

Critical examination of the inherent-structure-landscape analysis of two-state folding proteins

Johannnes-Geert Hagmann, Naoko Nakagawa, Michel Peyrard

▶ **To cite this version:**

Johannnes-Geert Hagmann, Naoko Nakagawa, Michel Peyrard. Critical examination of the inherent-structure-landscape analysis of two-state folding proteins. Physical Review E : Statistical, Nonlinear, and Soft Matter Physics, 2009, 80 (6), pp.061907-1-11. 10.1103/PhysRevE.80.061907 . ensl-00426353v2

HAL Id: ensl-00426353

https://ens-lyon.hal.science/ensl-00426353v2

Submitted on 22 Dec 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Critical examination of the inherent-structure-landscape analysis of two-state folding proteins

Johannes-Geert Hagmann,[1] Naoko Nakagawa,[2] and Michel Peyrard[1]

[1] *Université de Lyon; Ecole Normale Supérieure de Lyon, Laboratoire de Physique, CNRS, 46 Allée d'Italie, 69364 Lyon, France*
[2] *College of Science, Ibaraki University, Mito, Ibaraki 310-8512, Japan*

(Dated: December 22, 2009)

Recent studies attracted the attention on the inherent structure landscape (ISL) approach as a reduced description of proteins allowing to map their full thermodynamic properties. However, the analysis has been so far limited to a single topology of a two-state folding protein, and the simplifying assumptions of the method have not been examined. In this work, we construct the thermodynamics of four two-state folding proteins of different sizes and secondary structure by MD simulations using the ISL method, and critically examine possible limitations of the method. Our results show that the ISL approach correctly describes the thermodynamics function, such as the specific heat, on a qualitative level. Using both analytical and numerical methods, we show that some quantitative limitations cannot be overcome with enhanced sampling or the inclusion of harmonic corrections.

PACS numbers: 87.15.A-, 87.15Cc, 05.70.-a

I. INTRODUCTION

The biological and physical properties of proteins are compelling for many reasons. While just a small amount of the nowadays hundreds of thousands known protein sequences are experimentally characterized, the variety of their functions is overwhelming. Though the structure has been resolved only for a subset of these sequences, the number of stable folds that are expressed in nature is seemingly small compared to the number of sequences. The relationship between fold and function is far from obvious, and examples such as intrinsically unstructured proteins and multi-functional folds resist simple schemes for classification. The question what really makes a protein functional hence needs to be addressed in the context of its specific biological environment.

From a physical point of view, an attempt to find some unifying concepts for the interpretation of dynamics and thermodynamics is the description of proteins in term of energy landscapes[2], in which the evolution of the system is related to the dynamics on a high-dimensional rugged energy surface. The existence of local minima, connected by saddles of different barrier height and rank, leads to a distribution of timescales that are reflected in the dynamics of the proteins.

Although the energy landscape provides a reduced description, the complex interactions in proteins and their interaction with the environment, which involve multi-body interactions and subtle effects of charges, make its complete characterization neither experimentally nor theoretically conceivable. This situation is somewhat reminiscent of other complex systems such as glasses which, though being more homogeneous systems, share the property of displaying a high-dimensional landscape leading to complex dynamics. The energy landscape picture is useful for a qualitative analysis of protein properties, but for quantitative studies, an exhaustive sampling and and the full characterization are practically infeasible for this high-dimensional representation. Therefore, in order to obtain comprehensive quantitative predictions on generic protein properties from the information on the landscape, the picture needs to be simplified. A recent work[1] has shown that a reduced description of the energy landscape, originally devised for the analysis of super-cooled liquids by Stillinger and Weber[3], can successfully capture the essential thermodynamic aspects of folding in the context of a simplified protein model. In particular, it was shown that the density of states constructed from the local minima of the energy landscape, called inherent structures, can be used to compute the most important thermodynamic observables. This finding is important because it provides a general scheme for theoretical studies of protein thermodynamics, showing how the relevant information can be quantitatively accessed from its imprint on the potential energy surface. The approach has consequently been used in the context of studies of the folding properties of a β-barrel forming protein[4], the construction of the free energy landscape by mechanical unfolding[5], and the network of native contacts[6]. Other earlier works including inherent structure analysis but not necessarily seeking to characterize the full thermodynamics of folding can be found in [7, 8].

However the validity of an analysis based on the inherent structure landscape (ISL) must be critically examined because the method involves a fundamental assumption which could be questioned: The vibrational free energy within the basin of attraction of an inherent structure is assumed to be independent of the basin. A recent study [6] tried to go beyond this approximation by assuming that the vibrational free energy can depend on the energy of the inherent structures. Still, the question is subtle as we show in the present work that, even when the vibrational free energy

depends on the inherent structure energy, the derivation of thermodynamic quantities such as the specific heat from the ISL can be validly carried without any change in the procedure. Therefore an understanding of the limits of the ISL approach requires a deeper analysis. This is the aim of the present work.

We proceed in two steps. In a first step (Section III), after briefly summarizing the ISL formalism and introducing the protein model used in Section II, we test the validity of the ISL approach by comparing its results to the data obtained from equilibrium molecular dynamics for a set of structures. We selected four previously unstudied two-state folding proteins of varying size and secondary structure elements. In a second step (Section IV), we critically revisit the major hypotheses of the ISL approach, as well as its practical limitations, such as the sampling of the phase space, and suggest routes for improvements. Finally, we summarize our findings in Section V and give an outlook on possible future studies that stem from our results.

II. METHODS

A. Inherent structure analysis

In this subsection, we briefly review the major results of [1, 9] on obtaining reduced thermodynamics from an analysis of the inherent structures.

The method is general and not bound to a specific protein model, provided the phase space of the protein can be explored by molecular dynamics, and that the energies of the visited states can be calculated. From simulations at fixed temperature close to the folding temperature T_f, which insures that the system evolves in a large part of the configuration space, the local potential minima, labelled by α_i, are determined by conjugate gradient minimizations performed at fixed frequency along a molecular dynamics trajectory. The global minimum α_0 is defined as the reference ground state with zero energy. Let $\{x_i\},\ i = 1, 2, ..., 3N$, denote the $3N$ Cartesian coordinates of the N-particle system, and $V(\{x_i\})$ its potential energy function. The probability to find a particular minimum α_i with potential energy e_{α_i} can be written as

$$p(\alpha_i, T) \ = \ \frac{1}{Z(T)} \int_{B(\alpha_i)} d^{3N}x \ e^{-\beta V(\{x_i\})} = \frac{1}{Z(T)} e^{-\beta e_{\alpha_i}} \int_{B(\alpha_i)} d^{3N}x \ e^{-\beta \Delta V_{\alpha_i}(\{x_i\})} \quad , \tag{1}$$

where $\Delta V_{\alpha_i} = V - e_{\alpha_i}$, Z denotes the configurational part of the partition function and $B(\alpha_i)$ is the basin of attraction of the minimum α_i. With the definition

$$e^{-\beta F_v(\alpha_i, T)} \ := \ \int_{B(\alpha_i)} d^{3N}x \ e^{-\beta \Delta V_{\alpha_i}(\{x_i\})} \quad , \tag{2}$$

the unknown integral over the complex landscape of the basin of attraction $B(\alpha_i)$ is summed in an free-energy like function $F_v(\alpha_i, T)$ which in principle depends both on the nature of the basin and temperature. Notice that although we use the index v like "vibrational", $F_v(\alpha_i, T)$ is obtained from the full nonlinear integral over $B(\alpha_i)$, and not from its harmonic approximation.

The inherent structure landscape approach makes two key assumptions[1] which enable to considerably reduce the amount of information needed on the landscape while keeping its most important features.

- (A1) The function $F_v(\alpha, T)$ for two minima α_1, α_2 that are distinct but close in energy, $e_{\alpha_1} \approx e_{\alpha_2}$, is the same for both minima: $F_v(\alpha_1, T) \approx F_v(\alpha_2, T)$. Consequently, $F_v(\alpha, T) \approx F_v(e_\alpha, T)$.
- (A2) The function $F_v(e_\alpha, T)$ does not vary significantly for different minima, i.e. $F_v(e_\alpha, T) \approx F_v(T)$.

