The Journal of Chemical Physics

Ratcheted molecular-dynamics simulations identify efficiently the transition state of protein folding

Cite as: J. Chem. Phys. **137**, 235101 (2012); https://doi.org/10.1063/1.4769085
Submitted: 05 July 2012 . Accepted: 07 November 2012 . Published Online: 17 December 2012

Guido Tiana, and Carlo Camilloni

Ratcheted molecular-dynamics simulations identify efficiently the transition state of protein folding

Guido Tiana[1,a)] and Carlo Camilloni[2,b)]
[1]*Department of Physics, University of Milano, and INFN, via Celoria 16, 20133 Milan, Italy*
[2]*Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom*

(Received 5 July 2012; accepted 7 November 2012; published online 17 December 2012)

The atomistic characterization of the transition state (TS) is a fundamental step to improve the understanding of the folding mechanism and the function of proteins. From a computational point of view, the identification of the conformations that build out the transition state is particularly cumbersome, mainly because of the large computational cost of generating a statistically sound set of folding trajectories. Here we show that a biasing algorithm, based on the physics of the ratchet-and-pawl, can be used to approximate efficiently the transition state. The basic idea is that the algorithmic ratchet exerts a force on the protein when it is climbing the free-energy barrier, while it is inactive when it is descending. The transition state can be identified as the point of the trajectory where the ratchet changes regime. Besides discussing this strategy in general terms, we test it within a protein model whose transition state can be studied independently by plain molecular dynamics simulations. Finally, we show its power in explicit-solvent simulations, obtaining and characterizing a set of transition-state conformations for Acyl-Coenzyme A-Binding Protein (ACBP) and Chymotrypsin Inhibitor 2 (CI2). © *2012 American Institute of Physics*. [http://dx.doi.org/10.1063/1.4769085]

I. INTRODUCTION

The transition state of biomolecular processes is particularly important because it is the main determinant of the associated rate. Unfortunately, being the most unstable state of the process, its characterization is difficult. In the case of protein folding, a perturbative technique where one measures the effect of amino-acid mutations on folding/unfolding rates, has been successful in providing a structural characterization of this evanescent state.[1] Although this procedure is experimentally rather demanding, we have now information about the structure of the transition state of tens of proteins.

With the improvement of the molecular force fields[2,3] that describes the interaction in proteins, it becomes more interesting the attempt to characterize the folding transition state without employing experimental information.[4–7] Within this context, the determination of the transition state implies two challenging problems, namely the generation of folding trajectories and the identification of the transition state along each of them. Concerning the former, the most straightforward way is simply to perform molecular dynamics (MD) simulations solving the equation of motion of the system. In the case of proteins of realistic size and using realistic force fields in explicit solvent, generating a statistically sound number of folding trajectories is not trivial even if one can use the fastest computers available.[8,9] Smarter techniques, comprising transition path sampling,[10] milestoning,[11] bias annealing,[12] and dominant reaction pathways[13,14] exploit the fact that only a small subset of all possible trajectories is statistically relevant, but these methods are computationally efficient when the total number of atoms is not large (typically in implicit-solvent models). Even if one can generate efficiently folding trajectories, the problem of identifying the transition state is still hard. The transition state between two (meta)stable states is built out of the set of conformations for which the probability of falling down to each of them is 1/2. Consequently, the most direct way to identify the transition state is to start several MD simulations from each of the conformations selected from a folding trajectory, and to count the fraction of such trajectories which meet the native state before meeting the denatured state (or vice versa), until this fraction is exactly 1/2.[15] This procedure is very time consuming, but is the only safe way to identify the transition state.[16]

Some years ago, Marchi and Ballone introduced the idea of biasing MD simulations to generate efficiently trajectories between conformations of a system, using an algorithm based on the physics of the ratchet-and-pawl.[17] It consists in defining a ratcheting coordinate y and dumping the thermal fluctuations along the direction of y opposite to the wished target. The algorithm was later used to enhance the thermal unfolding of proteins interacting with an implicit-solvent force field.[18,19] Recently, ratcheted MD simulations were used to repeatedly simulate the folding of single-domain proteins in explicit solvent.[20] Using a simplified protein model, a Beccara *et al.*[21] employed a Onsager–Machlup functional and showed that the ratcheted MD algorithm produces trajectories that are overall statistically relevant, thus validating the approach of Ref. 20.

In what follows, we will investigate whether it is possible to use ratcheted MD simulations to obtain directly and efficiently a good approximation of the conformations which build out the transition state of protein folding. The basic idea is that while climbing the main free-energy barrier

a)Electronic mail: tiana@mi.infn.it.
b)Electronic mail: cc536@cam.ac.uk.

which separates the denatured from the folded state, the ratchet exerts work, while descending on the other side it is essentially off. The transition between the two regimes marks the transition state.

Although the whole goal of this work is to develop a method that can be used for realistic systems in explicit solvent, we first validate it using a model whose folding trajectories can be generated by plain MD and whose transition state can be obtained exactly using the committors method of Ref. 15.

II. THE MODEL AND THE SIMULATIONS

A model which is suitable for developing a computational strategy and validating it against the transition state obtained with the exact method is a modified all-atom Gō model, where a non-specific interaction between hydrophobic atoms is added on the top of the native structure. The Gō model assures that folding can be simulated repeatedly also without the ratcheting algorithm, in order to be able to obtain reference trajectories, while the hydrophobic interaction makes the energy landscape more roughed, and thus more realistic. The Gō implementation is that of Ref. 22, in which pairs of atoms building native contacts interact with a Lennard–Jones potential whose minimum lies at $\epsilon_0 = -0.62$ (in arbitrary units). The hydrophobic potential has also the Lennard–Jones form and acts between side chain carbons of ALA, VAL, LEU, ILE, PHE, and TRP. The minimum of the potential lies at a distance of 0.35 nm, where the depth is $\epsilon_{hy} = -0.3$. This value of ϵ_{hy} has been chosen because it is the lowest which guarantees the folding of the proteins studied within an root mean square deviation (RMSD) of 0.3 nm from the experimental native conformation.

The simulations were carried out with a modified[23,24] version of Gromacs,[25] using the topologies generated with the SMOG web server.[26] The time step used is 0.002 ps (time units are merely nominal).

The specific heat, calculated with parallel-tempering simulations,[27] is displayed in Fig. 1 and compared with that for a plain Gō model. As expected, the two-states character of the denaturation transition diminished. However, the stability of the native state increased, suggesting that the hydrophobic interaction introduced in the model favors the native conformation, where hydrophobic packing is optimized, more than the denatured state. On the basis of this specific-heat plot, we used the trajectory obtained at $T = 1$ to generate 10 uncorrelated unfolded conformations to be used as initial states of the folding simulations. The folding simulations were carried out at $T = 0.91$, which is regarded as room temperature.

From each of the 10 unfolded conformations we carried out 10 simulations at $T = 0.91$ for 6 ns each. The average folding time, defined as the time needed to reach a RMSD of 0.4 nm, is $\tau_f = 1505$ ps.

