The protein folding transition state: Insights from kinetics and thermodynamics

Rui D. M. Travasso,[1,a)] Patrícia F. N. Faísca,[2,b)] and Antonio Rey[3,c)]
[1]*Departamento de Física, Centro de Física Computacional, Universidade de Coimbra, Coimbra 3004-516, Portugal*
[2]*Centro de Física da Matéria Condensada, Universidade de Lisboa, Av. Prof. Gama Pinto 2, Lisboa 1649-003, Portugal*
[3]*Departamento de Química Física I, Facultad de Ciencias Químicas, Universidad Complutense, Madrid E-28040, Spain*

(Received 17 June 2010; accepted 12 August 2010; published online 23 September 2010)

We perform extensive lattice Monte Carlo simulations of protein folding to construct and compare the equilibrium and the kinetic transition state ensembles of a model protein that folds to the native state with two-state kinetics. The kinetic definition of the transition state is based on the folding probability analysis method, and therefore on the selection of conformations with $0.4<P_{\mathrm{fold}}<0.6$, while for the equilibrium characterization we consider conformations for which the evaluated values of several reaction coordinates correspond to the maximum of the free energy measured as a function of those reaction coordinates. Our results reveal a high degree of structural similarity between the ensembles determined by the two methods. However, the folding probability distribution of the conformations belonging to our definition of the equilibrium transition state ($0.2<P_{\mathrm{fold}}<0.8$) is broader than that displayed by the kinetic transition state. © *2010 American Institute of Physics*. [doi:10.1063/1.3485286]

I. INTRODUCTION

Understanding protein folding, the process according to which a linear chain of amino acids acquires its three-dimensional native structure, remains a challenging problem despite more than 70 years of dedicated research.[1,2] A significant part of the current knowledge on protein folding is based on experimental (i.e., real-world) and computer investigations of small (<100 amino acids), single domain proteins, epitomized by the 64-residue protein Chymotripsin Inhibitor 2. These proteins have been serving as "role models" in fundamental studies of protein folding due to the two-state character of their folding transition.[3] Indeed, most of them fold from the unfolded ensemble (U) to the native state (N) without populating metastable intermediate states, a simplification which renders the analysis more tractable.

A two-state folding transition is generally represented via a mass-action model between the two relevant macrostates, $U \rightleftharpoons N$, each macrostate being an ensemble of microscopic conformations with a certain enthalpy.[4,5] The observation of two well defined thermodynamic macrostates implies the existence of a high free energy barrier between them, and two-state folding transitions are typically rationalized by means of the transition state theory (TST). Accordingly, the observed folding rate, k_f, is proportional to the free energy of activation (i.e., the free energy difference per mole between the unfolded and the transition states), $\Delta G^{\ddagger}$, through the Eyring equation $k_f \sim \exp(-\Delta G^{\ddagger}/RT)$, where R is the universal gas constant and T is the temperature.[6] To confirm the two-state character of the folding kinetics, experimentalists often analyze the so-called "chevron plots," displaying the folding rate in the presence of denaturant, k_f^D, against denaturant concentration $[D]$.[3,7] For two-state transitions, a linear relation must hold, $-RT \ln k_f^D = \Delta G^{\ddagger}_{[D]} = \Delta G^{\ddagger}_{[D]=0} - m[D]$, where m, a constant called the m-value, is related to the average fractional change in the degree of exposure of residues upon unfolding. When an intermediate becomes kinetically relevant, a "rollover" or downward curvature starts to develop in this otherwise linear dependence as $[D]$ is lowered. This rollover is the hallmark of the presence of an on-pathway intermediate.

Apart from assuming chemical equilibrium conditions, one standard extension of the TST also presumes that the reaction's progress can be captured by a reaction coordinate. In essence, an ideal reaction coordinate would be a degree of freedom that connects reactant(s) and product(s) along the lowest free energy continuous path on the free energy surface of the reaction; the highest free energy point along this path is the transition state (TS). Moreover, while in ordinary chemical reactions, such as bond breaking in the gas phase, the TS corresponds to a unique molecular structure, in protein folding the TS is formed by an ensemble of high free energy conformations; it is therefore referred to as the transition state ensemble (TSE).[8,9]

The structural characterization of the protein folding TSE has been a central issue in protein science, but since TSE conformers are short-lived they cannot yet be observed directly in real-world experiments. A protein engineering technique, termed ϕ-value analysis, was developed by Ferhst and co-workers back in the late 1980s with the purpose of solving the structure of the protein folding TSE.[10,11] In the

[a)]Electronic mail: rui@lca.uc.pt.
[b)]Electronic mail: patnev@cii.fc.ul.pt.
[c)]Electronic mail: jsbach@quim.ucm.es.

ϕ-value method, the degree of structure formation of individual residues in the TSE is actually inferred from analyzing the effect of single-site mutations on folding rates and stability. The ϕ-value analysis has been extensively used to investigate the TSE of many proteins,[12] though the implications of a particular ϕ-value on the TSE geometrical traits are still a matter of active research.[13]

Contrary to what happens in real-world experiments, it is straightforward to isolate and directly analyze individual TSE conformations in computer investigations of protein folding, provided that a suitable reaction coordinate for folding is defined. Indeed, given an equilibrium sampling of a free energy barrier, the equilibrium definition of the TSE identifies it as the set of conformations of highest free energy along the path of lowest free energy between the native and unfolded macrostates. However, because proteins have many degrees of freedom, the conformational changes proteins undergo may be not perfectly captured in one- or two-dimensional reaction coordinates.[14–16] As a matter of fact, in recent years several approaches have been proposed to try to define a proper reaction coordinate for the protein folding process, usually as complex optimized combinations not always suitable to be generally exported to a wide range of simulation models or methods.[17–19]

Nevertheless, several simpler structural or energetic reaction coordinates (e.g., volume,[20] the fraction of native contacts, Q,[21–23] total number of native contacts,[22] the root mean square deviation to the folded conformation, RMSD,[24] or the radius of gyration, R_G,[25] just to mention a few) have been proposed for protein folding, and their utility has been explored. For example, it has been claimed that Q is a good reaction coordinate for proteins with smooth free energy landscapes,[26–28] and even though it is recognized as not being a perfect reaction coordinate, it is still used, as it also happens with R_g or RMSD, specially in the context of topology-based models.[19,29–31]

In general, however, structural reaction coordinates may not necessarily measure how dynamically close one conformation is to the native one (i.e., the kinetic progress toward the native state).[32] Hence, difficulties in identifying the appropriate reaction(s) coordinate for folding creates an additional challenge for theorists and simulators aiming to study protein folding in light of the TST.[33] Therefore, it may not be straightforward to accurately determine the folding TSE from equilibrium sampling. In order to overcome this difficulty, a reaction coordinate termed folding probability, P_{fold}, has been proposed that encompasses a kinetic definition of the TSE.[14,15] For a strictly two-state transition, the TSE would lie on a single stochastic separatrix between the U and N states. Assuming that there is a way to classify conformations as being part of the U or N states, the TSE can be defined as the ensemble of conformations for which the probability to find the native state before reaching an unfolded conformation is 0.5.[14] In other words, for two-state folding transitions, the TSE is the ensemble of conformations for which the commitment probability P_{fold} is 0.5. Just like the ϕ-value analysis in experiments with real-world proteins, the folding probability analysis method has been extensively used to investigate the TSE in computer simulations of protein folding.[13,34–37] The P_{fold} method has one caveat though: it is computationally costly.[38]