Both assumptions were discussed in [9]. In section IV, we show that assumption (A2) can actually be relaxed to the less strong form $\beta F_v(e_\alpha, T) \approx f_v(e_\alpha) + \beta F_v(0, T)$ while most calculations remain feasible and some of the thermodynamic variables unchanged. With these assumptions, the contribution from the function F_v factorizes in the numerator and denominator of (1) so that it can be eliminated to give

$$p(\alpha_i, T) \ = \ \frac{1}{Z_{IS}(T)} e^{-\beta e_{\alpha_i}} \quad , \quad Z_{IS} = \sum_{\alpha=\alpha_0}^{\alpha_{max}} e^{-\beta e_\alpha} \quad , \tag{3}$$

where the sum in the partition function includes all inherent structures found from the global minimum α_0 to the minimum α_{max} having the highest energy. Here, the energy scale is shifted such that the energy of the global

minimum α_0 is zero. Introducing an energy density function for the inherent structures $\Omega_{IS}(e) = \sum_{\alpha=\alpha_0}^{\alpha_{max}} \delta(e - e_\alpha)$, the probability to find a minimum in the interval $[e_\alpha, e_\alpha + de_\alpha]$ at temperature T is

$$P_{IS}(e_\alpha, T)de_\alpha \;=\; \frac{1}{Z_{IS}}\Omega_{IS}(e_\alpha)e^{-\beta e_\alpha}\; de_\alpha \quad , \quad Z_{IS} = \int_{e_{\alpha_0}}^{e_{\alpha_{max}}} de_\alpha\; \Omega_{IS}(e_\alpha)e^{-\beta e_\alpha} \quad . \tag{4}$$

For the model used in this work, the low energy minima are in practice sparsely separated in energy. As the ground state is isolated, one obtains $Z_{IS}(T) = 1/p_0(T)$ with the probability of the ground state $p_0(T)$, so that the inherent structure density of states can be estimated from the probability to be in the basin of attraction of a minimum in a fixed temperature simulation at temperature T_{MD} as

$$\Omega_{IS}(e_\alpha) \;=\; \frac{e^{\beta_{MD} e_\alpha}}{p_0(T_{MD})} P_{IS}(e_\alpha, T_{MD}) \quad . \tag{5}$$

Though above we have chosen a continuous notation to simplify the equations, it should be noted that the density built from an estimate of the probability density function of sampled minima for the present model in practice always comprises discrete and continuous parts which can be integrated separately. Once the inherent structure density of states is known, one can compute the inherent structure partition function Z_{IS} from (4), from which all thermodynamics functions can be derived, including the free energy F_{IS} and the internal energy U_{IS}. Given that most states of the system can be sampled close to the folding temperature, it is sufficient to simulate the system at a single temperature T_{MD} to construct the inherent structure landscape, in contrast to the full thermodynamics where one needs to sample different ranges of temperatures. Therefore, the ISL approach can be computationally very efficient. In the following, we will restrict ourselves to the computation of specific heat $C_{V,IS}$ being a quantity of fundamental importance in a physical system, as it is sensitive to fluctuations and, for instance, shows a clear signature of phase transitions. It can be deduced from numerical derivatives of the partition function Z_{IS} through

$$C_V = T\left(\frac{\partial S}{\partial T}\right)_V \quad , \tag{6}$$

and hence

$$C_{V,IS} = T\left(\frac{\partial^2(\beta^{-1}\log(Z_{IS})}{\partial T^2}\right)_V \quad . \tag{7}$$

B. Model and selected proteins

Since our goal is to analyze the validity of the ISL approach and not to derive quantitative data for a particular protein, we decided to choose a simplified model, which allows the sampling of phase space at a reasonable computational cost. However the model must be rich enough to properly describe the complex features of its physics, and should be able to distinguish between proteins which differ, for instance, in their secondary structure. We use frustrated off-lattice Gō-models identical to the ones introduced in [1] because they provide a good compromise between all-atom simulations and simplified models that do not fully describe the geometry of a protein. These models provide a representation with a single particle per residue centered at the location of each C_α-atom. For details on the model and the parameters, we refer to [1, 9] and a brief review in the appendix A. Although the validity of such models to provide a faithful representation of protein folding is a recurrent subject of debate, off-lattice Gō-models have been successfully used to study folding kinetics [10] and the mechanical resistance of proteins [11]. From a physical point of the view, despite a strong bias towards the ground state, these models have a complex energy landscape with a large number of local minima well suited for the analysis in terms of inherent structures.

As the results of the ISL approach depend on the density of states of the inherent structures, for a reliable test of the method it is important to examine examples which could differ in their properties, i.e. to investigate proteins of different size and structure. To test the inherent structure approach beyond the previously employed immunoglobulin (IG) binding domain of protein G (2GB1), we selected four two-state folding proteins of varying size and folds from the PDB database[12]: the trp-cage mini-protein construct (1L2Y, 20 residues, α-helical), the ww domain FPB28 (1E0L, 37 residues, β-sheets), the src-SH3 domain (1SRL, 56 residues, β-sheets) and ubiquitin (1UBQ, 76 residues, $\alpha - \beta$-fold). The motivation for these choices is discussed in Section III for each protein (see insets of figures 1-4 for structural representations of these four proteins drawn with pyMol[13]). The positions of the $C\alpha$-atoms of the PDB files is chosen as a reference for the construction of the Gō-model. In the case of NMR resolved structures, the first structure is selected as the reference. The native contacts of the model were established according to the

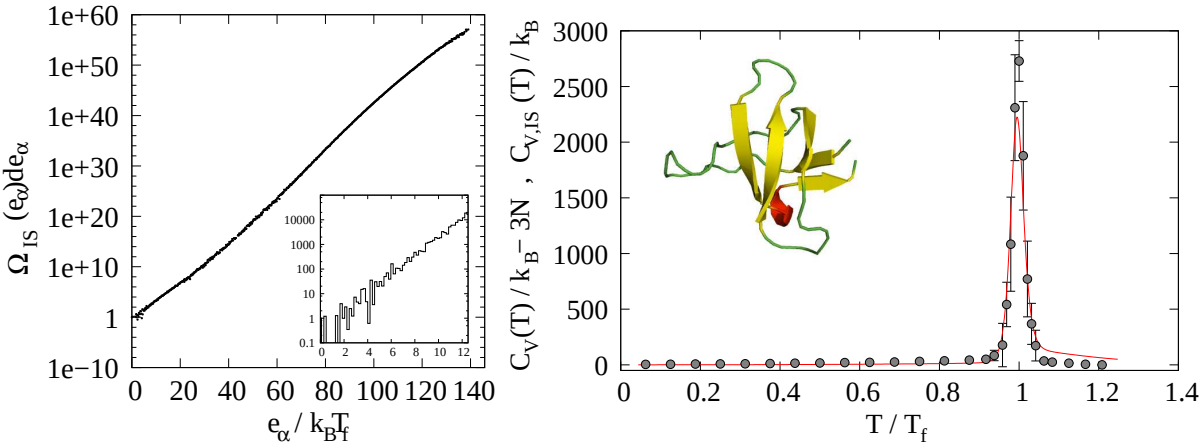


FIG. 1: Results for src-SH3. *Left:* Inherent structure density of states $\Omega_{IS}(e_\alpha)$. The inset shows a close up on the low-energy range. The size of the energy bins for the density estimate is $\Delta\ E = 0.2\ k_B T_f$. *Right:* Comparison of the specific heat from equilibrium trajectories $C_V(T)$ (points), from which the specific heat of a harmonic oscillator in $3N$ dimensions has been subtracted, and $C_{V,IS}(T)$ from inherent structure analysis (solid red line); see text for further explanations.

distances between atoms belonging to different residues. A native contact is formed if the shortest distance belonging to atoms of two different residues is smaller than $5.5\mathring{A}$. The number of native contacts according to this criterion are: $N_{nat} = 91$ for the ww domain, $N_{nat} = 225$ for ubiquitin, $N_{nat} = 216$ for src-SH3 and $N_{nat} = 36$ for trp-cage. This definition is simple and includes some arbitrariness. There exist other methods for probing contacts between side-chains, e.g. by invoking the van der Waals radii of residue atoms and solvent molecules[14]. Though using the latter method preserves the main structure of the contact map, it leads to quantitative differences in the the number and location of contacts along the sequence. Consequently, one can expect that the topology of the energy surface and key thermodynamic properties such as the folding temperature are also altered when the definition of the contact map is varied. For the purpose of the present study which does not attempt to give a quantitative description of side chain contacts, and focusses on global properties of the landscape rather than its detailed relation to the network of contacts, the cutoff-based approach is acceptable.
Molecular dynamics simulations were performed using the Brooks-Brünger-Karplus algorithm [15] with a time-step of $dt = 0.1$ and a friction constant of $\gamma = 0.01, 0.025$ (all units in this section are dimensionless, see [1] for details). To ensure equilibration, the system was thermalized starting from the native state (PDB coordinates) for $t = 2 \cdot 10^5$. The simulation time for a single temperature point and a single initial condition was $t = 2 \cdot 10^7$, and the data obtained for both inherent structure sampling (fixed temperature) and thermodynamic sampling (variable temperature) were averaged over various initial sets of velocities. Minimization was performed using the conjugate gradient method with the Polak-Ribière algorithm. To estimate the vibrational free energy at the minimum, mass weighted normal mode analysis was performed using LAPACK diagonalization routines. The second-order derivatives of the potential energy function at the minimum were calculated by numerical differentiation of the analytical first-order derivatives.

