Hierarchy of folding and unfolding events of protein G, CI2, and ACBP from explicit-solvent simulations

Carlo Camilloni, Ricardo A. Broglia, and Guido Tiana

Citation: The Journal of Chemical Physics **134**, 045105 (2011); doi: 10.1063/1.3523345
View online: http://dx.doi.org/10.1063/1.3523345
View Table of Contents: http://scitation.aip.org/content/aip/journal/jcp/134/4?ver=pdfcov
Published by the AIP Publishing

Hierarchy of folding and unfolding events of protein G, CI2, and ACBP from explicit-solvent simulations

Carlo Camilloni,[1,a)] Ricardo A. Broglia,[2,3,4] and Guido Tiana[2,3]
[1]*Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, United Kingdom*
[2]*Department of Physics, Universita' degli Studi di Milano, via Celoria 16, 20133 Milan, Italy*
[3]*INFN, Milan Section, via Celoria 16, 20133 Milan, Italy*
[4]*The Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17, DK-2100 Copenhagen, Denmark*

(Received 29 July 2010; accepted 11 November 2010; published online 31 January 2011)

The study of the mechanism which is at the basis of the phenomenon of protein folding requires the knowledge of multiple folding trajectories under biological conditions. Using a biasing molecular-dynamics algorithm based on the physics of the ratchet-and-pawl system, we carry out all-atom, explicit solvent simulations of the sequence of folding events which proteins G, CI2, and ACBP undergo in evolving from the denatured to the folded state. Starting from highly disordered conformations, the algorithm allows the proteins to reach, at the price of a modest computational effort, nativelike conformations, within a root mean square deviation (RMSD) of approximately 1 Å. A scheme is developed to extract, from the myriad of events, information concerning the sequence of native contact formation and of their eventual correlation. Such an analysis indicates that all the studied proteins fold hierarchically, through pathways which, although not deterministic, are well-defined with respect to the order of contact formation. The algorithm also allows one to study unfolding, a process which looks, to a large extent, like the reverse of the major folding pathway. This is also true in situations in which many pathways contribute to the folding process, like in the case of protein G.
 [doi:10.1063/1.3523345]

I. INTRODUCTION

Protein folding was initially described as a deterministic sequence of molecular events along specific pathways. As suggested by Levinthal in 1968, *a pathway of folding means that there exist a well-defined sequence of events which follow one another so as to carry the protein from the unfolded random coil to a uniquely folded metastable state*.[1] More recently, the focus has moved on the energy landscape underlying folding,[2–4] and the idea of a single folding pathway has been replaced by that of walks on such energy landscapes. As already noted,[5,6] the two perspectives are not mutually contradictory, provided that the Levinthal's pathways are intended in a statistical sense.

Experimentally, the search for folding pathways is a difficult task. The points of a protein free-energy which can be characterized are those corresponding to the minima associated with the native, the denatured, the intermediate (if present), and the transition states. Recently, single molecule experiments aimed at characterizing transition paths in protein folding have been reported in the literature.[7,8] For some proteins, like the IgG-binding domain of streptococcal protein G (GB1; 56 residues)[9] and acyl-coenzyme A binding protein (ACBP; 89 residues),[10] the transition state is structurally homogeneous, while for others, like chymotrypsin inhibitor 2 (CI2; 64 residues),[11] it corresponds to a more diverse set of conformations (within this context see also Ref. 12). This difference has been often interpreted in terms of two different folding mechanisms, a nonhierarchical nucleation model for CI2 and a hierarchical diffusion-collision model for GB1 and ACBP. Within this context, Baldwin and Rose[13] suggested that the folding of proteins is always hierarchic, and that the difference between the two observed behaviors is not qualitative, but is merely determined by the degree of stability of the secondary structures formed along the folding hierarchy.

Most of the computational study of protein folding has been carried out within the framework of simplified models. The results of these models suggest that folding is hierarchic and evolves from local to nonlocal structuring.[14–18] Realistic models in explicit solvent have been employed so far only to study, at great computational cost, the folding of few, extremely small proteins, like the 27-residue villin headpiece[19] and the 39-residue NTL9(1-39).[20] These simulations indicate a much larger degree of heterogeneity in the folding trajectories than what a simple model would suggest.

The purpose of the present work is to study to which extent the folding is a hierarchic process making use of a realistic, explicit-solvent model and a set of proteins of length slightly inferior to that of typical single-domain proteins.[21] Of course, this would be an essentially impossible computational task, would one use plain molecular dynamics (MD) simulations. On the other hand, we are not, within the present context, particularly interested in learning about folding times, but want to concentrate our attention on the less ambitious, but nonetheless important goal of finding out what the sequence of conformational events associated with folding is. For this purpose we feel justified to use the powerful biasing technique developed by Marchi and Ballone.[22] This algorithm is based on the introduction of a biasing potential which is

a)Author to whom correspondence should be addressed. Electronic mail: cc536@cam.ac.uk.

zero when the system is moving towards the desired arrival point and which damps the fluctuations when the system attempts at moving in the opposite direction. As in the case of the ratchet-and-pawl system, the algorithm is designed in such a way that the external field does not exert work in directing the system towards a specific direction. If the biasing potential is sufficiently soft, the resulting set of folding trajectories contain the correct sequence of events and can be simulated in a few days on a PC. By analyzing a statistically significant set of such folding trajectories for GB1, CI2, and ACBP, we aim at assessing to which extent protein folding is a hierarchical phenomenon.

For this purpose one needs to have a clear picture on two important issues, namely, (1) what is exactly meant by hierarchic folding and, (2) how to extract this information from the huge amount of data generated by multiple all-atom simulations.

A simple way to picture what one means by hierarchicity in a temperature-coupled molecular system is through a simple model developed by Hansen and co-workers.[23] If Ψ_i is a binary variable indicating the state of the ith native contact of the protein ($\Psi_i = 1$ means that the ith contact is formed, $\Psi_i = 0$ that it is not), a system controlled by the potential

$$U(\{\Psi_i\}) = -\epsilon(\Psi_1 + \Psi_1\Psi_2 + \cdots + \Psi_1\Psi_2\Psi_3 \ldots \Psi_N), \quad (1)$$

is perfectly hierarchic, in the sense that the formation of the ith contact is necessary for the formation of the $(i+1)$th contact. This model displays a first-order phase transition from the denatured state $\Psi = \{0, 0, 0, \ldots\}$ to the native state $\Psi = \{1, 1, 1, \ldots\}$ at temperature $\epsilon/\log 2$ (ϵ being the only energy scale of the system). At lower temperatures, folding takes place through a strictly ordered sequence of events ($\Psi_1 \to 1$, $\Psi_1 \to 2$, etc.). In other words, there is a set conditional probabilities associated with contact formation (i.e., in this case $p(i+1|i)$) which displays values close to 1.