Similar simulations were carried out using the ratcheting algorithm. The ratchet is implemented as in Ref. 20, that is adding to the molecular potential a ratcheting term

$$V_{rat}(\rho(t)) = \begin{cases} \frac{k}{2}\left(\rho(t) - \rho_m(t)\right)^2, & \rho(t) > \rho_m(t) \\ 0, & \rho(t) \le \rho_m(t), \end{cases} \quad (1)$$

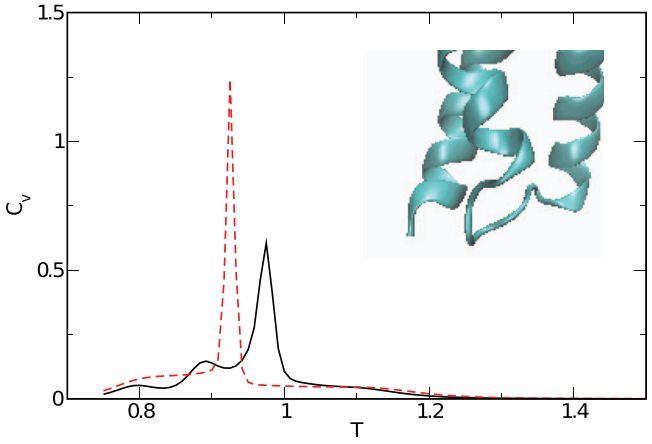

FIG. 1. The specific heat of ACBP (whose structure is displayed in the inset) as a function of temperature for the model interacting through the modified Go model (solid curve) and through a standard Go model (dashed curve). The modified Go model displays a transition between the native and the denatured state which is less cooperative than the standard Go model. The temperature is expressed in energy units.

where

$$\rho(t) = (y(t) - y_{target})^2 \quad (2)$$

and

$$\rho_m(t) = \min_{0\le\tau\le t} \rho(\tau). \quad (3)$$

The ratcheting coordinates $y(t)$ used in the present work are either the distance d_{CM} of the contact map of a given protein conformation from the native contact map, or the RMSD (in both cases $y_{target} = 0$). The distance d_{CM}, introduced by Bonomi *et al.*,[29] is defined as

$$d_{CM} = \|C - \tilde{C}\| = \left(\sum_{j>i+2}^{N} (C_{ij} - \tilde{C}_{ij})^2\right)^{1/2}, \quad (4)$$

were C_{ij} is the i,j element of a $N \times N$ matrix defined as

$$C_{ij}(r_{ij}) = \begin{cases} \frac{1-\left(\frac{r_{ij}}{r_0}\right)^p}{1-\left(\frac{r_{ij}}{r_0}\right)^q}, & r_{ij} \le r_{cut} \\ 0, & r_{ij} > r_{cut}, \end{cases} \quad (5)$$

r_{ij} is the distance between atom i and j and $\tilde{C}$ is the defined on the native state (corresponding to the experimental conformation, minimized according to the model potential). The parameters used in these simulations are $p = 6$, $q = 10$, $r_0 = 0.35$ nm, and $r_{cut} = 1.23$ nm.

Both in the case of plain MD and ratcheted simulations, the sequence of events along the folding trajectories under each set of conditions were studied calculating the matrix $M_{ij} = \theta(t(i,k) - t(j,k))$, where $t(i,k)$ is the time at which the ith contact is stably formed in the kth simulation and θ is the Heaviside's step function. This matrix satisfies $M_{ij} + M_{ji} = 1$ and each element M_{ij} assumes the value 1 if the formation of the ith contact precedes the formation of the jth, 0 if it follows it, and 1/2 if the two are uncorrelated. The average matrix

$$\overline{M_{ij}} = \frac{1}{n_s}\sum_{k=1}^{n_s} M_{ij}, \quad (6)$$

where $n_s = 100$ is the number of trajectories, is interpreted as the probability that the formation of the ith contact precedes the formation of the jth. A quantity related to $\overline{M_{ij}}$ is the probability $A_j = \sum_{i \neq j} \overline{M_{ij}}/(n_s - 1)$ that the jth contact is formed after any other contact.

The order of contact formation in two trajectories was compared using the distance

$$d(M, M') \equiv \frac{1}{n_s^2} \sum_{ij} [1 - \delta(M_{ij}, M'_{ij})], \tag{7}$$

between the associated matrices, where δ is the Kronecker symbol.

III. RATCHETED TRAJECTORIES

A necessary condition for the ratcheting algorithm to identify the correct transition state of folding is to generate statistically relevant trajectories. Failure of this condition would lead to the identification of free-energy saddle points not corresponding to the main transition state of the folding process. Ratchet-generated trajectories are not expected to be associated—as they are—to a large statistical weight, because the corresponding folding time lies in the low–probability initial region of the folding-time distribution. In fact, such trajectories run, in the space defined by the RMSD and d_{CM}, on the edge of the region where unbiased trajectory run (see Fig. S2 in the supplementary materials[28]). They are somewhat distorted along the direction in which the ratchet is acting, similarly to what found in Ref. 12. However, as suggested in Ref. 20 and validated in this section in the case of two model proteins, ratcheted MD simulations can provide the most probable sequence of contact formation if carried out in appropriate conditions. In this respect, ratcheted trajectories can be regarded as a coarse graining over time of the actual trajectories, which run parallel to the actual trajectories.

As a reference we generated 100 trajectories with plain MD simulations. The average folding time was 1505 ps and all trajectories reached the native conformations in the 10 000 ps made available for each of them. The mean distance $\overline{d}$ between each pair of matrix M_{ij} (cf. Eq. (7)) is 0.37, indicating that the sequence of events along the different folding trajectories are rather homogeneous (cf. Ref. 20). Briefly, this sequence implies first the formation of most contacts in the two terminal helices, than in the central helices and then the tertiary contacts.

Similar simulations were carried out starting from the same set of unfolded conformations, ratcheting the simulation along the distance d_{CM} of the contact map to the native one with different values of the ratcheting constant k. Not all the trajectories folded to the native conformation, but some of them got stuck, reducing drastically the diffusivity of the different parts of the protein and, essentially, freezing to non-native conformations. These are excluded from the analysis that follows. The fraction of stuck trajectories, displayed in the upper panel of Fig. 2, increases with k. The same figure also displays the average folding time, which decreases as the effect of the ratchet is increased. The average folding time of

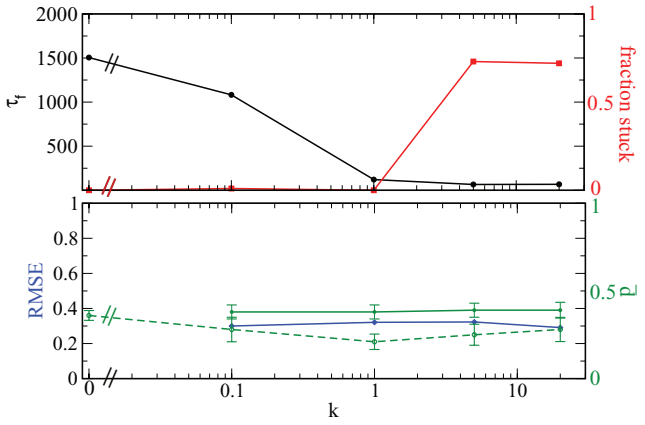

FIG. 2. Comparison of the folding simulations of ACBP ratcheted along d_{CM} at different values of k with those generated by plain MD, corresponding to $k = 0$. (upper panel) The average folding time (circles) and the fraction of stuck trajectories which are not able to reach the native state (squares), as a function of the ratcheting constant k. The latter is displayed in a logarithmic scale, except in the case of the points marked as 0, which identify the simulation carried out without ratcheting. (lower panel) The RMSE between the matrix $\overline{M_{ij}}$ calculated at k and that calculated at $k = 0$ (diamonds), the average distance $\overline{d}$ between the matrices M_{ij} calculated at k and those calculated at $k = 0$ (filled circles, the error bars indicate the standard deviation), and within the matrices M_{ij} calculated at k (empty circles).

ratcheted simulations has the only purpose of measuring the computational time needed to generate a folding trajectory, and has no physical meaning. The figure indicates that there is a range of values of k around unity where simulations generate fast trajectories to the native conformations.

