By assigning to the distance between residues a sign, equal to that of the triple (scalar) product of the three vectors defining the virtual torsion angle, local chirality is introduced in the definition of these interactions. This makes it possible to discriminate between a conformation and its mirror image. $w_{ij} \neq 0$ for any of these interactions, as well.
- Long range interactions: They are defined between amino acids i and $j \geq i+4$. In this case, $w_{ij} \neq 0$ only for residues i and j that are forming a contact in the native structure. Two residues are considered to be in contact in the native state if the shortest distance from all the possible pairs among heavy atoms belonging to both residues is smaller than or equal to 4.5 Å. This cutoff distance value is used because it is slightly larger than twice the average of van der Waals radii for heavy atoms in proteins[45] (taking into account the fact that hy-

drogen atoms are not explicitly represented in most of the PDB structures, but are included in these radii when considering a united atom model). Finally, $w_{ij}=0$ in these long range interactions for those pairs of residues that are not in contact in the native state.

In this work, interactions are classified as long range (already defined) and local interactions, which include both virtual bond angle and virtual torsion angle interactions. The values of w_{ij} control the depth of the harmonic well, which means a control of the strength of the interactions. In a previous work, we studied how the change in the relative strengths of local and long range interactions, keeping the total energy of the native state constant, has a direct influence on the thermodynamics of folding.[41] In that study, we showed how a proper balance of local (w_{ij}^{l}) and long range (w_{ij}^{lr}) interactions is necessary in order to correctly define the cooperative folding transition of a small protein.

The method by which we study the thermodynamic characteristics of the folding/unfolding transition for our model is a parallel tempering[46] Monte Carlo simulation algorithm, already described.[39] For the chain model, we use end moves for the vectors defining the virtual bonds for the two extreme units. For internal units, spike, "shifting", and pivot moves are included to properly sample the conformations accessible to the system at different temperatures. After every individual move, the resulting conformation is checked to see that it does not violate the hard sphere condition. Afterwards, if no overlappings exist, the energy is calculated according to the potential defined. Also, exchanges of conformations sampled at neighbor temperatures are occasionally tried. Whether a new conformation is accepted or not is decided by an adapted Metropolis test involving both temperatures.[46]

The number of temperatures for every parallel tempering simulation depends on the transition characteristics, which vary from one protein to another. It ranges from 13 to 20 for the cases considered in this work, and described in the next section. Every simulation consists of 5×10^6 Monte Carlo cycles, after 3×10^6 equilibration cycles (in both cases, at every temperature). One cycle implies the possibility of individually moving all the units in the model, according to the set of moves mentioned above. Simulation times depend on the number of temperatures and protein size. They range from 1 h for the smallest protein to 10 h for the largest ones. Simulations are run in single processor machines and, in order to warrant a correct sampling, five independent simulations are computed for every system. Each one of these simulations starts from a different seed number for random number generation.

The thermodynamic analysis of our numerical results has been carried out by using the weighted histogram analysis method (WHAM).[47–49] The numerical results presented in Section IV correspond to statistical averages over the sampling at every temperature and over the five independent runs.

III. PROTEINS STUDIED

In order to study the influence of the native topology on the thermodynamic behavior of the Gō-type model described before, we have chosen eight small proteins, whose folding transition characteristics have been experimentally determined and are available in the literature.

The first of them is the inmunoglobulin binding domain of streptococcal protein G (PDB code: 2GB1). This is a protein that has been previously analyzed with the potential we are using here.[39,41] Mainly for this reason, we include it among other proteins in this study. Moreover, this protein has been the subject of numerous detailed analysis, both experimentally[50–54] and theoretically.[55–58] The main conclusion from all these works is that this protein folds in a cooperative, all-or-none manner. Its folding is, thus, a first order thermodynamic transition, with a free-energy barrier.

Next, we have considered the peripheral subunit binding domain of the dihydrolipoamide succinyltransferase from the 2-oxoglutarate dehydrogenase multienzime complex of *Escherichia coli* (PDB code: 1BBL). This protein has been claimed to fold in a downhill manner,[11] although there is no full agreement about the validity of this statement.[14,24–30] To show that this protein does not fold in a downhill manner, Ferguson *et al.* have studied different mutants and structural homologs. One of these is the E3BD F166W mutant (PDB code: 1W4E), which is claimed to have a first order transition.[24–26] Therefore, we have included it among the proteins treated in the present paper.

Also, Ferguson *et al.* had previously shown that chymotrypsin inhibitor 2 (PDB code: 2CI2) folds by overcrossing a free energy barrier; this is why we have considered it appropriate to be one of the proteins to study with our model.

Naganathan *et al.* have measured the heights of the folding barriers for different proteins, among which we find 1W4E and 1BBL, already mentioned. The former is said to fold in a two-state manner, according to Ferguson *et al.*, while the latter is said to do it following a barrierless transition.[20] In the set of proteins studied by these authors, they have also included the peripheral subunit-binding domain of dihydrolipoamide acetyltransferase from the piruvate dehydrogenase multienzime complex of *Bacilius stereothermophilus* (PDB code: 2PDD) and the *Bacilus subtilis* major cold shock protein (PDB code: 1CSP). 2PDD is a structural homolog of 1BBL and the wild-type structure of mutant 1W4E. These two proteins are said to fold *via* a first-order thermodynamic transition. We consider both in this work.

Another protein included in our present work is the headpiece subdomain of protein villin (PDB code: 1VII). The experimental unfolding data for this protein can be fitted by using a two-state model.[59] However, the high folding rate of this protein suggests that, although it requires crossing a free energy barrier to fold, it is a very small one.[60] Theoretical studies have proposed that this protein is a weakly cooperative folder.[33,38]

Finally, we have included in the set of proteins the designed 20 residue Trp-cage miniprotein (PDB code: 1L2Y), which has been experimentally reported to be a two-state protein.[61] Theoretically, however, it has been claimed to fold with no free energy barrier.[38]

In this paper, we will refer to all these proteins by their PDB code.