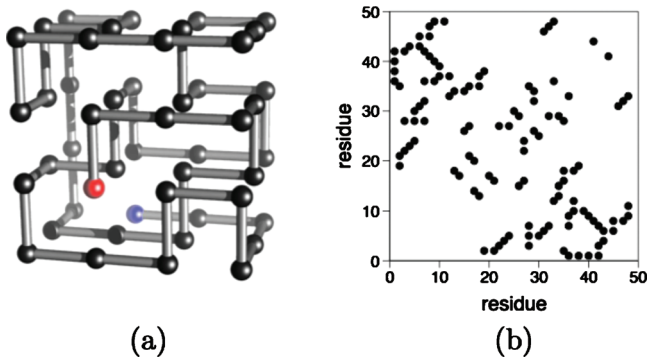


FIG. 1. Three-dimensional representation (a) and contact map (b) of the model system investigated in this study. Each circle in the contact map represents a native contact. Circles that are close to the main diagonal represent local contacts (i.e., contacts between beads that are close to each other along the chain), while nonlocal contacts are represented by circles that are located far away from the main diagonal. For maximally compact cuboids with $N=48$ beads (numbered from 0 to 47) there are 57 native contacts. The (relative) contact order (Ref. 44) of the selected target is 0.45, which translates in a predominance of nonlocal contacts. In the three-dimensional representation, the C- and N-terminal beads are colored red and blue, respectively.

Here we perform extensive Monte Carlo simulations of a simple lattice protein model to construct and compare two ensembles of conformations representative of the TSE. In one case the selection of conformations is based on the equilibrium definition of the TSE and the resulting ensemble is termed *equilibrium TS*. In the other case the kinetic definition of TS (i.e., $P_{\text{fold}}=0.5$) is employed to construct the ensemble of conformations which we label the *kinetic TS*. In the following section we describe in detail the computational procedures adopted in each situation. Afterwards we present the simulation results focusing our analysis on a direct comparison between the equilibrium and the kinetic TS. In doing so we search for similarities and differences between their average energies and structures, as well as their kinetic traits (including the folding probabilities).

II. MODELS AND METHODS

A. The simple lattice model and the Gō potential

We consider a simple three-dimensional lattice representation of a protein molecule with chain length $N=48$. In such a minimalist model, amino acids, represented by beads of uniform size, occupy the lattice vertices and the peptide bond that covalently connects amino acids along the polypeptide chain is represented by sticks with uniform (unit) length corresponding to the lattice spacing. Figure 1 displays the three-dimensional structure (a) and the corresponding contact map (b) of the model system investigated here.

To mimic protein energetics, we use the Gō potential.[39] In the Gō potential, the energy E of a conformation, defined by the set of bead coordinates $\{\vec{r}_i\}$, is given by

$$E(\{\vec{r}_i\}) = \sum_{i>j}^{N} \epsilon \Delta(\vec{r}_i - \vec{r}_j), \quad (1)$$

where ϵ is the (uniform) interaction energy parameter and the contact function $\Delta(\vec{r}_i - \vec{r}_j)$ is unity only if beads i and j form a noncovalent native contact, i.e., a contact between a pair of beads that is present in the native structure and is zero otherwise.

B. Folding simulation details

In order to mimic the protein's relaxation toward the native state we use the Metropolis Monte Carlo (MC) algorithm,[40] together with a local move set that includes corner-flips and end-moves (i.e., displacements of one single bead) and the crankshaft move (which involves the displacement of two beads at the same time).[41–43] At each MC step, the probability of applying the Metropolis criterion to a particular chain displacement is $0.2/(N+6)$, if the displacement involves moving only one bead, or $0.8/(2N-3)$ if it involves the simultaneous movement of two beads. Attempted random moves which are discarded by excluded volume constraints or fail the Metropolis rule are counted as MC steps. For the calculations leading to the dynamic definition of the TSE and the P_{fold} analysis, a MC simulation starts from a randomly generated unfolded conformation and the folding dynamics is monitored by following the evolution of the fraction of the established native contacts Q. The number of MC steps required to fold to the native state (i.e., to achieve $Q=1.0$) at a given temperature is the first passage time (FPT) and the folding time is computed as the mean FPT of 500 simulations. The temperature (T), which is kept constant in this type of simulations, is measured in units of ϵ/k_B, where k_B is the Boltzmann constant.

C. Equilibrium sampling: Parallel tempering and WHAM analysis

In addition, we want to study the thermodynamic characteristics of the folding-unfolding transition for our model. Thus, we have also used a simulation method which should be able to properly reproduce the equilibrium distribution of states for the system at every temperature under consideration. Therefore, we have used a parallel tempering (also called replica exchange) Monte Carlo simulation algorithm.[45] Parallel tempering allows surmounting free energy barriers as the replicas sample through the different temperatures. Since, as already mentioned, we have chosen to study the folding transition of a two-state protein, a relatively large number of temperatures has to be used in order to warrant a proper overlap of the energy distributions at neighbor temperatures around the transition midpoint, and the adequate sampling of the replicas through the different temperatures. Thus, we have run 27 temperatures in parallel (from $T=0.5$ to $T=1.04$, in reduced units), each with 10^{10} MC steps (after 10^6 MC steps of relaxation). Every 10^5 MC steps, the exchange of two conformations at neighboring temperatures has been attempted. The set of MC moves is exactly the same we have used at the single temperature simulations.

The free energy as a function of different possible reaction coordinates has been computed using the weighted histogram analysis method (WHAM).[46,47] The WHAM uses data from all the conformations sampled at different temperatures to estimate the populations of the different energetic states of the protein model. In this way, it is possible to obtain statistically reliable results for the thermodynamic properties, specially the free energy profile of the system at a given temperature.

D. Kinetic analysis: Folding probability calculation

The folding probability, P_{fold}, of a conformation at a given temperature is defined as the fraction of single temperature MC runs which, starting from that conformation, fold before they unfold.[14] Because a P_{fold} calculation amounts to a Bernoulli trial, the relative error resulting from using M runs scales as $M^{-1/2}$.[48] Thus, in order to accurately compute P_{fold} we consider 500 MC runs equally divided into five sets of 100 folding simulations. The average value of P_{fold} is computed for each set and the mean of all five sets, together with its standard deviation, is evaluated. Each MC run in this calculation stops when either the native fold ($Q=1.0$) or an unfolded conformation is reached. A conformation is deemed unfolded when its fraction of native contacts Q is smaller than a cutoff value Q_U, which is calculated from the free energy profile of the system at the desired temperature, as determined by the WHAM method. This fraction of native contacts ($Q_U \sim 0.09$, see below) is considerably low and therefore identifies states with minimal residual native structure.