To assess the physical meaning of such trajectories, we compared the order of native-contact formation to that of the unbiased simulations. The mean distance $\overline{d}$ between each pair of ratcheted trajectory is around 0.3 for any value of k and for the unbiased trajectories (see lower panel in Fig. 2), indicating that the sequence of events in the ratcheted simulations is as homogeneous as that of the MD trajectories. Also the mean distance between the matrices associated to ratcheted trajectories and those associated with plain MD trajectories is 0.39 at all the values of k considered. Comparing the average matrices $\overline{M_{ij}}$, one obtains that the root-mean-square error (RMSE) between the matrix $\overline{M_{ij}}$ generated ratcheting the simulations and with plain MD is around 0.3 for all values of k (cf. Fig. 2).

Summing up, the difference between ratcheted and plain MD trajectories is comparable with the (small) differences between pairs of plain MD trajectories. Even when the ratchet is strong, although the fraction of folding trajectories drops drastically, the sequence of events in the few folding trajectories results correct.

A similar analysis has been carried out ratcheting the simulation through a different coordinate, that is the RMSD with respect to the native conformation. Usually this is regarded as a bad reaction coordinate.[30] In fact, attempts to fold small proteins in explicit solvent ratcheting along the RMSD coordinate at different values of the ratcheting constant have failed.[20] The results of such simulations are displayed in Fig. 3. Also in this case folding simulations display a sequence of events that is similar to the one generated by

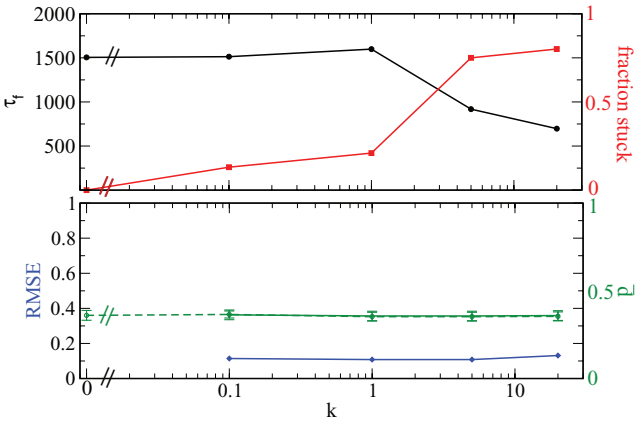

FIG. 3. Comparison of the folding simulations of ACBP ratcheted along the RMSD with those generated by plain MD. (upper panel) The average folding time (circles) and the fraction of stuck trajectories which are not able to reach the native state (squares), as a function of the ratcheting constant k. The latter is displayed in a logarithmic scale, except in the case of the points marked as 0, which identify the simulation carried out without ratcheting. (lower panel) The RMSE between the matrix $\overline{M_{ij}}$ calculated at k and that calculated at $k = 0$ (diamonds), the average distance $\overline{d}$ between the matrices M_{ij} calculated at k and those calculated at $k = 0$ (filled circles, the error bars indicate the standard deviation), and within the matrices M_{ij} calculated at k (empty circles). Here, the values of k are given in energy units divided by nm.

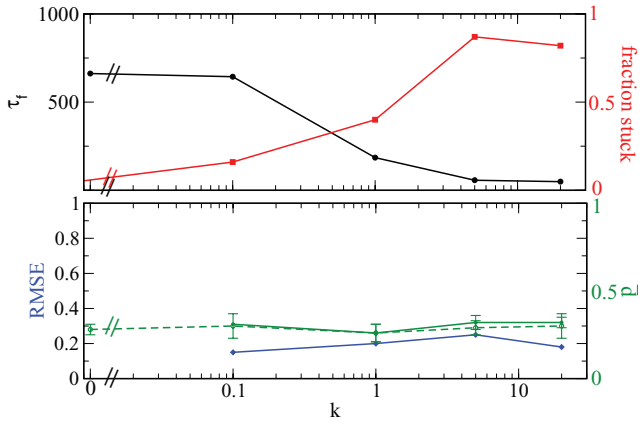

FIG. 4. Comparison of the folding simulations of CI2 ratcheted along d_{CM} with those generated by plain MD. (upper panel) The average folding time (circles) and the fraction of stuck trajectories which are not able to reach the native state (squares), as a function of the ratcheting constant k. The latter is displayed in a logarithmic scale, except in the case of the points marked as 0, which identify the simulation carried out without ratcheting. (lower panel) The RMSE between the matrix $\overline{M_{ij}}$ calculated at k and that calculated at $k = 0$ (diamonds), the average distance $\overline{d}$ between the matrices M_{ij} calculated at k and those calculated at $k = 0$ (filled circles, the error bars indicate the standard deviation), and within the matrices M_{ij} calculated at k (empty circles).

unbiased simulations. The main difference with the data obtained ratcheting along d_{CM} is that in the present case there is not a range of values of k at which ratchet is efficient. At small values of k the folding time τ_f is essentially identical to that of the unbiased MD simulations. Only using values of k larger than 10 one can observe a relevant decrease of τ_f, but here the fraction of folding sequences has become negligible. However, the sequence of events results similar to that of unbiased simulations, as shown in Fig. 3 by the values of RMSE and average distance to the unbiased matrices. Consequently it seems that, at least within this simplified model, ratcheting along the RMSD is less efficient, but results in trajectories which are not worse than those obtained ratcheting on d_{CM}.

ACBP is considered to fold according to a hierarchical diffusion-collision model,[31] where first elements of secondary structure are formed, then diffuse around until they bind together to form native tertiary contacts. This pattern, which is also observed in the present simulations, could favor the applicability of the ratchet. To check the generality of the above results we have tested it with another case, namely CI2, which is considered the prototype of proteins which fold according to a nucleation model, without populating consistently secondary structures prior to the transition state. The results are displayed in Fig. 4. Also in this case there is a range of k where the ratchet is efficient, that is both the folding time and the fraction of stuck trajectories are small. The efficiency is smaller than in the case of ACBP, probably because in this case the contact-map distance d_{CM} is not a reaction coordinate as good as for a protein folding through a diffusion-collision scenario. Anyway, the sequence of events results to agree with that of a plain MD simulation, within the range of variability of the latter (which is somewhat small than that of ACBP).