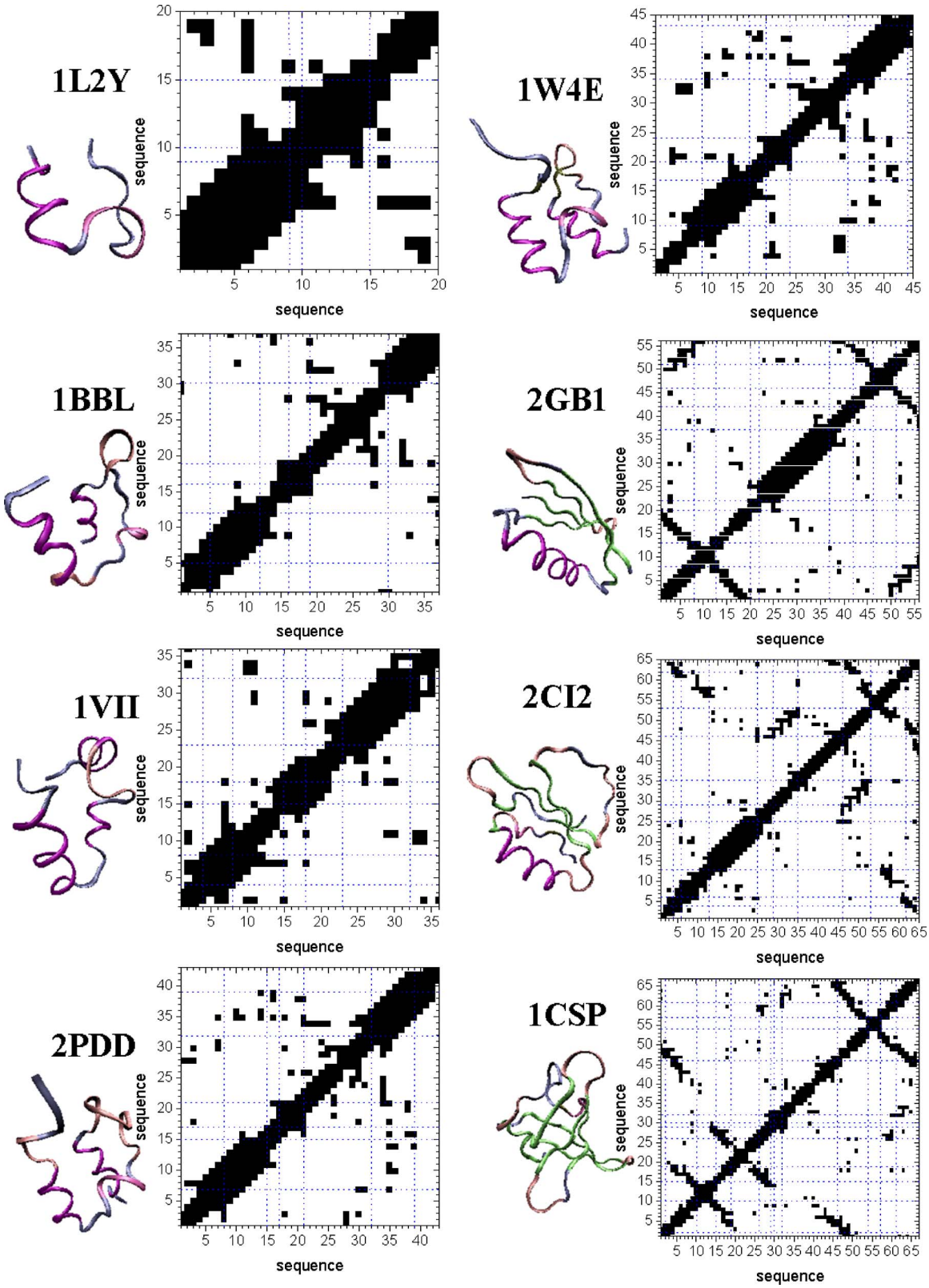

FIG. 1. (Color online) Set of eight proteins used in the present study. The ribbon structure for each protein lays at the left side of the corresponding contact map. Every spot in the latter indicates the presence of a native contact between residues i and j. Dotted lines separate elements of secondary structure, according to the PDB file headers.

The structures and contact maps of the proteins introduced above can be seen in Fig. 1. In this figure, the three-dimensional structure is displayed at the left side of the corresponding contact map for every protein. In the maps, the native contacts are represented as spots between the residues i and j numbered at the axes (which also serve to indicate the size of each protein). A spot that appears near the diagonal means that the contact happens between residues close along the protein sequence. Contrarily, when it appears far from the diagonal, two residues distant in the sequence, usually be-

longing to different secondary structure elements, are in contact in the native conformation. In these maps, secondary structure elements (as defined in the PDB file headers) are separated by dotted vertical and horizontal lines. When a contact spot is into one of the boxes along the main diagonal, the contact belongs to one of these local structure elements. Spots in off-diagonal boxes are indicative of contacts that stabilize the tertiary structure of the protein. It can be seen that homologous structures such as 1BBL and 1W4E have very similar contact maps along their main diagonals. As 1W4E is a larger protein, it has a higher number of tertiary contacts. In addition, the contact map of 2PDD is quite similar to that of 1W4E, as the former is a mutant of the latter. The contacts seen in Fig. 1 are those considered by the Gō-type potential described above to contribute to the stability of the protein, according to the simulation model we are using here.

IV. RESULTS AND DISCUSSION

As it has already been said, the aim of this work is to determine which influence the native topology has on the folding thermodynamics of small proteins, by using a Gō-type model as the one described above. In previous studies about the model and the potential employed here, we used the structure of 2GB1 to evaluate the different parameters of the potential.[39,41] As it was shown in these works, a cutoff parameter of $a=0.7$ Å and all the interactions having the same strength [which means that the weights in Eq. (1) have the values $w_{ij}^{l}=w_{ij}^{lr}=1/a^2$, where the superindices l and lr stand for "local" and "long range" interactions, respectively] give an appropriate description of the transition taking place for this protein. That means that the simulation results we obtain are quite reasonably comparable to the experimental data for the folding transition.

With the same parametrization of the potential, we have simulated in the first place the folding of 1BBL and studied the thermodynamic characteristics of the process. This protein, which mainly possesses secondary structure, with a rather small number of tertiary contacts (as it may be appreciated in Fig. 1) has been proposed to be a downhill folder,[11] as mentioned before. It has only 36% of long range contacts (as defined in this work) over the total number of contacts in the native state. Thus, only 36% of the native energy is due to the action of long range interactions in our simulation model. When making the same thermodynamic analysis for 1BBL as the one previously done for 2GB1,[39,41] it is seen that, with our model, the folding transition for 1BBL is described as a barrierless process. We explain this conclusion now.