III. REDUCED AND FULL THERMODYNAMICS OF A SET OF MODEL PROTEINS

In this section, the validity of the ISL approach is tested by comparing the equilibrium thermodynamics deduced from molecular dynamics simulations to the reduced thermodynamics from inherent structure sampling. As discussed in Section II, we evaluate the specific heat C_V as a function of temperature as a representative example of the thermodynamic observables.

A. src-SH3

The src-SH3 domain was chosen since it has the same number of residues as IG binding domain of protein G studied in [1, 9], but contrasts to the latter in terms of structure. The src-SH3 domain is mostly composed of β-sheets and does not contain an α-helical-secondary structure element (only five residues form a small right-handed helix segment). The inherent structure density of states, shown on the left-hand side of figure 1, was obtained from various simulations close to the folding temperature ($T_{MD} \approx T_f$) and built from ≈ 72000 minima according to (5). After computing an

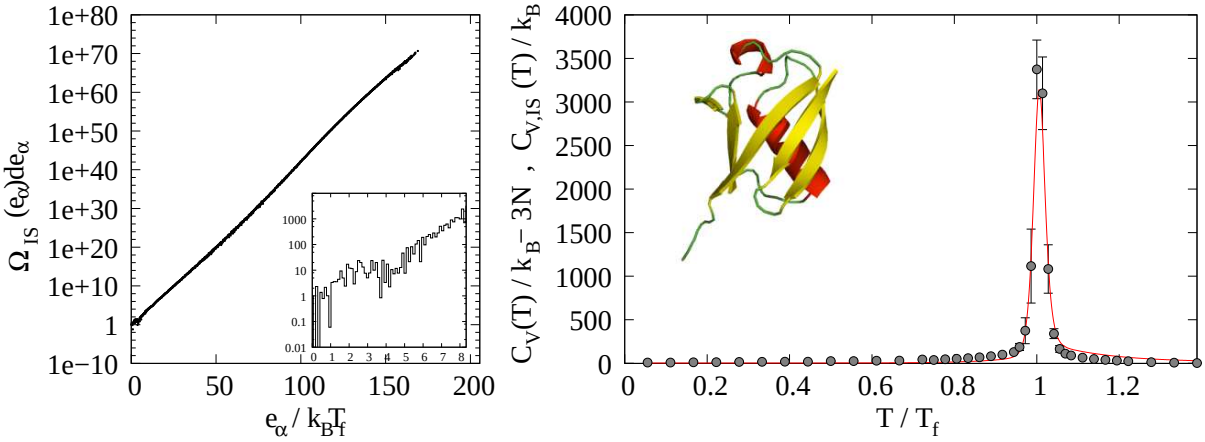


FIG. 2: Results for ubiquitin, see caption of figure 1 for annotation.The size of the energy bins for the density estimate is $\Delta\ E = 0.24\ k_B T_f$.

energy histogram using 1000 bins to yield an estimate of the inherent structure probability density, energy bins with only a single count have been discarded from the analysis to avoid a bias that could be introduced by insufficiently sampled isolated minima. The right hand side of figure 1 shows a comparison between the temperature dependence of the specific heat calculated from inherent structures, $C_{V,IS}(T)$, and the temperature dependence of the specific heat calculated from equilibrium molecular dynamics simulations at variable temperature $C_V(T)$. The equilibrium thermodynamics has been determined by averaging the results of 10 initial conditions per temperature step except for points close to the transition where the results of 20 initial conditions have been used. Despite this averaging, the variance, indicated by error-bars on the y-axis, is large in the vicinity of the folding transition as the waiting time for a transition to occur becomes comparable to the simulation time. For a harmonic system, $C_V(T) = N_{dof} k_B/2$ where N_{dof} is the number of degrees of freedom. At low temperatures $T \ll T_f$, harmonic contributions are dominating, and the difference between $C_V(T)$ and $C_{V,IS}(T)$ is approximately $3Nk_B$, which is subtracted from $C_V(T)$ in figure 1.

Figure 1 shows that the ISL approach is able to capture the main features of the thermodynamics of the Gō-model of the src-SH3 domain. The value of the folding temperature is correctly determined by the ISL approach, but $C_{V,IS}$ underestimates the maximum by more than 20% if the highest point of $C_V(T)$ is selected as a reference. This discrepancy at the maximum had previously also been observed for the inherent structure analysis of the IG binding domain of protein G [1]. On the other hand, towards higher temperatures, $C_{V,IS}(T)$ decays slower than $C_V(T)$.

B. ubiquitin

Ubiquitin, with its 76 amino-acids, is a protein with a fairly rich secondary structure since it contains a α helix and 5 β sheets. Similarly to the src-SH3 domain, the ubiquitin Gō-model presents a sharp folding transition associated with a large peak in the specific heat (see right-hand side of figure 2). The specific heat $C_V(T)$ was estimated from averages on 8 initial conditions per temperature step, and $\Omega_{IS}(e_\alpha)$ was obtained using ≈ 79000 minima from several independent trajectories close to the folding temperature using a histogram of 2000 bins. The agreement between the full thermodynamics and the ISL approach is better than for src-SH3, though similar trends of discrepancies can be observed.

C. ww domain

To contrast with sharp two-state transitions of protein G (56 residues), src-SH3 (56 residues) and ubiquitin (76 residues), we selected smaller structures, the ww domain (37 residues) and the trp-cage (20 residues), to examine the performance of the inherent structure approach for less structured proteins, showing a broader transition. For such small protein domains, the validity of the Gō-model can be questioned as the model is built from the geometrical structure of the folded state. For small molecules the discrimination between folded and unfolded states becomes subtle due to fluctuations covering a large part of the accessible configurational space in a broad range of temperatures. The point in selecting these structures is not to asses the validity of the Gō-model itself, but to test the ISL approach in very stringent cases to highlight possible limitations. The density of states $\Omega_{IS}(e_\alpha)$ was obtained from ≈ 79000

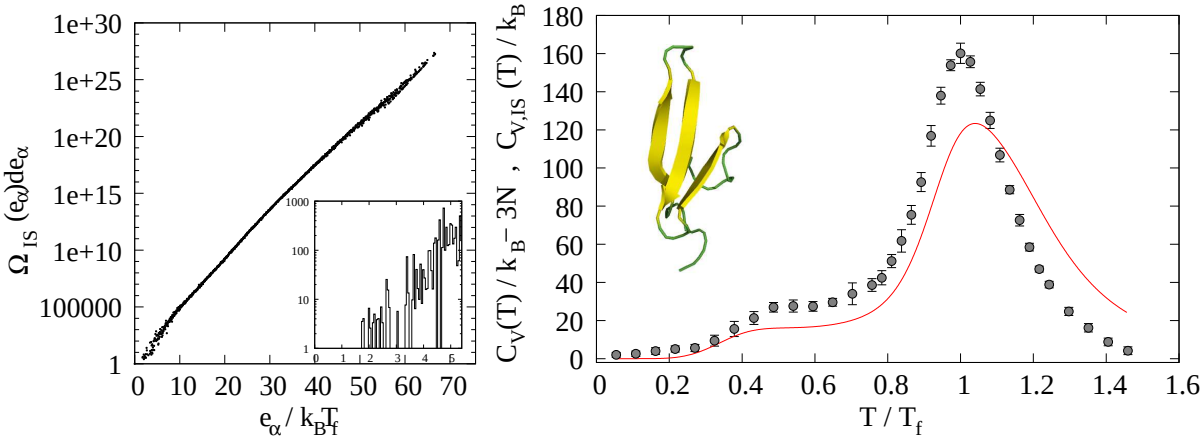


FIG. 3: Results for the ww domain, see caption of figure 1 for annotation. The size of the energy bins for the density estimate is $\Delta\ E = 0.05\ k_B T_f$.