IV. IDENTIFICATION OF THE TRANSITION STATE: THE STRATEGY

The analysis of the time-dependence of the degrees of freedom associated with the ratchet can provide some information to localize the transition state of the system. The basic idea is that, as the system climbs the free energy barrier whose top is the transition state, the ratchet is very active and thus V_{rat} is well above zero. When the system crosses the transition state and descends the free-energy barrier, the ratchet is essentially inactive and V_{rat} small. The point of the trajectory where V_{rat} drops is hypothesized to be the transitions state.

Before verifying this hypothesis, we attempt to formalize the above idea, in a simple scenario where the molecular force can be approximated in an elementary form. Assuming that the dynamics of the degrees of freedom $\vec{x}$ of the system can be described by an over-damped dynamics

$$\frac{d\vec{x}}{dt} = \frac{1}{\gamma}[\vec{f}(\vec{x}) - k\Delta\rho \cdot \vec{u}_\rho + \vec{\eta}], \tag{8}$$

where γ is the friction coefficient; η is the thermal noise satisfying $\langle\vec{\eta}(t) \cdot \vec{\eta}(t')\rangle = (2NT\gamma)\delta(t - t')$; $\vec{u}_\rho$ the versor that defines the direction of the ratcheting coordinate ρ; $\Delta\rho(t) \equiv \rho - \rho_m$ is the difference between the value of the ratcheting coordinate and its minimum; and Boltzman's constant is set to 1. Let us assume that ρ is a good reaction coordinate, that is it moves according to the slowest time scale of the system,[32] and that the associated diffusion constant is approximately equal to that of the microscopic degrees of freedom. Then, the dynamics of ρ can be written as

$$\frac{d\rho}{dt} = \frac{1}{\gamma}[f_\rho - k\Delta\rho + \eta], \tag{9}$$

where f_ρ is the effective force which moves the one-dimensional degree of freedom ρ (i.e., minus the gradient of the free energy). By virtue of its definition, ρ_m follows

the dynamics

$$\frac{d\rho_m}{dt} = \frac{d\rho}{dt}\delta(\Delta\rho)\theta(-d\rho/dt), \tag{10}$$

where θ is a step function that is 1 if its argument is positive and 0 otherwise. Consequently the quantity $\Delta\rho$ which measures the activity of the ratchet follows:

$$\frac{d(\Delta\rho)}{dt} = \begin{cases} \frac{1}{\gamma}[f_\rho - k\Delta\rho + \eta] & \text{if } \Delta\rho > 0 \text{ or } d\rho/dt > 0 \\ 0 & \text{if } \Delta\rho = 0 \text{ and } d\rho/dt < 0. \end{cases} \tag{11}$$

If the molecular force $\vec{f}$ pushes the system downhill towards its target state and it is overwhelming with respect to the typical diffusive force (i.e., $f_\rho \ll -(2T\gamma/\Delta t)^{1/2}$), then $\Delta\rho$ is approximately zero along the associated part of trajectory.

A more common scenario is that where the system is running downhill, the diffusive term is not negligible, but the molecular force is overwhelming with respect to the ratchet (i.e., $f_\rho \ll -k\Delta\rho$), so that we can neglect the term proportional to k in Eq. (11). To make things simple, let's focus on a fraction of the trajectory that is short enough that the force f_ρ can be approximated as constant. In this case $\Delta\rho$ experiences a diffusion biased by a constant force, and by a trap at 0. In fact, when $\Delta\rho = 0$ the system can move away only if $d\rho/dt > 0$ (cf. the condition controlling the first line of Eq. (11)); thus the exit rate w_{exit} from the trap is proportional to the probability that $\eta > |f_\rho|(\Delta t/2T\gamma)^{1/2}$. This case is analogous to that of a massive particle diffusing on a slope with a trap at the bottom. Since the stochastic noise η is normally distributed, w_{exit} is proportional to $\mathrm{erfc}(|f_\rho|(\Delta t/2T\gamma)^{1/2})$. We can assign to the trap an effective energy $-T\log(\frac{1}{2}\mathrm{erfc}[|f_\rho|(\frac{\Delta t}{2T\gamma})^{1/2}])$, so that w_{exit} is equal to Kramers escape rate. In the neighborhood of 0, $\Delta\rho$ will soon populate a distribution given by

$$p(\Delta\rho) = \begin{cases} \frac{2}{Z}\mathrm{erfc}\left[|f_\rho|\left(\frac{\Delta t}{2T\gamma}\right)^{1/2}\right]^{-1} & \text{if } 0 < \Delta\rho < \epsilon \\ \frac{1}{Z}\exp\left[-\frac{|f_\rho|\Delta\rho}{T}\right] & \text{if } \Delta\rho > \epsilon, \end{cases} \tag{12}$$

where ϵ is the (small) length which defines the trap and

$$Z = 2\cdot\mathrm{erfc}\left[|f_\rho|\left(\frac{\Delta t}{2T\gamma}\right)^{1/2}\right]^{-1} + \frac{T}{|f_\rho|\epsilon}. \tag{13}$$

The average value of $\Delta\rho$ expected in this regime is then

$$\langle\Delta\rho\rangle = \frac{T^2}{\epsilon f_\rho^2\,\mathrm{erfc}[|f_\rho|(\Delta t)/2T\gamma)^{1/2}]^{-1} + |f_\rho|T} \xrightarrow{\epsilon\to 0} \frac{T}{|f_\rho|}. \tag{14}$$

On the other hand, if the system is climbing the free-energy barrier (i.e., $f_\rho \gg (2T\gamma/\Delta t)^{1/2}$), the conditions $\Delta\rho = 0$ and $d\rho/dt < 0$ in Eq. (11) are never satisfied simultaneously. Consequently,

$$p(\Delta\rho) = \frac{1}{Z}\exp\left[-\frac{\frac{k}{2}\Delta\rho^2 - f_\rho\Delta\rho}{T}\right], \tag{15}$$

where

$$Z = \left(\frac{\pi T}{2k}\right)^{1/2}\exp\left[\frac{f_\rho^2}{2kT}\right]\left(1+\mathrm{erf}\left[\frac{f_\rho}{(2kT)^{1/2}}\right]\right), \tag{16}$$

giving the average

$$\langle\Delta\rho\rangle = 2\frac{(2/\pi)^{1/2}Tk\exp\left[-\frac{f_\rho^2}{2kT}\right] + f_\rho(Tk)^{1/2}\left(1+\mathrm{erf}\left[\frac{f_\rho}{(2kT)^{1/2}}\right]\right)}{(k^3T)^{1/2}\left(1+\mathrm{erf}\left[\frac{f_\rho}{(2kT)^{1/2}}\right]\right)} \xrightarrow{f_\rho^2\gg 2kT} \frac{2f_\rho}{k}. \tag{17}$$

