Modeling protein thermodynamics and fluctuations at the mesoscale

Naoko Nakagawa
Department of Mathematical Sciences, Ibaraki University, Mito, Ibaraki 310-8512, Japan and Laboratoire de Physique, ENS-Lyon, 46 allée d'Italie, 69364 Lyon Cedex 07, France

Michel Peyrard
Laboratoire de Physique, ENS-Lyon, 46 allée d'Italie, 69364 Lyon Cedex 07, France
(Received 7 December 2005; published 25 October 2006)

We use an extended Gō model, in unfrustrated and frustrated variants, to study the energy landscape and the fluctuations of a model protein. The model exhibits two transitions, folding and dynamical transitions, when changing the temperature. The inherent structures corresponding to the minima of the landscape are analyzed and we show how their energy density can be obtained from simulations around the folding temperature. The scaling of this energy density is found to reflect the folding transition. Moreover, this approach allows us to build a reduced thermodynamics in the inherent structure landscape. Equilibrium studies, from full molecular dynamics (MD) simulations and from the reduced thermodynamics, detect the features of a dynamical transition at low temperature and we analyze the location and time scale of the fluctuations of the protein, showing the need of some frustration in the model to get realistic results. The frustrated model also shows the presence of a kinetic trap which strongly affects the dynamics of folding.

DOI: 10.1103/PhysRevE.74.041916 PACS number(s): 87.15.Ya, 87.15.He, 87.15.Cc

I. INTRODUCTION

Proteins are fascinating molecules which perform a large variety of functions in biological systems. The most remarkable property of some proteins is their ability to work as molecular motors, i.e., to turn chemical energy into mechanical motion. How this occurs is not understood but even the basic properties of proteins, such as their folding or their glass transition, raise many questions yet unanswered. For instance, for a long time folding has been assumed to occur in successive stages, first the formation of secondary structures and then their positioning in space. This is now questioned because studies show that the α helix may require the environment of the protein to be stable, so that folding has to be global [1]. On another hand, the "glass transition," which is the low-temperature freezing of the large conformational changes which exist at biological temperature, is still the subject of many investigations presently, in particular to determine to what extent it is slaved to a transition in the solvent [2], and a consensus has yet to be reached concerning the form and time scales of the protein motions activated when this dynamical transition [3] is passed by raising temperature.

The specific properties of proteins are closely related to their fluctuations, which have been the object of numerous investigations. For instance, NMR spectroscopy [4] which can precisely study the fluctuations of the side chains of the amino acids that compose a protein, neutron scattering [5], or dielectric measurements [6] have observed a large increase of the fluctuations above approximately 200 K corresponding to the dynamical transition. But experiments are rather crude tools to study fluctuations [7] because they provide ensemble averages such as root-mean-square quantities. Recently subtle properties of the fluctuations have been investigated with single molecule experiments [8] but even when ensemble averages are eliminated by such studies, the experiments are observing time averages because they measure over time domains which are much longer than the time scale of the fluctuations. This is why many theoretical studies using protein models, which can study individual motions with high resolution, have been performed. Here too various views are possible. It is tempting to use all-atom molecular dynamics simulations to directly investigate protein dynamics with a high accuracy. This has been done successfully for small proteins [3,9,10]. Molecular dynamics is perfect to study subnanosecond dynamics but some conformational changes occur on much longer time scales which can extend to μs or more. Therefore even when low temperature simulations observe some features characteristic of the glassy behavior such as the Boson peak [10], they only study the properties of fast fluctuations and not the conformational changes which, at these temperatures, are rare events. Folding is an even slower process and the first numerical experiment to "fold a protein with a computer" with an all-atom model was a challenge that required a huge computer power [11].

To overcome these difficulties, another limit is to consider highly simplified models. Two-dimensional dynamical models have given useful information [12,13] and even simpler models, lattice models, where the motion of the protein is restricted to positions on a discrete lattice [7,14] have given interesting insights on protein fluctuations.

There are, however, some questions that these simple models cannot answer, it is all those which are related to the complex geometrical structure of a protein. To address these questions, one needs a model which describes the full three-dimensional structure of an actual protein, while being simple enough to allow studies on long time scales. Such a model, proposed a long time ago by N. Gō and co-workers [15], became popular as the "Gō model." The idea is to describe only the backbone of the protein and design the potentials for the interaction between its elements to make sure that the ground state of the model corresponds to the native state of the protein of interest. Although the original paper

shows that the model can describe the fluctuations of the protein configuration between different conformations, the model has essentially been used to study folding. The properties of a protein cannot be reduced to its folding transition, and therefore one should consider a model which can reproduce not only protein folding but other typical features. One of them, which is particularly important and typical of proteins, is their dynamical transition, which is rather universally observed for various proteins. Because the two transitions are major features of proteins, it is important to study them in conjunction with each other. This is the aim of our study and to our knowledge it is the first time that these two basic features of proteins are studied theoretically in the same framework. The Gō model, provided it is enriched by some minimal frustration, allows such a study, which on one hand tells us more about proteins by comparing the results provided by the model with known experimental facts, and, on another hand, brings results on the requirements that a minimal protein model has to fulfill. The questions that we want to address are the following:

(i) *How is the energy landscape of a protein?* The notion of energy landscape, which designates the potential energy of a protein as a function of its numerous degrees of freedom [16] has been a very fruitful tool for the reasoning on protein properties because it leads to a conceptual image of the energetic configurational space. However, it is a highly multidimensional object which cannot be drawn or even computed although some of its properties have been determined for very simple protein models [17,18], small molecules [19], or secondary structure elements of a protein [20]. We introduce here a method which allows us to fully characterize the density of states of a reduced version of the energy landscape which corresponds to the minima of the full energy landscape. It can be obtained for a three-dimensional model of an actual protein and, moreover, the ability to precisely get its density of states allows us to exhibit scaling properties in the landscape, which are strongly correlated with the folding transition. Moreover, we show that the knowledge of the reduced energy landscape can be used to derive a reduced thermodynamics of the protein, which characterizes the properties of its conformational fluctuations to a high accuracy.

(ii) *How does the protein explore its energy landscape?* This question is related to the dynamical transition of the protein. Below the transition temperature, the fluctuations of the protein are restricted and very local, and the protein can hardly explore its global energy landscape. While experiments clearly demonstrate that the solvent plays a role in this transition [21], one may wonder whether the complexity of the landscape of an actual protein would be sufficient to lead to such a dynamical transition. Molecular dynamics simulations of isolated proteins answer this question positively [22–24], but these calculations are limited to a time scale of the order of 1 ns and thus cannot probe a possible slowing down when the transition is approached from above. Besides the effect of the solvent, the conformational change of side chains can contribute to the dynamical transition [25,26]. Such local conformational changes would in turn induce changes on a larger scale, showing up in fluctuations of the backbone itself. In simple models the effect of solvent is implicitly included in the effective potential for the backbone, which does not depend on temperature. A possible effect of a temperature dependent dynamics of the solvent is precluded and the occurrence of the conformational changes of the side chains is omitted. Thus the only dynamical effects that can occur come from the way the protein explores its energy landscape. A full molecular dynamics study of our Gō model for time scales longer than 1 μs indicates that the motion of the backbone can exhibit the dynamical transition, without any explicit driving by the solvent or the conformational changes of the side chains. We think that the complexity of an actual protein, with its full chemical structure and solvent effects is not a prerequisite for a physical system to exhibit both a folding and a dynamical transition, and this suggests that minimal models can be used in a meaningful way to understand some important properties of proteins.

(iii) *What features are required in a "minimal" protein model?* In spite of its importance, taking into account the complex geometry of the protein backbone in the modeling is not sufficient to lead to a realistic model. Our study allows us to precisely see how the observations depend on the features of the model by showing that a minimal frustration is necessary to lead to realistic results, in agreement with the conclusions of statistical physics studies [27,28].

The organization of the paper is the following. We first introduce the model in Sec. II. We discuss two versions of the model, a simple unfrustrated Gō model, and a model in which frustration is introduced through the dihedral angles. They are then tested against various actual properties of proteins in order to determine the conditions that an appropriate model should meet. Examining folding and unfolding gives a first hint that frustration is necessary.

Section III investigates the equilibrium properties of the model. In the spirit of studies performed to study glasses we analyze the numerous metastable states which exist in the energy landscape of the model, which correspond to the so-called "inherent structures" of glasses. We show that a thermodynamics of the inherent states can be built and that it describes the equilibrium properties associated to the conformational changes of the protein.

Section IV investigates the dynamical properties of the protein in equilibrium. The temperature evolution of its fluctuations, their location within the structure, and their time scales are computed and discussed in comparison with known experimental properties. This section strengthens the need of some frustration in the model.

Section V investigates the out of equilibrium properties of the protein, in particular the dynamics of its folding when temperature is abruptly reduced from a high value to a value below the folding temperature. It allows us to characterize the energy landscape further by extracting some information on the barriers between minima. All these results allow us to discuss the modeling of proteins at mesoscopic scale in Sec. VI and to examine some of their properties, particularly the dynamical transition.

II. MODEL

A. The models

We chose to study a small protein containing all types of typical secondary structure elements (α helix, β sheets, and

loops), protein G (the B_1 domain of immunoglobulin binding protein [29], Protein Data Bank ID code 2GB1). It contains 56 residues, with one α helix and four β strands forming a β sheet. Its NMR structure is shown in Fig. 1.

As discussed in Sec. I we want a model able to describe the actual geometrical structure of the protein, but nevertheless sufficiently simple to allow us to investigate slow processes, such as the conformational changes at low temperature, or folding. To study the fluctuational dynamics of the protein we need an off-lattice model. All those requirements naturally lead to the choice of an off-lattice Gō model [31,15], which is rather simple but nevertheless accurate enough to reproduce experimental properties of the transition state for several proteins [32].

The design of this model, which only describes the backbone of the protein defined by the chain of C^α carbons, is tailored to the correct description of the geometry, but the expression of the potential energy only uses a small number of parameters and does not intend to be quantitatively correct, as it is the case for the all-atom models. The main idea of the model is to classify all possible contacts as either "native" (i.e., present in the native conformation) or "non-native." The potential energy is then constructed so that native contacts are favorable, and non-native contacts are less favorable, neutral, or repulsive. It is given by

$$V = \sum_{i=1}^{N_r-1} \frac{K_b}{2}(b_i - b_{0i})^2 + \sum_{i=1}^{N_r-2} \frac{K_a}{2}(\theta_i - \theta_{0i})^2 + \sum_{i>j-3}^{\text{native}} \epsilon\left\{5\left(\frac{r_{0ij}}{r_{ij}}\right)^{12} - 6\left(\frac{r_{0ij}}{r_{ij}}\right)^{10}\right\} + \sum_{i>j-3}^{\text{non-native}} \epsilon\left(\frac{C}{r_{ij}}\right)^{12} + \sum_{i=1}^{N_r-3} \frac{K_d}{2}\left[1 - \cos\left(2\phi_i - \frac{\pi}{2}\right)\right], \quad (1)$$

where N_r is the number of residue, b_i, θ_i, and ϕ_i are the ith bond length, ith bond angle, and ith dihedral angle, respectively. r_{ij} is the distance between ith and jth residues. The index 0 denotes the parameters determined from the native structure. The model includes local and nonlocal interactions. The local interactions are the bonds which define the C^α chain through stiff elastic potentials describing the covalent bonds, angular potentials between adjacent bonds, and dihedral potentials. The natural bond length or bond angles are determined from the native structure. The nonlocal bonds connect C^α carbons which are not adjacent along the peptide chain, but nevertheless close in space in the folded geometry of the protein. Such carbons are said to form a "native contact" and interact with a Lennard-Jones potential in the model. They are determined from the experimental structure of the protein in its folded state: two C^α carbons are said to form a native contact if the distance between them or between atoms of the side chains attached to them is less than a critical distance, here chosen as 5.5 Å. A repulsive interaction is introduced between all C^α carbons which do not form a native contact to describe the steric repulsion. For $K_d=0$ all the interactions are at their minimum energy in the native state. This is the choice made in the original version of the Gō model and models having this property are generally good folders, i.e., their folding transition is sharp and leads to a well defined structure, because the native state does not result from a compromise between competing interactions. The model with $K_d=0$ is *unfrustrated* and will be henceforth denoted as "model U."