The opposite kind of behavior is that of the model controlled by the potential

$$U(\{\Psi_i\}) = -N\epsilon\Psi_1\Psi_2\Psi_3 \ldots \Psi_N, \quad (2)$$

the so-called golf-course system. This system displays a phase transition at the same temperature of that associated with the potential (1), but this time the transition is much sharper. This is because we have now to deal with an exact two-state system and, consequently, the corresponding thermodynamics is highly cooperative. Of note that calorimetry experiments involving a large number of single-domain proteins indicate that folding thermodynamics is indeed highly cooperative.[24] On the other hand, at low temperature folding takes place through a purely random sequence of events that is nonhierarchicall. The conditional probabilities are all $\approx 1/2$.

The two models encoded by Eqs. (1) and (2) thus describe two limiting cases: the former hierarchic and mildly cooperative, the latter nonhierarchic and strongly cooperative. Their behavior can be interpolated by a potential of the type

$$U(\{\Psi_i\}) = -\epsilon \sum_i \left(\prod_j \mathcal{M}_{ij}\Psi_j \right) \Psi_i, \quad (3)$$

where $\mathcal{M}_{ij}$ is a matrix which indicates to which extent the formation of the ith contact depends on the state of the jth contact. A triangular matrix gives Eq. (1) while a matrix $\mathcal{M}_{ij} = 1$ gives Eq. (2). Different choices of the matrix give different balances between hierarchicity and cooperativity. Anyway, it was shown in Ref. 25 that it is not difficult to design a system displaying both a strong hierarchicity and a cooperative transition. The dynamics encoded by $\mathcal{M}_{ij}$ can result to be rather hierarchical even if it does not follow a deterministic sequence of events like the model of Eq. (1), for example in the case of a protein displaying different pathways towards the native state (i.e., a $\mathcal{M}_{ij}$ containing triangular blocks), or complicated relations among contacts (i.e., a $\mathcal{M}_{ij}$ containing binary elements generated at random). In Sec. II we shall use this model as a benchmark to quantify the degree of hierarchicity of protein folding and to evaluate the effect of the ratchet on MD simulations.

When modeling folding through simplified models as those introduced above, displaying a limited number of degrees of freedom, it is quite easy to elucidate the possible hierarchy of events associated with the process. For atomic models in explicit solvent such a hierarchy can be difficult to highlight, especially in the case in which it does not correspond to a deterministic sequence of events. The method of analysis to be used to extract such information from a very large ensemble of results of ratcheted simulations will be discussed in Sec. II B. Of note is that such a strategy of analysis can also be used for other scopes than that of characterizing ratcheted simulations. In particular, to analyze the large amount of data arising from distributed folding simulations.[19,26–28]

II. THE PHYSICS OF RATCHETED SIMULATIONS

A. Ratcheted molecular dynamics

Adiabatic biased molecular dynamics[22,29] is an algorithm developed to connect any two points in the conformational space of a given system. The method is based on the introduction of a biasing potential, which is a function of a chosen coordinate of the system and which is zero when the system is moving toward the desired target point, while disfavoring motions in the opposite direction. This is similar to what happens in a ratchet-and-pawl system, which undergoes random thermal fluctuations, while the pawl allows the ratchet to move only in one direction. Here the chosen coordinate plays the role of the ratchet and the biasing potential that of the pawl. In this respect, the ratcheting potential does not exert any work to direct the system towards the target conformation, as it happened when pulling the system with a force; on the contrary the system makes moves toward the target conformation under the driving effect of the potential of the force-field alone. Consequently, if the chosen coordinate were the actual reaction coordinate of the system, the most probable sequence of events generated by the ratcheted molecular dynamics would coincide with the minimum free-energy path independently on the damping constant. In fact, if the system enters a high-free-energy pathway, the system moves (by definition) in a direction which is normal to the reaction coordinate, and then the ratcheting potential does not apply,

nor does it change. Thus, the system can explore the high-free-energy region, eventually returning to the minimum-free-energy pathway. The outcome of a set of ratcheted molecular dynamics simulations is then a set of trajectories following the free-energy minimum of the system, but in which time and the statistical weight of the conformations along the trajectories are unphysically modulated.

Of course one does not know what is the actual reaction coordinate, or even if it exists. Ratcheting the dynamics with a wrong reaction coordinate can result in the selection of some highly unlikely pathways (as illustrated in Fig. S3 of the supplementary material[30]), without any possibility to return to more natural folding trajectories. In fact, if the ratcheting coordinate is the wrong one, the biasing potential can increase also when the system moves orthogonally to the lowest-free-energy pathway, follow the system in such byways,and trap it in dead-ends. Using a soft ratcheting potential (i.e., a harmonic potential with a small harmonic constant) helps to compensate the poor knowledge of the actual reaction coordinate. The softer is the potential, the higher is the probability that the system can come back if it enters some unlikely pathway. In the limit of very soft harmonic potential, the simulation tends to a plain molecular dynamics and the role of the ratcheting coordinate becomes immaterial.

The biasing potential is implemented as

$$V(\rho(t)) = \begin{cases} \frac{\alpha}{2}\left(\rho(t) - \rho_m(t)\right)^2, & \rho(t) > \rho_m(t), \\ 0, & \rho(t) \le \rho_m(t), \end{cases} \tag{4}$$

where

$$\rho(t) = \left(S(t) - S_{\text{target}}\right)^2, \tag{5}$$

is the distance along the coordinate S of the actual configuration of the system with respect to a target value S_{target}, and

$$\rho_m(t) = \min_{0 \le \tau \le t} \rho(\tau), \tag{6}$$

is the minimum distance reached until time t. The algorithm is thus defined by the choice of the coordinate S and of the damping constant α. In what follows we shall look for a value of α small enough so as to provide the correct sequence of events associated with protein folding and unfolding, but large enough to make the proteins to fold and unfold in a computationally reasonable time.

B. Analyzing the sequences of events

The quantitative analysis of such a large amount of data generated from multiple folding trajectories of a realistic protein requires an algorithmic scheme. The basic information ratcheted simulations can provide is the sequence of events along the calculated trajectories. Within this context, in what follows we shall focus our attention on the formation of native contacts between the amino acids of the protein, investigating which contacts repeatedly precede (or follow) the formation of some other contacts.

A native contact between the ith and the jth amino acids is defined as a contact in which, in a 20 ns of a plain MD simulation at 300 K, the average value of their relative minimum distance—calculated taking into account all of the atoms of the two residues—is less than or equal to 3 Å. Residues i and j must be separated by at least two residues. A native contact thus characterized is said to be stable along a folding simulation if, once formed, the associated minimum distance does not exceed at any (nominal) time the contact distance defined above plus three times its standard deviation (as obtained from the plain MD simulation). Following this definition, we found $n_c = 98$, 123, and 189 native contacts for proteins GB1, CI2, and ACBP, respectively (see Tables S1, S2, and S3 of the supplementary material[30]).