III. RESULTS

A. The simulation temperatures

In this study, folding is investigated at two different temperatures: the optimal folding temperature, T_{opt}, and the transition temperature, T_f. The optimal folding temperature is the temperature that maximizes folding speed, and for the target model investigated here $T_{\text{opt}}=0.65$.[49–51] The folding transition temperature (also termed as melting temperature, T_m, in the experimental literature) is defined as the temperature at which the denatured and native states are equally populated at equilibrium.[52,53] Experimentalists usually estimate this transition temperature as the temperature at which the heat capacity attains its peak value. Here, the heat capacity is computed from the energy fluctuations at each temperature considered in the parallel tempering simulations, $C_v=(\langle E^2\rangle - \langle E\rangle^2)/T^2$. The peak value of the heat capacity curve is attained at $T=T_f=0.79$.[54] In the temperature range below the folding transition temperature the native state is stabilized; this is precisely what happens at T_{opt}, which is lower than T_f (Table I). Indeed, the conformational distribution (which is equivalent to the energy distribution for the Gō potential) of our model protein is clearly bimodal at T_f but becomes strongly shifted toward the native state at the lower temperature T_{opt}.[54] Table I reports the kinetic information at the two temperatures considered in this study.

TABLE I. Summary of kinetic and thermodynamic properties of the model protein studied. The optimal folding temperature T_{opt} is the temperature of fastest folding, while the transition temperature T_f is the temperature at which the heat capacity attains its maximum value. The folding time is measured as the mean first passage time (MFPT) of 500 folding runs.

T_{opt}	$\log_{10}$(MFPT) at T_{opt}	T_f	$\log_{10}$(MFPT) at T_f
0.65	7.11 ± 0.02	0.79	8.42 ± 0.03

B. The equilibrium determination of the transition state ensemble

To define the equilibrium transition state (ETS), we start by analyzing the dependence of the free energy on three putative folding reaction coordinates, namely, the energy E (measured in units of ϵ), the radius of gyration R_g, and the root mean square deviation with respect to the native conformation, RMSD (both measured in lattice units).

At T_f a clear peak is observed in the corresponding free energy curves, indicating that for this model these particular parameters are suitable (though not necessarily perfect, as mentioned in the Introduction) reaction coordinates for folding at the transition temperature (Fig. 2, top). Our choice of the energetic and geometrical coordinates is motivated by the fact that they are the most extensively used in simulation studies of protein folding.[29–31] The minima corresponding to the folded and the unfolded states do not have the same values for the free energy. This is due to the constraints imposed by the lattice, which create an extremely narrow minimum for the folded state, where even the "wall" framing this minimum from the left is missing, in comparison with the wide and shallow unfolded state.

In the graphs at the left column in Fig. 2 we can also appreciate the minimum for the unfolded state at an energy value of about $E=-5$, indicating a small number of residual native contacts in the conformations belonging to the unfolded state. This minimum corresponds to a fraction of native contacts $Q=5/57 \simeq 0.09$, thus justifying the Q_U value mentioned previously for the calculation of the folding probability, P_{fold}.

From these profiles, we can select the range of values that identify the conformations belonging to the TS. Therefore, by using the equilibrium definition, a conformation is deemed a member of the equilibrium TS (ETS) at T_f if its reaction coordinates fall in the vicinity of the peak of the corresponding free energy curves, i.e., $-29 < E < -23$, $2.05 < R_g < 2.35$, and $1.1 < \text{RMSD} < 1.55$. This definition is similar, although not quantitatively identical, to that used previously.[26] The definition of the ETS is robust with regard to changing these intervals provided they will enclose the free energy peak and comprise the region where P_{fold} varies rapidly with E, R_g, and RMSD (see Fig. 6 top, for the energy case, and its discussion later in the text).

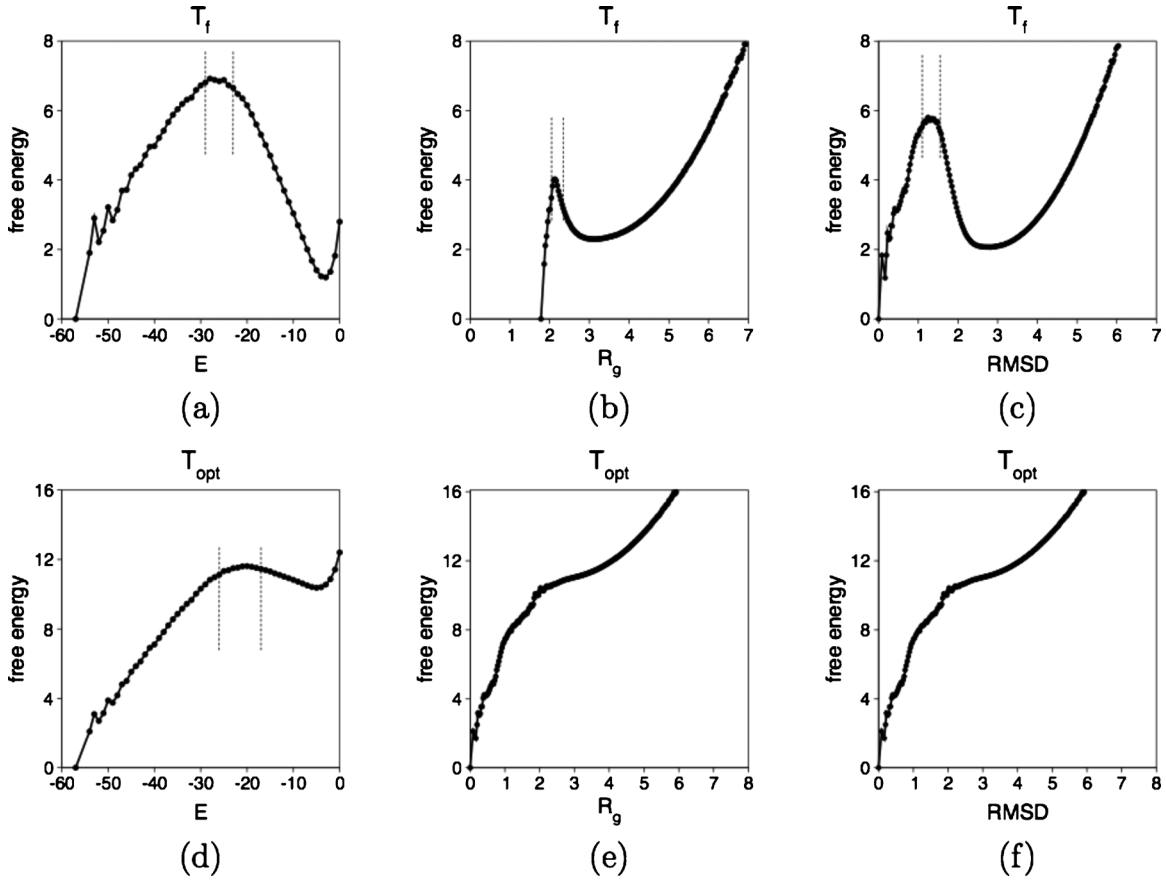


FIG. 2. Free energy as function of energy (left), radius of gyration (middle), and root mean square distance (right) at the transition temperature T_f (top) and at the optimal folding temperature T_{opt} (bottom). The regions enclosed between the dotted vertical lines indicate the values taken by each reaction coordinate in the equilibrium transition state conformations, according to the definition used in this work.