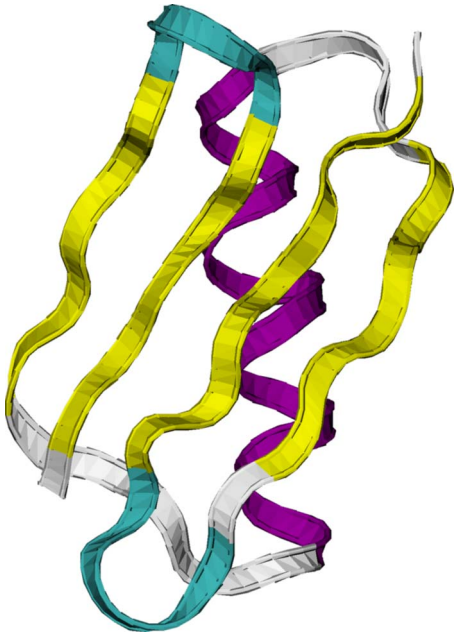

FIG. 1. (Color online) Representation of the main structural elements of the B_1 domain of protein G (drawn with VMD [30]). The purple part corresponds to the α helix, while the yellow ribbons are β sheets. Residue number 1 is the end of the chain at the top right of the figure. The main structural elements of this protein are two β strands comprising residues 1–7 and 14–19, which form a β sheet denoted by β_N, one α helix made of residues 23–35, and two β strands comprising residues 42–46 and 51–55, which form the β sheet called β_C. Other residues belong to loops.

The choice of an unfrustrated model becomes questionable if we want to investigate the fluctuations of the protein because the structure of actual proteins results from a compromise between different interactions. This introduces frustrations which are likely to play an important role in the protein dynamics. Proteins have been described as "minimally frustrated systems" [16], and this is why we have introduced the possibility to add a frustration coming from competitive dihedral angle interactions [31]. It is provided by the last term of Eq. (1). When $K_d \neq 0$ ($K_d=0.3$ in our case), the last term tends to favor dihedral angles equal to $\pi/4$ (mod π) which are not exactly equal to the value of the dihedral angles in the native structure. This energy terms competes with the other contributions. This leads to a ground state slightly distorted with respect to the native structure, but the geometrical influence of the dihedral potential terms stays small, while it has a significant influence on the dynamical properties of this version of the model, called *frustrated model* (or "model F"), as shown below. Such frustrated models are still simple but nevertheless accurate enough to reproduce experimental properties of the transition state for several proteins. The introduction of the frustration through a constraint on the dihedral angle might sound arbi-

trary. First one should realize that, as pointed out above, the potential energy of Gō models does not claim to be quantitatively correct. This is a remarkable feature of Gō models that they are able to give results which agree with experiments, for instance for the structural details of the transition state and intermediates for folding [32] although their potential energy is highly simplified. In the same spirit, introducing frustration through the dihedral angles can be understood as a way to include a physical effect, frustration, through the simplest description that is qualitatively correct. Second, an analysis of the folding of different proteins described by frustrated Gō models where the frustration was introduced through the dihedral angles showed the validity of this approach [31] because, for small fast-folding proteins, the energetic roughness is not the dominant factor determining a sequence's foldability. Topological frustration, which can be described by the dihedral angle contribution, is even more important [32].

The dimensionless parameters used in our calculations are $K_b=200.0$, $K_a=40.0$, $K_d=0.3$, $\epsilon=0.18$, and $C=4.0$. Time is also a dimensionless variable, measured in arbitrary time units.

Our investigations have been carried by molecular dynamics simulations using underdamped Langevin simulations [33], where the mass of all the residues is assumed to be equal to 10 and the time step used in the simulation is equal to 0.1 time unit. They show that the model has the expected qualitative properties of a good folder: at low temperature it folds to the global minimum of the potential energy, while at high temperature it denatures. If the model is cooled down quickly from a high temperature unfolded state, the folding is, however, very slow, and may never reach the global minimum within the time scale accessible in a simulation. This suggests that this model protein exhibits the glassy properties observed for actual molecules.

However, except in Sec. V, our goal is not to explore these out-of-equilibrium states, but rather to analyze the equilibrium properties of the protein. Therefore to avoid long nonequilibrium transients which occur on cooling, most of our simulations have been carried by starting from the ground state, and heating up to the temperature of interest. As discussed below, the formalism that we introduce to study the equilibrium properties allows us to avoid the difficult issue of the ergodicity of simulations at low temperature so that the pathway that we use to reach a particular temperature is not an issue which could affect the results.

B. Folding properties

The folding transition of the two models can be detected by comparing the shape of the protein with the fully folded native state. The comparison can be quantified by introducing the *dissimilarity factor* [34] which is a weighted distance map between two conformations. Let A and B be two conformations of the protein (for our purpose B will be the native structure) and a_{ij}, b_{ij} the distance between residues i and j in these conformations. The dissimilarity $d(A,B)$ between the two conformations is defined by

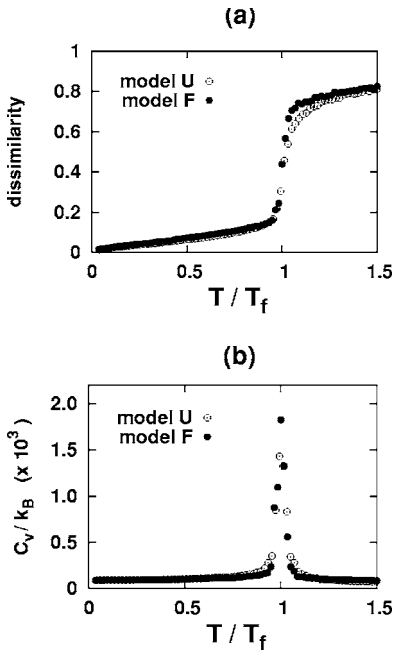

FIG. 2. (a) Temperature evolution of the dissimilarity between the average structure of the protein model and the native structure. (b) Temperature evolution of the specific heat of the protein model.

$$d(A,B)=\frac{\sum_{i=1}^{N_r}\sum_{j=i+2}^{N_r}|a_{ij}^{-p}-b_{ij}^{-p}|}{\frac{1}{2}\left(\sum_{i=1}^{N_r}\sum_{j=i+2}^{N_r}a_{ij}^{-p}+\sum_{i=1}^{N_r}\sum_{j=i+2}^{N_r}b_{ij}^{-p}\right)}, \quad (2)$$

where a_{ij} and b_{ij} are the distances between residues i and j in the A and B conformations and p is an integer which determines how much residues which are far apart contribute. For large p only the closest neighbors contribute to the weighted map, while for the value $p=2$ that we are using, residues separated in space by a greater distance also contribute. The dissimilarity factor is a quantitative measure of the deviation between the geometries of two conformations.

For our applications conformation B will always be chosen as the native structure of the protein, so that when we speak henceforth of the "dissimilarity factor" we mean the dissimilarity between the conformation of interest and the native structure. With this definition the dissimilarity factor, that we shall denote by D, vanishes when the conformation is identical to the native structure and increases when the geometry of the conformation of interest deviates from the native structure.

Figure 2(a) shows that, when temperature is lowered from a high value, the dissimilarity factor drops sharply in a narrow temperature range. The corresponding temperature can be identified as the folding temperature of the protein model, henceforth denoted by T_f. The variation of the specific heat, plotted in Fig. 2(b), shows a peak which confirms that T_f can be viewed as a thermodynamic transition between two states, the unfolded state at high temperature and the folded state below T_f. This transition is sharp as expected for Gō models

which are unfrustrated (model U) or weakly frustrated (model F). There are, however, differences between the two cases, $T_f(U) < T_f(F)$, and the transition is slightly sharper for model F than for model U.

III. EQUILIBRIUM PROPERTIES

The complete energy landscape of the protein is very hard to determine but a partial picture can be obtained by looking for all its metastable states. They correspond to the so called "inherent structures" in the language of glasses [36]. The interest is that, although it contains less information than the free energy landscape, the inherent structure landscape contains nevertheless a rich set of data on the protein and, moreover, it can be obtained from simulations at temperatures well above the temperature at which the protein shows a dynamical freezing which avoids possible ergodicity problems. In this section we would like to show how it can be obtained, and then how it can be used to build a thermodynamics of the conformational motions of a protein.

A. Obtaining the inherent structure landscape

Inherent structures are local minima of the energy landscape of the protein. Each one, labeled by an index α, corresponds to a basin of attraction in the phase space which is defined as the set of conformations which are connected to the same local energy minimum. In practice they are obtained by sampling the phase space with molecular dynamics (MD) trajectories at constrained temperature, and then quenching instantaneous configurations by calculating a steepest-descent path from these points of the trajectory. The MD trajectories involve 3×10^8 steps after the desired temperature is reached, and for each temperature 20 simulations are performed with different sets of random numbers in the Langevin equations used to thermalize the system. For each temperature, quenchings are made for 6×10^4 instantaneous configurations separated by 10^5 steps.

In order to characterize the inherent structure landscape we need not only the values of the energies e_α but also their density of states $\Omega_{IS}(e_\alpha)$. A realistic protein has a huge number of inherent states so that $\Omega_{IS}(e_\alpha)$ cannot be determined from a systematic search. Instead it can be deduced from a sampling of the inherent structures which gives the probability $P_{IS}(e_\alpha,T)$ to be in the basin of the inherent structure α at temperature T.