From each trajectory, the order of formation of the native contacts of the protein is defined by the quantity $t(i,k)$ that is the (nominal) time at which the ith contact is stably formed in the kth simulation. From this quantity, one can define the matrix

$$M_{ij}(k) = \theta\left(t(i,k) - t(j,k)\right), \tag{7}$$

where θ denotes Heaviside's step function. This matrix satisfies the relation $M_{ij} + M_{ji} = 1$ 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 they take place exactly at the same time. The average of M_{ij} over the n_r trajectories is

$$\overline{M_{ij}} = \frac{1}{n_r}\sum_{k=1}^{n_r} M_{ij}(k), \tag{8}$$

whose elements indicate the frequency for the ith contact to be formed before the jth. We shall interpret $\overline{M_{ij}}$ in a probabilistic sense. Thus, values of $\overline{M_{ij}}$ close to 1 indicate that the formation of the ith contact always precedes the formation of the jth contact and thus that their formation is hierarchically ordered. A value of $\overline{M_{ij}}$ close to $1/2$ shall be interpreted as the lack of a well-defined sequence of events or as the presence of few different well-defined sequences of events.

The degree of heterogeneity of the sequence of events associated with folding is investigated with the help of a trajectory distance defined as

$$d(k,k') \equiv \frac{1}{n_c(n_c-1)}\sum_{i \ne j}\delta(M_{ij}(k) - M_{ij}(k')), \tag{9}$$

where δ is the Kronecker function, and studying the distribution $p(d) \equiv \sum_{kk'}\delta(d - d(k,k'))$ of pair distances. In analogy to the case of the order parameter in the thermodynamics of complex systems,[31] a distribution displaying a single peak with a centroid located at low values of d reflects an homogeneous sequence of folding events, while a bimodal distribution indicates heterogeneity.

A quantity related to $\overline{M_{ij}}$ is the probability $A_j = \sum_{i \ne j}\overline{M}_{ij}/(n_c - 1)$ that the jth contact is formed after any other contact. The plot of the A_j, ordered from the smallest to the largest values (cf. Fig. 2(a)), can be used to study to which extent the formation of contacts during folding is hierarchic. If, during the process, native contacts are formed along a deterministic hierarchy of events, as in a chain of chemical reactions, the ordered A_j values will lay on the diagonal of the plot. On the contrary, if folding is fully cooperative, in the sense that all native contacts are formed simultaneously at the

transition state, the ordered A_j values will lay on a horizontal line. One can thus define a parameter hi to measure the degree of hierarchicity of the folding process, as the angular coefficient of the ordered A_j values, ascribing to it the value 1 in the case of a deterministic hierarchy.

One should take notice of the fact that the above scheme of analysis is not only limited to ratcheted simulations, but can also be used in the analysis of the results of generic simulations, in particular in the case of distributed folding simulations.[19,26–28]

C. Hierarchicity in minimal protein models

The models described in connection with Eqs. (1) and (2) can be used to get some insight into the analysis strategy of folding data and into the validity of ratcheted folding simulations. The quantity ϵ is the only energy scale of the system. The dynamics of the model is carried out making n_r simulations starting from the $\Psi = 0, 0, 0, \ldots$ state using Kubo dynamics[32] at temperature $T = \epsilon$. In the case of the hierarchical model of Eq. (1) without the use of ratcheted dynamics and with large statistics ($n_r = 200$) one obtains $hi = 0.99$ (cf. black dots in Fig. S1 in the supplementary material[30]). Decreasing the statistics to $n_r = 10$ does not affect significantly the result, giving $hi = 1.0$ (cf. black curve in Fig. S1). If the ratchet is switched on with (a) energy constant much lower than the energy scale of the system, the result is again unaffected (e.g., $\alpha = \epsilon/100 \ll \epsilon$ gives $hi = 1.0$, blue curve in Fig. S1). As the energy constant of the ratchet approaches the energy scale of the system, the value of hi results is underestimated (e.g., $hi = 0.73$ at $\alpha = \epsilon/2$ and $hi = 0.62$ at $\alpha = \epsilon$).

The same kind of calculation can be carried out with the nonhierarchical model defined by Eq. (2). The "true" value, obtained with a large statistics and without the use of the ratchet gives $hi = 0.16$, which increases to 0.23 if the statistics is reduced to $n_r = 10$. The reason for this increase is the following: in the non-hierarchical model one expects that the average of A_j is zero with an error decreasing as $1/\sqrt{n_r}$. As the values of A_j are sorted, this produces a slope which increases as the error. The ratchet does not affect substantially this result, leading to $hi = 0.27$ for $\alpha = \epsilon/2$ and $hi = 0.31$ for $\alpha = \epsilon$. If one increases the ratcheting potential to many times the energy scale of the protein ($\alpha = 5\epsilon$), a value $hi = 0.41$ is obtained.

The results for two models displaying intermediate degree of hierarchicity are displayed in Fig. S2 in the supplementary material.[30] A model which includes three parallel folding pathways (i.e., controlled by a $\mathcal{M}_{ij}$ containing three triangular blocks) results in a value $hi = 0.90$; a value which is not affected if statistics are reduced down to $n_r = 10$. The effect of the ratchet ($\alpha = \epsilon/2$) is to reduce hi to 0.69. The model defined by a $\mathcal{M}_{ij}$ containing binary variables generated at random (but kept fixed) give $hi = 0.53$. Reducing the statistics to $n_r = 10$ leads to $hi = 0.64$, while ratcheting with $\alpha = \epsilon/2$ give $hi = 0.59$.

The lesson to be learned from these tests is that the ratchet, if applied with a constant α smaller than the intrinsic energy scale of the system, slightly underestimates the hierarchicity of hierarchical processes and slightly overestimates that of nonhierarchal processes, but without allowing confusion between the two. The lack of statistics instead, even in the absence of a biasing potential, has little effect on simulations which are already hierarchic, but overestimates the value of hi for those whose hierarchicity is essentially zero. It is possible to gauge the value of hi for noncooperative processes to $1/\sqrt{n_r}$, which is 0.32 in the case $n_r = 10$.

D. Ratcheted simulations provide the lowest-free-energy pathways in simple molecular models in explicit solvent

The study of minimal models suggest that the damping constant of ratcheted simulations has to be smaller than one-half of the typical energy scale of the system. For biological molecules, such energy scale is, at room temperature, kT (i.e., 2.5 kJ/mol). Consequently, we make the hypothesis that $\alpha = 1$ kJ/mol is a good choice to obtain the minimum free-energy pathways of this class of systems. To test such a hypothesis, use is made of two molecules of different size, that is alanine dipeptide (ACE-ALA-NME) and the 20mer helix of GB1, both described in explicit solvent. These two systems are good benchmarks because they display the same interactions which stabilize larger proteins, and their free-energy profiles have been extensively characterized.[33–35]

In Fig. 1(a) the free-energy surface of alanine dipeptide as a function of the dihedral angles ϕ and ψ is show. The movement of ϕ across 0 is a slow process, involving a free-energy barrier of $\sim$10 kT. A typical trajectory obtained ratcheting the system along the reaction coordinate ϕ with a ratchet constant of $\alpha = 1$ kJ/mol is displayed with a yellow curve. The trajectory follows the minimum-free-energy pathways, going through the lowest saddle point at $\phi = 0$, $\psi = 1.4$. If a larger ratcheting constant $\alpha = 20$ kJ/mol along the coordinate ϕ is used, the trajectories cross the free-energy at the higher-free-energy saddle point located at $\phi = 0$, $\psi = -1.6$ (see Fig. S3 in the supplementary material[30]). The reason is that ϕ is not a correct reaction coordinate of the system: as the system reaches the local minimum at $\phi = -0.8$, $\psi = -0.8$, in order to go on the lowest-free-energy pathway, ϕ (i.e., the ratcheting coordinate) should go back to -1, something that it cannot do if the ratcheting potential is too hard (i.e., $\alpha \gg 1$ kJ/mol).