First of all, from the energy fluctuations along the simulation we have computed the heat capacity of our modeled protein at every temperature. It should be said that, although we are obtaining the constant volume heat capacity, instead of the constant pressure heat capacity measured in differential scanning calorimetry experiments,[7,62] this does not make any difference to our conclusions,[9] given the qualitative character of this study. The heat capacity for each temperature for 1BBL is represented in Fig. 2. When a transition

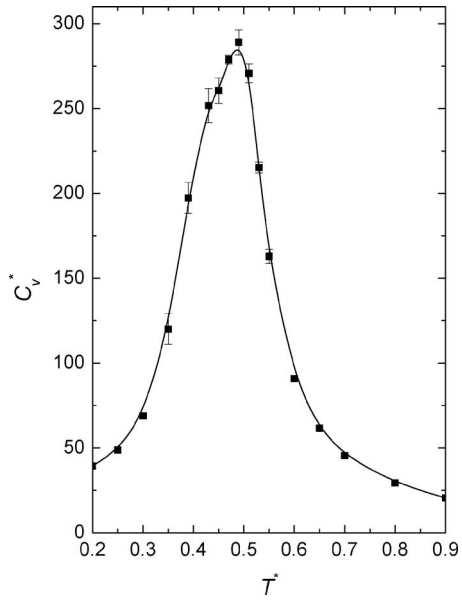

FIG. 2. Simulation results for the constant volume heat capacity (C_v^*) as a function of temperature for protein 1BBL.

takes place, the heat capacity has a maximum. In Fig. 2 a typical denaturation curve appears with only one peak. We consider the temperature at which this maximum occurs as the folding temperature (T_m^*, in the standard reduced unit we use along this work). From the denaturation curve itself, the type of thermodynamic transition that underlies the folding process cannot be extracted, though the peak is considerably wider than that previously found for 2GB1 for the same potential.[39] Experimentally, as said, several methodologies, such as the calorimetric criterion, have been applied to address this question. Nevertheless, it has been stated that an arbitrary division of a system into two states automatically satisfies the calorimetric criterion. Furthermore, satisfaction of this criterion does not necessarily mean that intermediate states may be at least marginally populated.[9] Other criteria have been proposed to distinguish the type of transition as well.[33] Simulation, however, directly allows one to study the microstates present in the system. Therefore, we can readily determine the type of transition from the distribution of microstates along the energy scale sampled during the simulation (for a Gō-type model, this energy scale can reasonably represent a folding/unfolding coordinate). For this purpose, we calculate the energy histogram at T_m^*. In fact, from these energy histograms at all temperatures included in the parallel tempering simulation, it is possible to calculate free energy profiles at any temperature by using the WHAM technique.[47–49] When the energy distribution at the folding temperature presents only one peak, the transition would be considered as downhill. This would imply the presence of only one minimum in the free energy profile. The energy histogram and the free energy profile calculated from our simulations for 1BBL at T_m^* and other temperatures above and below (but close to) it are shown in Fig. 3. The curves show the distinct characteristics of a downhill transition: a single maximum in the energy histogram (or a single minimum in the free energy plot) that smoothly shifts toward larger, less negative energies as the temperature increases

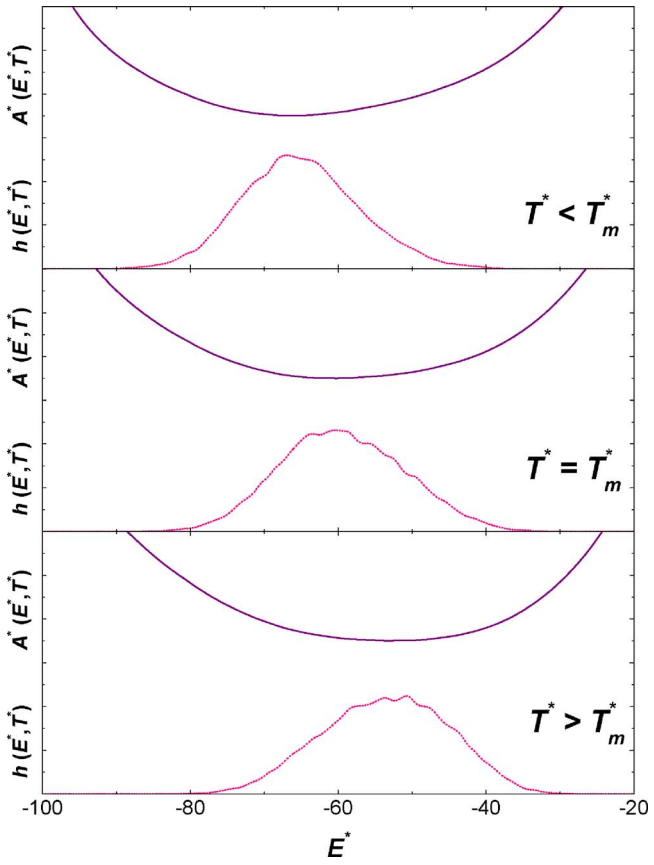

FIG. 3. (Color online) Free energy profiles (solid lines) and energy histograms (dotted lines) for protein 1BBL, at its folding temperature T_m^*, and at temperatures below and above it.

from below to above T_m^*. The folding of 1BBL, simulated with our Gō-type model, happens following a continuous process involving a single macrostate at every temperature, in agreement with the results from experimental data by Garcia-Mira *et al.*[11]

In a previous paper from our group,[41] we showed that most proteins have about the same proportion of local and long range contacts among the total contact number in the native state. In a set of 1590 proteins, the percentage of long range contacts for most of them roughly goes from 50% to 60%. From our previous simulation results, this would explain why most of the proteins fold in a two-state manner. However, small proteins present a proportion of long range contacts that is frequently below this major trend.[41] The proteins studied here have percentages of long range contacts that span from small values (as already seen in 1BBL) to the most abundant ratio, where both local and long range contributions are about the same. These values can be seen in Table I, as well as the length of the chain, for all the proteins considered in this work. The three largest proteins of the set, 2GB1, 2CI2, and 1CSP, have long range contact percentages corresponding to the most common range. From our previous work, they are expected, thus, to behave as two-state folders. On the other hand, the smallest proteins, 1L2Y and 1BBL, are clearly outside the major range, with only 36% of long range contacts. The questions then are: What happens to the intermediate cases 1VII, 2PDD, and 1W4E, and to the remaining proteins included in the set? Could it be stated that all proteins that do not have a percentage of long range interactions to the native state energy around or slightly above 50% fold via a barrierless process?