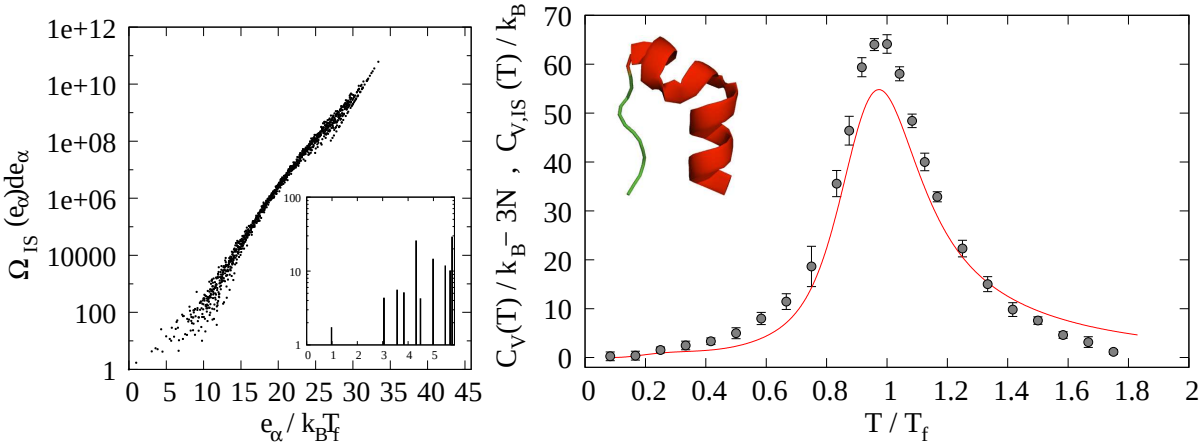


FIG. 4: Results for trp-cage, see caption of figure 1 for annotation. The size of the energy bins for the density estimate is $\Delta\ E = 0.02\ k_B T_f$.

minima from several independent trajectories slightly above the folding transition (see figure 3) using a histogram of 2000 bins. In contrast to the two previous cases, the histogram of minima does not show a clear separation of basins of local minima associated to the folded/unfolded state. We observe a difference in the apparent shape of the density of states (left-hand side of figure 3), which is globally concave in contrast to the convex densities obtained for src-SH3 and ubiquitin. The same shape was also found for protein G in [1], for which the relation between the concave shape and the two-hump structure of $P_{IS}(e_\alpha, T)$ was discussed in the vicinity of the folding temperature. Moreover, by comparing the insets of the left-hand side of figs. 1-3, one observes that the low energy range of $\Omega_{IS}(e_\alpha)$ is more discrete, and states tend to lie less densely packed.

The temperature dependence of $C_V(T)$ was obtained by averaging over 12 initial conditions. An interesting feature of the curve is the shoulder in the low temperature range which indicates a partially unfolded structure associated to the breaking of a small number of contacts. Comparing the results of $C_V(T)$ and $C_{V,IS}(T)$, it is apparent that though the specific heat reconstructed from inherent structure thermodynamics correctly captures the global shape of $C_V(T)$, including the existence of the shoulder, important deviations can be observed. Similarly to the cases analyzed above, $C_{V,IS}(T)$ underestimates $C_V(T)$ at lower temperatures while giving an overestimation at high temperatures. In contrast to the results for larger proteins, we also observe a significant shift of the transition temperature.

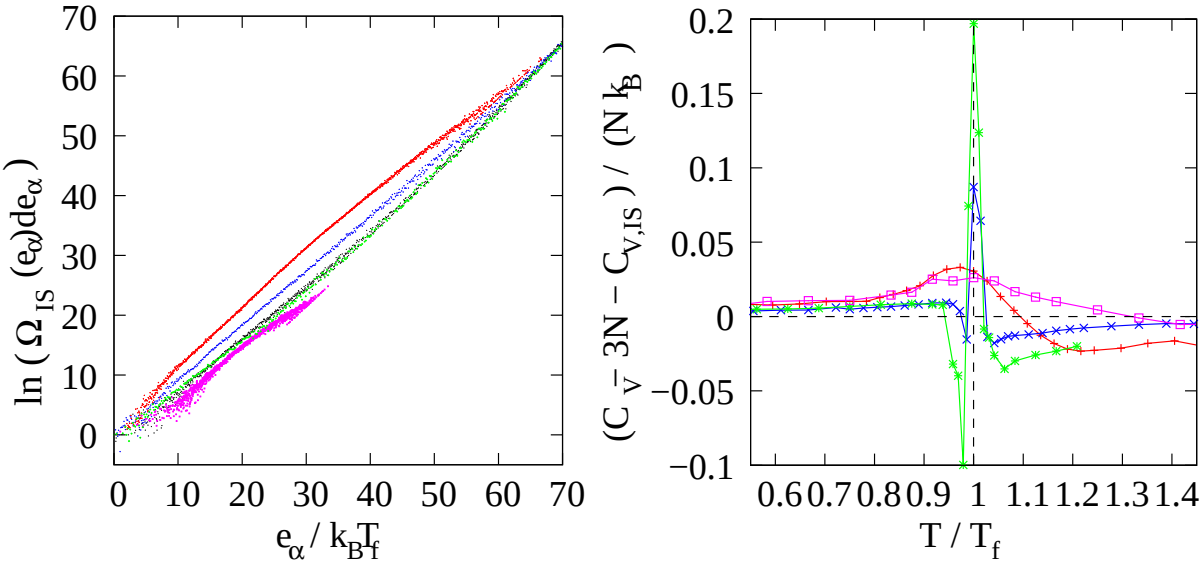


FIG. 5: *Left:* Comparison between the different inherent structure density of states $\Omega_{IS}(e_\alpha)$ including the data for the IG binding domain of protein G from [1]; the color coding is: protein G (black), src-SH3 (green), ww domain (red), ubiquitin (blue), trp-cage (magenta). *Right:* Deviation between the specific heat obtained from full thermodynamics and inherent structure analysis.

D. trp-cage

With only 20 residues, the last protein fragment studied in this series is also the smallest, and mainly consists of a single α-helix. Its inherent structure density of states $\Omega_{IS}(e_\alpha)$ was estimated from ≈ 90000 minima sampled from several independent trajectories close to the folding temperature using a histogram of 2000 bins. For such a small system, the low lying energy states are largely separated from each other, and the resulting density of states presents large gaps in a relatively broad range of energies. The continuum representation assumed for the inherent structure landscape is certainly questionable in such a case. The equilibrium thermodynamics were constructed from averages on 10 runs over different initial conditions. It is interesting to notice that the value of the specific heat computed with reduced thermodynamics is still fairly close to the actual specific heat, although we notice again that the peak is underestimated and a high temperature tail is observed as in the previous cases. In contrast to the results for the ww domain, the folding temperature of the trp-cage protein domain is correctly found by the analysis of inherent structures.

E. Discussion

Our studies of four proteins, combined with the earlier results on protein G [1, 9], allow us to describe some trends in the inherent structure analysis of Gō-model proteins.

For the density of states given by (5), a general exponential dependence, $\Omega_{IS}(e_\alpha) \propto \exp(-e_\alpha/k_BT_0)$ is observed for all proteins, with slightly different slopes for the low energy states, corresponding to states occupied in the folded configuration, and for the high energy states, occupied in the unfolded configuration. The value of T_0 associated to the low energy range is a good estimate of the folding temperature, as previously reported for the case of protein G [1]. Figure 5 (left-hand panel), which compares the density of states of the inherent structures for the four proteins shows that, when being presented in reduced units as a function of e_α/k_BT_f, the functional form of these densities is highly similar. For the large proteins that we studied, src-SH3 and ubiquitin (as well as protein G), the slope is slightly larger in the high energy range than in the low energy range. The converse is true for the small protein domains ww and trp-cage. A formal calculation of the reduced specific heat $C_{V,IS}$ from a bi-exponential density of inherent structure energies shows that this property is related to the sharpness of the folding transition. A density of states that is curved downwards for the energies associated to unfolded configurations leads to the broad folding transition expected for small protein domains.