In practice, in order to build an ensemble representative of the ETS, we randomly select conformations (one conformation per run) from the final part (i.e., from the last 5×10^5 MC steps) of independent folding runs at the desired temperature. We keep only those for which E, R_g, and RMSD simultaneously fall in the above ranges. By selecting conformations from the last stage of the folding process one is able to obtain a relatively large ensemble of transition state conformations, and by selecting only one conformation per run one guarantees that the selected conformations are not correlated. At this temperature we have inspected a total of 15 115 conformations. Of these, 315 conformations meet all of the cutoffs and thus they constitute the ETS, according to our definition.

A different scenario is observed at $T=T_{\rm opt}$ (Fig. 2, bottom) where a peak in the free energy profile is only observed for the dependence of the latter on the energy. As expected, at a low temperature where the native state is clearly stabilized,[54] the free energy barrier from the unfolded state to the TS is much lower than at T_f, and it only appears along the energy coordinate. Thus, at $T_{\rm opt}$ a conformation is considered part of the ETS if its energy E is such that $-26<E<-17$. As a result of this "less strict" definition, the ETS ensemble obtained at $T_{\rm opt}$ contains a large number of conformations. Indeed, 1194 out of the 8000 independent conformations inspected at this temperature belong to the ETS. The fact that the free energy peak is broader at this low temperature also influences the larger number of conformations selected as part of the ETS at $T_{\rm opt}$. It probably reflects that the energy is not a perfect reaction coordinate, especially at this temperature, as we have mentioned before.

Finally, we observed that the structural properties of the TSE at each considered temperature reflect the gross topological features of the conformations forming the pool from which they originally belonged.[54] This observation should be taken into account in order to correctly construct an ensemble of conformations representative of the TSE.

C. The kinetic determination of the transition state ensemble

As mentioned in the Introduction, we use the $P_{\rm fold}$ analysis method to construct the kinetic transition state (KTS) at T_f and at $T_{\rm opt}$. The results at $T_{\rm opt}$ have been already presented in a previous paper.[55] For each conformation we have evaluated $P_{\rm fold}$ and the time (mean FPT of 500 MC runs) it needs to achieve the native structure without passing through an unfolded conformation (i.e., any conformation for which $Q<0.09$); we term this quantity the *forward folding time*, t_{fw}, in order to distinguish it from the folding time as defined before. Results plotted in Fig. 3 show the existence of a considerably large stripe, formed by fast folding conformations, that makes a sharp turn downward at $P_{\rm fold}\approx0.95$ (Fig. 3). This observation suggests the formation of a postcritical folding nucleus which guarantees that folding is fast and certain.[56] At low values of $P_{\rm fold}$ this stripe widens up due to the significantly lower number of conformations that are able to find the native state prior to visiting an unfolded conformation. Outside the main stripe there are a few clusters of conformations which take a distinctively long time to find the native fold; in particular, those with high $P_{\rm fold}$ are trapped states. These will naturally be disregarded in our subsequent analysis because they have particular structural traits (such as nonlocal non-native contacts formed with high probability, which have been the aim of previous work)[13,55] that may mask the structural characterization of the TSE. Anyhow, these states represent a very small fraction (<5%) of the full set of conformations considered, and therefore their real effect in the analysis reported below would be minimal if included.

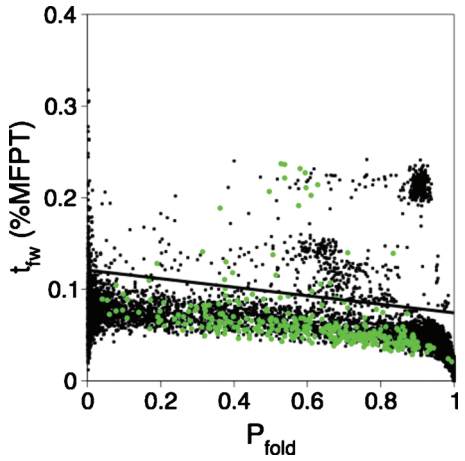


FIG. 3. The time necessary to reach the native structure without passing through some unfolded conformation, t_{fw}, for each conformation within the ensemble of 15 115 conformations analyzed in this study at T_f. The green dots highlight the 315 conformations belonging to the ETS at T_f.

The main folding pathway contains all the conformations which fulfill the inequality $t_{fw}<(2.75-0.75P_{\rm fold})\times10^5$ MC steps (i.e., those conformations located below the solid line in Fig. 3). Although these conformations have been selected from the final steps of the simulation runs, they display all the possible values of $P_{\rm fold}$. Therefore, according to the adopted kinetic definition of the TSE, we have selected 453 conformations with $0.4<P_{\rm fold}<0.6$ that lie on the main folding pathway.

The KTS obtained by the same method at $T=T_{\rm opt}$ comprises 401 conformations which were selected from a total of 8000 analyzed conformations. Figure 3 also reports the $P_{\rm fold}$ values, and the corresponding t_{fw}, of the 315 conformations representing the ETS at T_f. All of the referred ensembles will be analyzed in detail in the following sections.

D. The transition state structure and the folding nucleus

In order to compare the structures of the TSEs based on the kinetic and the thermodynamic definitions, we have computed the probability maps displaying the frequency of occurrence of the 57 native contacts in each considered ensemble of conformations. Results reported in Fig. 4 (top) show that at $T=T_f$ the structure of the ETS is remarkably similar, on average, to that of the KTS. Indeed, except for the contact between beads 3 and 22 in the KTS (which is not displayed in the corresponding probability map), all the other native contacts have a probability higher than 20% of being formed in TS conformations. As expected, due to the native-