In order to address these questions, we have performed parallel tempering simulations spanning a wide temperature range for all the proteins in Table I, and calculated from them the free energy profiles at their corresponding transition temperatures T_m^*, calculated again from the maximum in the corresponding heat capacity curves. These temperatures are also collected in Table I. In addition, the total (both local and long range) number of contacts in the native state is included for every protein. It can be seen that, as expected for this model, the folding temperature roughly increases as the total number of contacts becomes higher, though the trend is far from being smooth.

The free energy profiles for the full set of proteins simulated are represented in Fig. 4. As expected, the larger proteins 2GB1, 2CI2, and 1CSP fold in a two-state manner, as it is clear from the presence of a free energy barrier. On the contrary, 1L2Y and 1BBL, which have a long range energy contribution to native stability very different from the major trend, fold via a downhill process. These results agree with most of the experimental data commented on in Sec. III.

On the other hand, the intermediate cases of 1VII, 2PDD, and 1W4E fold differently from one another. The folding simulation of the two first proteins results in a barrierless transition. In particular, we consider the folding of 1VII as downhill, although there is a tiny free energy barrier in the simulation data shown in Fig. 4. Experimental studies have stated that, although this protein may require crossing a free energy barrier to fold, it seems to be a very small one.[60] Our simulation results quite nicely coincide with this observation. However, this barrier is not high enough to make the

TABLE I. Structural and folding transition temperature (T_m^*) values for the proteins of the set (Fig. 1). N^raa is the number of amino acids, $\%E^{\mathrm{lr}}$ is the proportion of long range energy in the native state, N^rcon is the number of total contacts in the native structure, and N_N^Z and N_N^P are the number of long range contacts per residue in the native state calculated under Zuo *et al.* (Ref. 38) and our contact classification, respectively.

PDB	1L2Y	1BBL	1VII	2PDD	1W4E	2GB1	2CI2	1CSP
N^raa	20	37	36	43	45	56	65	67
$\%E^{\mathrm{lr}}$	36	36	43	42	47	50	51	52
N^rcon	55	108	117	140	161	215	257	270
T_m^*	0.480	0.490	0.525	0.535	0.625	0.630	0.670	0.683
N_N^Z	0.40	0.73	0.67	0.88	1.07	1.79	1.86	2.13
N_N^P	1.00	1.05	1.39	1.37	1.67	1.93	2.02	2.10

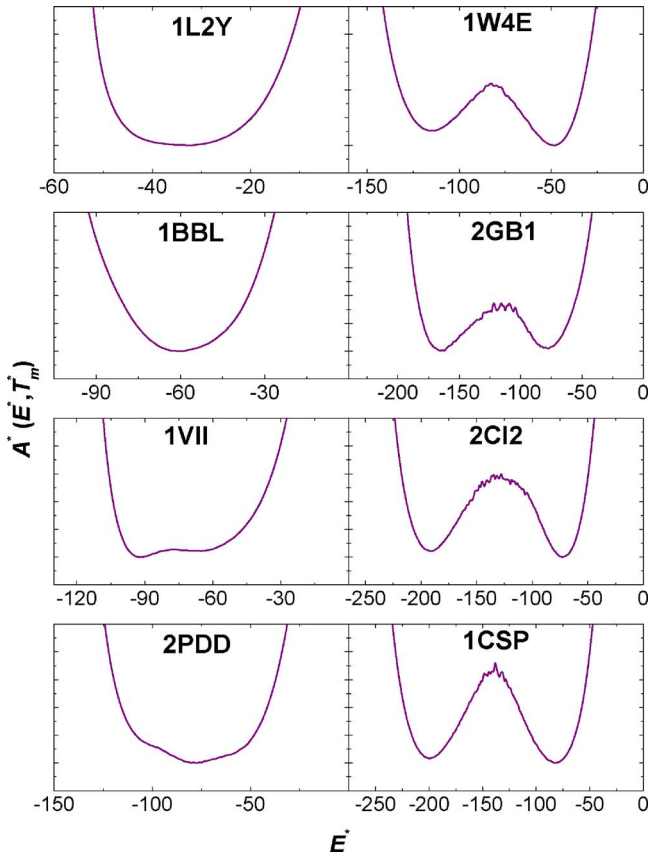

FIG. 4. (Color online) Free energy profiles for all the proteins in Fig. 1 at their folding temperatures, obtained through WHAM analysis of the simulation results. The (numerically meaningless) scales for the free energy axes are the same in all the plots.

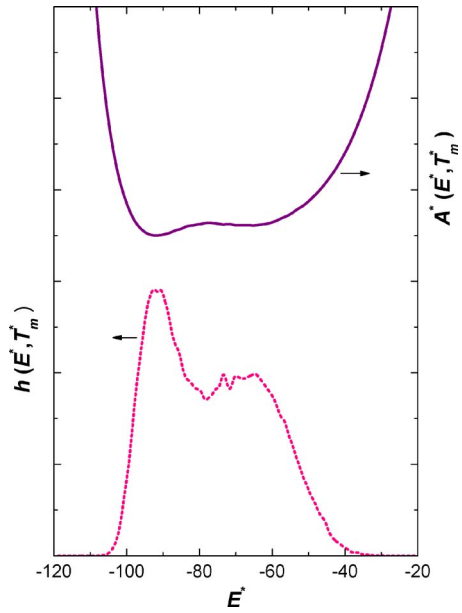

FIG. 5. (Color online) Free energy profile and energy histogram for protein 1VII at its folding temperature.

population of states with intermediate energy negligible (or, at least, rather low) at T_m^* (the detailed behavior of this protein and its energy histogram at its T_m^* is shown in Fig. 5). This is why we do not consider that this protein folds in a two-state manner, according to our model. In a much more clear way, our simulation results undoubtedly state that the folding transition for 2PDD is a barrierless process.

Contrary to these two proteins, 1W4E, which has 47% of long range energy in the native state, has to cross in our model a neat free energy barrier to acquire the native structure. It folds, thus, in a two-state manner. The percentage of long range contacts over the total number of contacts in the native structure of this protein is very close to the most common range. 1VII and 2PDD, however, differ in a higher degree from this major tendency.

All together, we have shown how the folding behavior predicted with our model is qualitatively comparable to many experimental results. Therefore, it is clear that topology has a great influence on the thermodynamic properties of the folding/unfolding transition of at least small proteins. Although the type of transition has to be, ultimately, demonstrated by experiment, Gō-type potentials may provide some explanation about the deviations of the usually considered two-state behavior that appear in some natural proteins.