The transition state is the intermediate scenario where f_ρ vanishes. Assuming $|f_\rho| \ll (2T\gamma/\Delta t)^{1/2}$, one can neglect the molecular force in Eq. (11), obtaining

$$p(\Delta\rho) = \begin{cases} 2/Z & \text{if } 0 < \Delta\rho < \epsilon \\ \frac{1}{Z}\exp\left[-\frac{k\Delta\rho^2}{2T}\right] & \text{if } \Delta\rho > \epsilon, \end{cases} \tag{18}$$

where $Z = 2 + (\pi k/2T)^{1/2}/\epsilon$. This is an ideal distribution, because it is unlikely that the system spends enough time at the transition state to populate it. However, it can be useful to obtain the average $\Delta\rho$ which separates the rising from the descending regime. In fact, we get

$$\langle\Delta\rho\rangle = \frac{2T}{4k\epsilon + (2\pi kT)^{1/2}} \xrightarrow{\epsilon\to 0} \left(\frac{2T}{\pi k}\right)^{1/2}. \tag{19}$$

The behavior of $\Delta\rho$ in a typical ratcheted simulation is displayed in the middle panel of Fig. 5. Although it is difficult to distinguish *a priori* where the system is climbing and where it is descending the folding free-energy barrier, it is reasonable to argue that in part of the trajectory in the range $30 < t < 220$ the system is climbing, while in the range $220 < t < 250$ it is descending. The distribution $p(\Delta\rho)$ associated with these two parts of the trajectory are displayed in Fig. 6 with solid black and red curves, respectively. The black curve is fitted by Eq. (15), the correlation coefficient being 0.958. The red curve displays a sharp peak at low values of $\Delta\rho$ as predicted by the first line of Eq. (12), allowing to obtain $\epsilon = 0.03$, while the remaining part is fitted by the second line of Eq. (12), with a correlation coefficient of 0.965. This means that, although the molecular force f_ρ certainly depends on the specific point of the trajectory, the system crosses the free-energy barrier experiencing an effective force of $f_\rho = 0.75$ and descend it pushed by an effective force $f_\rho = -1.45$.

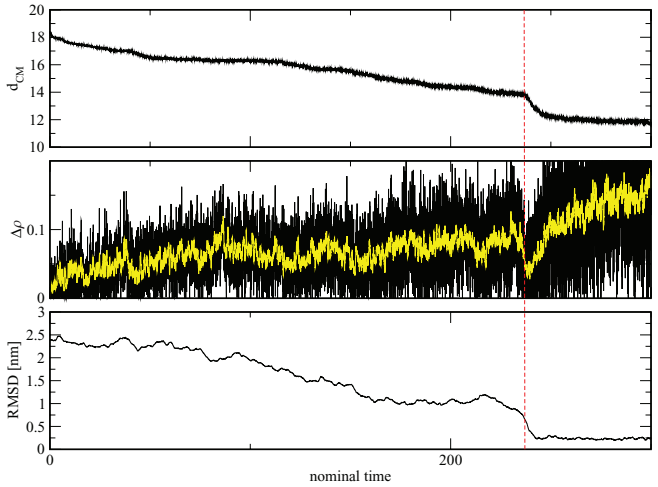

FIG. 5. The typical behavior of the contact-map distance to the native conformation d_{CM}, the ratchet displacement $\Delta\rho$ and the RMSD to the native conformation in a folding trajectory ratcheted with $k = 1$. The light curve in the middle panel is a 0.8-ps running average of the underlying curve. The vertical dashed bar marks the TS identified in they trajectory.

The value of $\langle\Delta\rho\rangle$ obtained in Eq. (19) can be used to estimate the order of magnitude of the threshold to distinguish the regime where the system is climbing the free-energy barrier from that in which it is descending, that is the transition state. For example, in the simulation we performed with $T = 0.91$ and $k = 1$, we obtain $\langle\Delta\rho\rangle = 0.076$ (cf. Fig. 5).

V. IDENTIFICATION OF THE TRANSITION STATE: RESULTS

Before applying the strategy discussed above, the actual TS was identified through a commitment analysis[15] on 10 plain MD folding trajectories of ACBP. From each of them we extracted a variable number (from 5 to 10) of conformations chosen in the region where the value of d_{CM} displays a rapid

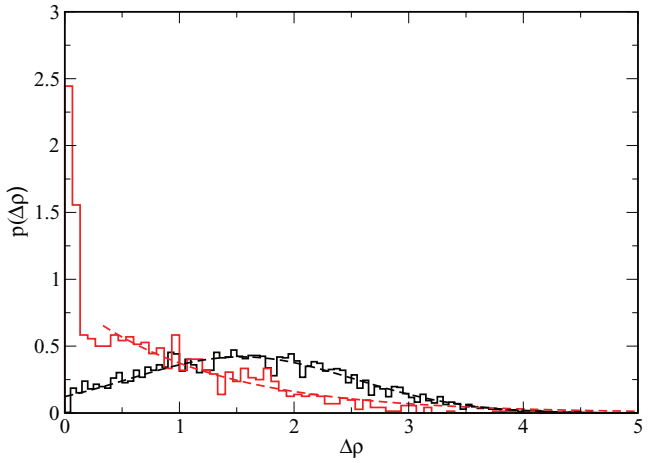

FIG. 6. The histogram of $\Delta\rho$ obtained from the parts of the trajectory of Fig. 5 which ranges between nominal times 30 and 220 ps, presumably corresponding to the climbing of the folding free-energy barrier (black solid curve) and between 220 and 250 ps, presumably corresponding to the descent to the native state (red solid curve). The fit obtained by Eq. (15) (dashed black curve) and (12) (dashed red curve) are also indicated.

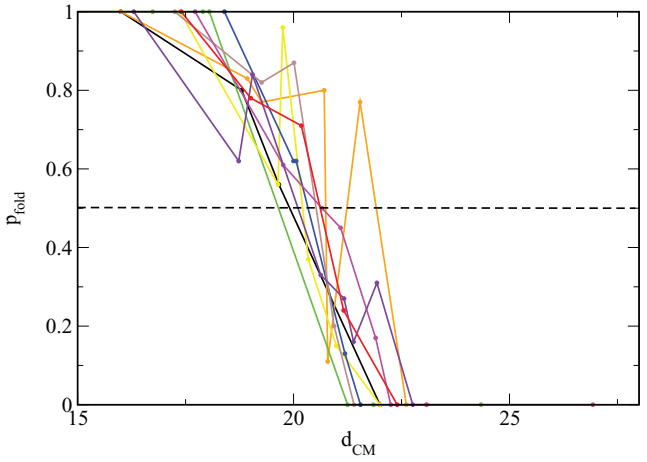

FIG. 7. The folding probability p_{fold}, that is the probability of reaching the native state before reaching the denatured state, calculated initiating unbiased simulations from a set of conformations extracted from 10 folding simulations. The conformations displaying $p_{fold} = 0.5$ build out, by definition, the folding transition state.

decrease to low values. From each of them we started 100 plain MD simulations, calculating the probability p_{fold} that the simulation reaches the native basin (operatively defined from $d_{CM} < 19$) before reaching the denatured basin (operatively defined from $d_{CM} > 25$). The conformations displaying $0.4 < p_{fold} < 0.6$ are defined as TS conformations. The behavior of p_{fold} with respect to the value of d_{CM} of the associated conformation is displayed in Fig. 7. The associated conformations are displayed in Fig. 8(a). They are remarkably native-like, displaying an average RMSD to the native conformation of 0.68 ± 0.17 nm, and fairly homogeneous, their mutual average RMSD being 0.85 ± 0.17 nm.