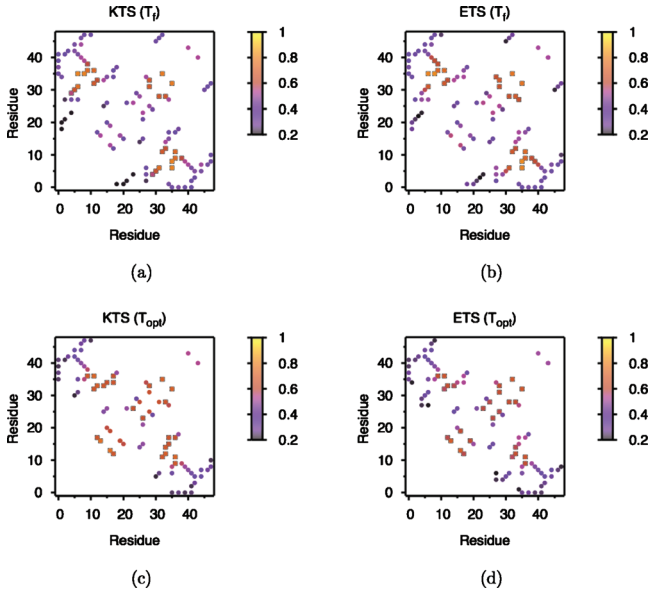


FIG. 4. Probability maps of the KTS (left) and ETS (right) at $T=T_f$ (top) and at $T=T_{opt}$ (bottom) showing the probability of occurrence of each contact in the corresponding transition state ensemble. Only contacts with a probability of occurrence higher than 20% are displayed. The contacts forming the putative folding nucleus obtained from each TSE determination (i.e., the 13 most probable contacts with $p>0.5$ in each case) are marked with squares (see text).

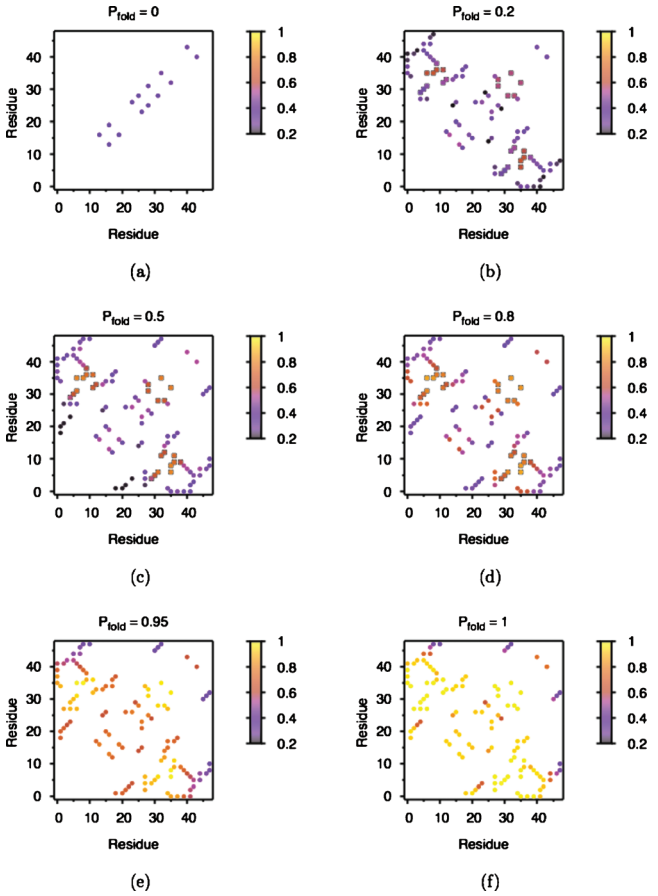


FIG. 5. Probability maps showing the native structure evolution along the reaction coordinate P_{fold} at T_f. The contacts forming the FN are marked with squares. Interestingly, 77% of the FN contacts are already formed with a considerably high probability ($p>0.49$) at $P_{fold}=0.2$.

centric nature of the interaction potential employed, there are not non-native contacts formed with a probability of occurrence higher that 20%.

The comparison of the TSEs at T_{opt} also reveals a high degree of structural similarity between both ensembles (Fig. 4, bottom). At T_{opt}, however, the number of native contacts with probability of occurrence higher than 20% (44 in the KTS and 47 in the ETS) is lower than that observed at $T=T_f$. The native contacts that are formed with very low probability in these ensembles are nonlocal contacts, showing that at $T=T_{opt}$ the TSE is "more local" than the TSE at the higher temperature $T=T_f$. As before, there are not non-native contacts formed with a probability of occurrence higher that 20%.

We now focus our analysis on a particular subset of native contacts, which is that of the critical folding nucleus (FN). To this end we follow Shakhnovich and co-workers[57] and define the FN as the set of native contacts forming with high probability in the TSE. To favor the comparison among the different ensembles considered in the manuscript, we consider the FN to be constituted always by the same number of the most populated contacts in the TSE (13 contacts, corresponding to about 20% of the native structure).[56] This set of contacts is marked with squares in the probability maps.

The very high degree of structural similarity between the kinetic and equilibrium TSEs that is observed at T_f allows us to anticipate a very large overlap (85%) between the kinetic and the equilibrium folding nuclei at this temperature. Indeed, at T_f the TSE is stabilized by a subset of native contacts that are predominantly nonlocal and long-ranged (e.g., 6:31, 6:35, 8:35, 9:36, 11:32, 11:36, and 12:33).

At the optimal folding temperature T_{opt} there is also an overlap of about 77% between the kinetic and the equilibrium folding nuclei. However, in this case 40%–60% of the contacts that make up the FN are local contacts (e.g., 13:16, 28:33, and 32:35). The fact that the FN is more local at T_{opt} than at T_f explains the general observation that the TSE is considerably more consolidated (i.e., overall it has more native contacts formed with high probability) at T_{opt} than at T_f.

It is also interesting to observe how the native structure, and the FN in particular, develops as the folding "reaction" evolves. We have done that by comparing the probability maps corresponding to ensembles of conformations with different values of P_{fold} at T_f (Fig. 5). For very low values of P_{fold}, i.e., at the very early stage of the folding process, it is only the very local native contacts that are formed in more than 20% of the analyzed conformations. Surprisingly, however, when P_{fold} increases up to 0.2, there are eight contacts (five of which are nonlocal) formed with a relatively high probability ($p>0.5$); these contacts belong to the FN. At this stage of the folding process there is already a large overlap (77%) between the 13 most probable contacts at $P_{fold}=0.2$ and the FN (i.e., the 13 most probable contacts at $P_{fold}=0.5$). From $P_{fold}=0.5$ onward, the protein's native contacts display a continuously increasing probability of being formed and those forming the FN are definitely the most probable contacts.

We conclude that ensembles of conformations with P_{fold} in the range $0.2 < P_{\text{fold}} < 0.8$ share a common structural trait: the contacts forming the FN (or at least a very significant part of these contacts) are formed in these conformations with very high probability. In light of this observation it is not surprising that conformations with $0.2 < P_{\text{fold}} < 0.8$ have similar t_{fw} (Fig. 3).

At the very final stages of the folding process ($P_{\text{fold}} \sim 1$) most of the native contacts form with very high probability, $p > 0.8$, although there is some structural disorder on the surface of the protein and close to the terminal residues.