In the basin of attraction of inherent structure α, the potential energy of the protein can be written as $V(\mathbf{r})=e_\alpha+\Delta V(\mathbf{r})$, where $\mathbf{r}$ designates all the degrees of freedom of the protein. Let us assume that the inherent structures can be split into a set of discrete states $\alpha_0,\alpha_1,\ldots,\alpha_K$ (which includes at least one state, the ground state) and a continuum of higher energy states. With this notation, the configurational contribution of the partition function of the protein is

$$Z(T)=\sum_{\alpha_i=\alpha_0}^{\alpha_K} e^{-\beta e_{\alpha_i}}e^{-\beta F_v(\alpha_i,T)}+\int_{e_{\alpha_K}}^{e_{\max}}\Omega_{IS}(e_\alpha)e^{-\beta e_\alpha}e^{-\beta F_v(\alpha,T)}de_\alpha, \tag{3}$$

where $e_{\max}$ is the highest inherent structure energy. Moreover, we have introduced

$$e^{-\beta F_v(\alpha,T)}=\int_{B(\alpha)} e^{-\beta\Delta V(\mathbf{r})}d\mathbf{r} \tag{4}$$

which is the vibrational free energy in the basin of attraction $B(\alpha)$ of the inherent structure α. The probability to be in the basin of attraction of a discrete inherent structure is

$$p_{\alpha_i}(T)=\frac{1}{Z(T)}\exp(-\beta e_{\alpha_i})\exp[-\beta F_v(\alpha,T)] \tag{5}$$

and, for the continuum of inherent structures, the probability density of inherent structures is

$$P_{IS}(e_\alpha,T)=\frac{1}{Z(T)}\Omega_{IS}(e_\alpha)\exp(-\beta e_\alpha)\exp[-\beta F_v(\alpha,T)]. \tag{6}$$

If we can assume that the free energy $F_v(\alpha,T)$ does not depend on the inherent structure, in Z the term $\exp[-\beta F_v(T)]$ can be factorized so that p_{α_i} and P_{IS} simplify into

$$p_{\alpha_i}(T)=\frac{e^{-\beta e_{\alpha_i}}}{Z_{IS}(T)},\quad P_{IS}(e_\alpha,T)=\frac{1}{Z_{IS}(T)}\Omega_{IS}(e_\alpha)e^{-\beta e_\alpha} \tag{7}$$

with

$$Z_{IS}(T)=\sum_{\alpha_i=\alpha_0}^{\alpha_K} e^{-\beta e_{\alpha_i}}+\int_{e_{\alpha_K}}^{e_{\max}}\Omega_{IS}(e_\alpha)e^{-\beta e_\alpha}de_\alpha. \tag{8}$$

Z_{IS} can be viewed as an inherent structure partition function. It can be expressed in terms of the probability $p_{\alpha_0}(T)$ that the protein is in the basin of attraction of the ground state, which is henceforth denoted as $p_0(T)$. This allows us to rewrite $P_{IS}(e_\alpha,T)$ as

$$P_{IS}(e_\alpha,T)=p_0(T)\Omega_{IS}(e_\alpha)e^{-\beta e_\alpha}, \tag{9}$$

if we chose the ground state energy as the reference state of the energies ($e_{\alpha_0}=0$). Equation (9) provides a method to compute $\Omega_{IS}(e_\alpha)$ from the probability density $P_{IS}(e_\alpha,T)$ deduced from the sampling of MD trajectories.

The validity of the method relies on the assumption that the vibrational free energy in a basin of attraction $F_v(\alpha,T)$ does not vary significantly from a basin to another. For harmonic basins of attraction, we have

$$F_v(\alpha,T)=k_BT\sum_q \ln\left(\frac{\hbar\omega_q}{k_BT}\right), \tag{10}$$

where the values ω_q are the frequencies of the vibrational modes of the protein when it is in the configuration corresponding to the inherent structure α. Of course these frequencies change individually from one inherent state to another, but they keep the same order of magnitude, and the sum has an averaging effect that keeps $F_v(\alpha,T)$ only weakly dependent of the conformation. This is verified by our calculations because $\Omega_{IS}(e_\alpha)$ deduced from $P_{IS}(e_\alpha,T)$ at different temperatures gives the same result in all the range of inherent structure energies which are accessed with a sufficient probability to allow a correct sampling at the temperature consid-

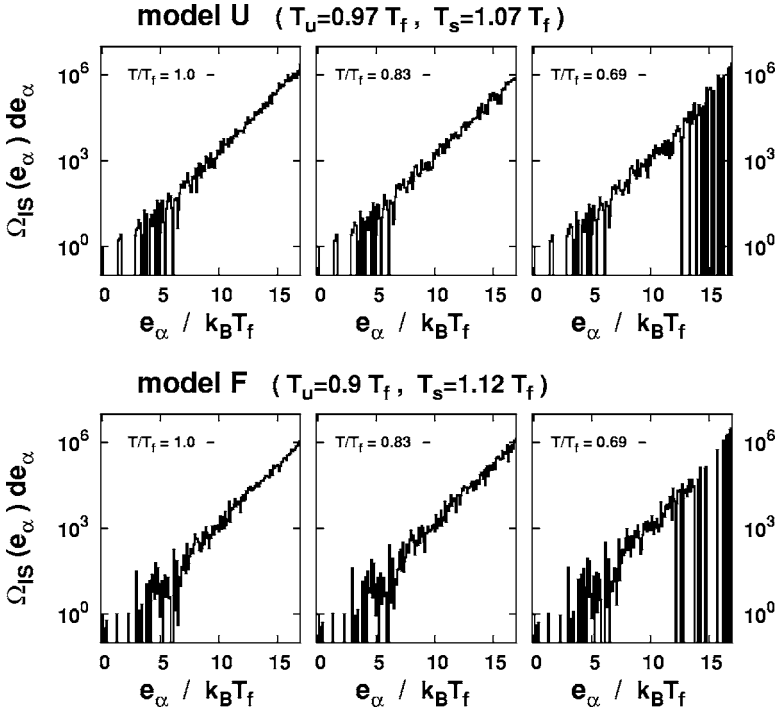

FIG. 3. Comparison of the density of states $\Omega_{IS}(e_\alpha)$ computed from the probability densities $P_{IS}(e_\alpha,T)$ deduced from molecular dynamics trajectories at three different temperatures $T=T_f$, $T=0.83T_f$, and $T=0.69T_f$ for the unfrustrated model U (top part) and the frustrated model F (bottom part) ($de_\alpha=0.02$).

ered. Figure 3 shows the density of states $\Omega_{IS}(e_\alpha)$ computed from the sampling of molecular dynamics trajectories at three temperatures. The figures are remarkably similar, demonstrating that the assumption that $F_v(\alpha,T)$ can be eliminated in the calculation is a good approximation, however, the three figures are obviously not identical. The figure deduced from the calculation at the lowest temperature shows some values of e_α where the density of states appears to vanish, while it does not vanish if $\Omega_{IS}(e_\alpha)$ is computed from a trajectory at higher T. This is particularly noticeable for high values of e_α and it points out the limit of the numerical calculation. As shown from Eq. (9), the probability to be in the basin of attraction of an inherent structure decreases as $\exp(-e_\alpha/k_BT)$. Therefore if we use a molecular dynamics trajectory at a low temperature to sample the basins of attraction of the inherent structures, it may happen that we completely miss some inherent structures due to insufficient sampling. This is even more likely if the basin of attraction of a given inherent structure is narrow because, besides the temperature factor, its small size also reduces the chance to sample it properly. Studies performed at higher temperature reduce this problem and $T=T_f$ is particularly appropriate because we expect a molecular dynamics trajectory to sample both structures corresponding to folded and unfolded states. Once $\Omega_{IS}(e_\alpha)$ has been determined from Eq. (9) at a temperature where the sampling is particularly efficient, $Z_{IS}(T)$ can be computed at any temperature with Eq. (8), including low temperatures for which MD calculations might become incorrect due to ergodicity problems.

Figure 4 displays $\Omega_{IS}(e_\alpha)$ in a large range of energies. Although it has the same general shape for models U and F, there are some differences in the low energy range (see the insets), near the ground state, which is separated from the others. For model U there is a large gap between the ground state and the next state. On the contrary, model F shows a few states which have an energy very close to the ground state, separated from higher energy states by a large gap. The calculation of the dissimilarity between the native state and the cluster of states near the ground state gives very small values ($D\approx0.02$), well below dissimilarity values that correspond to a real change of the protein shape when it unfolds. This indicates that, although the ground state is not isolated from others by a large gap, model F is nevertheless a good folder [35] because the cluster of states near the ground state can be considered as belonging to the same fold as the native state [37]. Therefore the protein does evolve towards a unique geometrical structure when it folds. This points out an additional richness of the off-lattice models with respect to lattice models for which a ground state isolated from all the others by a large gap appears to be a prerequisite for good folders [35].

For larger energies $\Omega_{IS}(e_\alpha)$ shows an exponential growth, with two different slopes depending on the energy range:

$$\Omega_{IS}(e_\alpha) = \Omega_0 \exp[g(e_\alpha)] \tag{11}$$

with

$$g(e_\alpha) = \begin{cases} \dfrac{e_\alpha}{k_BT_s} + A & (e_\alpha < 45k_BT_f), \\ \dfrac{e_\alpha}{k_BT_u} & (e_\alpha > 50k_BT_f), \end{cases} \tag{12}$$

where Ω_0 and A are positive constants, and $T_s=1.07T_f$, $T_u=0.97T_f$ for model U and $T_s=1.12T_f$, $T_u=0.9T_f$ for model F. It is interesting to notice that a similar exponential dependence of the density of metastable state was found for spin glasses [38] although the energy dependence of $g(e_\alpha)$ is more complex for such systems.

The density of states in the inherent structure landscape is related to the probability density of the occupation of the

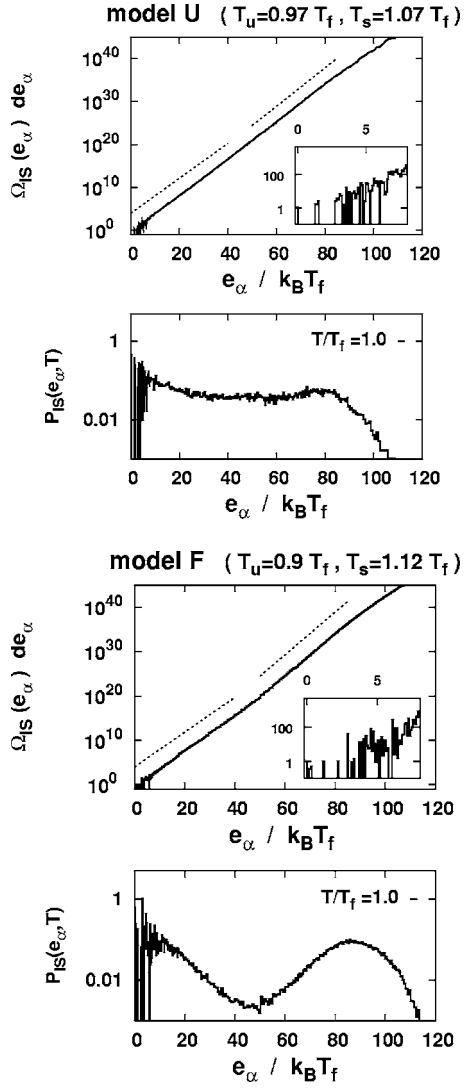

FIG. 4. Top figures: density of state $\Omega_{IS}(e_\alpha)$ for the inherent structures, which is obtained from $P_{IS}(e_\alpha,T_f)$ according to Eq. (9). The inset shows a magnification of the curve for the energies close to the ground state. The dotted lines are two distinct exponential functions with exponent corresponding to Eq. (12). Bottom figures: probability distribution (in logarithmic scale) of the inherent state energies at temperature T_f.

different basins of attraction versus temperature through Eq. (9) which gives

$$P_{IS}(e_\alpha,T)=\frac{1}{Z_{IS}(T)}\Omega_0\exp\left[\frac{e_\alpha}{k_B}\left(\frac{1}{T_s}-\frac{1}{T}\right)+A\right] \quad \text{for } e_\alpha<45k_BT_f, \tag{13}$$

$$P_{IS}(e_\alpha,T)=\frac{1}{Z_{IS}(T)}\Omega_0\exp\left[\frac{e_\alpha}{k_B}\left(\frac{1}{T_u}-\frac{1}{T}\right)\right] \quad \text{for } e_\alpha>50k_BT_f. \tag{14}$$

In the low temperature range, the contribution to $P_{IS}(e_\alpha,T)$ given by Eq. (13) dominates that given by Eq. (14). As a result, in this temperature range, $P_{IS}(e_\alpha,T)$ takes its largest values for low e_α, corresponding to the folded state. On the contrary, at high temperature $P_{IS}(e_\alpha,T)$ is dominated by the term coming from Eq. (14), so that high-energy inherent structures are the most likely, corresponding to the unfolded state. And, in the range $T_u<T\approx T_f<T_s$ none of the two terms dominates and $P_{IS}(e_\alpha,T)$ has two humps, one of them at low e_α corresponding to folded structures, the other one at high e_α, corresponding to unfolded structures, as shown in Fig. 4 (bottom figures). This is particularly true for model F while Fig. 4 shows that at $T=T_f$ model U occupies a continuum of structures ranging from folded to unfolded states because the two temperatures T_u and T_f are close to each other and the humps overlap. This picture is consistent with a first order transition, $T_u<T<T_s$ being the coexistence region, so that the two temperatures exhibited by the analysis of $\Omega_{IS}(e_\alpha)$ appear as the limit of the spinodals. These characteristics are consistent with the behavior of the specific heat C_v.