The calculation of the specific heat $C_{V,IS}$ shows that the ISL approach is able to determine the specific heat of a protein with reasonable accuracy, including the overall shape of the folding transition. To obtain such an agreement

the ground state probability $p_0(T_{MD})$ must be sufficiently well sampled to ensure that the density of inherent states is correctly normalized.

However, our studies of several proteins shows that limitations exist since systematic deviations from the full thermodynamics are apparent. At low temperatures, and up to the transition temperature, $C_{V,IS}(T)$ underestimates the specific heat. The peak of $C_{V,IS}(T)$ is less pronounced than expected from the equilibrium trajectories, and tends to broaden towards higher temperatures ($T > T_f$) where $C_{V,IS}$ is larger than $C_V(T)$. These deviation are shown in a comparative illustration in the right-hand side panel of figure 5 for the four proteins. Our data also reveals that the Gō model, with its strong bias towards the native state, does not yield qualitatively differences depending on the secondary structure of the protein under consideration.

Owing to the results found for various proteins which show systematic deviations from the results of equilibrium thermodynamics, it is important to examine the assumptions made in approximating the inherent structure of states, which we do in the following section.

IV. THE LIMITATIONS OF THE ISL APPROACH

A. Local normal mode analysis

In Section II we introduced the major assumptions (A1) and (A2) of the ISL approach. The derivation of thermodynamic quantities such as the specific heat is carried out as if the free energy contribution within a basin of attraction, defined by (2), did not depend on the particular inherent structure α_i. It is difficult to test this assumption as it would in principle require the determination of the complete basin on the energy landscape, including the calculation of all the saddle points that determine the frontier of the basin as well as the shape of the basin within this frontier. Still, one can at least compute $F_v(\alpha)$ *in the harmonic approximation,* as done also in [6].

Assuming that the contribution to the integral (2) can be approximated by local normal modes in the vicinity of the energy minimum, the effective free energy can be written as

$$\begin{aligned} \beta F_{NMA}(\alpha, T) &= \sum_{q=1}^{3N-6} \log\left(\frac{\hbar\omega_q(\alpha)}{k_B T}\right) = \sum_{q=1}^{3N-6} \log(\omega_q(\alpha)/\omega_q(0)) - \sum_{q=1}^{3N-6} \log(k_B T/\hbar\omega_q(0)) \\ &= f_{NMA}(\alpha) + \beta F_{NMA}(0, T) \quad , \end{aligned} \tag{8}$$

where $\omega_q(0)$ are the normal mode frequencies at the ground state $\alpha = 0$. This expression allows us to calculate a harmonic approximation to $F_v(\alpha, T)$ by calculating the normal modes for each minimum that we sample, and subsequently summing their different contributions according to (8). Figure 6 shows the α-dependent part $f_{NMA}(\alpha)$ as a function of the energy minimum for the two examples of src-SH3 domain (left) and ww domain (right). The minima were obtained along a single trajectory close to the folding temperature. The distribution of the minima along the energy axis reflects the character of the probability distribution function for the two proteins, one being divided into two basins (src-SH3 domain), the other being a single distribution (ww domain).

As a first important observation, we note that the variation of $f_{NMA}(\alpha)$ with e_α cannot be assumed to be negligible as assumed in (A2), according to which F_{NMA} should be approximately constant in e_α. Both proteins show the same trend toward a decreasing effective free energy with increasing e_α. For the src-SH3 domain, a nonlinear dependence can be observed in the high energy range in agreement with previously reported results on protein G [6].

A second point to be noticed is that, for a given e_α, a distribution of values of $f_{NMA}(\alpha)$ can be found (variance on the y-axis in figure 6). While such a variance (or fluctuation) could be attributed to the limited numerical accuracy of the normal modes at high energies, this is also true for the low energy part for which the numerical scheme provides accurate results, as can be checked on the vanishing six lowest eigenmodes. Consequently, it appears that the first approximation (A1), i.e. $F(\alpha, T) \approx F(e_\alpha, T)$ does not hold as assumed, leaving the possibility that the observed deviations in section III could be caused by this simplification.

B. Free energy correction

The results of the previous subsection seem to challenge the validity of the ISL approach performed under the assumptions (A1) and (A2). At a first glance, the main problem seems to arise from the approximation (A2) that $F_v(\alpha, T)$ does not depend on the particular basin considered, which is obviously untrue. On the other hand, although there is clearly a variance associated to F_v for the same energy range e_α in disagreement with approximation (A1),

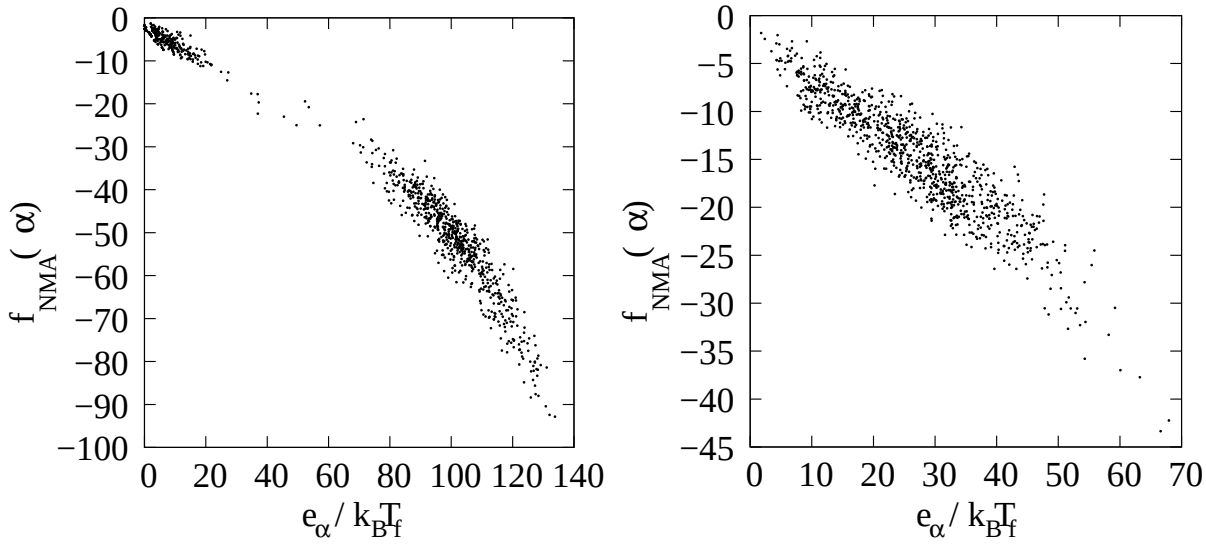


FIG. 6: The frequency dependent component $\sum_{q=1}^{3N-6} \log(\omega_q)$ of the vibrational free energy βF_{NMA} in the harmonic approximation, calculated from a sample of 1000 minima. *Left:* Src-SH3 domain. *Right:* ww domain. An arbitrary offset was added to shift the origin of the ordinate.

one can observe a general evolution of $F_v(\alpha, T)$ with e_α, which suggests that approximating $F_v(\alpha, T)$ by $F_v(e_\alpha, T)$ may be acceptable. However, we show below that, if the free energy within a basin can be written as a sum

$$\beta F_v(\alpha, T) \;=\; f_v(e_\alpha) + \beta F_v(0, T)\,, \tag{9}$$

with $f_v(e_\alpha = 0) := 0$ at the ground state, the calculation of $C_{V,IS}$ *can be carried out without any change,* so that approximation (A2) appearing to be to particularly bad at a first glance may not be the decisive one. It should be noticed that, as discussed in the previous section, the property (9) is verified if the motion in each basin of attraction can be described by a combination of harmonic vibrations.
Starting again from (1), and proceeding as in section II A, we can eliminate the part of $F_v(\alpha, T)$ that depends on temperature only in the expression for the partition function and the probability distribution function, keeping only the α-dependent part. One gets

$$Z(T) \;=\; e^{-\beta F_v(0,T)} \int \Omega_{IS}(e_\alpha) e^{-\beta e_\alpha} e^{-f_v(e_\alpha)} \quad de_\alpha \qquad , \tag{10}$$

$$Z_{IS}(T) \;=\; \int \Omega_{IS}(e_\alpha) e^{-\beta e_\alpha} e^{-f_v(e_\alpha)} de_\alpha \quad . \tag{11}$$