For each trajectory generated with the ratcheting algorithm, we have looked for the TS in the region where the RMSD to the native conformation was in the range between 0.2 nm and 1 nm. The putative TS is the conformation such that the average value of $\Delta\rho$ in the preceding 8 ps is larger than that predicted by Eq. (19) and in the following 0.8 ps is smaller. In this way, we could identify a conformation in 64% of trajectories at $k = 0.1$, in the 86% of the trajectories at $k = 1$ and in the 49% of trajectories at $k = 20$. In no cases more than one conformation is identified. Although the fluctuations in $\Delta\rho$ are large, the fact that the TS, as identified by these algorithmic rules, separates a region where the average $\Delta\rho$ displays a local maximum to that in which it displays a minimum is apparent to the naked eye (cf. the yellow curve, indicating a running average, in Fig. 5. See also Fig. S3 in the supplementary material[28]).

After having overcome the TS and reached the native state (i.e., in the region where the RMSD is smaller than ≈0.3 Å), the values of $\Delta\rho$ climbs again to large values because the systems tries to reach the very conformation which we set as target of the ratchet, which is structurally similar, but not precisely the free-energy minimum of the system at the temperature at which we carry out the simulations (cf. Fig. S1 in the supplementary materials[28]).

Note that there are also other points of the trajectory, besides the TS, in which the average $\Delta\rho$ is close to zero. These

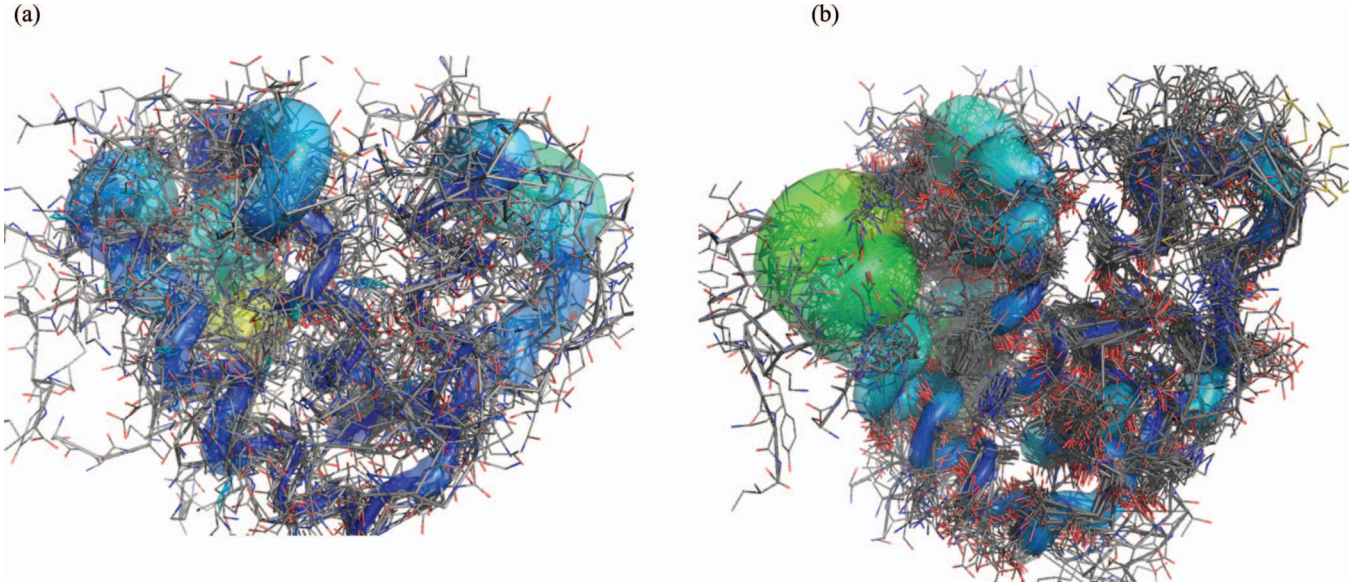

FIG. 8. Comparison between the Go-model transition-state conformations for ACBP obtained by plain MD simulations through the commitment analysis (a) and those obtained by ratcheted simulations as explained in the text (b). The width and the color of the average conformations denote the RMS fluctuations.

corresponds to crossings of free-energy barriers in the high-energy region of the rough energy landscape of the protein. However, these points correspond to large values of RMSD and d_{CM}, and consequently cannot be confused with the main TS.

The structural properties of the conformations identified by the above criteria are summarized in Fig. 9. The average contact–map distance is comparable to that of the actual TS at all values of k. The structural homogeneity of the TS conformations is slightly decreasing with the increasing of k, the mutual average RMSD going from 0.85 nm at $k = 0$ to 0.61 nm at $k = 20$. The average similarity of the TS conformations

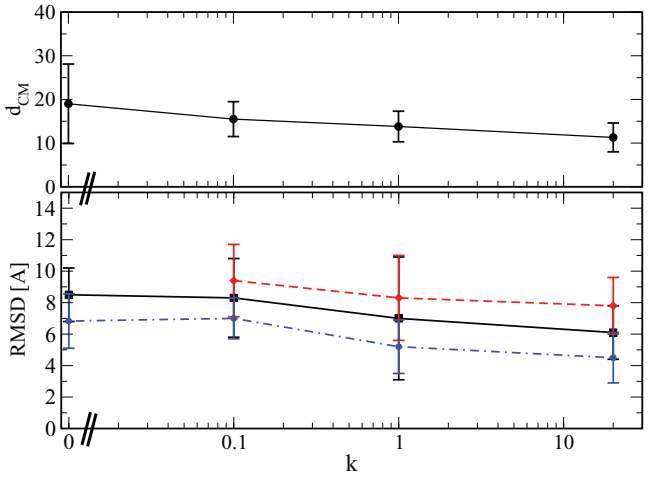

FIG. 9. Structural properties of transition-state conformations obtained at various values of k and obtained from plain MD simulations ($k = 0$). (upper panel) The average distance between the contact map of TS conformations and that of the native state. (lower panel) The average RMSD between pairs of TS conformations at each value of k (black squares), the average RMSD between TS conformations obtained at different values of k and those obtained by plain MD simulations (red diamonds) and average RMSD to the native conformation (blue circles). The error bars indicate the standard deviation.

obtained from ratcheted simulations to the actual TS conformations is within the error bars σ associated with the intrinsic variability of the TS conformations (the difference between the two averages being $\approx 0.2\sigma$; black error bars in the figure). Also the RMSD to the native conformation displays a slight decrease from 0.68 nm at $k = 0$ to 0.46 nm at $k = 20$. Summing up, at all values of k analyzed, ratcheted MD simulations can identify TS conformations which are structurally similar to the actual TS conformations, becoming slightly more native-like at increasing k.

A representation of the protein in the TS obtained at $k = 1$ is displayed in Fig. 8(b). The main differences between the actual TS and that obtained by ratcheted conformation at $k = 1$ involves the terminals of the protein. The actual TS displays large fluctuation in the C-terminal part of the chain and, to a smaller extent, in the N-terminal and in the loop region. The ratcheted TS overestimates the fluctuations in the C-terminal region, while it slightly underestimates those involving the loop. Anyway, the two sets are remarkably similar.