B. Thermodynamics in the inherent structure space

The density of inherent structures energies and the associated partition function $Z_{IS}(T)$ are not only useful to characterize the folding transition. They can also tell us a lot on the thermodynamics of the protein. We can build a thermodynamics only based on the configurations of the inherent structures by the usual scheme of statistical mechanics,

$$U_{IS}(T)=\langle e_\alpha\rangle=\int e_\alpha P_{IS}(e_\alpha,T)de_\alpha, \tag{15}$$

$$F_{IS}(T)=-k_BT\ln[Z_{IS}(T)], \tag{16}$$

$$S_{IS}(T)=\frac{U_{IS}-F_{IS}}{T}, \tag{17}$$

$$C_{IS}(T)=\frac{\langle\Delta e_\alpha{}^2\rangle}{k_BT^2}=\frac{\langle e_\alpha{}^2\rangle-\langle e_\alpha\rangle^2}{k_BT^2}, \tag{18}$$

where U_{IS}, F_{IS}, S_{IS}, or C_{IS} are the internal energy, Helmholtz free energy, entropy, and specific heat, respectively, deduced only from the thermal fluctuations between conformations, so that they can be called conformational energy, conformational free energy, etc. They are shown in Fig. 5.

The conformational specific heat is very similar to the full specific heat of Fig. 2, although the peak at T_f is smaller than the peak obtained from MD simulations, which is not surprising because some aspects are lost when one only considers the inherent structures instead of the exact conformations of the proteins. The detailed analysis of Fig. 5 (middle figures) which displays the vibrational specific heat per residue $C_W/N_r=(C_v-C_{IS})/N_r$ and conformational specific heat C_{IS} shows that the vibrational contribution of the specific heat is very close to the value $3k_B/2$ expected for harmonic oscillators. This suggests that the thermalization in the basin of attraction of each inherent structure allows a good separation between a vibrational part and a conformational part of the specific heat. When comparing the values of C_W and C_{IS} one

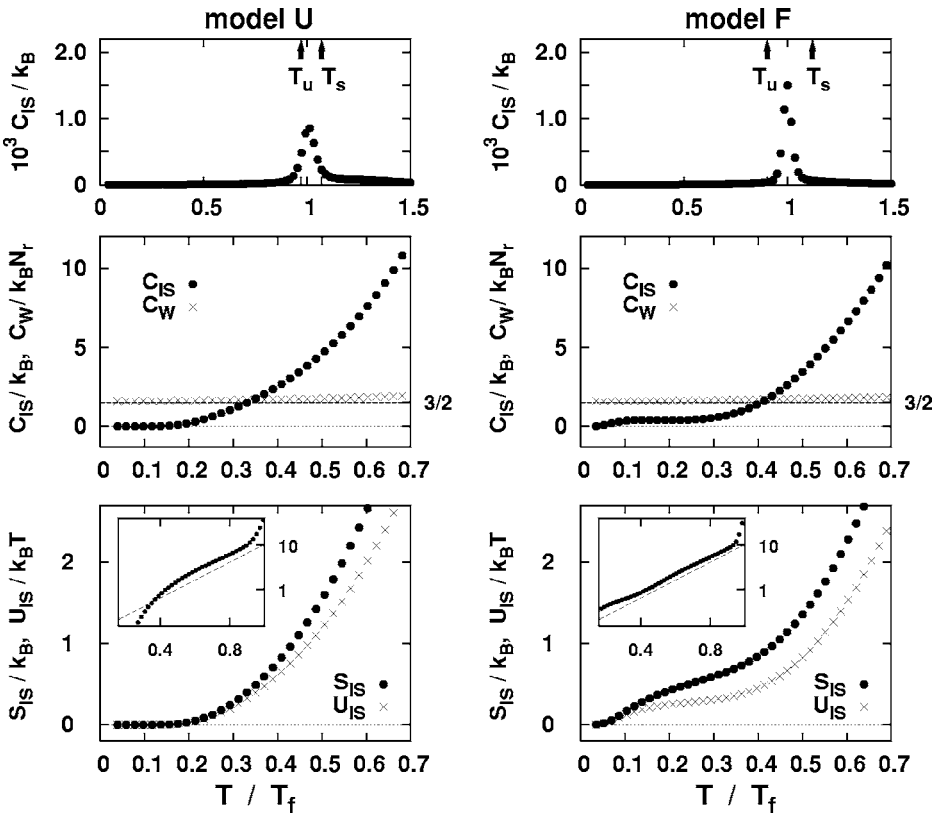

FIG. 5. Thermodynamic quantities defined from the inherent structures. Top figure: conformational specific heat C_{IS}/k_B. The arrows on the top axis point to the values of the temperatures T_u and T_s which enter in the expression of the density of inherent states [Eq. (12)]. Middle figure: magnification of the variation of the specific heat in the low temperature range, showing also the contribution $C_W=C_v-C_{IS}$, which is expected to reflect the properties of thermalization inside each well. Notice that the figure shows the value C_W/N_r, i.e., the vibrational specific heat per residue, C_{IS}, which is expected to reflect a global property of the protein, is shown for the full molecule. The dashed line indicates the harmonic limit $C_W/N_r=3k_B/2$. Bottom figure: conformation entropy S_{IS} and conformation energy U_{IS}. Inset: conformational entropy in logarithmic scale.

should notice an essential difference between the two. C_W is a vibrational contribution, which comes from all the residues and, as it is well approximated by the harmonic value, it indicates that the contributions of all vibrational degrees of freedom are independent, so that C_W/N_r has an actual physical meaning. On the contrary, the structural fluctuations entering in C_{IS} may be either a flip of a residue for instance, and thus can be very local, or a slow structural change extending to a significant part of the protein, such as the motion of the whole α helix. This is why the quantity C_{IS}/N_r would not have a physical meaning because a protein is not a self-similar structure. A meaningful comparison must compare C_{IS} to C_W/N_r: structural fluctuation will start to play a role in the physics of the protein when $C_{IS}>C_W/N_r$. As shown in Fig. 5, C_{IS} starts to rise significantly around $0.3T_f$ and quickly takes over for $T>0.4T_f$, while the fluctuations are dominated by the vibrational contribution in the low temperature range.

The behaviors of the conformational entropy and energy (Fig. 5, bottom plots) show a similar rise clearly visible for $T>0.4T_f$. The increase of entropy which appears to accelerate and become exponential for $T>0.4T_f$ (see inset) can be understood as coming from the structural variety observed in the thermal fluctuations. This would be consistent with the behavior of C_{IS} and C_W.

The results in the low temperature range show that models U and F have different behaviors for $T<0.3T_f$. Model F shows a small increase of specific heat and entropy around $T=0.05T_f$ which does not appear for model U. This may be related to the existence of inherent structures with an energy very close to the ground state for model F, which do not exist for model U (see insets in Fig. 4).

IV. DYNAMICAL ASPECTS IN EQUILIBRIUM

The inherent structure landscape gives a static view of the energy surface of the protein and provides us with some information on the states which are accessible to the molecule. We want now to examine the fluctuations of the protein in equilibrium at a given temperature because they are important for its function and have been extensively studied experimentally [6]. One of the issues is the existence and properties of the so-called "dynamical transition" of proteins, which is observed at around 200 K for many proteins.

A. Onset of the fluctuations

Figure 6 shows the temperature evolution of the fluctuations of the protein model in the low temperature range $T<0.7T_f$, measured by

$$\Delta u^2 = \frac{1}{N_r}\sum_i^{N_r} (\langle d_i^{02}\rangle - \langle d_i^0\rangle^2), \tag{19}$$

where d_i^0 is the distance between residue i and the center of mass of the protein. It is a quantity which is sensitive to the global fluctuations of the protein in a way which is very similar to the Debye-Waller factor, which only has a meaning for a protein crystal and cannot be defined for the single-protein model that we study here. Figure 6 clearly exhibits two regimes for the thermal fluctuations of the model protein, in a striking similarity with the experimental observations made, for instance, by Mösbauer absorption spectroscopy [6] or neutron scattering [5]. Below $T_D \approx 0.4T_f$, the calculation shows a linear increase which can be expected from the harmonic vibrations of the residues, but around

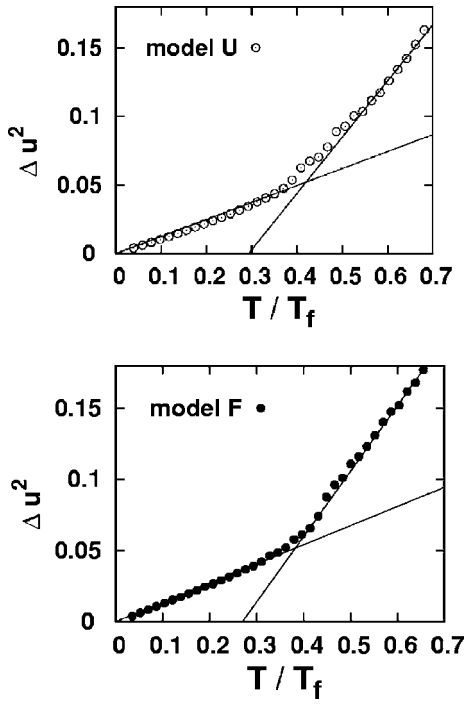

FIG. 6. Temperature variation of the average fluctuations of the positions of residues for models U and F, measured by Δu^2 [Eq. (19)].

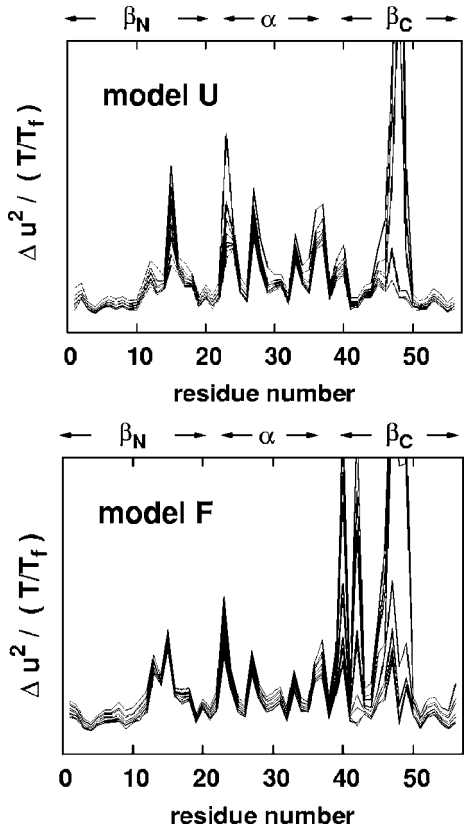

FIG. 7. Fluctuations of the position of the residues as a function of the index of the residue along the protein chain, at different temperatures, for model U and model F. As indicated in the figures, residues with the lower indices correspond to the β_N terminal part of the protein, residues with intermediate indices belong to the α helix, while the residues with the higher indices are part of the β_C terminal region. The different curves show $\Delta u^2/(T/T_f)$ at temperatures $T=0.20-0.60$ by steps of 0.05, the curves being in order of increasing temperatures from bottom to top.