Figure 1(b) displays the free-energy of the 20mer helix of protein GB1 with respect to the number of i-$(i+4)$ backbone hydrogen bonds and the radius of gyration of the chain. The yellow line indicates a typical folding trajectory obtained ratcheting the system along the distance d_{CM} of the contact map of the system to that of the native conformation[36]

$$d_{CM} = \|C - \tilde{C}\| = \left(\sum_{j>i+35}^{N} (C_{ij} - \tilde{C}_{ij})^2 \right)^{1/2}, \qquad (10)$$

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

$$C_{ij}(r_{ij}) = \frac{1 - \left(\frac{r_{ij}}{r_0}\right)^6}{1 - \left(\frac{r_{ij}}{r_0}\right)^{10}}, \qquad (11)$$

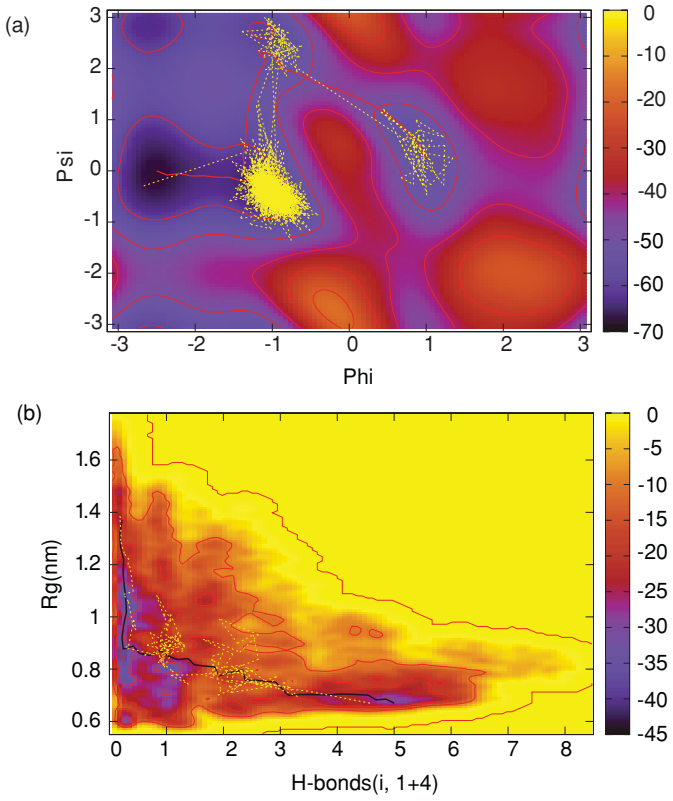

FIG. 1. The free-energy landscapes of alanine dipeptide (a) and of the helix of GB1 (b), expressed in kJ/mol. The yellow lines indicate a ratcheted trajectory generated using a damping constant $\alpha = 1$ kJ/mol. The lowest free-energy pathway, derived from a nudge elastic band calculation, is indicated by a red curve in the case of alanine dipeptide and by a black curve in the case of the GB1 helix.

r_{ij} is the distance (in nm) between atomes i and j, $r_0 = 0.35$ nm, $\tilde{C}$ is the defined on the target state, and N includes all the atoms of the system. This function has been used in order to have a differentiable definition of the contact map, and the small value of r_0 is justified by the fact that we are considering the contacts between all the atoms of the system. Again, the ratcheting potential is defined by $\alpha = 1$ kJ/mol, a choice which allows the system to follow the lowest-free-energy pathway. If α is raised to 20 kJ/mol, the system reaches the native conformation following a high-free-energy pathway (cf. Fig. S4 in the supplementary material[30]).

III. FOLDING AND UNFOLDING OF PROTEINS IN EXPLICIT SOLVENT

A. Model systems and simulation details

The ratcheting algorithm allows one to obtain complete trajectories in computational times of the order of days on a PC in the case of the folding of proteins of realistic length. We have studied the folding of protein GB1 (pdb code 1pgb), which is a 56 residues globular protein with a β-sheet formed by the N-terminal and the C-terminal β-strands opposed to an α helix; of CI2 that is a 64-residue protein (pdb code 2ci2), characterized by a large β-sheet opposed on one hand to a helix and on the other to the active site loop; and of ACBP that is a 4 helix protein composed of 89 residues (pdb code 2abd). While these proteins have been extensively characterized both experimentally as well as theoretically, their folding has never been simulated by means of an all-atom, explicit solvent molecular dynamics.

All the simulation we have carried out were performed with the help of a modified version of GROMACS.[37,38] The interactions used are the Amber 2003 all-atom force-field.[39,40] The proteins were enclosed in dodecahedron boxes displaying a volume ≥ 261 nm^3. Periodic boundary conditions were used throughout. Solvation was implemented making use of at least 8325 TIP3P water molecules. Van der Waals interactions were cut-off at 1.4 nm and the long-range electrostatic interactions were calculated by the particle-mesh-Ewald algorithm, using a mesh space of 0.125 nm. The systems were coupled to a Nosé–Hoover thermal bath.

For each protein studied, ten unfolded initial conformations were generated from high-temperature unfolding simulations (600 K for 20 ns) and a subsequent thermalization at 300 K. All the starting conformations are at least 1 nm in RMSD from the native state and 0.5 nm from each other. The presence of residual contacts is less than 7% and the overlapping of the residual structure between them is in the range of 0%–4%.

From each initial conformation a molecular-dynamics trajectory at 300 K is generated, ratcheting the system, along d_{CM}, as described in Sec. II D for 5×10^6 steps. In all cases a conformation with RMSD lower than 1.3 Å is reached with respect to the pdb structures.

B. The folding of the three proteins is rather hierarchic

The 30 folding simulations associated with the three proteins under study were analyzed as described in Sec. II B. The ordered values of A_j are displayed in Fig. 2. The resulting values of the hierarchicity parameters hi, calculated from the linear fit of A_j weighted by its standard deviations σ_j, are 0.91 for ACBP, 0.80 for CI2, and 0.72 for GB1. The analysis of the hierarchicity of the models discussed in Sec. II C, suggests that in the case of ten simulations a $hi = 1$ indicates a perfectly hierarchic process, while $hi \sim 0.3$ indicates a nonhierarchic process. Accordingly, the folding of ACBP and CI2 appears rather hierarchic, while that of GB1 seems only mildly hierarchic.