Overall, the TS conformations obtained from the modified Go model are strikingly native-like. This is not unrealistic, since a similar result was obtained for a number of proteins by Paci *et al.*[6,7] using not a Go model, but a direct modelization of experimental phi-values.

The folding probabilities p_{fold} obtained from 10 unbiased simulation starting from some of the putative TS find above are displayed in Table I. The intermediate values found for p_{fold} indicate that these conformations behave as actual TS. However, a closer look at the unbiased folding trajectories (see Fig. S5 in the supplementary materials[28]) show that the protein, starting from the putative TS, first moves closer to the actual TS (i.e., those found by the very definition of TS) and then moves towards either the native or the denatured state. Consequently, putative TS are structurally very similar to the actual TS, but lie at slightly higher energy, in the direction perpendicular to the main folding path (cf. also Fig. S2 in the supplementary materials[28]).

TABLE I. The result of commitment analysis carried out on TS conformations obtained from the trajectories generated ratcheting along d_{CM} and along the RMSD.

Ratchet on d_{CM}			Ratchet on RMSD		
d_{CM}	RMSD	p_{fold}	d_{CM}	RMSD	p_{fold}
14.1	0.80	0.4	13.8	0.69	0.6
13.4	0.71	0.3	13.1	0.61	0.4
13.1	0.74	0.6	12.4	0.51	0.6
13.3	0.57	0.6	14.3	0.47	0.7
12.5	0.40	0.7	13.1	0.40	0.7

A similar procedure can be applied on trajectories generated ratcheting along the RMSD, and provides TS conformations which are comparable to those obtained ratcheting along d_{CM} (cf. Fig. S4 in the supplementary materials[28] and Table I). The similarity between the TS conformations obtained ratcheting along two different directions further testifies to the correctness of the result.

Another example is the structural characterisation of the TS of ACBP under an external force,[33] where the transition state in explicit solvent found with the present algorithm is consistently less native-like.

VI. AN EXPLICIT-SOLVENT CASE: THE TRANSITION STATE OF ACBP AND CI2 SIMULATED WITH THE AMBER FORCE FIELD

The very goal of the strategy discussed above is not the identification of the transition state with simplified protein descriptions, but in realistic explicit-solvent models. In order to test the algorithm we analyzed the folding and unfolding trajectories generated using the ratcheting algorithm in Ref. 20. Using the Amber03 force field,[34] we simulated 10 folding and 10 unfolding trajectories of ACBP and CI2 in a dodecahedron box of 261 nm^3 solvated with $\sim 10^4$ TIP3P water molecules, ratcheting along d_{CM} with a ratcheting constant $k = 1$ kJ/mol for 50 ns at $T = 300K$. All trajectories folded within 0.25 nm from the native conformation.

The transition state is identified with the same strategy used in the Gō-model simulations, requiring that the average of $\Delta\rho$ in the preceding 8 ps is larger than 1 and in the following 8 ps is smaller than 2 (this is a somewhat looser condition than for the Gō model, but guarantees the identification of a unique TS for each trajectory), while the RMSD to the native state should range between 0.3 and 1 nm. The conformation thus obtained are displayed in Fig. 10. They are less homogeneous than those obtained by means of the Gō model, the average mutual RMSD being 0.82 ± 0.19 nm in the case of ACBP and 0.76 ± 0.14 nm in the case of CI2. Their RMSD to the native state is 0.68 ± 0.19 nm in the case of ACBP and 0.70 ± 0.17 nm in the case of CI2.

In order to validate the TS without carrying out a commitment analysis which is extremely time-consuming in explicit solvent,[35] we have compared the TS conformations obtained from the folding trajectories to the TS conformations obtained by unfolding trajectories under the same conditions. According to the principle of detailed-balance, under the same conditions the two TS must be identical.[36] The TS conformations obtained in this case are slightly more native-like, displaying a RMSD to the native conformation of 0.43 ± 0.13 nm for ACBP and 0.62 ± 0.07 nm for CI2. In order to compare the set of TS conformations obtained from folding and from unfolding trajectories, we have calculated the average pairwise RMSD of conformations across the two sets, which is 0.81 ± 0.10 nm for ACBP and 0.76 ± 0.13 nm for CI2.

The average similarity between the folding and the unfolding TS is compatible, within the error bars, to the intrinsic heterogeneity of each set (their difference is 0.05σ in the case of ACBP and 0 in the case of CI2), and so guarantees that the two TS can be regarded as approximatively identical.

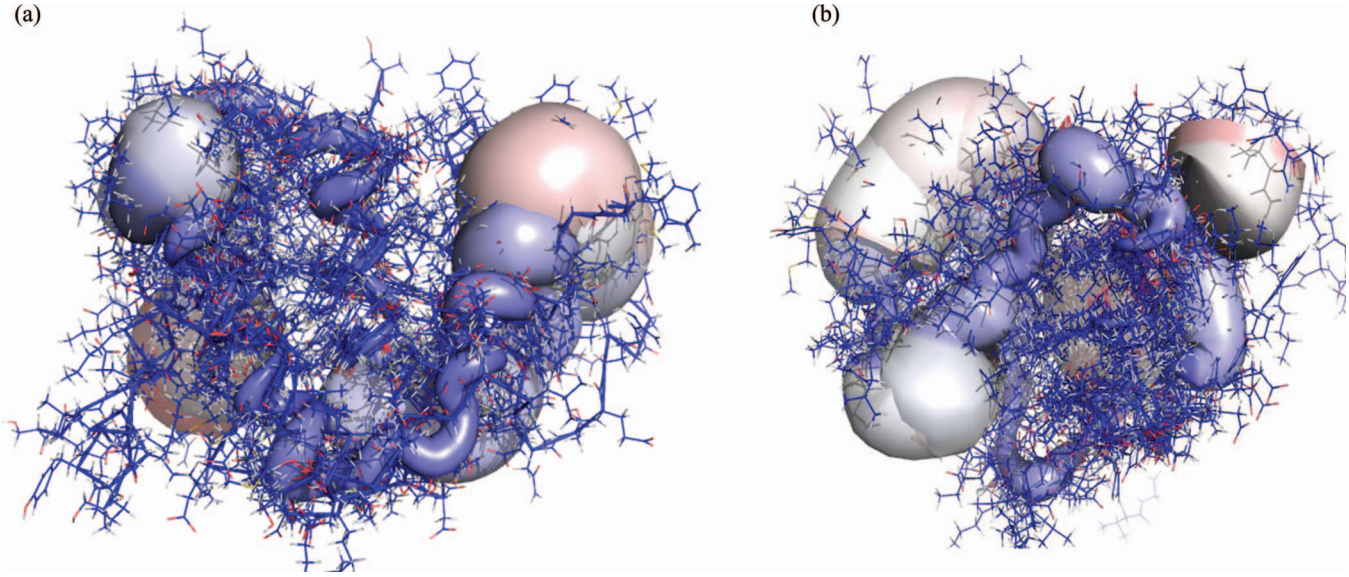

FIG. 10. The conformations corresponding to the transition state of ACBP (a) and CI2 (b) from the explicit-solvent ratcheted simulations. The thickness of the surface indicates the standard deviation associated to the average structure.