In our previous work, we showed how changing the relative weights of local and long range interactions to the total energy of the native state modifies the thermodynamic properties of the simulated transition.[41] In that work, using 2GB1 alone, we proposed that, to have an appropriate thermodynamic description of the simulated folding transition as a two-state process, it is necessary to have a proper balance of local and long range interactions. For both 1L2Y and 1BBL, due to the description of contacts that we are using, the percentage of nonlocal contacts among the total native contacts is 36% without any artificial reweighting of the native contacts (i.e., $w_{ij}^{\mathrm{l}}=w_{ij}^{\mathrm{lr}}=1/a^2$). This value lays outside the usual interval of relative contribution of local versus nonlocal interactions we had found before.[41] For us, then, it is not a surprise that the folding transition for this protein has no free energy barrier, as it also happens with 2PDD and 1VII. In 2GB1, we were able to "tune" the thermodynamic properties of the folding transition by modifying the local/nonlocal ratio.[41] To check the robustness of our results, we have also tried the same effect here by simulating the case in which the percentage of long range energy in the native state is artificially enlarged (by increasing w_{ij}^{lr} and decreasing w_{ij}^{l}). With these new simulations, we want to see if the agreement between the results shown above and the experimental ones from Garcia-Mira *et al.* is due to the native topology itself or to a particular parametrization of the potential where, changing the values of w_{ij}^{l} and w_{ij}^{lr}, other types of folding transitions could be attributed to these proteins. The results of the simulations we have computed with this new reweighted contributions of the potential are summarized in Fig. 6. We show the free energy profiles at T_m^* (its value inferred as always from the corresponding heat capacity curves) for the native proportion of long range contacts, two increased values in the case of 1L2Y and 1BBL, and values above and below the native one for 1VII and 2PDD. Although some properties of the folding transition slightly change, such as the position of the free energy minimum along the energy scale, the thermodynamic properties of the transition stay essentially the same for the four proteins: They remain to

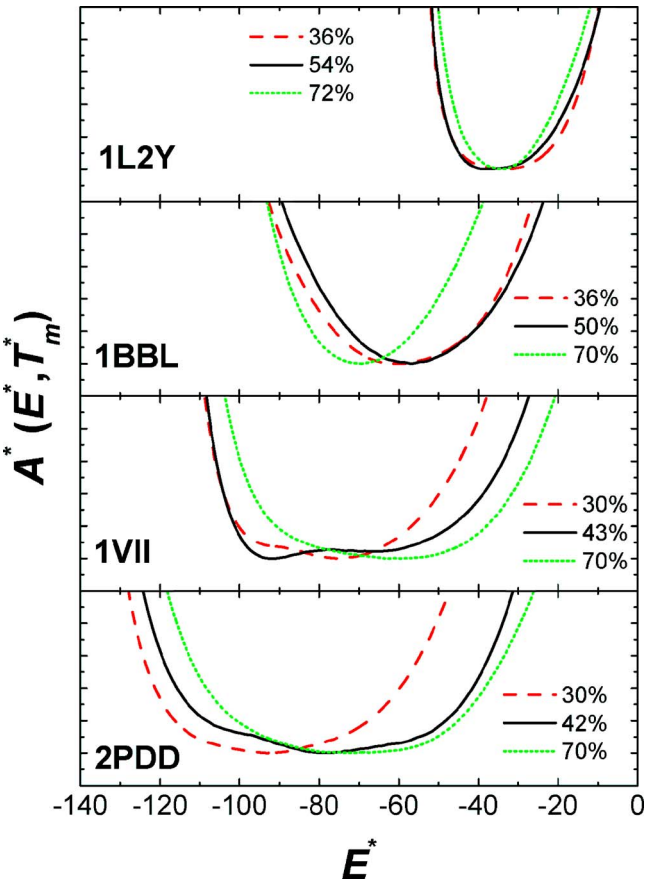

FIG. 6. (Color online) Free energy profiles at the folding temperature for the indicated proteins, with native and reweighted contributions of long range interactions, as indicated in every panel.

behave as downhill folders, even when the weight of long range interactions represents up to 70% of the total energy of the native state. For the four proteins which, according to our results, show a cooperative folding (1W4E, 2GB1, 2CI2, and 1CSP) we have done the same procedure (data not shown). The two-state character of the transition is kept in the range of reweighting analyzed, about the same that is shown for the other four proteins in Fig. 6. Therefore, the downhill or two-state behavior of the folding transition for proteins under a Gō-type model such as the one we are using is not just due to the numerical weight of local and long range interactions, but it must be caused by other distinct characteristics of the native topology.

Until now, we have established a relationship between the type of transition taking place and the contribution of long range interactions to the energy in the native sate. In a recent work,[38] Zuo *et al.*, using a similar model to the one described here and Langevin dynamics simulations, stated that there is a correlation between cooperativity (two stateness) and the number of long range contacts per residue,

$$N_N = \frac{\text{number of nonlocal contacts}}{\text{number of residues}}. \quad (2)$$

They consider N_N as an important topological parameter and define $N_N=0.9$ as a crossover to separate two-state and downhill folders.[38] Although the specific numerical value of the limit for N_N depends on the contact classification, the important point is that there is a correlation between the type of transition and the contribution of long range interactions, also related to the length of the polypeptide chain. In Table I the values of N_N are calculated under Zuo *et al.* contact classification (N_N^Z) and under ours (N_N^P). Though both values of N_N roughly grow with the protein size, there is no linear dependence among them, since the distribution of long range contacts depends on the particular structure, and not only on the sequence length. It is clearly seen that $N_N^Z=0.9$ separates in our simulation results downhill folders from two-state proteins. With our definition of long range contacts a different threshold value of N_N^P can be defined (about 1.4–1.5 in our case), supporting anyway the treatment by Zuo *et al.*[38]