E. P_{fold} analysis of the equilibrium transition state

We have shown that the (average) structure of the kinetic TSE is remarkably similar to that of the equilibrium TSE. Therefore, it is pertinent to ask if such structural similarity is underlined by a kinetic similarity between the two ensembles. In other words, are the P_{fold} values of the conformations representing the equilibrium TSE close to 0.5? In order to address this question we have evaluated P_{fold} for each conformation belonging to the ETS at $T = T_f$ (represented by the green dots in Fig. 3). Clearly, there is no shift toward $P_{\text{fold}} \sim 0.5$ and, indeed, the distribution is rather uniform within the range $0 < P_{\text{fold}} < 1$ (with an averaged mean $P_{\text{fold}} = 0.59$ and a standard deviation $= 0.22$).

In order to better interpret this result, we have computed the mean average number of native contacts (for the Gō potential this is equivalent to the mean averaged energy) formed in each ensemble of conformations characterized by some value of P_{fold}, and we have mapped the mean energy thus computed onto the corresponding free energy curve (Fig. 6, top). We verify that conformations with $P_{\text{fold}} = 0$, which have a very small number of native contacts formed, are mapped onto the unfolded state basin of the free energy curve, while an opposite scenario holds for conformations with $P_{\text{fold}} = 1$, which are mapped onto the native state. At the very top of the free energy barrier there is the ensemble of conformations with $P_{\text{fold}} = 0.5$. However, we can readily appreciate that conformations with folding probability in the range $0.2 < P_{\text{fold}} < 0.8$ lie remarkably close to the maximum of the free energy.

To check whether this observation is an effect of the averaging process, we have analyzed the histograms for the energy distribution of the ensembles of conformations with $P_{\text{fold}} = 0, 0.3, 0.5, 0.7$ and $P_{\text{fold}} = 1$, at T_f (Fig. 6, bottom). While the histogram associated with $P_{\text{fold}} = 0.5$ is readily distinguished from the histograms corresponding to $P_{\text{fold}} = 0$ and $P_{\text{fold}} = 1$, there is a considerably large superposition between the energy histograms corresponding to $P_{\text{fold}} = 0.3$, 0.5, and 0.7. Thus both the average values of the energy and the energy distributions show an impressive similarity. Altogether, these observations suggest that the energy alone is not able to distinguish among conformations with folding probabilities in a rather wide range around $P_{\text{fold}} = 0.5$.

We have carried out the same analysis for the other two order parameters used in this work to identify the TSE, namely, the radius of gyration and the RMSD, and the situation is even worse, in the sense described above (data not shown). As expected, it turns out that for the target model studied here the energy is the parameter which "better" distinguishes conformations with different values of P_{fold}, despite the shortcomings mentioned in the previous paragraph. Hence, it is now understandable that ETS conformations, which are chosen based on their geometrical traits, may have different P_{fold} values, as observed in Fig. 3.

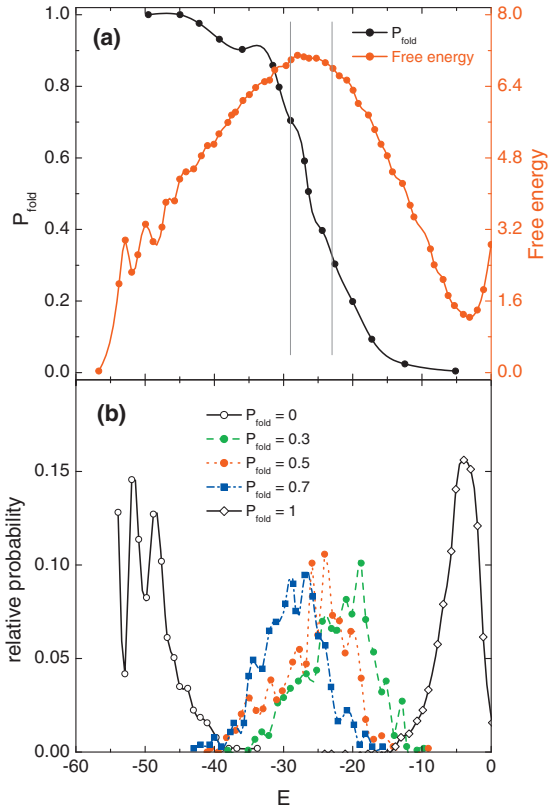


FIG. 6. (Top) The (mean averaged) energy of conformations within each P_{fold} ensemble at T_f and its mapping onto the free energy curve. The energy of conformations with $0.2 \leq P_{\text{fold}} \leq 0.8$ lies very close to the region near the peak of the free energy barrier (enclosed by the thin vertical lines). (Bottom) Energy distribution within the ensembles of conformations with $P_{\text{fold}} = 0$, 0.3, 0.5, 0.7, and 1. There is a very large superposition between the histograms corresponding to $P_{\text{fold}} = 0.3$, 0.5, and 0.7.

IV. CONCLUSIONS

We have performed extensive Monte Carlo simulations to investigate the TSE of a small lattice protein that folds to the native state with single exponential (i.e., two-state) kinetics. More precisely, we have used two different computational methodologies, the P_{fold} analysis method and the WHAM analysis, to construct two ensembles of conformations representative of the TSE at different simulation temperatures. One ensemble, which we call kinetic transition state ensemble, is characterized for having conformations with folding probability $0.4 < P_{\text{fold}} < 0.6$, while the other, the equilibrium transition state ensemble, contains conformations sampled from the top of the free energy barrier (i.e., conformations for which certain evaluated properties, the so-called reaction coordinates, correspond to a maximum of the free energy as shown by the WHAM method). For the latter, we choose simple energetic and geometrical coordinates

which not being perfect reaction coordinates are still widely used, specially for simple interaction models as that used in this work.

The comparison of the two TSEs thus obtained reveals a remarkably large structural similarity between them. Indeed, the most structured region of the TSE (i.e., the folding nucleus) in the two ensembles of conformations examined at each temperature displays a large overlap (~80%). We stress, however, that this observation may, in part, be due to the simplicity of the lattice model investigated here and that it may not hold for atomistic models based on real interactions.

We have also computed the folding probability of each conformation forming the equilibrium transition state and we have found a rather broad distribution of P_{fold} values for these conformations. In order to interpret this surprising result (we were expecting to find a shift toward 0.5 in the P_{fold} values), we have determined the mean average energy of the conformations with a certain P_{fold} and mapped it onto the WHAM free energy versus the E curve. Not surprisingly, we have found that the mean energy of the set of conformations with $P_{\text{fold}}=0.5$ corresponds to the maximum of the free energy. However, we have as well found that conformations with P_{fold} as different as 0.2 and 0.8 lie remarkably close to the top of the free energy curve. It is important to mention, however, that the average P_{fold} for the conformations belonging to our equilibrium definition of the TSE is 0.59 ± 0.01, in spite of the large variability shown in Fig. 3 for the values of the individual conformations. This result is similar to that reported in Ref. 26, where the relation between the fraction of native contacts and the folding probability was investigated in the context of an off-lattice Gō model. It was found that conformations with fraction of native contacts $Q=0.5$ have folding probabilities in a rather wide range $0.2 < P_{\text{fold}} < 0.8$.