The principle of the calculation is to compute $Z_{IS}(T)$ form a *measurement* of the probability density function $P_{IS}(T)$ estimated from MD simulations at a given temperature T_{MD}. This can be achieved through the intermediate calculation of a density of states of inherent structures $\Omega_{IS}(e_\alpha)$, which is temperature independent and from which $Z_{IS}(T)$ can be obtained at all temperatures. In equation (11), we notice that the inclusion of the term $e^{-f_v(e_\alpha)}$ is equivalent in definition an "effective" density of states $\Omega_{IS}(e_\alpha)e^{-f_v(e_\alpha)}$ from which the classical thermodynamic expression can be derived as shown below. In the calculation of section II A, the density of states is given by equation (4). This density $\Omega_{IS}^{(0)}$ and all other observable calculated in section II A will henceforth denote with and index (0). In the new scheme including the α-dependent part of the free energy in the harmonic approximation, the probability density $P_{IS}(e_\alpha, T)$ becomes

$$P_{IS}(e_\alpha, T) \;=\; \frac{1}{Z_{IS}(T)} \Omega_{IS}(e_\alpha) e^{-\beta e_\alpha} e^{-f_v(e_\alpha)} \quad . \tag{12}$$

yielding the inherent structure density

$$\Omega_{IS}(e_\alpha) \;=\; \frac{P_{IS}(e_\alpha, T_{MD})}{p_0(T_{MD})} e^{\beta_{MD} e_\alpha} e^{f_v(e_\alpha)} \quad . \tag{13}$$

The latter expression shows that if the variation of $F_v(\alpha, T)$ with e_α cannot be ignored, the previously derived density of states $\Omega^{(0)}_{IS}(e_\alpha)$ is not the correct one. The two are related by

$$\Omega_{IS}(e_\alpha) \;=\; \Omega^{(0)}_{IS}(e_\alpha) e^{f_v(e_\alpha)} \quad . \tag{14}$$

A similar result was also reported in [6] which considered the particular case of a piecewise linear dependence on e_α. Though the densities differ, substituting (14) in equations (11) or (4), we immediately have $Z_{IS}(T) = Z^0_{IS}(T)$, and the inherent structure observables such as $U_{IS} = \langle e_\alpha \rangle$ and $C_{V,IS} = \left(\langle e^2_\alpha \rangle - \langle e_\alpha \rangle^2 \right) / k_B T^2$ are unchanged, i.e. $U_{IS} = U^{(0)}_{IS}$ and $C_{V,IS} = C^{(0)}_{V,IS}$ even when $f_v \neq 0$.
As a consequence, using equation (10), the full free energy $F(T)$ of the protein can be written as

$$\begin{aligned} F(T) \;&=\; -k_B T \log(Z(T)) = -k_B T \log(Z_{IS}(T)) + F_v(0,T) \\ &=\; -k_B T \log(Z^{(0)}_{IS}(T)) + F_v(0,T) = F^{(0)}_{IS} + F_v(0,T) \quad . \end{aligned} \tag{15}$$

Therefore, in the ISL formalism, taking into account the variation of $F_v(\alpha, T)$ as in equation (9) does not alter the free energy and cannot be expected to be at the origin of the quantitative differences between C_V derived from the ISL formalism and the full numerical results presented in section III. We conclude that the origin of these discrepancies is likely to be found in the non-separability between the α and the T dependency within the basins. Such an non-separability can be expected as soon as the anharmonicity of the different basins is taken into account. This is certainly relevant for proteins, in particular as the denaturation involves frequent transitions between basins of different shape and volume associated to the semi-rigid folded and the highly flexible unfolded state.
When $F_v(\alpha, T)$ cannot be separated into α- and T-dependent contributions to simplify the calculation of the partition function, the remaining possible approximation that can tackle the computational difficulty would be the saddle point approximation of the free energy [3]. This approximation is acceptable in the thermodynamic limit for large systems, but cannot be justified in the present problem as the number of particles involved is still small and the interactions between particles are heterogeneous. While the harmonic approximation seems to be invalid for the present problem which involves large conformational changes due to the denaturation transition, results on supercooled liquids [17] indicate that the correction of the heterogeneity of the basins at low temperatures is small and the decoupling approximation of vibrational and inherent structure contributions appears to be possible at least in these temperature regimes.

C. Effect of limited sampling efficiency

The main practical difficulty of the ISL method comes from the need to properly sample all the inherent structures in order to get a meaningful density of states $\Omega_{IS}(e_\alpha)$ in all energy ranges. In the present scheme of inherent structure sampling, a single temperature is selected to simulate the dynamics of the protein for a finite period of time. The choice of a temperature close to the folding temperature is natural as the protein samples both the folded and the unfolded configurations at this temperature. There are however two questions that have to be answered: *i)* How long should we follow a protein MD trajectory to get a sufficient sampling? *ii)* Is it possible to combine data from simulations at a few different temperatures instead of keeping T fixed?
Let us first analyze the effect of the sampling time. It should be chosen long enough to cover the slowest intrinsic timescale of the system, and various trajectories with different initial conditions or realizations of the thermostat should be used to ensure that the order of events does not alter the shape of the distribution. An estimate of the time range that sampling must cover is provided by the folding/unfolding time of the protein, to guarantee that the protein explores both configurational subspaces. A possible check of the choice of the simulation time is to compute the specific heat with an increasing number of samples, and stop when the improvement brought by additional samples is negligible. Still, as computer time is limited, it cannot be ensured that all relevant states are sufficiently well sampled to yield a converged probability density. In particular, the high energy minima are sampled only with low probability, such that the high energy cut-off in the density of states is likely to be underestimated in finite time sampling. In this section, we analyze the impact of this cutoff on a model density to see how the inherent structure specific heat is possibly affected.

To analyze the effect of the sampling independently of a particular case, let us assume a "model" inherent structure density of states taken as a single exponential

$$\Omega^{(0)}_{IS}(e_\alpha) = \begin{cases} e^{e_\alpha / a} & 0 \leq e_\alpha \leq e_{max} \\ 0 & e_{max} < e_\alpha \end{cases} \quad ,$$

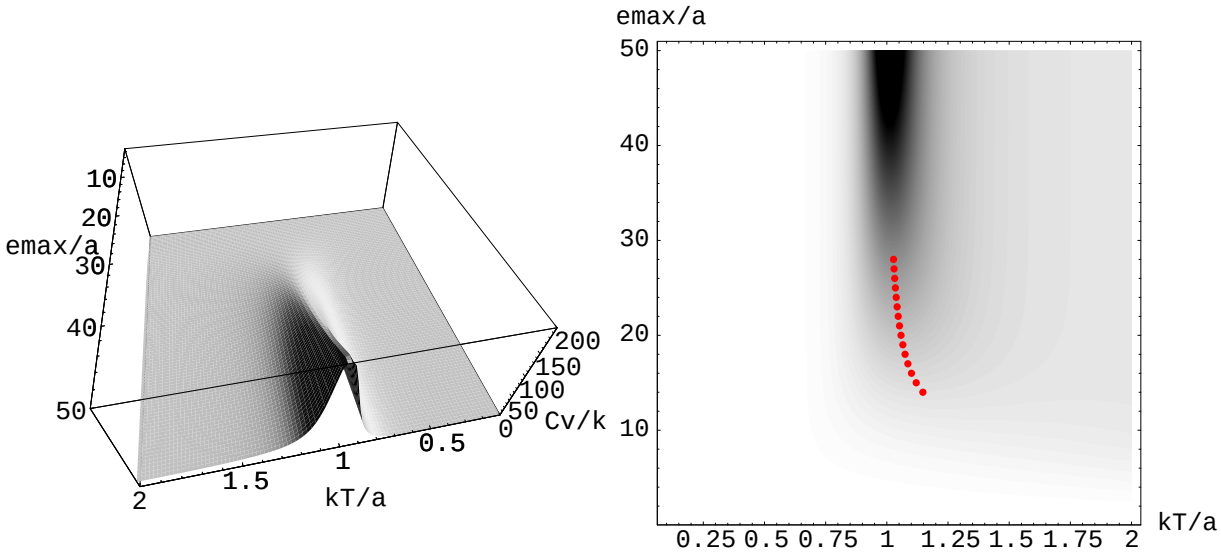


FIG. 7: *Left:* The specific heat $C_{V,IS}$ as a function of temperature and the cutoff parameter e_{max}. *Right:* Location of the maxima of $C_{V,IS}$ for different values of the cutoff e_{max}.