Nevertheless, global properties as the sequence length, or the total number of long range contacts, are too coarse definitions of the protein native state. We still have to try to answer the questions: How does the *specific* native topology affect the thermodynamic behavior of folding? What *detailed* structural factors determine proteins to fold in one way or the other? Even at first glance, Fig. 1 provides some information about the structural differences of proteins folding in different ways. It can be seen that the proteins that fold in a two-state manner have a better defined tertiary structure that the proteins that do not have to cross a free energy barrier to fold, a fact that is partially mirrored by smaller values in the N_N parameter for the latter. To what extent does the tertiary structure determine the type of folding? If we focus on 2PDD and 1W4E, we see that they are structurally very similar. This is completely reasonable, considering that 2PDD is the wild-type structure of the 1W4E mutant. However, 1W4E has a two-state transition while 2PDD folds in a downhill manner under the conditions of our model. In spite of the structural similarity, they slightly differ in their length (1W4E is two amino acids longer) and especially in the number of contacts (see Table I): 2PDD has 81 local contacts and 59 long range contacts, while 1W4E has 85 and 76 local and long range contacts, respectively, according to our definition. The higher number of local contacts of 1W4E is just due to the presence of two additional virtual bond angle contacts and two additional virtual torsion angle contacts. How do the 17 extra long range contacts present in 1W4E change the topology to make this protein fold differently from 2PDD? From the contact maps in Fig. 1, it is not easy to appreciate significant structural differences between these two proteins. Therefore, to accomplish a more detailed analysis of the tertiary structure of the proteins, we have computed the histograms of the distances along the sequence between pairs of residues in contact, i.e., the histograms of their absolute contact orders,[63] which can be seen in Fig. 7 for the eight proteins considered in this work. In this figure, each vertical bar represents the fraction of native contacts between pairs of residues that are at a distance $|j-i|$; i.e., there are $|j-i|-1$ residues between them. From these representations, we can appreciate additional differences between the structure of proteins that have a barrier-controlled process and those that do not. All but one of the proteins show a large peak at $|j-i|=4$. These are contacts characteristic of α helices. They could not probably be strictly considered as tertiary contacts, but are included in the long range category according to our

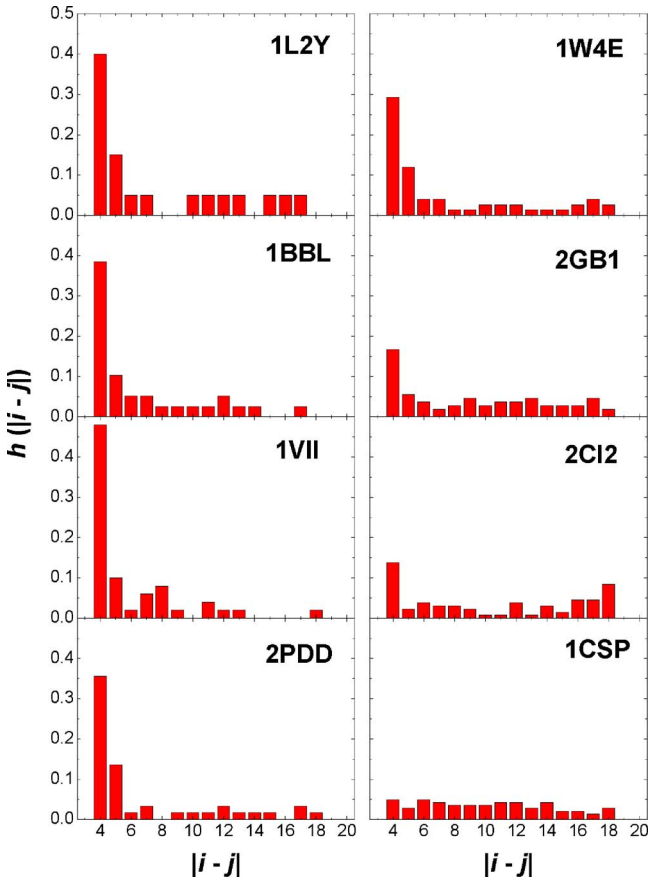

FIG. 7. (Color online) Native contact order histograms for all the proteins in Fig. 1, with i and j being the residue positions along the sequence which define a contact. The values of the histogram scale are divided by the total number of long range contacts for every protein.

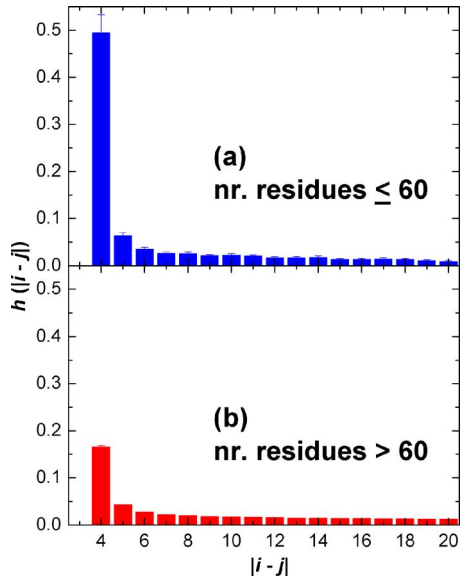

FIG. 8. (Color online) Average contact order histograms for a large set of monomeric protein structures: (a) With up to 60 residues, and (b) with more than 60 residues.

definition. This continues to prove that an excessively high degree of local contacts (or helical secondary structure) in the native state disfavors cooperativity.[37,63,64] In the figure, it can be readily appreciated how this peak represents well above 30% of the total long range contacts in downhill folders, while in two-state proteins the peak is below this threshold. This does not mean, however, that the presence of α helices in the structure automatically implies a downhill transition, although a high percentage of them has been already pointed out as a possible reason for low or negligible barriers.[65] A lot of proteins have a well defined helical secondary structure in spite of their length. However, only those that have enough tertiary structure to balance local interactions may have the necessary energy/entropy compensation to give rise to a free energy barrier (see Ref. 41).

Proteins that fold following a downhill transition have a higher degree of helical contacts than of any other type of contacts. On the other hand, proteins that have a more homogeneous distribution of the different values for the contact order fold via a two-state process. In the particular case of the mutants 2PDD and 1W4E, they have a quite similar distribution of long range energy in the native state. However, the extra long range contacts present in 1W4E but nonexistent in 2PDD appear between residues with $|j-i|>4$, as shown in Fig. 7. That means that the presence of α-helix content relative to the chain length in 1W4E is smaller than that for 2PDD, as the number of contacts with $|j-i|=4$ is the same for both structures but 1W4E has a larger number of residues and, more importantly, of contacts with larger contact orders.