When folding amounts to surmounting a considerably high free energy barrier we expect to observe large variations in the P_{fold} reaction coordinate near the maximum of the free energy. The relationship between the free energy profile and the behavior of P_{fold} close to the top of the free energy peak will be the scope of future work.

ACKNOWLEDGMENTS

R.D.M.T and P.F.N.F thank Fundação para a Ciência e Tecnologia (FCT) for financial support through the *Ciência 2007* program. P.F.N.F and R.D.M.T thank Conselho de Reitores das Universidades Portuguesas for financial support thought grant "Acção Integrada Luso-Espanhola E-16/09." A.R. thanks financial support from the Spanish Ministerio de Ciencia e Innovación (Acción Integrada Hispano-Portuguesa HP-2008-0065 and Grant No. FIS2009-13364-C02-02).

[1] K. A. Dill, B. Ozkan, T. R. Weikl, J. D. Chodera, and V. A. Voelz, Curr. Opin. Struct. Biol. **17**, 342 (2007).
[2] R. F. Service, Science **321**, 784 (2008).
[3] S. E. Jackson, Folding Des. **3**, R81 (1998).
[4] K. A. Dill, Protein Sci. **8**, 1166 (1999).
[5] J. Schonbrun and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **100**, 12678 (2003).
[6] S. Berry, S. A. Rice, and J. Ross, *Physical and Chemical Kinetics* (Oxford University Press, Oxford, 2002), Chap. 30.
[7] H. S. Chan, S. Shimizu, and H. Kaya, Methods Enzymol. **380**, 350 (2004).
[8] J. N. Onuchic, N. D. Socci, Z. Luthey-Schulten, and P. G. Wolynes, Folding Des. **1**, 441 (1996).
[9] J. N. Onuchic and P. G. Wolynes, Curr. Opin. Struct. Biol. **14**, 70 (2004).
[10] A. Fersht, *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*, 3rd ed. (Freeman, San Francisco, 1998).
[11] L. S. Itzhaki, D. E. Otzen, and A. R. Fersht, J. Mol. Biol. **254**, 260 (1995).
[12] P. F. N. Faísca, J. Phys.: Condens. Matter **21**, 373102 (2009).
[13] P. F. N. Faísca, R. D. M. Travasso, R. C. Ball, and E. I. Shakhnovich, J. Chem. Phys. **129**, 095108 (2008).
[14] R. Du, V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. S. Shakhnovich, J. Chem. Phys. **108**, 334 (1998).
[15] C. Snow and V. S. Pande, Biophys. J. **91**, 14 (2006).
[16] D. K. Klimov and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **97**, 2544 (2000).
[17] R. B. Best and G. Hummer, Proc. Natl. Acad. Sci. U.S.A. **102**, 6732 (2005).
[18] S. Muff and A. Caflisch, J. Chem. Phys. **130**, 125104 (2009).
[19] R. B. Best and G. Hummer, Proc. Natl. Acad. Sci. U.S.A. **107**, 1088 (2010).
[20] E. I. Shakhnovich and A. V. Finkelstein, Biopolymers **28**, 1667 (1989).
[21] E. Shakhnovich, G. Farztdinov, A. M. Gutin, and M. Karplus, Phys. Rev. Lett. **67**, 1665 (1991).
[22] A. Sali, E. Shakhnovich, and M. Karplus, Nature (London) **369**, 248 (1994).
[23] P. E. Leopold, M. Montal, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **89**, 8721 (1992).
[24] R. Brüschweiler, Proteins **50**, 26 (2003).
[25] M. Doi, *Introduction to Polymer Physics* (Oxford University Press, Oxford, 1997).
[26] S. S. Cho, Y. Levy, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **103**, 586 (2006).
[27] H. Nymeyer, N. D. Socci, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **97**, 634 (2000).
[28] C. Clementi, M. Cheung, H. Nymeyer, and J. N. Onuchic, Biophys. J. **78**, 46A (2000).
[29] H. Kaya and H. S. Chan, J. Mol. Biol. **326**, 911 (2003).
[30] C. Clementi, A. E. Garcia, and J. N. Onuchic, J. Mol. Biol. **326**, 933 (2003).
[31] W. Guo, S. Lampoudi, and J. E. Shea, Biophys. J. **85**, 61 (2003).
[32] H. S. Chan and K. A. Dill, Proteins **30**, 2 (1998).
[33] B. Peters, J. Chem. Phys. **125**, 241101 (2006).
[34] F. Ding, N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley, and E. I. Shakhnovich, Biophys. J. **83**, 3525 (2002).
[35] I. A. Hubner, M. Oliveberg, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **101**, 8354 (2004).
[36] I. A. Hubner, K. A. Edmonds, and E. I. Shakhnovich, J. Mol. Biol. **349**, 424 (2005).
[37] F. Ding, W. Guo, N. V. Dokholyan, E. I. Shakhnovich, and J.-E. Shea, J. Mol. Biol. **350**, 1035 (2005).
[38] D. A. C. Beck and V. Daggett, Biophys. J. **93**, 3382 (2007).
[39] N. Gō and H. Taketomi, Proc. Natl. Acad. Sci. U.S.A. **75**, 559 (1978).
[40] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[41] D. P. Landau and K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics* (Cambridge University Press, Cambridge, 2000).
[42] H. S. Chan and K. A. Dill, J. Chem. Phys. **99**, 2116 (1993).
[43] H. S. Chan and K. A. Dill, J. Chem. Phys. **100**, 9238 (1994).
[44] K. W. Plaxco, K. T. Simmons, I. Ruczinski, and D. Baker, Biochemistry **39**, 1117783 (2002).
[45] U. H. E. Hansmann, Chem. Phys. Lett. **281**, 140 (1997).
[46] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, J. Comput. Chem. **13**, 1011 (1992).
[47] J. D. Chodera, W. C. Swope, J. W. Pitera, C. Seok, and K. A. Dill, J. Chem. Theory Comput. **3**, 26 (2007).
[48] I. A. Hubner, J. Shimada, and E. I. Shakhnovich, J. Mol. Biol. **336**, 745 (2004).