similar to the shape of the densities that can be found in limited ranges of energies for the numerical results in section III. The partition function than can be readily calculated as

$$Z_{IS}^{(0)} = \int_0^{e_{max}} de_\alpha \; \Omega_{IS}(e_\alpha) \; e^{-\beta e_\alpha} = \frac{e^{e_{max}(a^{-1}-\beta)} - 1}{a^{-1} - \beta} \quad . \tag{16}$$

Likewise, we can calculate the first two moments $\langle e_\alpha \rangle$ and $\langle e_\alpha^2 \rangle$ to find the specific heat as a function of the temperature, the parameter a and cutoff in energy e_{max}

$$C_{V,IS}^{(0)}(T; a, e_{max}) = \frac{\langle e_\alpha^2 \rangle - \langle e_\alpha \rangle^2}{k_B T^2} \quad , \tag{17}$$

and analyze the result graphically as a function of the energy cutoff e_{max} in figure 7. As can be inferred from left-hand side of figure 7, $\max_T \left[C_{V,IS}(T; a, e_{max}) \right]$ increases with higher cut-off e_{max}. Using symbolical computation [18], we can further inspect the result to find the local maxima of the the specific heat for fixed cutoff e_{max}. On the right-hand side of figure 7, we observe that a lower cutoff in the density of states shifts the maximum of the specific heat towards higher temperatures. In addition to the shift, the curve becomes broader and the value at the maximum decreases. This situation is similar to the physical scenario when protein folding is altered by confinement (see e.g. [19]). The high energy states disappear from the density of states as the system is prohibited to explore these by external forces. For the present purpose of the inherent structure analysis, although derived for a highly idealized model density, the results indicate that insufficient sampling at high energies can significantly alter the global shape of the transition. We have checked that these conclusions remain unchanged for a piecewise constant density of states. For instance, for the ww domain, one finds $e_{max}/a \approx 60$ which is higher than the range of values for which a shift of the maximum can be expected from figure 7. Consequently, the origin of this shift cannot be attributed to inefficient sampling in the high energy range.

Because the high energy minima are less frequently visited, it is tempting to try to sample the minima from a high temperature molecular dynamics trajectory. On the other hand, as the method relies on the probability to occupy the ground state which determines $1/Z_{IS}(T)$, it is also necessary to properly sample the ground state, i.e. to select a simulation temperature which is below T_f. To reconcile these two exclusive conditions, one solution is to combine results sampled at two different temperatures to calculate $\Omega_{IS}(e_\alpha)$, which should be temperature independent. This is possible because, according to (3)

$$\frac{P_{IS}(e_\alpha, T_1)}{P_{IS}(e_\alpha, T_2)} = \frac{Z_{IS}(T_2)}{Z_{IS}(T_1)} e^{-(\beta_1 - \beta_2) e_\alpha} \quad , \tag{18}$$

with $\beta_{1,2} = 1/(k_B T_{1,2})$, so that the ratio of $Z_{IS}(T_2)/Z_{IS}(T_1)$ can be calculated from the probabilities to occupy a basin at temperatures T_1 and T_2. A molecular dynamics trajectory obtained at a temperature $T_1 < T_f$ can be used to

determine $Z_{IS}(T_1)$ from the probability to occupy the ground state, and subsequently a second simulation at a higher temperature T_2 can sample high energy basins more efficiently. For all basins which are properly sampled in both molecular dynamics runs, the ratio $Z_{IS}(T_2)/Z_{IS}(T_1)$ can be evaluated with (18). Although it should not depend on the particular basin that was used for its calculation, this ratio actually fluctuates around a mean value which can be used to determine $Z_{IS}(T_2)$ from $Z_{IS}(T_1)$. Then (3), applied at the higher temperature T_2, can be used to compute $\Omega_{IS}(e_\alpha)$ in the high energy range. Moreover, in the intermediate energy range, the basins are properly sampled by the two trajectories at temperatures T_1 and T_2, which gives two ways to evaluate $\Omega_{IS}(e_\alpha)$ for those basins, and thus provides a way to check the consistency of the method.

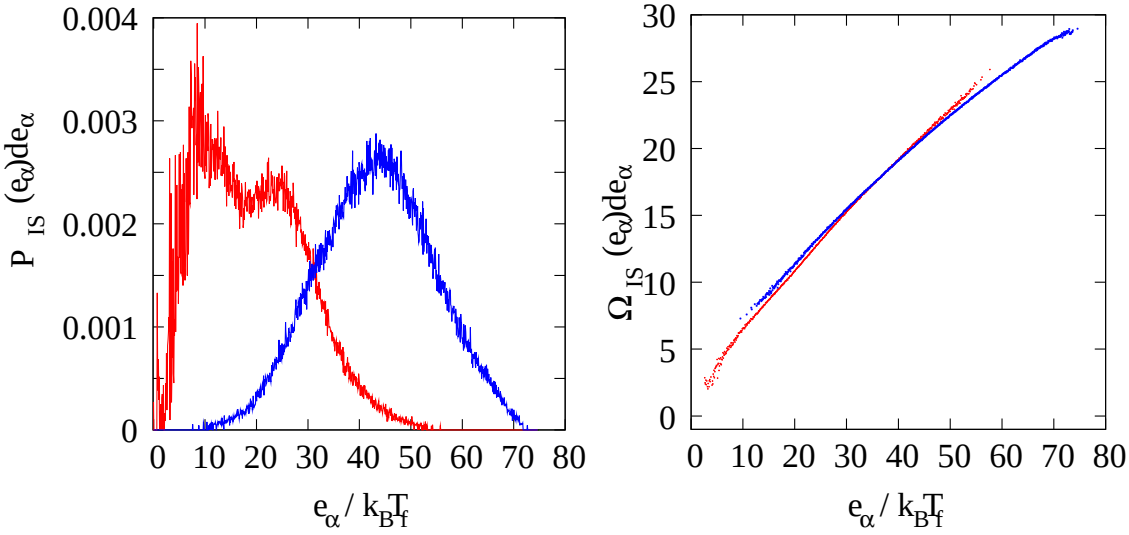


FIG. 8: Results from the sampling of inherent structures of the ww domain at two different temperatures $T_1 = 1.03\ T_f$ ($\approx$ 79000 minima, red) and $T_2 = 1.41\ T_f$ ($\approx$ 18000 minima, blue) *Left:* Probability of occupation of the inherent structures versus e_α. *Right:* Density of inherent states energy (in logarithmic scale), calculated from the simulation at T_1 (red) and calculated at T_2 using the ratio $Z_{IS}(T_2)/Z_{IS}(T_1)$.

Figure 8 shows the results for the ww domain from different molecular dynamics simulations, simulated respectively at $T_1/T_f = 1.03$ and $T_2/T_f = 1.41$. For the simulations at T_2, the ground state is not sampled in the finite interval of time of the simulations, but Ω_{IS} can nevertheless be obtained through the evaluation of $Z_{IS}(T_2)$ deduced from $Z_{IS}(T_1) = 1/p_0(T_1)$ for all values of e_α for which the histograms of the left-hand side of figure 8 overlap. The right-hand side of figure 8 shows that the values of the density of inherent state energies computed from data at T_1 and T_2 are in rather good agreement in the whole energy range where the basins are sampled at the two temperatures. There is however a discrepancy between the two results, with a systematic deviation towards lower values of $\Omega_{IS}(e_\alpha)$ for high inherent state energies when the density of states is calculated with data sampled at high temperatures. This is counterintuitive as one could tend to believe that sampling the basins at low temperatures would, on the contrary, underestimate the density of basins at high energy. This systematic deviation, which has been observed in all our calculations could point out a limitation of the ISL method as presently applied, i.e. with a calculation which is only valid if $\beta F_v(\alpha, T)$ is the sum of a term that depends on e_α and a term that depends on temperature (see (9)). When the specific heat is calculated with the value of $\Omega_{IS}(e_\alpha)$ extended in the high energies range with this method, the function is still very close to the value obtained with only the data at temperature T_1 and the agreement with the numerical value of $C_V(T)$ is not improved. This shows that the discrepancy between the exact results and those deduced from the ISL approach are not due to technical difficulties such as an insufficient sampling in some energy range, but that they are rather inherent to the method itself together with its assumptions.