In Ref. 41 we used the results from our simulation model to state that a certain balance between local and long range interactions is necessary for the presence of a thermodynamic barrier in protein folding. In this work, we can see that this balance is also necessary among the long range contacts themselves. In the group of long range contacts (according to our definition), there is always a certain, maybe important, fraction that appears between pairs of residues that are close in sequence. These are the contacts that mainly define the distribution in the case of proteins without a free energy barrier. These proteins have also contacts with higher values of $|j-i|$, but their proportion is low, and the distribution is rather inhomogeneous, with gaps for several values of the abscissa, as it may be seen in the graphs at the left column in Fig. 7. On the other hand, in the structures of the proteins that fold through a free energy barrier, there is a more homogeneous distribution of the distances between residues in contact, with all the values of $|j-i|$ up to a certain limit populated for these proteins (graphs at the right column in Fig. 7). This may be the main difference between the native structures of 2PDD and 1W4E which determines the distinct thermodynamic characteristics of their folding processes.

As a final check, we have carried out this type of contact order analysis for a large set of 1590 nonredundant and well-defined monomeric protein structures. The results, first normalized for every individual protein and then averaged over the full set, are collected in Fig. 8. To favor a fair comparison, we have separated in this figure the proteins with a sequence of up to 60 residues (a total of 87 structures) from the larger ones. Although the averaging process may somehow blur the details of individual proteins, several facts are readily evident: The average proportion of "helical" contacts

is almost three times larger in small proteins than in larger ones; and both types of proteins show, on average, a smooth distribution of contact order values. Since most proteins experimentally studied up to now, regardless of their size, seem to follow a two-state folding process, the average results in this figure would be indicative of the contact order features for this type of transition to happen. Small proteins with large helical content and uneven distributions of long range contacts, as those considered in this work (1L2Y, 1BBL, 1VII, and 2PDD), depart from this typical behavior. This reason, according to our model, leads to a different barrierless process.

V. SUMMARY AND CONCLUSIONS

In this work, we have studied how the topology of different small proteins constitutes a determinant factor for the type of folding transition they follow, under the conditions of a Gō-type potential and a coarse grained simulation model. The results we got are compatible with previous experimental data about the folding transition for the same proteins. Therefore, the conclusions of the present study can be helpful to understand the different features of the folding behavior that proteins may present, especially in those cases which deviate from the customarily accepted one.

As the model we use has been previously proven to correctly reproduce the behavior of 2GB1 as a two-state folder, we have used it to simulate folding of 1BBL, which, not without controversy, has been said to fold in a continuous manner without crossing a free energy barrier. The results from our simulation coincide with this experimental observation.

We have extended our simulation of the folding process to a set of eight different small proteins. Following previous studies from our group, we have checked that when the proportion of long range and local interactions are about the same in the native structure, the proteins fold, indeed, in a two-state manner. Furthermore, 1L2Y and 1BBL, which deviate clearly from this general trend with a long range contact percentage of only 36%, behave in our simulations as downhill folders. The same happens with 1VII and 2PDD. Although there is no universal cutoff value for the long range contact percentage to discriminate between theoretical two-state and downhill folders, the fact is that a certain degree of equivalence between the two types of interactions considered in the native topology is necessary to describe the transition as a first order thermodynamic process. Interestingly, an artificial change in the relative weights of both types of interactions does not change the folding type for these proteins. It seems to be the native topology itself that regulates the presence or not of a folding free energy barrier.

Moreover, it is not only a balance between local and long range interactions but also among the long range contacts themselves, which creates a two-state scenario. When we have plotted the histograms of distances along the sequence for pairs of residues in contact, we have seen that those proteins which present a more homogeneous distribution along all values of distances are those that are described to fold overcrossing a significant barrier. This type of distribution also corresponds to the average contact order distribution computed over a large set of different proteins, both short and large. The downhill folders, on the contrary, have their contact distance distributions shifted to a great extent to values of $|j-i|=4$, which corresponds to α-helix contacts. In these cases, there is a great contribution of helical structure in our definition of long range interactions, with other values of $|i-j|$ having zero contacts in the native state, resulting in a less smooth distribution of contact orders. Proteins that fold in a cooperative way tend to distribute their long range contacts in a more homogeneous way along all values of the contact order parameter.

If there is a homogeneous distribution of long range contacts, our results indicate that cooperativity is favored, while a predominance of short distance contacts leads to a downhill process. It can thus be said that, when all the possible interactions contribute more or less in the same way to the native protein stability, the folding/unfolding process results to be cooperative. In other words, a cooperative transition seems to be related to an equitable distribution of interactions along the sequence. If some type of interactions, especially short range ones, predominate over the others, a noncooperative process is favored, giving rise to what is called a downhill folding.

Although our simulations may not exactly reproduce the experimental features of the proteins simulated, the results of this paper might explain to some degree the higher or smaller cooperative behavior that these, and maybe also other, proteins have.

ACKNOWLEDGMENTS

This work was partially supported by Spanish Ministerio de Educación y Ciencia (Grant No. FIS2006-12781-C02-02) and by Comunidad Autónoma de Madrid/Universidad Complutense de Madrid (Grant to Consolidated Groups 910068). One of the authors (L.P.) acknowledges a Scholarship from Spanish Ministerio de Educación y Ciencia.

[1] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins **21**, 167 (1995).
[2] P. L. Privalov, Adv. Protein Chem. **33**, 167 (1979).
[3] B. Ibarra-Molero and J. M. Sanchez-Ruiz, in *Advanced Techniques in Biophysics*, edited by J. L. R. Arrondo and A. Alonso (Springer, Berlin, 2006).
[4] C. Tanford, Adv. Protein Chem. **23**, 121 (1968).
[5] W. N. Jackson and J. F. Brandts, Biochemistry **9**, 2294 (1970).
[6] J. A. Schellman, Annu. Rev. Biophys. Biophys. Chem. **16**, 115 (1987).
[7] J. M. Sturtevant, Annu. Rev. Phys. Chem. **38**, 463 (1987).
[8] S. E. Jackson and A. R. Fersht, Biochemistry **30**, 10428 (1991).
[9] Y. Zhou, C. K. Hall, and M. Karplus, Protein Sci. **8**, 1064 (1999).
[10] B. Ibarra-Molero and J. M. Sanchez-Ruiz, in *Protein Folding and Misfolding*, edited by V. Munoz (Royal Society of Chemistry, London, in press).
[11] M. M. Garcia-Mira, M. Sadqi, N. Fisher, J. M. Sanchez-Ruiz, and V. Munoz, Science **198**, 2191 (2002).
[12] F. Y. Oliva and V. Munoz, J. Am. Chem. Soc. **126**, 8596 (2004).
[13] V. Munoz and J. M. Sanchez-Ruiz, Proc. Natl. Acad. Sci. U.S.A. **101**, 17646 (2004).
[14] A. N. Naganathan, R. Perez-Jimenez, J. M. Sanchez-Ruiz, and V. Munoz, Biochemistry **44**, 7435 (2005).
[15] K. A. Dill and D. Shortle, Annu. Rev. Biochem. **60**, 795 (1991).
[16] M. Gruebele, C. R. Biol. **328**, 701 (2005).
[17] R. B. Dyer, Curr. Opin. Struct. Biol. **17**, 1 (2007).
[18] A. Akmal and V. Munoz, Proteins **57**, 142 (2004).