The reason why the linear fit is carried out weighting the values of A_j by their standard deviation is connected with the detailed shape of the A_j curves. In the case of ACBP and CI2 one can detect three regions characterized by different slopes. The initial and final regions are steeper and display a smaller standard deviation, indicating that the associated contacts forms more hierarchically and more homogeneously in all ten simulations. The large central region displays larger standard deviation ($\sigma_j \approx 0.2$) and a slope (~ 0.7) similar for the two proteins and also similar to the overall slope of GB1. Had one done an unweighted fit, the result would have been essentially determined by the numerous contacts associated with the central regions of the folding events. The weighted fit highlights the importance of the initial and final contacts, whose order of formation is more homogeneous among the

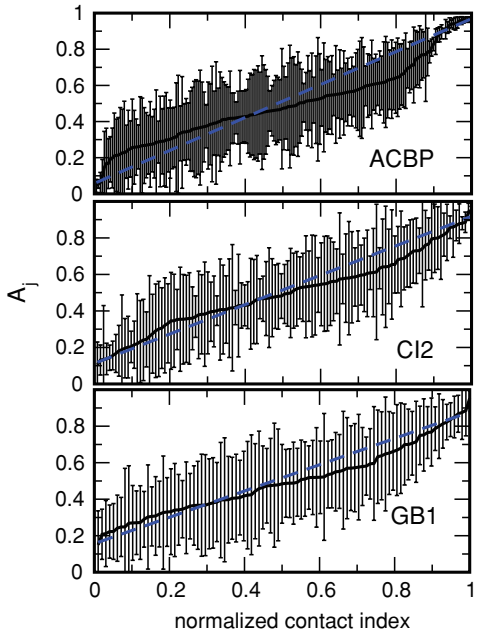

FIG. 2. The average and standard deviation of the fraction of native contacts A_j which precede the formation of contact j, ordered according to its increasing value, calculated for the three proteins under study. The contact index j is parametrized with a real number ranging from 0 to 1. The dashed line is the linear fit of A_j, weighted by the associated standard deviation. The slopes of the fit *hi* is 0.91 for ACBP, 0.80 for CI2, and 0.72 for GB1.

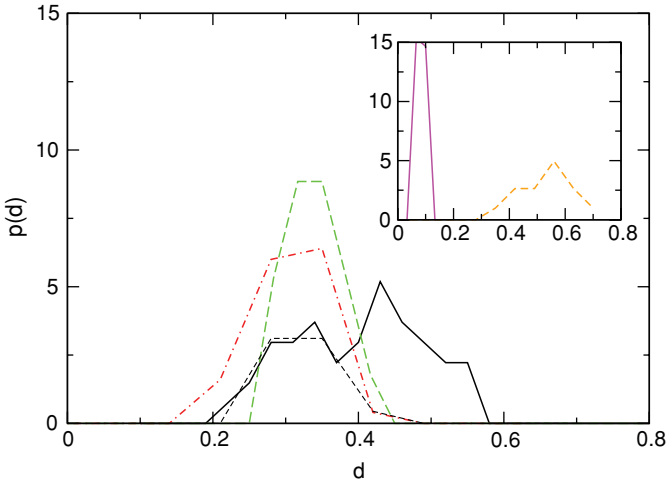

FIG. 3. The distribution of distances between pairs of matrices M_{ij} calculated from the ten simulations of GB1 (black solid curve), CI2 (green dashed curve), and ACBP (red dotted–dashed curve). The black dashed curve indicates the distribution of distances calculated on the matrices of GB1 belonging to the same pathway, keeping the same normalization constant as the full distribution associated with GB1 in order to facilitate the comparison. In the inset, the distribution of distances for the hierarchic model defined by Eq. (1) (solid magenta curve) and for the nonhierarchic model defined by Eq. (2) (dashed orange curve).

different simulations than those associated with the central region, are.

The low-A_j contacts of ACBP ($A_j < 0.2$), that is those which are formed before all others in all the simulations ($\sigma_j \approx 0.08$) are all local and involve mainly the two terminals of helix $\alpha 4$ (regions 67–71 and 78–84), two contacts in helix $\alpha 2$ (25–29 and 28–32) and one in helix $\alpha 3$ (48–52). On the other hand, the contacts which are formed after all the others ($A_j > 0.82$) homogeneously in all the simulations ($\sigma_j = 0.02$) involve essentially the packing of the side chains of helix $\alpha 4$ with helices $\alpha 1$ and $\alpha 2$ (residues 8–73, 11–77, 15–81, 15–86, 20–84, 27–76, 33–69, and 35–66) and the local contacts in the central regions of helix $\alpha 4$. In Ref. 41 the authors identify, by a combination of experimental techniques, a subset of residues which are critical for the correct folding of ACBP. These are: F5, A8, V12, M24–Y28, D56, N59, D68–D75, V77, E78, and L80–K82. These residues match quite well the residues involved in early contacts of the ratcheted simulations.

The first contacts formed in CI2 ($A_j < 0.35$) display a standard deviation larger than ACBP ($\sigma_j \sim 0.15$) and involve mainly the formation of the α-helix (hydrogen bonds 13–17, 14–18, 16–20, 17–21, 18–22, 20–24 and side chain contacts 16–19, 17–20, 18–21, 20–23, 22–25) except for some defects at its C-terminal and the docking of the N-terminal strand to the helix. Interestingly, among the early contacts there are few nonlocal contacts between the structure involving the N-terminal strand and the α-helix, and the C-terminal (5–63, 10–58, 14–58). The contacts formed after all the others were formed ($A_j > 0.73$, $\sigma_j \approx 0.1$) involve the sheet built out of strands $\beta 3$ and $\beta 4$ and various elements distributed throughout the protein. The curve associated with GB1 in Fig. 2 seems quite different from that associated with the other two proteins, in that it does not display marked variations in the slope and in the value of σ_j.