V. DISCUSSION

We applied the inherent structure landscape (ISL) approach to four different proteins of varying size and secondary structure elements using a coarse-grained off-lattice protein model, and calculated their inherent structure density of states. Using these densities, we derived the specific heat from the reduced inherent structure thermodynamics, and compared it to the value obtained from equilibrium molecular dynamics as a function of temperature. Our results show that the ISL approach can correctly capture the shape of the temperature dependence of the specific

heat, including some characteristic features such as the hump observed at $T \approx 0.4T_f$ for the ww domain. This is remarkable since the result is deduced from molecular dynamics simulations at a single temperature, close to T_f, and nevertheless predicts the main features of the specific heat in a large temperature range, including low temperatures for which very long simulations would be necessary to reach exhaustive sampling of the phase space. This shows that *many features of protein thermodynamics are encoded in the inherent structure landscape.* Still, the approach is not perfect as we observed quantitative differences between $C_{V,IS}(T)$ and $C_V(T)$ which are particularly significant for small protein domains. The deviations show a systematic trend, the specific heat being underestimated below T_f and overestimated above. This lead us to reexamine the approximations that enter the construction of the reduced thermodynamics from inherent structures for the model that we considered. The first approximation assumes that the correction to the density of states due to the structure of the basin associated to a minimum $f_v(\alpha)$ depends on the energy level of the minimum only, and not of the individual minimum. An evaluation of $f_v(\alpha)$ using a harmonic approximation based on local normal modes (section IV A) shows that this is only approximately correct. For a given e_α, the values of $f_v(\alpha)$ are actually distributed around an average value, with fluctuations that grow for higher values of e_α and that appear to be larger for small proteins. This could explain some of the discrepancies between the ISL results and the equilibrium data. Moreover, the calculation of $f_v(\alpha)$ indicates that the second assumption that the correction can be considered to be α-independent is certainly not valid. However, we showed that if $\beta F_v(\alpha)$ splits into a temperature-dependent and an α-dependent part, which is the case in the harmonic approximation, most of the thermodynamic results deduced from a direct application of the ISL approach are not affected. This is true in particular for the the specific heat. In view of the persisting quantitative differences between reduced inherent structure and equilibrium thermodynamics, we therefore conclude that the correction of the free energy in term of a harmonic approximation is not sufficient. It is likely that the nonlinear terms in the free energy associated to a basin cannot be ignored, and play a significant role. This is not surprising because, especially in the high temperature range, the protein fluctuates by exploring many basins, and consequently cannot be assumed to be well described by a harmonic approximation.

In future studies, it would be useful to analyze the role of the structure of the full basin on the thermodynamic results beyond the approximation by local normal modes around the minima. This is a true challenge owing to the complexity of the energy landscape. A starting point for such a study might be the examination of the distribution of first rank saddle points associated to the different minima on the potential energy surface. A second aspect which is suggested by the present work is to apply the ISL approach for protein folding in the context of more complex energy landscapes that arise in more realistic potential energy functions. The results on the small proteins analyzed in this study show that the global separation of the probability density into two basins associated to folded and unfolded states is not a necessary requirement to construct the reduced thermodynamics. It appears therefore likely that the formalism remains useful in cases where the energy landscape is less biased towards the ground state than in the Gō-model. An application of the ISL method to other protein models therefore appears to be desirable and promising.

Acknowledgements: We thank the anonymous referee for valuable suggestions. J-G.H. acknowledges financial support from the Collège Doctoral Franco-Japonais and the Région Rhône-Alpes. The simulations were partially performed at the Pôle Scientifique de Modélisation Numérique (PSMN) in Lyon.

APPENDIX A: MODEL HAMILTONIAN AND PARAMETERS

In this work, we analyze the properties of an off-lattice Gō-type [1, 9] in which the smallest building unit is a single amino acid. Effective interactions between amino acids are based on the reference positions of the C_α-carbons of each residue. These interactions are "color-blind" in the sense that they do not distinguish between the type of amino acids. The potential energy of the system comprises five terms:

$$\begin{aligned} V \;=\; & \sum_{i=1}^{N-1} \frac{1}{2} K_h (d_i - d_{i0})^2 \;+\; \sum_{i=1}^{N-2} \frac{1}{2} K_b (\theta_i - \theta_{i0})^2 \\ & + \sum_{i>j-3}^{native} \epsilon \left[5 \left(\frac{r_{0ij}}{r_{ij}} \right)^{12} - 6 \left(\frac{r_{0ij}}{r_{ij}} \right)^{10} \right] + \sum_{i>j-3}^{non-native} \epsilon \left(\frac{C}{r_{ij}} \right)^{12} \\ & + \sum_{i=1}^{N-3} K_d \left(1 - cos \left(2\phi_i - \frac{\pi}{2} \right) \right) \end{aligned} \tag{A1}$$

Here, r_{ij} denotes the Euclidean distance between residues i, j. The $3N-6$ degrees of freedom of the system are most conveniently expressed via internal coordinates: $N-1$ Euclidean bond distances d_i along the backbone, $N-2$ bond

angles formed between two consecutive bond directions, and $N-3$ dihedral angles measured between the normal vectors of planes spanned by atoms $(i, i+1, i+2)$ and $(i+1, i+2, i+3)$. Bond distance and bond angle interactions are modelled through harmonic forces with coupling strengths K_h and K_a respectively. The zero indices indicate that the quantities (angles, distances) are evaluated in the reference state, i.e., the position from the NMR/crystallographic structure. The dihedral angle potential does not assume a minimum in the reference position defined by the experimentally resolved structure: it favors angles close to $\pi/4$ and $3\pi/4$ irrespective of the secondary structure element (helix, sheet, turn) the amino acid belongs to. While such values can be found in α-helices, they statistically do not appear largely in other secondary structure elements. As a consequence, the reference state defined by an NMR or crystallographic structure can give rise to a competition between the helix for which the reference state is close to its minimum energy, and other parts which experience forces due to the angular constraint. The constraint was introduced as a source of additional "frustration" affecting the dynamics and thermodynamics of the model towards a more realistic representation [1]. Attractive non-bonded native interactions are modelled by steep (12,10)-Lennard-Jones potentials accounting for non-local contacts between residues due to side chains. The database of contacts is established as described in section II. In addition, non-native repulsive interactions are added to those residues which do not form a contact and which lie at least four residues apart. The dimensionless parameters used in this study are [1]: $K_b = 200.0$, $K_a = 40.0$, $K_d = 0.3$, $\epsilon = 0.18$, $C = 4.0$.

[1] N. Nakagawa and M. Peyrard; Proc. Natl. Acad. Sci. USA **103**, 5279 (2006)
[2] P.W. Fenimore, H. Frauenfelder, B.H. McMahon and R.D. Young; Proc. Nat. Acad. Sci. USA **101**, 14408 (2004)
[3] F.H. Stillinger and T.A. Weber; Phys. Rev. A **25**, 978 (1982)
[4] J. Kim and T. Keyes; J. Phys. Chem. B **111**, 2647 (2007)
[5] A. Imparato, S. Luccioli, A. Torcini; Phys. Rev. Lett. **99**, 168101 (2007)
[6] D.M. Ming, M. Anghel and M.E. Wall; Phys. Rev. E **77**, 021902 (2008)
[7] Z. Guo and D. Thirumalai; J. Mol. Biol. **263**, 323 (1996)
[8] A. Baumketner, J.-E. Shea and Y. Hiwatari; Phys. Rev. E **67**, 011912 (2003)
[9] N. Nakagawa and M. Peyrard; Phys. Rev. E **74**, 041916 (2006)
[10] J. Karanicolas and C.L. Brooks III; Proc. Nat. Acad. Sci. USA **100**, 3954 (2003)
[11] D.K. West, D.J. Brockwell, P.D. Olmsted. S.E. Radford and E. Paci; Biophys. J. **90**, 287 (2006)
[12] http://www.rcsb.org
[13] W.L. DeLano (2002); *The PyMOL Molecular Graphics System.*; http://www.pymol.sourceforge.org
[14] V. Sobolev *et al.*; Bioinformatics **15**, 327 (1999)
[15] A. Brünger, C.L. Brooks III and M. Karplus; Chem. Phys. Lett. **105**, 495 (1984)
[16] N. Nakagawa; Phys. Rev. Lett. **98**, 128104 (2007)
[17] F. Sciortino, W. Kob and P. Tartaglia; Phys. Rev. Lett. **83**, 3214 (1999)
[18] Mathematica 5.0, Wolfram Inc.
[19] N. Rathore, T.A. Knotts, J.J. de Pablo; Biophys. J. **90**, 1767 (2006)