$T_D=0.4T_f$ the fluctuations start to rise much faster with temperature. The change is sharp for model F and smoother for model U but clearly noticeable in both cases.

B. Location of the fluctuations

The quantity Δu^2 can be used to determine the location of the fluctuations if, instead of summing over all the residues as in Eq. (19), we compute it for a single residue. Figure 7 shows the result at different temperatures for models U and F. First one notices that the fluctuations are not homogeneous along the protein, some residues appearing as peaks over a low background, in agreement with experimental results obtained in actual proteins [40]. This is not surprising because it simply reflects the inhomogeneities of proteins. For instance, the largest peak in Fig. 7 is obtained for residue 47 belonging to a loop connecting two β strands, which corresponds to a highly flexible part of the protein. What is more interesting is the significant difference between model U and model F for these large fluctuations. For model U the largest fluctuation appears abruptly between $T=0.35T_f$ and $T=0.40T_f$ and it is highly localized on one residue. Several other sites also show fluctuations that become large at high temperature, but they are still isolated and scattered along the protein chain. On the contrary, model F shows more cooperativity. The growth of the largest amplitude fluctuation when temperature increases is more gradual and involves several sites around the most mobile residue. Moreover, the fluctuations are very different from one region to another. In the β_N and α helix region of the protein chain the fluctuations do not grow significantly with temperature, whereas the β_C region shows a large growth distributed over many of its sites. This higher cooperativity of model F is consistent with several other observations in our calculations: the folding transition (Fig. 2) and the change in slope in the variation of Δu^2 vs T (Fig. 6) are both sharper for model F than for model U.

Another way to assess the distribution of the fluctuations within the protein is to examine distances between atoms belonging to the main structural elements of the protein, i.e., the two β sheets and the α helix. Relative fluctuations of different structural elements can be defined by

$$\Delta r_{SS'}^2 = \left(\frac{1}{N_p} \sum_{p=1}^{N_p} \frac{\langle d_p^2 \rangle - \langle d_p \rangle^2}{\langle d_p \rangle^2} \right) \times \left(\frac{1}{N_p} \sum_{p=1}^{N_p} \langle d_p \rangle^2 \right), \quad (20)$$

where the index p designates a pair of residues, the first one belonging to structural element S and the second one to structural element S'. For given S and S', the sum extends over the N_p possible native pairs. For $S=S'$, Δr_{SS}^2 provides a measure of the rigidity of the structural element, while, for $S \neq S'$, $\Delta r_{SS'}^2$ measures the fluctuations in the relative motions of the two structural elements.

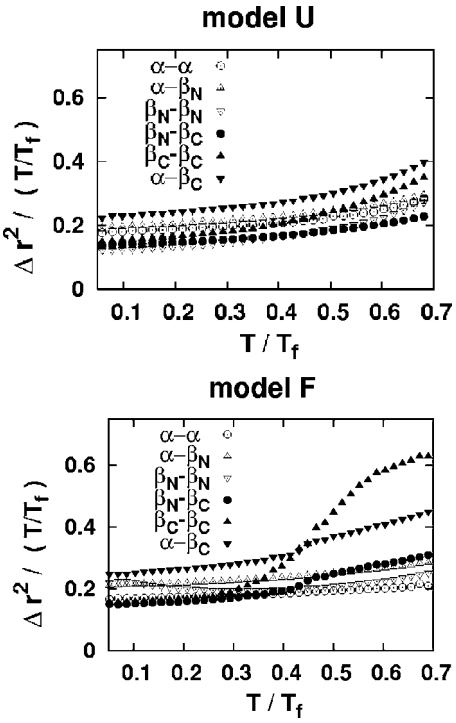

FIG. 8. Temperature evolution of the fluctuations of the distances between different structural elements of the protein [defined by Eq. (20)] for model U and model F. The residues belonging to each structural element are listed in the caption of Fig. 1.

Figure 8 shows a significant difference between models U and F. An increase of the amplitude of the fluctuations around $T_D=0.4T_f$ is visible for both models, in agreement with the global results provided by Δu^2 (Fig. 7) but the rise is much more significant for model F, and, as noticed above it mainly concerns one part of the protein, attesting of a greater cooperativity. The fluctuations inside the β_C sheet rise very significantly, and this also affects the relative motions $\beta_N-\beta_C$ and $\alpha-\beta_C$. The large fluctuations inside the β_C sheet come from a relative motion of the two β strands, showing again that model F exhibits some collective effects.

Thus, for the model that we study, the dynamical transition temperature T_D appears as the onset of fluctuations on a large scale. This is consistent with the properties of the statistical quantities studied in Sec. III B. As discussed for Fig. 5, the comparison of C_{IS} with C_W/N_r implies that conformational fluctuations begin to dominate the protein's fluctuations above $0.4T_f$. This is also suggested by the beginning of growth for the conformation entropy S_{IS}.

C. Time scale of the fluctuations

In order to characterize the fluctuations of the protein, it is also important to determine their characteristic time scale. Let us consider a quantity $x(t)$ that fluctuates in time such as the dissimilarity factor $x(t)=d(A,B)(t)$. We can compute the time-averaged quantity

$$\bar{x}(t;\tau)=\frac{1}{\tau}\int_t^{t+\tau}x(t')dt' \tag{21}$$

and then evaluate its fluctuations according to the usual expression,

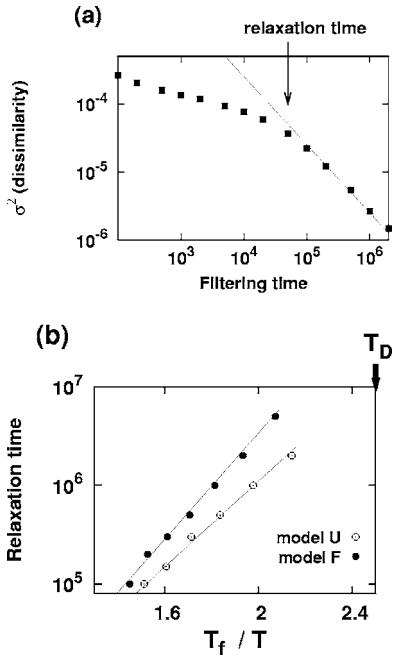

FIG. 9. (a) Variation of the fluctuations of the dissimilarity D between the model conformations and the native structure averaged over a time τ vs τ for model U at $T=0.7T_f$. The solid line shows the decay given by the law of large numbers. (b) Arrhenius plot of the relation time τ_r deduced from the fluctuations of the dissimilarity factor for models U and F. The solid lines are fits to $\tau_r \propto \exp(E_B/k_BT)$ with activation energies $E_B=6.2k_BT_f$ for model F and $E_B=5k_BT_f$ for model U.

$$\Delta\bar{x}^2(\tau)=\langle\bar{x}^2(t;\tau)\rangle_t-\langle\bar{x}(t;\tau)\rangle_t^2, \tag{22}$$

where the averages $\langle\cdot\rangle_t$ are time averages. Obviously, if $\tau\to\infty$, $\Delta\bar{x}^2(\tau)\to 0$, and generally $\Delta\bar{x}^2(\tau)$ decays with an increase of τ. But the way it decreases tells us about the time scales of the fluctuations of $x(t)$ because as long as τ is shorter than the typical time scale of the variation of $x(t)$, $\Delta\bar{x}^2(\tau)$ is not significantly different from the fluctuations Δx^2 of the original quantity $x(t)$, while it decays quickly when τ becomes longer than the typical time scale of the fluctuations of $x(t)$.

Figure 9 shows the result of such a calculation where $x(t)=D(t)$ is the dissimilarity between the instantaneous structure and the native state, at a given temperature. It clearly shows the existence of a characteristic time scale, that we denote by τ_r ($\tau_r\approx 5\times 10^5$ time units). For $\tau>\tau_r$, $\Delta\bar{x}^2(\tau)$ decreases significantly faster than for lower values of τ which indicates that the fluctuations of the dissimilarity D include a slow component having a period at least equal to τ_r. This appears clearly in Fig. 10 which shows the time evolution of the dissimilarity factor at $T=0.5T_f$. The results are consistent with a splitting of the fluctuations in two components, small-amplitude fast fluctuations within the basin of attraction of a given inherent structure and large jumps from basin to basin as discussed in Sec. III.

Similar calculations can be performed for various quantities, and in particular for the fluctuation of the energy of the

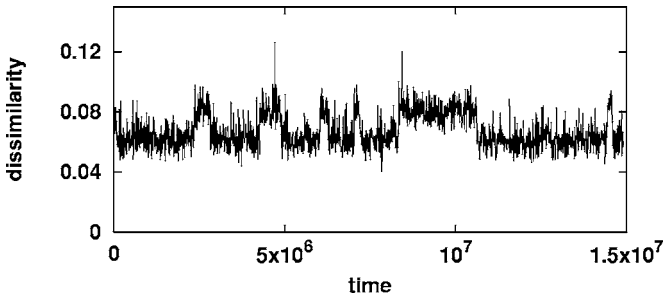

FIG. 10. Example of the time evolution of the dissimilarity factor D at $T=0.5T_f$. The presence of two types of fluctuations is visible: fast small amplitude variations and large changes occurring from time to time.

model, which determine the specific heat. The same behavior is observed, although the change of slope at $\tau=\tau_r$ may be smaller for some quantities. The value of τ_r does not depend on the quantity that is considered, showing that the relaxation time τ_r is an intrinsic property of the system.

If our analysis that τ_r is related to jumps from one inherent structure to another is correct, it should depend on temperature. As shown in Fig. 9(b), it exhibits an Arrhenius behavior in the range $0.45T_f \leqslant T \leqslant 0.75T_f$. For lower temperatures the relaxation time becomes so large that it cannot be measured in a simulation, and, for higher temperatures, approaching T_f, the separation of time scales between the oscillations within the basin of an inherent structure and the diffusion from basin to basin becomes blurred. Figure 9(b) shows that the relaxation time is significantly larger for model F than for model U. The activation energy is $6.2k_BT_f$ for model F and $5k_BT_f$ for model U. It can be viewed as the typical barrier to move from the basin of an inherent structure to another, suggesting that the energy landscape of model U is smoother than that of model F. This point is confirmed by the folding studies of Sec. V.

The longer time scales associated to the fluctuations of model F with respect to model U can also be observed in another aspect of our study, the calculation of the density of inherent states by sampling the phase space. As discussed in Sec. III the derivation of $\Omega_{IS}(e_\alpha)$ from $P_{IS}(e_\alpha,T)$ should be independent of the temperature T at which the phase space is sampled to determine $P_{IS}(e_\alpha,T)$. We noticed, however, that this is not exactly true because, at low temperature, the finite time of the simulation could lead to insufficient sampling and some inherent states can be missed. This is more critical when the switch from basin to basin is slow, which is the case for model F. Figure 11 illustrates this point by comparing $\Omega_{IS}(e_\alpha)$ obtained by sampling at two different temperatures for model U and model F: while for model U, even the low temperature calculation can find most of the inherent states, for model F, on the contrary, many states are missed when the temperature is too low. It should, however, be noticed that this is not only a matter of time scale, but also a matter of the accessibility of the states. As discussed above and in Sec. V the energy landscape of model F is rougher than for model U so that some basins of attraction may lie behind high barriers or be very narrow.