C. ACBP and CI2 follow a single folding pathway, GB1 follows three different pathways

While so far we have analyzed the quantity A_j, which relates the formation of a contact to the average formation of the other contacts, it is interesting to study the order of formation of each specific pair of contacts. Figure 3 displays the distribution $p(d)$ of the distances between the matrices M_{ij} associated with different trajectories (see Sec. II B). The distributions $p(d)$ calculated from the minimal models defined by Eqs. (1) and (2) are displayed in the inset and provide the reference curves for a perfectly hierarchic (solid magenta curve) and non-hierarchic (orange dashed curve) processes. ACBP (red dotted–dashed curve) and CI2 (green dashed curve) display a unimodal distribution centred at $d = 0.32$ and $d = 0.34$, respectively. This means that, approximately, 70% of pairs of contacts display the same order of formation (see Sec. II C). We can interpret the unimodal distribution associated with ACBP and CI2 in terms of a single folding pathway, built out of sequences of events which typically are 70% similar. In fact, inspection of the folding trajectories of these proteins (cf. Fig. 4) show that ACBP always formed first the terminal parts of helix $\alpha 4$, while the first nonlocal contacts forms are between helices $\alpha 4$ and $\alpha 2$, the last local contacts formed are in the N-terminal of $\alpha 1$ and at the centre of $\alpha 4$ and the last non-local event is the docking of $\alpha 1$ to the rest of the protein. CI2 forms first the helix and the contacts between the helix and the N-terminal strand, followed by the nonlocal contacts between the N- and the C-terminal strands, the

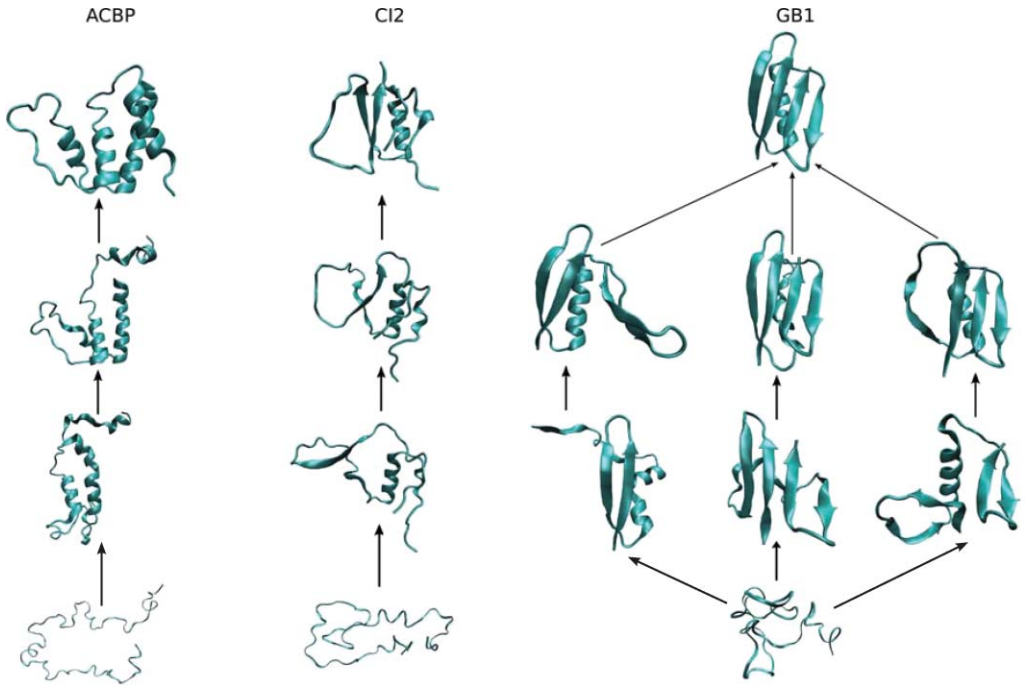

FIG. 4. Schematic representation of the folding pathways of the three proteins from unfolded (below) to the native conformation (above). The conformations displayed are taken from the actual trajectories and are chosen to picture the main steps towards the native conformation. The resulting picture of the ACBP protein displays an overall agreement with that observed in Fig. 4 of Ref. 44.

formation of the β-sheet between $\beta 3$ and $\beta 4$ and the docking of this to rest of the protein (cf. Fig. 4).

Of notes that the formation of the helix $\alpha 4$ as the first event in folding of ACBP has been suggested on the basis of NMR analysis of its denatured state.[41,42] While simulations show that the formation of most of $\alpha 4$ is the very first event along the folding pathway, the formation of few contacts at the centre of the helix results to be one of the last events to take place. The presence of a kink in the solution structure of the protein[43] corresponding to the late-forming contacts suggests that, in this case, the interactions which stabilise the helix are poorly optimised. The order of contact formation resulting from ratcheted MD simulations indicate that the native interactions between $\alpha 1$ and the complex $\alpha 2 \alpha 4$ are the folding rate-limiting step of ACBP (cf. Sec. III B). This is in agreement with the results of secondary chemical shift analysis of mutated proteins[44] and of paramagnetic relaxation enhancement studies.[45] The overall folding picture which arises from the simulations is then remarkably similar to that described in Ref. 44.

Also in the case of CI2, NMR experiments in the denatured state indicate the formation of some contacts in the helix,[46] contacts which simulations indicate as the first event in folding. Moreover, the fact that the last-formed contacts involve all parts of the protein agree with the idea of an extended folding nucleus arising from the analysis of φ-values.[11]

GB1, on the other hand, displays a bimodal distribution, with peaks centered at $d = 0.32$ and 0.47. This bimodal distribution is interpreted in terms of more folding pathways (nothing can be said yet about their number). Each pathway is built out of sequences of events $\approx 70\%$ similar, while sequences of events belonging to different pathways are $\approx 50\%$ similar. Inspection of the folding trajectories of GB1 show that in seven cases out of ten folding starts from the binding of the helix to the first hairpin (in only one of these cases the second hairpin is formed together with the first one, but binds to it only later). In two cases it starts from the binding of the two hairpins and in one single case from the binding of the helix to the second hairpin. As a matter of fact, if one calculates the distribution $P(d)$ only within each pathway, one obtains an unimodal peak centered around $d = 0.31$ (dashed black curve in Fig. 3). The early contacts of the principal pathway matches well the residual structure obtained in acid-denatured studies of GB1.[47] Moreover, phi-values analysis indicates that the rate-limiting step towards folding is that associated with the formation of native contacts between the hydrophobic residues of the chain and the second hairpin.[9] This is in agreement with the order of contact formation observed in all the three pathways obtained in the simulations.

The presence of different pathways agrees with the results obtained from Gō-model all atom simulations,[48] which highlight three pathways, the most probable (59%) starting from the binding of the first hairpin to the helix. Within this context, the value of hi calculated for GB1 in Fig. 2 is not meaningful, as the associated matrix M_{ij} is not self-averaging. What can be done instead is to evaluated the degree of hierarchicity of the different folding pathways, which for the most visited pathway is $hi = 0.83$.

D. Non-native contacts along the folding trajectories

Non-native contacts have been shown to play an important role in protein folding.[49–51] Operatively, the non-native

contact is considered established when two amino acids, lying more than three residues apart along the chain and which in the native conformation are more than 5 Å apart, come closer than 4 Å. Non-native contacts present in the starting structures are not considered. As the ratcheted trajectories end up in the native conformation, the formation of non-native contacts is necessarily transient, and the duration of the non-native contact has no physical sense. To be statistically sound, we shall focus on those non-native contacts which are formed in at least six over ten trajectories.