[19] A. N. Naganathan and V. Munoz, J. Am. Chem. Soc. **127**, 480 (2005).
[20] A. N. Naganathan, J. M. Sanchez-Ruiz, and V. Munoz, J. Am. Chem. Soc. **127**, 17970 (2005).
[21] W. Eaton, Proc. Natl. Acad. Sci. U.S.A. **96**, 5897 (1999).
[22] H. K. Nakamura, M. Sasai, and M. Takano, Proteins **55**, 99 (2004).
[23] M. Sadqi, D. Fushman, and V. Munoz, Nature (London) **442**, 04859 (2006).
[24] N. Ferguson, P. J. Schartau, T. D. Sharpe, S. Sao, and A. R. Fersht, J. Mol. Biol. **344**, 295 (2004).
[25] N. Ferguson, R. Day, C. M. Johnson, M. D. Allen, V. Dagget, and A. R. Fersht, J. Mol. Biol. **347**, 855 (2005).
[26] N. Ferguson, T. D. Sharpe, P. J. Shartau, S. Sato, M. D. Allen, C. M. Johnson, T. J. Rutherford, and A. R. Fersht, J. Mol. Biol. **353**, 427 (2005).
[27] N. Ferguson, T. D. Sharpe, C. M. Johnson, and A. R. Fersht, J. Mol. Biol. **356**, 1237 (2006).
[28] N. Ferguson, T. D. Sharpe, C. M. Johnson, P. J. Schartau, and A. R. Fersht, Nature (London) **445**, E14 (2007).
[29] Z. Zhou and Y. Bai, Nature (London) **445**, E16 (2007).
[30] M. Sadqi, D. Fishman, and V. Munoz, Nature (London) **445**, E17 (2007).
[31] S. J. Hagen, Proteins **50**, 1 (2003).
[32] J. Kubelka, J. Hofrichter, and W. A. Eaton, Curr. Opin. Struct. Biol. **14**, 76 (2004).
[33] M. Knott and H. S. Chan, Proteins **65**, 373 (2006).
[34] A. Irbäck, Acta Phys. Pol. B **34**, 4867 (2003).
[35] H. Taketomi, Y. Ueda, and N. Gō, Int. J. Pept. Protein Res. **7**, 445 (1975).
[36] N. Gō, J. Stat. Phys. **30**, 413 (1983).
[37] H. S. Chan, Nature (London) **392**, 761 (1998).
[38] G. Zuo, J. Wang, and W. Wang, Proteins **63**, 165 (2006).
[39] L. Prieto, D. de Sancho, and A. Rey, J. Chem. Phys. **123**, 154903 (2005).
[40] H. Kaya and H. S. Chan, J. Mol. Biol. **326**, 911 (2003).
[41] L. Prieto and A. Rey, J. Chem. Phys. **126**, 165103 (2007).
[42] M. Knott, H. Kaya, and H. S. Chan, Polymer **45**, 623 (2004).
[43] A. Liwo, M. Khalili, and H. A. Scheraga, Proc. Natl. Acad. Sci. U.S.A. **102**, 2362 (2005).
[44] S. Kmiecik and A. Kolinski, Proc. Natl. Acad. Sci. U.S.A. **104**, 12330 (2007).
[45] J. Tsai, R. Taylor, C. Chotia, and M. Gerstei, J. Mol. Biol. **290**, 253 (1999).
[46] U. H. E. Hansmann, Chem. Phys. Lett. **281**, 140 (1997).
[47] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **61**, 2635 (1988).
[48] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **63**, 1195 (1989).
[49] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollmar, and J. M. Rosenberg, J. Comput. Chem. **13**, 1011 (1992).
[50] P. Alexander, S. Fhnestock, T. Lee, J. Orban, and P. Bryan, Biochemistry **31**, 3597 (1992).
[51] P. Alexander, J. Orban, and P. Bryan, Biochemistry **31**, 7243 (1992).
[52] F. J. Blanco, M. A. Jimenez, A. Pineda, M. Rico, J. Santoro, and J. L. Nieto, Biochemistry **33**, 6004 (1994).
[53] K. Ding, J. M. Louis, and A. M. Gronnenborn, J. Mol. Biol. **335**, 1299 (2004).
[54] D. Idiyatullin, I. Nesmekova, V. A. Daragan, and K. H. Mayo, J. Mol. Biol. **325**, 149 (2003).
[55] E. L. McCallister, E. Alm, and D. Backer, Nat. Struct. Biol. **7**, 669 (2000).
[56] J. Shimada and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **99**, 11175 (2002).
[57] P. Derremaux, J. Chem. Phys. **119**, 4940 (2003).
[58] S. Y. Lee, Y. Fujitsuka, D. H. Kim, and S. Takada, Proteins **55**, 128 (2004).
[59] J. Kubelka, W. A. Eaton, and J. Hofrichter, J. Mol. Biol. **329**, 625 (2003).
[60] J. Kubelka, T. K. Chiu, D. R. Davies, W. A. Eaton, and J. Hofrichter, J. Mol. Biol. **359**, 546 (2006).
[61] L. Qui, S. A. Pabit, A. E. Roitberg, and S. J. Hagen, J. Am. Chem. Soc. **124**, 12952 (2002).
[62] E. Freire, in *Energetics of Biophysical Macromolecules*, edited by M. L. Johnson and G. K. Ackers (Academic, San Diego, 1996).
[63] K. W. Plaxco, K. T. Simmons, and D. Baker, J. Mol. Biol. **277**, 985 (1998).
[64] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, J. Mol. Biol. **252**, 460 (1995).
[65] A. N. Naganathan, U. Doshi, A. Fung, M. Sadqi, and V. Muñoz, Biochemistry **45**, 8466 (2006).