At this stage of our study, although model U and model F show rather similar properties, we have exhibited significant differences:

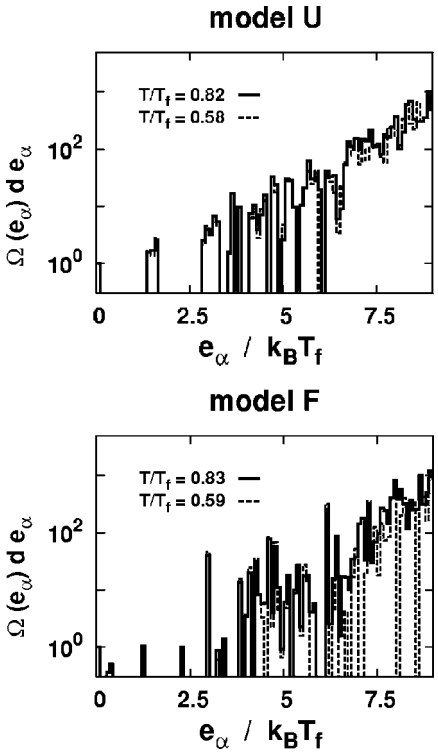

FIG. 11. Density of inherent states of models U and F obtained at different temperatures, showing the difficulty to get a complete sampling of the phase space for model F at low temperature.

(i) Both models show a qualitative change in their fluctuations which occurs in a narrow temperature range around $T_D=0.4T_f$, which appears to be the temperature of the dynamical transition observed in proteins. This transition is sharper and more cooperative for model F than for model U.

(ii) The probability of occupation of the inherent structures at the folding temperature $P_{IS}(e_\alpha,T_f)$ (Fig. 4) clearly shows two humps for model F, which corresponds to the coexistence of two different kinds of states, folded and unfolded. On the contrary for model U, there is an overlap of the probabilities of these two kinds of states, with many intermediate states.

(iii) The fluctuations in model F are much more cooperative, while model U can exhibit very large fluctuations of a single residue in a loop.

(iv) The time scale of the fluctuations between the basin of attraction of one inherent structure to another is about one order of magnitude longer for model F than for model U. The energy landscape is rougher for model F than for model U.

All these points suggest that model F is more appropriate to describe an actual protein, i.e., a minimal frustration has to be introduced in the Gō model to make it sufficiently realistic. This is confirmed by the study of the dynamics of the folding of Sec. V.

V. OUT OF EQUILIBRIUM PROPERTIES: DYNAMICS OF THE FOLDING

In Sec. II we showed that both model U and model F fold in the sense that the cooling of an unfolded initial state at

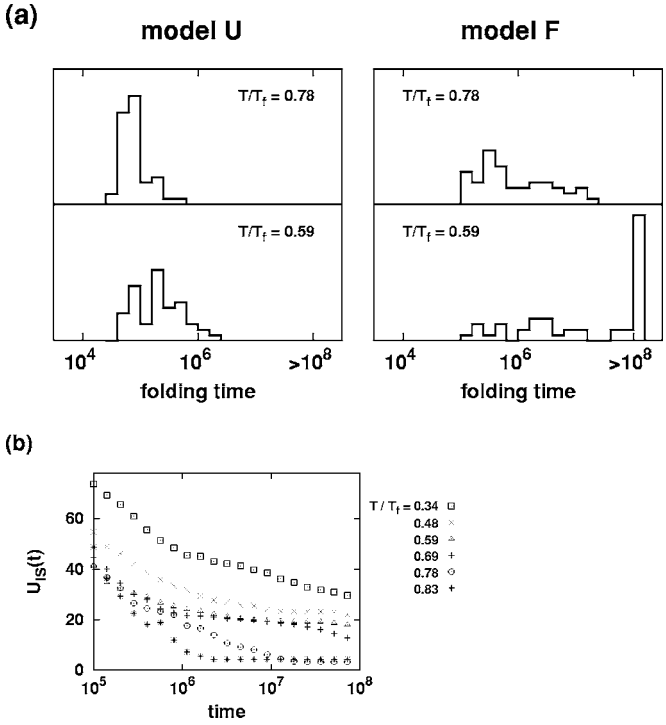

FIG. 12. (a) Histogram of the folding time of model U and model F at two temperatures. The large peak observed at low T for model F occurs because most of the initial states do not reach the native state within the simulation time of 10^8 time units. (b) Time evolution of the folding of model F at different temperatures. The time points t_i selected to calculate $U_{IS}(t_i)$ evolve with a logarithmic scale according to $t_{i+1}=\sqrt{2}t_i$ and for each calculation the time interval Δt_i for the integration is $\Delta t_i=(1-\sqrt{2})t_i$ so that the time domains which are analyzed are adjacent to each other.

high temperature leads to a well defined state at low temperature. However, the dynamics of the folding is very different for the two models. To test this aspect, we have performed a series of out-of-equilibrium studies in which a thermalized initial state which is a random coil at temperature $T_1=1.73T_f$ is suddenly cooled to a temperature T_2 lower than T_f. For each value of T_2, 30–50 MD simulations are performed allowing us to get an ensemble average of the dynamics of the folding. The folding time t_f is the time between the temperature jump and the instant at which the model reaches the basin of attraction of the native state.

Figure 12(a), shows histograms of t_f for the two models at two temperatures T_2, which both are above the temperature of the dynamical transition T_D. When T_2 is sufficiently high ($T_2=0.78T_f$ in Fig. 12), both models fold within the duration of our MD simulations, i.e., in less than 10^8 time units, but the folding time of model F extends to values about two orders of magnitude larger than the largest folding time of model U. For lower T_2 ($T_2=0.59T_f$ in Fig. 12), many of the initial configurations of model F do not achieve folding in 10^8 time units while the largest folding time of model U hardly exceeds 10^6 time units. Although we work in dimensionless variables and our parameters are not fitted to reproduce experimental results, an estimate of the time scales of our simulations can be made. The period for the dynamics of the vibrations of two residues forming a native contacts is about ten time units for our model. Experimentally such modes are observed in a range 10–100 cm^{-1}. Taking an estimate of 20 cm^{-1} this corresponds to a period of 1.7 ps, so that our time unit corresponds approximately to 0.2 ps. The folding time of model U, of the order of 10^5–10^6 time units, corresponds thus to 0.02–0.2 μs, which is very small for an actual protein, while the values of 10^7 to more than 10^8 time units of model F, corresponding to 2 to more than 20 μs, are more realistic for a small protein although they are still small. This is an additional hint that model F is more appropriate than model U to describe a protein in a broad temperature range, and particularly at temperatures well below T_f.

Figure 12(b) completes the view of the dynamics of the folding of model F because it shows the time evolution at different folding temperatures T_2 of $U_{IS}(t,T_2)$ defined by

$$U_{IS}(t,T_2)=\frac{1}{\Delta t}\left\langle \int_t^{t+\Delta t} e_\alpha(t)dt \right\rangle, \tag{23}$$

which measures the average inherent structure energy of the protein in a small time interval Δt. The value of Δt, measured with the time units of the simulation, is always well below the characteristic time of folding of the models. Very slow (logarithmic) relaxations are observed, even for $T_2>T_D$.

Although folding is an out-of-equilibrium process, some of its properties can be obtained from equilibrium studies. It is possible to build an equilibrium free energy profile along the folding pathway, i.e., a one-dimensional picture of the free energy landscape plotted by using the dissimilarity factor D as the reaction coordinate. The most straightforward approach is to sample the phase space by MD simulations, and to compute the dissimilarity D with the native state for each of the sample points. A histogram of the number of events for which the dissimilarity lies in the range $[D,D+dD]$ gives a probability distribution $P(D)$ which can be used to build an effective free energy $F(D)$ defined by $P(D)=\exp[-\beta F(D)]$. This calculation can be made for the actual points of the MD trajectories, which gives a full effective free energy. But for each sample point one can also determine the corresponding inherent structure by quenching, and then compute D for this inherent structure. This gives a probability density $P_{IS}(D)$ and an inherent structure free energy $F_{IS}(D)$. These free energies $F(D)$ and $F_{IS}(D)$ are plotted in Fig. 13 for model F at three temperatures in the vicinity of the folding transition.

The quantities $F(D)$ and $F_{IS}(D)$ have the shape that one could expect: below the folding temperature they show a deep minimum with a low value of D, which corresponds to the folded state, and a second, higher and shallower, minimum corresponding to an unfolded molten globule state. Exactly at the folding temperature, the two minima have equal effective free energies, and above T_f the minimum corresponding to the unfolded state becomes the deepest. The curve $F(D)$ is shifted to higher values of D with respect to $F_{IS}(D)$ because fluctuations around the minima of the free energy landscape bring additional contributions to the dissimilarity with the native state, which are suppressed in $F_{IS}(D)$. Apart from this systematic shift, Fig. 13 shows that

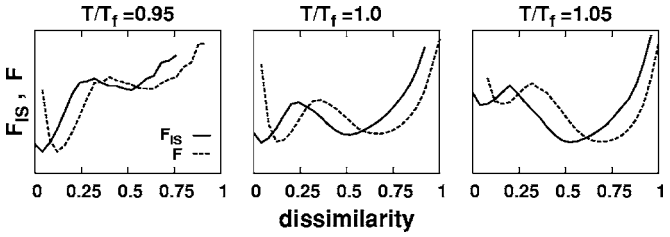

FIG. 13. Effective free energy $F(D)$ (dotted line) and inherent structure free energy $F_{IS}(D)$ (full line) of model F at three temperatures in the vicinity of the folding temperature. The graphs have been shifted along the vertical axis to get the same values for the left minimum.

studying inherent structures can give almost the same results as the points of the MD trajectories because the curves for $F(D)$ and $F_{IS}(D)$ are very similar. This is an additional proof of the interest of the inherent structure analysis. However, the calculation of $F_{IS}(D)$, as it has been done to compute the results shown in Fig. 13, does not exploit the full power of the analysis in the inherent structure landscape because it relies on MD trajectories at the temperature at which we wish to obtain $F_{IS}(D)$. In the vicinity of the folding temperature, this does not introduce any difficulty, but if we wish to obtain $F_{IS}(D)$ at very low temperatures, MD simulations will become highly inefficient to sample the phase space and may suffer from ergodicity problems.