Within this scenario the folding of ACBP is characterized by the formation of non-native salt bridges involving residues 11–16 (helix $\alpha 1$), 18–23 (between $\alpha 1$ and $\alpha 2$), and 62–66 (between $\alpha 3$ and $\alpha 4$). Also non-native hydrophobic contacts 73–80 and 80–86 are formed within helix $\alpha 4$. Further non-native contacts are formed between the loop and $\alpha 3$ (40–55) and, interestingly, between the N- and the C-terminal of the protein (4–70), as reported by the paramagnetic relaxation enhancement analysis of Ref. 45.

CI2 displays a group of non-native contacts involving residues which are distant along the sequence, namely contacts 7–20, 7–63, 8–61 between the C- and the N-terminal regions, contacts 22–29 between the helix and the β-sheet, and contacts 32–38, 33–40, and 41–47 in the regions involving the β-sheet and the active-site loop.

In GB1 contacts 19–26 and 21–26 stabilize a non-native turn which has been observed in a NMR study of the GB1 helix[52] (see also Ref. 35). This turn is disrupted when the N-terminal turn of the helix is formed. Residues 31–40 form a non-native salt-bridge which stabilizes the C-terminal segment of the helix.

It is of note that the ratchet algorithm, which biases the folding trajectories towards the native conformation, does not prevent the formation of non-native contacts.

E. Unfolding

In order to study the hierarchicity of the unfolding process we have performed ten simulations for each protein starting from the native states and ratcheting the system towards the initial conformation used for each of the folding simulation. The analysis of Sec. II B has been repeated for these simulations and the associated hierarchicity curves A'_j are shown in Fig. 5.

The resulting values of the hierarchicity parameters hi, calculated from the linear fit of A'_j weighted by its standard deviations σ_j, are 0.93 for ACBP, 0.84 for CI2, and 0.87 for GB1. Accordingly, the folding and unfolding of ACBP and CI2 display approximately the same degree of hierarchicity. On the other hand, the value of hi for the unfolding of GB1 is larger than that of folding, averaged on the three pathways, but comparable with that of the major pathway.

The low-A'_j contacts of ACBP ($A'_j < 0.2$), that is those which are broken before all others in all the simulations are mainly nonlocal contacts between helix $\alpha 4$ and both helices $\alpha 1$ and $\alpha 2$ (residues 8–74, 11–77, 12–77, 15–80, 15–81, 27–76, 30–72, 33–69, and 35–66); the local contacts that are broken first are all i-$(i+3)$ side chain contact within the helices $\alpha 1$ and $\alpha 4$ (2–5, 3–6, 75–78, and 76–79). Interestingly five

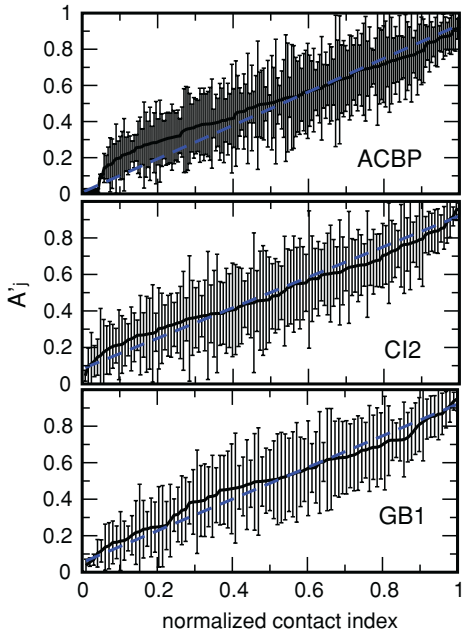

FIG. 5. The average and standard deviation of the fraction of native contacts A'_j which precede the breaking of contact j, calculated in ten unfolding simulations of each of the three proteins. The dashed line is the linear fit of A'_j, weighted by the associated standard deviation. The slopes of the fit hi is 0.93 for ACBP, 0.84 for CI2, and 0.87 for GB1.

out of nine nonlocal contacts that are broken first correspond to nonlocal contacts that are formed last in the folding trajectories (cf. Sec. III B). On the other hand the contacts which are broken after all the others ($A'_j > 0.82$) homogeneously in all the simulations involve essentially the C-terminal loops of helix $\alpha 4$ (regions 78–84) and single loops in the other helices (2–6 in $\alpha 1$, 26–30 in $\alpha 2$, and 48–52 in $\alpha 3$), it is of note the similarity found between these last contacts and the first contacts in the folding simulations.

The unfolding of CI2 begins ($A'_j < 0.35$) with the undocking of the two β-strands including the active loop from the helix and the C- and N-terminal strands (breaking of contacts between regions 28–30 and 46–50 and regions 6–11, 18–22, and 60–64). The last events in the unfolding are ($A'_j > 0.73$): the breaking of the N-terminal loops of the helix and the contact between these loops and the N-terminal strand (6–9, 8–12, 9–12, 12–16, 13–17, 14–17, 15–18, 16–20, and 17–21) and the breaking of some residual contacts within the β-sheet (31–50, 33–50, 40–49, 42–47, 47–65, and 49–65).

The distribution $p(d)$ of distances between the matrices associated with unfolding of GB1 displays a single peak centred at $d = 0.34$, suggesting that there is a single unfolding trajectory. This pathway is characterized by the early breaking of contacts involving the docking of a second hairpin with the rest of the protein (4–50, 26–52, 30–52, and 31–43) followed by the breaking of the helix and ending with the first hairpin (4–15, 5–16), few residual contacts within the helix (22–26, 33–37, 34–39), the turn of the second hairpin (46–50), and between first and second hairpin (8–56 and 11–56) indicating that the protein unfoldnes along a pathway similar to the most visited folding pathway.

IV. CONCLUSION

Folding and unfolding are stochastic events controlled by thermal motion. For this reason the understanding of these processes require a statistically meaningful collection of folding trajectories. This cannot be done with plain molecular dynamics simulations unless in the case of very small proteins and at the price of a very large computational effort. One can avoid such limitations by restricting the scope of the simulations to that of solely providing step sequence information. Such a program can be made operative with the help of a ratcheting algorithm which is able to generate the needed folding/unfolding trajectories at small computational cost, coupled to an analysis scheme which allows one to extract eventual regularities in the sequence of folding events which may lie in the complexity of the data. This scheme, which can also be applied to massive unbiased molecular dynamics simulations, was found to be instrumental in showing that proteins ACBP and CI2 fold following single pathways which are homogeneous, while protein GB1 folds through a small number (three) of different pathways. In all cases the different pathways are rather hierarchic, in the sense that most pairs of contacts take place in a sequential-like fashion. Furthermore, the results of the ratcheted simulations strongly indicate that unfolding can be viewed as the the sequence-reverse phenomenon of folding.

ACKNOWLEDGMENTS

The authors thank, in alphabetic order, S. A. Beccara, M. Bonomi, G. Bussi, G. Colombo, P. Faccioli, A. Laio, C. Micheletti, M. Parrinello, and F. M. Poulsen for useful discussions and acknowledge the computer resources provided by the Consorzio Interuniversitario Lombardo per l'Elaborazione Automatica (CILEA). C.C. was supported by a FEBS long-term fellowship.