VII. CONCLUSIONS

The complexity of the characterisation of biomolecular processes is driving a continuos improvement of the experimental and the computational techniques.[37] In particular, in the field of computer simulations, in the last few years we have observed a leap in the accessible time scale of plain MD simulations. Nonetheless, even these major improvements are not able to address the complexity of folding problem for realistic proteins.[35] This points to the necessity of carrying on with the development of both simplified model and advanced sampling methods. The present work further validates the use of the ratcheting algorithm in the study of protein folding and extends its use to the approximate identification, at an atomic level, of the transition state ensemble of a protein in explicit solvent.

Specifically, protein folding simulated by the ratcheting algorithm display the same sequence of contacts formation of unbiased simulations, generating trajectories that are "parallel" in conformational space to the unbiased ones, displaying a small shift in the direction of the ratcheting coordinate. The algorithm is also able to provide, at no additional cost, a good approximation (within an RMSD of few Angstrom) of the structure of the protein in the transition state. The putative transition state identified by the ratchet is slightly shifted high in energy with respect to the actual transition state, in the direction perpendicular to the main folding path.

The above results have been obtained with a simplified model and ratcheting along what seems a good reaction coordinate for this specific system, namely the fraction of native contacts. Our results suggest that worsening the quality of the ratcheting coordinate makes less efficient the generation of folding trajectories (many of the get stuck in denatured conformations), but does not affect consistently the identification of the folding sequence of events and of the transition state.

The ratcheting algorithm allows to generate efficiently folding simulations even in the case of explicit-solvent models. In this case, due to the complexity of the system, it is less obvious that the fraction of native contacts is a good reaction coordinate. Nonetheless, it allows to generate folding trajectories of small-size proteins in few days of calculations on a desktop computer, and to identify a putative transition state. Of course, it is not trivial to validate this transition state by a commitment analysis. Anyhow, the similarity between the TS found along folding and unfolding trajectories gives an indirect argument for the validity of the method also in this case.

ACKNOWLEDGMENTS

C.C. was supported by a Marie Curie Intra-European fellowship. We acknowledge the use of computing facilities provided by CamGrid.

[1] A. R. Fersht, *Structure and Mechanism in Protein Science* (W. H. Freeman and Co., 2002).
[2] K. A. Beauchamp, Y.-S. Lin, R. Das, and V. S. Pande, J. Chem. Theory Comput. **8**, 1409 (2012).
[3] K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw, PLoS ONE **7**, e32131 (2012).
[4] N. Calosci, C. N. Chi, B. Richter, C. Camilloni, Å. Engström, L. Eklund, C. Travaglini-Allocatelli, S. Gianni, M. Vendruscolo, and P. Jemth, Proc. Natl. Acad. Sci. U.S.A. **105**, 19241 (2008).
[5] C. D. Geierhaas, X. Salvatella, J. Clarke, and M. Vendruscolo, Protein Eng. **21**, 215 (2008).
[6] E. Paci, M. Vendruscolo, C. M. Dobson, and M. Karplus, J. Mol. Biol. **324**, 151 (2002).
[7] M. Vendruscolo, E. Paci, and C. M. Dobson, Nature (London) **409**, 641 (2001).
[8] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, Science **334**, 517 (2011).
[9] R. Best, Curr. Opin. Struct. Biol. **22**, 52 (2012).
[10] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, Annu. Rev. Phys. Chem. **53**, 291 (2002).
[11] A. K. Faradjian and R. Elber, J. Chem. Phys. **120**, 10880 (2004).
[12] J. Hu, A. Ma, and A. R. Dinner, J. Chem. Phys. **125**, 114101 (2006).
[13] P. Faccioli, M. Sega, F. Pederiva, and H. Orland, Phys. Rev. Lett. **97**, 108101 (2006).
[14] P. Faccioli, J. Phys. Chem. B **112**, 13756 (2008).
[15] P. Geissler, C. Dellago, and D. Chandler, J. Phys. Chem. B **103**, 3706 (1999).
[16] V. Pande, A. Grosberg, T. Tanaka, and E. Shakhnovich, J. Chem. Phys. **108**, 334 (1998).
[17] M. Marchi and P. Ballone, J. Chem. Phys. **110**, 3697 (1999).
[18] E. Paci and M. Karplus, J. Mol. Biol. **288**, 441 (1999).
[19] E. Paci and M. Karplus, Proc. Natl. Acad. Sci. U.S.A. **97**, 6521 (2000).
[20] C. Camilloni, R. A. Broglia, and G. Tiana, J. Chem. Phys. **134**, 045105 (2011).
[21] S. A. Beccara, T. Skrbic, R. Covino, and P. Faccioli, Proc. Natl. Acad. Sci. U.S.A. **109**, 2330 (2012).
[22] P. C. Whitford, J. K. Noel, S. Gosavi, A. Schug, K. Y. Sanbonmatsu, and J. N. Onuchic, Proteins **75**, 430 (2009).
[23] C. Camilloni, D. Provasi, G. Tiana, and R. A. Broglia, Proteins **71**, 1647 (2008).
[24] M. Bonomi, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Pietrucci, and R. A. Broglia, Comput. Phys. Commun. **180**, 1961 (2009).
[25] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, J. Chem. Theory Comput. **4**, 435 (2008).
[26] J. K. Noel, P. C. Whitford, K. Y. Sanbonmatsu, and J. N. Onuchic, Nucleic Acids Res. **38**, W657 (2010).
[27] U. Hansmann, Chem. Phys. Lett. **281**, 140 (1997).
[28] See supplementary material at http://dx.doi.org/10.1063/1.4769085 for the figures displaying the free energy surface of the Gō model; the effect of the ratchet on the trajectories; and more results about the transtition state of the Gō model.
[29] M. Bonomi, F. L. Gervasio, G. Tiana, D. Provasi, R. A. Broglia, and M. Parrinello, Biophys. J. **93**, 2813 (2007).
[30] R. Best and G. Hummer, Proc. Natl. Acad. Sci. U.S.A. **107**, 1088 (2010).
[31] B. B. Kragelund, P. Osmark, T. B. Neergaard, J. Schiødt, K. Kristiansen, J. Knudsen, and F. M. Poulsen, Nat. Struct. Biol. **6**, 594 (1999).
[32] H. Risken, *The Fokker-Planck Equation* (Springer, 1996).
[33] P. Heidarsson, I. Calpapuram, C. Camilloni, A. Impratao, G. Tiana, F. Poulsen, B. Kragenlund, and C. Cecconi, J. Am. Chem. Soc. **134**, 17068 (2012).
[34] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, and P. Kollman, J. Comput. Chem. **24**, 1999 (2003).
[35] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. K. Salmon, and W. Wriggers, Science **330**, 341 (2010).
[36] A. V. Finkelstein, Protein Eng. **10**, 843 (1997).
[37] A. Bartlett and S. E. Radford, Nat. Struct. Mol. Biol. **16**, 582 (2009).