These difficulties are completely avoided if one uses the inherent-structure landscape approach that we introduced in Sec. III A. The idea is to reproduce the method that we used earlier to derive the inherent structure density of states and then the inherent structure partition function, but instead of considering all the states at once, select the states having a dissimilarity factor in a small range $[D,D+dD]$. Thus we define a density of inherent states $\Omega_{IS}(D,e_\alpha)$ and a probability density of inherent structures $P_{IS}(D,e_\alpha,T)$, which are such that

$$\Omega_{IS}(e_\alpha) = \int_0^1 \Omega_{IS}(D,e_\alpha)dD,$$

$$P_{IS}(e_\alpha,T) = \int_0^1 P_{IS}(D,e_\alpha,T)dD. \tag{24}$$

Separating inherent structures in the same D range, Eq. (9) becomes

$$P_{IS}(D,e_\alpha,T)dD = p_0(T)\Omega_{IS}(D,e_\alpha)e^{-\beta e_\alpha}dD, \tag{25}$$

so that the numerical determination of $P_{IS}(D,e_\alpha,T)$ can be used to compute $\Omega_{IS}(D,e_\alpha)$, and then build an inherent structure partition function restricted to inherent states which have a dissimilarity factor in the range $[D,D+dD]$ by

$$Z_{IS}(D,T) = \int \Omega_{IS}(D,e_\alpha)e^{-\beta e_\alpha}de_\alpha. \tag{26}$$

The inherent structure partition function of Eq. (8) can be expressed as

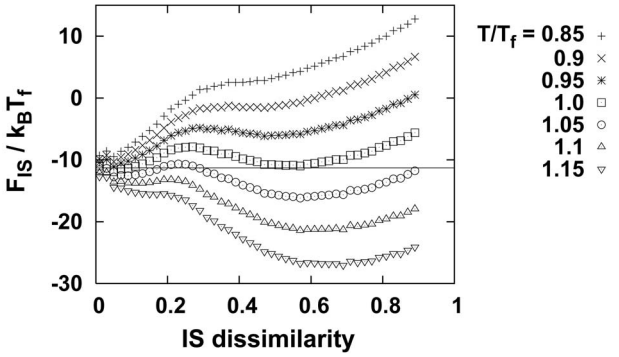

FIG. 14. Inherent structure free energy $F_{IS}(D,T)$ as a function of the dissimilarity factor D of the inherent structures, obtained from the analysis of folding pathways at different temperatures T_2 for model F.

$$Z_{IS}(T) = \int_0^1 Z_{IS}(D,T)dD, \tag{27}$$

from which an inherent structure free energy $F_{IS}(D,T) = -k_B T \ln[Z_{IS}(D,T)]$ can be derived.

Figure 14 shows the inherent structure free energy $F_{IS}(D,T)$ as a function of the dissimilarity factor D of the inherent structures obtained from probability densities $P_{IS}(D,e_\alpha,T)$ computed along folding pathways at different temperatures T_2 for model F. It qualitatively exhibits the shape that we have obtained from the straightforward analysis leading to Fig. 13, but the results are much more reliable, especially in the low temperature range because they do not rely on the sampling of a low temperature MD trajectory. For a folding temperature T_2 slightly below T_f, starting from the unfolded state (large value of D) one first meets a shallow well with a minimum around $D=0.5$, which can be understood as the first stage of the folding, i.e., the evolution towards a molten globule. This stage should be fast since it corresponds to a decrease of the free energy. Then, if D decreases further, the free energy raises again before dropping to the inherent structure free energy of the native state $F_{IS}(D=0,T)$. The maximum corresponds to the transition state. The second stage of the folding involves overcoming the maximum, and thus must be a slow step. In the same figure made for model U, the barrier height from the molten globule state to the transition state is very small. Therefore it is not surprising that model U leads to very fast folding.

Thus Fig. 14 for model F appears to show the behavior that we expect for an actual protein, but a more careful examination reveals, however, a feature that could seem puzzling: for decreasing values of T_2, the barrier E_B from the molten globule state to the transition state decreases very significantly, so that $E_B/(k_B T_2)$ actually decreases in spite of the decrease of T_2. This could lead to the conclusion that the folding time should decrease when T_2 decreases. This is not what Fig. 12 shows for model F since, on the contrary, the histogram shows the presence of many cases which do not succeed to fold in 10^8 time units. Actually the figure shows that, at low values of T_2, the histogram of the folding times for model F splits into two parts: there is a first set of folding

times in the range 10^5-10^7 time units, which is the same range as the range of folding times observed at a higher T_2, and there is a second set group of folding times around 10^8 time units or larger. The first set would be consistent with the properties of the inherent structure free energy of Fig. 14, while the second set is not. This suggests that *two different mechanisms* could contribute to determine the folding time of model F, as it is observed in studies of some proteins made with more complex models [41] or in experiments [42]: some folding pathways can evolve rather quickly to the native state, while others are trapped for a very long time in a kinetic trap. If the density of states $\Omega_{IS}(D,e_\alpha)$ in the free energy basin of the kinetic trap is small, the kinetic trap may not lead to any peculiarity in the inherent structure free energy landscape, but, if it exists in model F, it should appear in studies which explicitly analyze the time evolution of the folding. This is indeed the case as shown by Fig. 15, which displays the probability distribution of the inherent structures $P^t_{IS}(e_\alpha,T_2)$ derived from the sampling of folding trajectories at temperature T_2 in a given time range.

In each case $P^t_{IS}(e_\alpha,T_2)$ for the whole folding trajectory (except for a short initial transient) is compared to $P^t_{IS}(e_\alpha,T_2)$ in the last part of the simulation $(0.7\times10^8<t<1.0\times10^8$ time units), in order to show how the population of the different inherent structures evolves with time. The calculation of $P^t_{IS}(e_\alpha,T_2)$ is made by a statistical averaging over 30 or 50 different folding trajectories.

At the lowest value of T_2, $T_2=0.34T_f$, which is below the dynamical transition, the tendency of the protein to stay frozen in metastable states instead of approaching its ground state is clear. Even at the end of the simulation time, a broad range of inherent states are occupied, although a tendency to evolve towards the ground state is visible because the population of the inherent states with the highest energies is lower in the last part of the simulation than in the figure showing the average over the full simulation. For a slightly higher value of T_2 $(T_2=0.48T_f>T_D)$, the picture changes sharply. At the end of the simulations, in the low e_α range the populated states tend to concentrate towards the ground state although, in this range, $P^t_{IS}(e_\alpha,T_2)$ still looks random. But there is a group of inherent structures with a higher energy $[30<e_\alpha/(k_BT_f)<50]$ which stays populated in the long term. These states correspond to a kinetic trap. They appear as a set of states from which it is difficult to escape to evolve toward a lower energy state. On the time scale of the calculations, the region of the protein phase space corresponding to these states seems to be disconnected from the rest of the phase space. Actually there are phase space trajectories which can leave these states, but the decay of the population of the kinetic trap is, however, very slow, indicating that the trap is separated from the native state by a high barrier. The decay of its population is hardly visible in Fig. 15(b) but it can be detected from the evolution of $U_{IS}(t)$ in Fig. 12. Figure 15(c) shows that the same phenomenon persists at higher temperatures $(T_2=0.69T_f)$. The same range of inherent structure energies stay populated for a very long time, while the inherent structures with lower energies have a distribution which is smoother than at $T_2=0.48T_f$, the population of the low-energy states decreasing approximately exponentially with their energy.

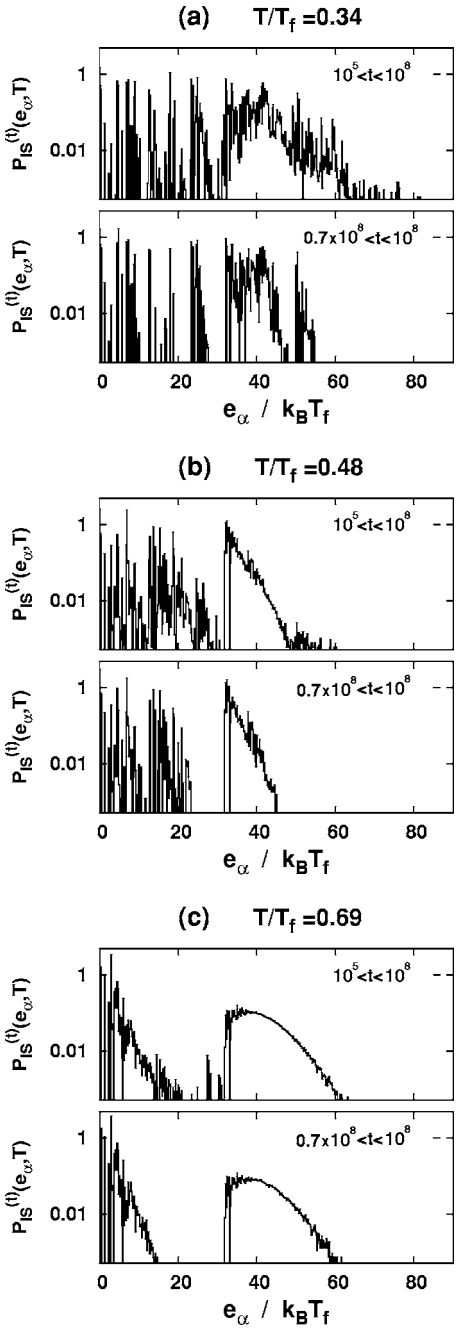

FIG. 15. Probability distribution of the inherent structures $P^t_{IS}(e_\alpha,T_2)$ derived from the sampling of folding trajectories at temperature T_2 in a given time range for various folding temperatures T_2. In each case the top figure shows $P^t_{IS}(e_\alpha,T_2)$ averaged for the whole duration of the simulation 10^8 time units, except for the first 10^5 time units which correspond to a short transient time during which the results might be influenced by the initial condition of each simulation. The lower figure shows $P^t_{IS}(e_\alpha,T_2)$ in the last part of the folding trajectory $(0.7\times10^8<t<1.0\times10^8$ time units). (a) $T_2/T_f=0.34$, (b) $T_2/T_f=0.48$, (c) $T_2/T_f=0.69$. A logarithmic scale is used for the probability distributions in all these figures.

Thus for model F a study of the time evolution of the occupation of the basins of attraction of the structures shows the existence of a set of states in which the protein may stay for a very long time. The roughness of the energy landscape

within this kinetic trap seems to be lower than in the vicinity of the native state because the probability distribution $P^t_{IS}(e_\alpha, T_2)$ for the states in the trap shows an exponential decay with the energy e_α well before the probability distribution of the states near the ground state reaches such a behavior. A plot of the structure of the protein when it is trapped in these long-lived metastable states shows that the α-helix has a "knee" and the β sheet is distorted with respect to the native structure. The presence of such a kinetic trap indicates that model F is rich enough to exhibit the complex properties which are observed in protein folding, contrary to model U which leads to an unrealistically fast folding.

VI. DISCUSSION

In the Introduction, we posed three questions about proteins. The study that we presented in this work puts us in a position to provide some answers.

(i) *How is the energy landscape of a protein*? Using an analysis based on the inherent structures of the protein, i.e., its metastable states, we have shown that it is possible to build an "inherent structure landscape," which is a kind of simplified view of the free energy landscape, which is accessible to numerical computations even for a rather complex protein model. It can be used to build a statistical physics analysis in terms of an inherent structure partition function, from which a reduced thermodynamics can be obtained. This simplified picture cannot describe all aspects of protein dynamics because it ignores the fluctuations inside the basin of attraction of an inherent structure, but it is nevertheless a useful tool to analyze the results of the computations, as shown, for instance, in our investigations of the dynamics of the folding. One of its major interests is that the inherent structure landscape is best obtained from molecular dynamics simulations around the folding temperature, which efficiently sample the full phase space and do not suffer from a possible lack of ergodicity which could appear in low temperature simulations. Once the inherent structure density of states is obtained it can be used at any temperature, including very low ones, to derive the reduced thermodynamics, which provides a lot of data on the protein properties although it does not include the small vibrational motions in the inherent structure basins.