[1]C. Levinthal, J. Chim. Phys. **65**, 44 (1968).
[2]R. L. Baldwin, Nature **369**, 183 (1994).
[3]A. Sali, E. I. Shaknovich, and M. Karplus, Nature **369**, 248 (1994).
[4]P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, Science **267**, 1619 (1995).
[5]T. Lazaridis and M. Karplus, Science **278**, 1928 (1997).
[6]V. S. Pande, A. Yu. Grosberg, T. Tanaka, and D. S. Rokhsar, Curr. Opin. Struct. Biol. **8**, 68 (1998).
[7]C. Cecconi, E. A. Shank, C. Bustamante, and S. Marqusee, Science **309**, 2057 (2005).
[8]H. Sung Chung, J. M. Louis, and W. A. Eaton, Proc. Natl. Acad. Sci. U.S.A. **106**, 11837 (2009).
[9]E. L. McCallister, E. Alm, and D. Baker, Nature Struct. Biol. **7**, 669 (2000).
[10]B. B. Kragelund, P. Osmark, T. B. Neergaard, J. Schioedt, K. Kristiansen, J. Knudsen, and F. M. Poulsen, Nature Struct. Biol. **6**, 594 (1999).
[11]L. S. Itzhaki, D. E. Otzen, and A. R. Fersht, J. Mol. Biol. **254**, 260 (1995).
[12]M. Karplus and D. L. Weaver, Nature **260**, 404 (1976).
[13]R. L. Baldwin and G. D. Rose, TIBS **24**, 77 (1999).
[14]K. A. Dill, K. M. Fiebig, and H. S. Chan, Proc. Natl. Acad. Sci. USA **90**, 1942 (1993).
[15]R. A. Broglia and G. Tiana, J. Chem. Phys. **114**, 7267 (2001).
[16]L. Sutto, G. Tiana, and R. A. Broglia, Protein Sci. **15**, 1638 (2006).
[17]S. B. Ozkan, G. A. Wu, J. D. Chodera, and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **104**, 11987 (2007).
[18]C. Camilloni, L. Sutto, D. Provasi, G. Tiana, and R. A. Broglia, Protein Sci. **17**, 1424 (2008).
[19]D. L. Ensign, P. M. Kasson, and V. S. Pande, J. Mol. Biol. **374**, 806 (2007).
[20]V. A. Voelz, G. R. Bowman, K. Beauchamp, and V. S. Pande, J. Am. Chem. Soc. **132**, 1526 (2010).
[21]D. Xu and R. Nussinov, Folding Des. **3**, 11 (1998).
[22]Marchi and P. Ballone, J. Chem. Phys. **110**, 3697 (1999).
[23]A. Hansen, M. H. Jensen, and K. Sneppen, Eur. Phys. J. B **10**, 193 (1999).
[24]P. L. Privalov and N. N. Khechinasvili, J. Mol. Biol. **86**, 665 (1974).
[25]A. Bakk, J. S. Høye, A. Hansen, K. Sneppen, and M. H. Jensen, Biophys. J. **79**, 2722 (2000).
[26]D. A. C. Beck, A. L. Jonsson, R. D. Schaeffer, K. A. Scott, R. Day, R. D. Toofanny, D. O. V. Alonso, and V. Dagget, Protein Eng. Design Sel. **21**, 353 (2008).
[27]P. L. Freddolino, F. L., M. Gruebele, and K. Schulten, Biophys. J. **94**, L75 (2008).
[28]R. O. Dror, M. Ø. Jensen, D. W. Borhani, and D. E. Shaw, J. Gen. Phys. **135**, 555 (2010).
[29]E. Paci and M. Karplus, J. Mol. Biol. **288**, 441 (1999).
[30]See supplementary material at http://dx.doi.org/10.1063/1.3523345 for the figures displaying the effect of the ratchet on simplified and realistic models and for the tables containing the list of the native contacts of the proteins studied.
[31]M. Mezard, G. Parisi, and M. A. Virasoro, *Spin Glasses and Beyond* (World Scientific, Singapore 1987).
[32]M. Suzuki and R. Kubo, J. Phys. Soc. Jpn. **24**, 51 (1968).
[33]P. E. Smith, J. Chem. Phys. **111**, 5568 (1999).
[34]A. Laio and M. Parrinello, Proc. Natl. Acad. Sci. U.S.A. **99**, 12562 (2002).
[35]C. Camilloni, D. Provasi, G. Tiana, and R. A. Broglia, Proteins **71**, 1647 (2008).
[36]M. Bonomi, D. Branduardi, F. L. Gervasio, and M. Parrinello, J. Am. Chem. Soc. **130**, 13938 (2008).
[37]B. Hess, C. Kutzner, D van der Spoel, and E. Lindahl, J. Chem. Theory Comput. **4**, 435 (2008).
[38]M. Bonomi, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Pietrucci, R. Broglia, and M. Parrinello, Comp. Phys. Comm. **180**, 1961 (2009).
[39]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. Comp. Chem. **24**, 1999 (2003).
[40]E. J. Sorin and V. Pande, Biophys. J. **88**, 2472 (2005).
[41]K. Modig, V. W. Jürgensen, K. Lindorff-Larsen, W. Fieber, H. G. Bohr, and F. M. Poulsen, FEBS Lett. **581**, 4965 (2007).
[42]K. Teilum, B. B. Kragelund, and F. M. Poulsen, J. Mol. Biol. **324**, 349 (2002).
[43]K. V. Andersen and F. M. Poulsen, J. Biomol. NMR **3**, 271 (1993).
[44]S. W. Bruun, V. Iesmantavicius, J. Danielsson, and F. M. Poulsen, Proc. Natl. Acad. Sci. U.S.A. **107**, 13306 (2010).
[45]S. Kristjansdottir, K. Lindorff-Larsen, W. Fieber, C. M. Dobson, M. Vendruscolo, and F. M. Poulsen, J. Mol. Biol. **347**, 1053 (2005).
[46]S. L. Kazmirski, K.-B. Wong, S. M. V. Freund, Y.-J. Tan, A. R. Fersht, and V. Daggett, Proc. Natl. Acad. Sci. USA **98**, 4349 (2001).
[47]N. Sari, P. Alexander, P. N Bryan, and J. Orban, Biochem. **39**, 965 (2000).
[48]J. Shimada and E. I. Shakhnovich, Proc. Natl. Acad. Sci U.S.A. **99**, 11175 (2002).
[49]E. Paci, M. Vendruscolo, and M. Karplus, Proteins **47**, 397 (2002).
[50]A. R. Vigura, C. Vega, and L. Serrano, Proc. Natl. Acad. Sci. U.S.A. **99**, 5349 (2002).
[51]C. Clementi and S. S. Plotkin, Protein Sci. **13**, 1750 (2004).
[52]F. J. Blanco, A. R. Ortiz, and L. Serrano, Folding Des. **2**, 123 (1997).