[49] A. Gutin, A. Sali, V. Abkevich, M. Karplus, and E. I. Shakhnovich, J. Chem. Phys. **108**, 6466 (1998).
[50] P. F. N. Faisca and R. C. Ball, J. Chem. Phys. **116**, 7231 (2002).
[51] P. F. N. Faisca and M. M. Telo da Gama, Biophys. Chem. **115**, 169 (2005).
[52] G. P. Privalov and P. L. Privalov, Methods Enzymol. **323**, 31 (2000).
[53] P. F. N. Faísca and K. W. Plaxco, Protein Sci. **15**, 1608 (2006).
[54] See supplementary material at http://dx.doi.org/10.1063/1.3485286 for Fig. S1 for heat capacity curve as function of temperature, Fig. S2 for the histogram for the different Q values, and a note for the practitioner for what should not be done in order to correctly construct an ensemble of conformations representative of the TSE.
[55] R. D. M. Travasso, M. M. T. Gama, and P. F. N. Faisca, J. Chem. Phys. **127**, 145106 (2007).
[56] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, Biochemistry **33**, 10026 (1994).
[57] L. Li, L. A. Mirny, and E. I. Shakhnovich, Nat. Struct. Biol. **7**, 336 (2000).

Supplementary Material

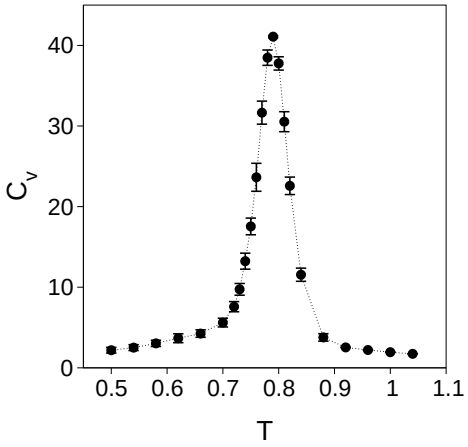
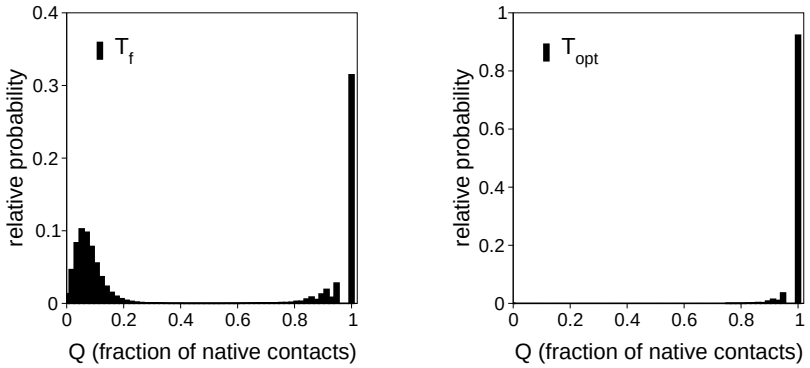


Figure S1: Heat capacity as a function of temperature, both in reduced units. The peak at $T = 0.79$ indicates the transition temperature T_f.



(a) (b)

Figure S2: The conformational distribution of our model protein is clearly bimodal at T_f (a) and becomes strongly shifted towards the native state at T_{opt} (b).

A note for the practitioner

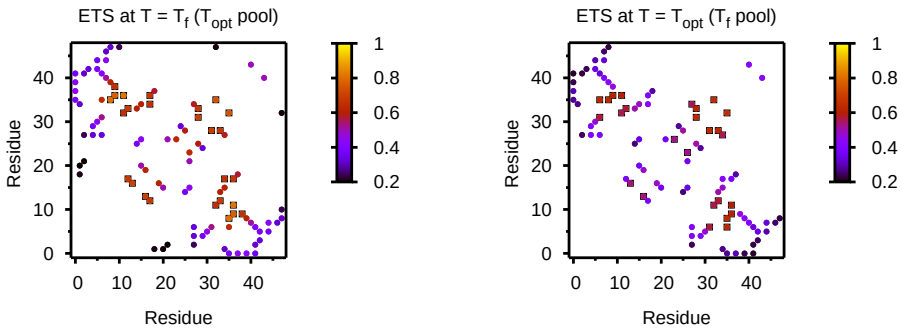
In order to find an ensemble of conformations representative of the equilibrium transition state (ETS) at a certain temperature T_0, it is necessary to firstly locate the peak of the free energy curve as a function of several reaction coordinates (e.g. energy, radius of gyration and RMSD), and then to perform many independent MC runs at T_0 from which conformations will be selected only if their reaction coordinates satisfy the requirement of falling on an interval centered around the peak of the corresponding free energy curve. Since the sampling of such conformations may be scarce - it is difficult to sample high free energy conformations - one would be tempted to apply the selection criteria established for temperature T_0 to conformations sampled from simulations carried out at a different temperature (e.g. from the different replicas in parallel tempering simulations) in order to get a larger sample, and therefore achieve better statistics. Here we show that this procedure leads to wrong results. Indeed, in order to obtain an ensemble of conformations representative of the ETS at a desired temperature T_0, one should *only* consider conformations sampled from MC runs at that particular temperature. In other words, conformations whose 'reaction coordinates' match the TSE criteria found from the free energy profiles computed at temperature T_0 but that are sampled from runs at different temperatures do not form a reliable representative of the ETS.

As an example, we have considered the following pools of conformations: i) the T_f pool, formed by ~ 15115 conformations sampled from the last 5×10^5

MCS of single temperature simulations at $T = T_f$, and ii) the T_{opt} pool, formed by ∼8000 conformations sampled from the last 5×10^5 MCS of single temperature simulations at $T = T_{opt}$. Figure S3a shows the frequency map of a putative ETS that is formed by the conformations of the T_{opt} pool whose reaction coordinates match the thermodynamic criteria established for T_f, termed ETS $T_f(T_{opt})$, and Figure S3b shows the frequency map of a putative ETS that is formed by the conformations of the T_f pool whose reaction coordinates match the criteria established for T_{opt}, termed ETS $T_{opt}(T_f)$. These frequency maps should be compared with the frequency maps for the 'correct' ETS at T_f and T_{opt} (Figures 4b and 4d).

We observe dramatic differences between the structure of the ETS constructed from different pools of conformations. The ETS $T_f(T_{opt})$ determined by using the T_{opt} pool (Figure S3a) shows considerably more local structure formed than the ETS at $T = T_f$ constructed from the T_f pool (Figure 4b). Indeed, the putative FN obtained from the T_{opt} pool does not exhibit the non-local geometrical features of the TSE at T_f, showing equal overlaps (69%) with the FNs obtained from the ETSs at $T = T_f$ and at $T = T_{opt}$ (Figure 4d).

On the other hand, the ETS $T_{opt}(T_f)$ obtained from the T_f pool (Figure S3b) displays a strong prevalence of non-local contacts, which is a structural trait typical of the TSE at the transition temperature T_f. Accordingly, the calculated overlap of 85% between its putative FN and the FN obtained from the ETS at $T = T_f$ (Figure 4b), is higher than the overlap (69%) with the FN determined at $T = T_{opt}$ (Figure 4d).



(a) (b)

Figure S3: Frequency map of a putative ETS at $T = T_f$ obtained by averaging over conformations selected from the pool at T_{opt} (a) and frequency map of a putative ETS at $T = T_{opt}$ obtained by averaging over conformations selected from the pool at T_f (b). The corresponding putative FNs are marked with squares.