Another interesting point which emerges from our results is the distribution of the energies of the inherent structures which shows an exponential scaling, with two slopes. The exponential scaling itself is known for some models of glasses [38] and it has also been observed in small Lennard-Jones clusters [39]. Its existence for the protein model too may be an indication of a deep similarity between proteins and these systems. For proteins, the existence of two slopes in the scaling [Eq. (12)], which is associated to two different regimes in the distribution of inherent structures [Eqs. (13) and (14)] when the temperature changes from $T<T_u$ to $T>T_s$, may be related to their folding transition. It will be interesting to test other protein models to determine whether these properties of the density of inherent structures states are general features of proteins. To our knowledge such a two-slope feature has not been observed in Lennard-Jones clusters, but this may be because the energy landscape of clusters has been determined in the temperature range of the solid and liquid phases. There are a few inherent structures relevant for the solid state, similarly to the few states that are populated when a protein is its native state or its vicinity. At higher temperatures clusters melt and in the range of the relevant inherent structure energies, an exponential scaling is found in their density of states, similarly to the proteins below the folding transition. Thus it is tempting to make the parallel between the three temperature domains that characterize a protein, (i) low temperature "frozen" state below T_D, (ii) folded state with its multiple conformations, (iii) unfolded state, and the three phases of a cluster (i) solid, (ii) liquid, (iii) gas. This could suggest that, for the high energy inherent structures of a cluster, another exponential scaling could exist, at least when the gas is sufficiently confined to allow enough interactions between the atoms, as for the molten globule of the protein. Of course the analogy is crude and speculative, but we think that the similarity of some general features of the inherent structure landscape of proteins and small atom clusters raises an interesting question: why are proteins special? Of course we know that their structure and the nature of their interactions, with a hierarchy of different interactions, has not much to do with the structure and simple interaction potentials of atomic clusters, but can we detect this specificity of proteins from their equilibrium inherent-structure energy landscape? The scaling of the density of inherent states might be fairly general, but the values of its slopes are probably specific of a protein, in connection with the temperature and sharpness of its folding transition. Therefore understanding the relation between the potentials that connect the residues and the scaling of the inherent state energies appears an interesting open question closely related to protein folding.

(ii) *How does the protein explore its landscape*? The study of the fluctuations of the model versus temperature shows a qualitative change at a temperature $T_D \approx 0.4 T_f$ where the amplitude of the fluctuations start to strongly raise above the linear increase versus temperature which is observed at $T<T_D$. The fluctuations in model F are more cooperative than for model U: the β_N sheet and α helix are more rigid than in the unfrustrated case but the β_C sheet is globally more flexible.

This increase in the fluctuations above a particular temperature is very reminiscent of the dynamical transition which has been observed around 200 K for many proteins [6] and has been the object of a renewed interest in the last few years [2]. In the experiments, this phenomenon is closely related to properties of the water which is in contact with the protein. First, water appears to be necessary for the existence of the dynamical transition, which is not seen in dry samples, and observed at higher temperatures for weakly hydrated samples. Second, the analysis of the experiment points out a clear correlation between a change in the fluctuation of the water as temperature varies and the evolution of the dynamics of the protein embedded in this water [21,43]. The role of water, which is sometimes described as a "lubricant" in protein dynamics because it allows fluctuations which are necessary for biological function, can be understood since water molecules are making hydrogen bonds with some groups of

the protein which, otherwise, would be directly linked by hydrogen bonds. Thus water alters the strength of the interactions within the protein.

Our calculations do not use explicit water but the role of water is nevertheless included through the effective potentials which link the amino acids. However, in our results, the transition cannot come from a change in the properties of water at a particular temperature because the model uses interaction potentials that do not depend on temperature. Nevertheless, the calculations show the same "dynamical transition" as in the experiments. We think that this is an interesting result because it indicates that the complexity of the energy landscape itself is sufficient to lead to such a "transition." This transition is not a true thermodynamic transition but, as shown in Sec. III, it corresponds to a temperature range around which the fluctuations from the basin of an inherent structure to another take over the fluctuations inside a single basin. It should be noticed that such a dynamical transition can be observed in a model *without side chains*. The complexity of the energy landscape of the C_α chain is sufficient to show such a behavior.

Of course in experiments the solvent and the side chains play a role to determine the quantitative properties of the dynamical transition, but our results suggest that they may not be the only driving force. This view could raise an objection: since the dynamical transition is observed approximately at the same temperature for all proteins, it is tempting to conclude that its origin should not be searched within the protein itself but rather within the common factor to all the experiments, the solvent. However, there may be another simple explanation to the fact that T_D is almost the same for all proteins although they may have very different structures. The dynamical transition is the onset of conformational fluctuations. Therefore it is not determined by global properties of the proteins but by local effects which depend on the short range interactions (native contacts in the terminology of the Gō model). These interactions are (on average) the same for all proteins, even if their global shapes can be very different from each other. Thus the dynamical transition can have a component which is intrinsic to proteins and nevertheless occur around the same temperature for many proteins. Actually the views that the dynamical transition is due to the solvent or that it is intrinsic to the protein are not exclusive. We show with a simple model that a dynamical transition intrinsic to the protein can exist, but it is certainly influenced by the properties of the solvent, which are themselves strongly dependent on the protein because hydration water, the water in contact with the protein, is very different from bulk water [44]. The protein and the solvent are actually deeply coupled, leading to a subtle interplay between the solvent and the protein fluctuations.

(iii) *What features are required in a "minimal" protein model*? Although this is a difficult question because the answer depends on the properties of the protein which are of interest, our studies show that a minimally frustrated Gō model is able to exhibit the two main features that characterize a protein, the folding to a well defined structure and the dynamical transition. The ability of the model to lead to folding is not surprising because its design has been tailored for that purpose, since the Gō model favors interactions corresponding to the native state. What is interesting is that this constraint based on the geometry of the native state also leads naturally to a model showing a dynamical transition, although the interaction potentials are not optimized to quantitatively match the potential energy of an actual protein. This points out the crucial role of the geometry of the protein backbone. We have shown here that a significant improvement towards results that match experimental observations on proteins can be obtained by adding dihedral angle frustration. With this additional feature, Gō models, which were used only for folding studies or to analyze small amplitude vibrations, appear to be able to describe conformational fluctuations as well. The addition of the frustration requires a minimal modification and it does not increase significantly the complexity of the simulations, while it makes the model much richer. Model *F*, which includes frustration, is able to show a complex folding pathway including a kinetic trap, and conformational fluctuations which exhibit some cooperativity. This suggests that the mesoscopic model *F* can be used in a reliable way to investigate the thermodynamics and dynamics of proteins at a qualitative level.

ACKNOWLEDGMENTS

We are very grateful to Fumiko Takagi for helpful advice. N.N. would like to thank JSPS-CNRS and MEXT KAKENHI (Grant No. 16740217) for support and M.P. would like to thank Hans Frauenfelder (Los Alamos) and Wolfgang Doster and Fritz Parak (Munich) for helpful discussions and CNRS for the support of a visit to Ibaraki University where this paper was completed.

[1] A. Fersht, *Structure and Mechanism in Protein Science* (Freeman, New York, 1999).
[2] P. W. Fenimore, H. Frauenfelder, B. H. McMahon, and R. D. Young, Proc. Natl. Acad. Sci. U.S.A. **101**, 14408 (2004).
[3] A. L. Tournier and J. C. Smith, Phys. Rev. Lett. **91**, 208106 (2003).
[4] W. T. Franks, D. H. Zhou, B. J. Wylie, B. G. Money, D. T. Graesser, H. L. Frericks, G. Sahota, and C. M. Rienstra, J. Am. Chem. Soc. **127**, 12291 (2005).
[5] W. Doster, S. Cusak, and W. Petry, Nature (London) **337**, 754 (1989).
[6] F. G. Parak, Rep. Prog. Phys. **66**, 103 (2003).
[7] Karen E. S. Tang and Ken A. Dill, J. Biomol. Struct. Dyn. **16**, 397 (1998).
[8] D. Collin, F. Ritort, C. Jarzynski, S. B. Smith, I. Tinoco, and C. Bustamante, Nature (London) **437**, 231 (2005).
[9] J. Wang, Z. Zhang, H. Liu, and Y. Shi, Phys. Rev. E **67**, 061903 (2003).

[10] V. Kurkal-Siebert and J. C. Smith, J. Am. Chem. Soc. **128**, 2356 (2006).
[11] A. E. Garcia and J. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **100**, 13898 (2003).
[12] A. Torcini, R. Livi, and A. Politi, J. Biol. Phys. **27**, 181 (2001).
[13] L. Bongini, R. Livi, A. Politi, and A. Torcini, Phys. Rev. E **68**, 061111 (2003).
[14] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan, Protein Sci. **4**, 561 (1995).
[15] H. Taketomi, Y. Ueda, and N. Gō, Int. J. Pept. Protein Res. **7**, 445 (1975).
[16] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins: Struct., Funct., Genet. **21**, 167 (1995).
[17] L. Bongini, R. Livi, A. Politi, and A. Torcini, Phys. Rev. E **72**, 051929 (2005).
[18] M. A. Miller and D. J. Wales, J. Chem. Phys. **111**, 6610 (1999).
[19] D. A. Evans and D. J. Wales, J. Chem. Phys. **119**, 9947 (2003).
[20] R. Zhou, B. J. Berne, and R. Germain, Proc. Natl. Acad. Sci. U.S.A. **98**, 14931 (2001).
[21] M. M. Teeter, A. Yamano, B. Stec, and U. Mohanty, Proc. Natl. Acad. Sci. U.S.A. **98**, 11242 (2001).
[22] J. Smith, K. Kuczera, and M. Karplus, Proc. Natl. Acad. Sci. U.S.A. **87**, 1601 (1990).
[23] J. A. Hayward and J. C. Smith, Biophys. J. **82**, 1216 (2002).
[24] C. Baysal and A. R. Atilgan, Biophys. J. **88**, 1570 (2005).
[25] A. L. Lee and A. J. Wand, Nature (London) **411**, 501 (2001).
[26] E. Kussel, J. Shimada, and E. L. Shakhnovich, Proteins: Struct., Funct., Genet. **52**, 303 (2003).
[27] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **84**, 7524 (1987).
[28] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins: Struct., Funct., Genet. **21**, 167 (1995).
[29] A. M. Gronenborn, D. R. Filpula, N. Z. Essig, A. Achari, M. Whitlow, P. T. Wingfield, and G. M. Clore, Science **253**, 657 (1991).
[30] W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33 (1996). VMD was developed by the Theoretical and Computational Biophysics Group in the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana-Champaign.
[31] J. Karanicolas and C. L. Brooks, J. Mol. Biol. **334**, 309 (2003).
[32] C. Clementi, H. Nymeyer, and J. N. Onuchic, J. Mol. Biol. **298**, 937 (2000).
[33] J. D. Honeycutt and D. Thirumalai, Biopolymers **32**, 695 (1992).
[34] D. P. Yee and K. A. Dill, Protein Sci. **2**, 884 (1993).
[35] A. Sali, E. Shakhnovich, and M. Karplus, Nature (London) **369**, 248 (1994).
[36] F. H. Stillinger and T. A. Weber, Phys. Rev. A **25**, 978 (1982).
[37] H. Frauenfelder, F. Parak, and R. D. Young, Annu. Rev. Biophys. Biophys. Chem. **17**, 451 (1988).
[38] A. J. Bray and M. A. Moore, J. Phys. C **14**, 1313 (1981).
[39] J. P. K. Doye and D. J. Wales, J. Chem. Phys. **102**, 9659 (1995).
[40] H. Frauenfelder, G. A. Petsko, and D. Tsernogolou, Nature (London) **280**, 558 (1979).
[41] J. Shimada, E. L. Kussell, and E. I. Shakhnovich, J. Mol. Biol. **308**, 79 (2001).
[42] W. Hoyer, K. Ramm, and A. Pluckthun, Biophys. Chem. **96**, 273 (2002).
[43] P. W. Fenimore, H. Frauenfelder, B. H. McMahon, and R. D. Young, Physica A **351**, 1 (2005).
[44] *Hydration Processes in Biology*, edited by M.-C. Bellissent-Funel (IOS Press, Amsterdam, 1999).