Mechanical unfolding of a simple model protein goes beyond the reach of one-dimensional descriptions

R. Tapia-Rojo,[1,a)] S. Arregui,[1] J. J. Mazo,[2] and F. Falo[1]
[1]*Instituto de Biocomputación y Física de Sistemas Complejos and Departamento de Física de la Materia Condensada, Universidad de Zaragoza, 50009 Zaragoza, Spain*
[2]*Instituto de Ciencia de Materiales de Aragón and Departamento de Física de la Materia Condensada, CSIC-Universidad de Zaragoza, 50009 Zaragoza, Spain*

(Received 16 July 2014; accepted 16 September 2014; published online 6 October 2014)

We study the mechanical unfolding of a simple model protein. The Langevin dynamics results are analyzed using Markov-model methods which allow to describe completely the configurational space of the system. Using transition-path theory we also provide a quantitative description of the unfolding pathways followed by the system. Our study shows a complex dynamical scenario. In particular, we see that the usual one-dimensional picture: free-energy vs end-to-end distance representation, gives a misleading description of the process. Unfolding can occur following different pathways and configurations which seem to play a central role in one-dimensional pictures are not the intermediate states of the unfolding dynamics. © *2014 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4896620]

I. INTRODUCTION

The characterization of folding and unfolding energy landscapes of biomolecules is a major problem in biophysics which sheds light onto biomolecules' role and function.[1–5] In this effort, the emergence of single-molecule techniques that let the manipulation of individual molecules has opened a new wide field, allowing to monitor unfolding processes by looking into a single specimen.[6–14]

In force-pulling experiments, the one-dimensional description is usually adopted, as force is considered to impose a preferred direction that appears as the slowest degree of freedom compared with the remaining ones. In this sense, optical tweezers,[6,7] magnetic tweezers[8,9] or AFM[10–12] experiments are usually analyzed considering the end-to-end distance as the proper reaction coordinate, with a well developed force spectroscopy theory[15–18] that allows stating predictions grounded on this hypothesis. Also, recent studies of single molecule Foester resonant energy transfer fluorescence study thermal unfolding by tracking the radius of gyration of individual molecules.[13,14] Computational works similarly take advantage of this simple description, choosing reaction coordinates such as the fraction of native contacts Q,[19–22] the RMSD from the native structure[23] or the Principal Components.[24–27] Nevertheless, this tempting approach must be used with great care, as some energy minima which represent relevant metastable conformations and the barriers connecting such states may be hidden when projecting the actual large-dimensional free energy landscape onto a low-dimensional subspace. Besides, one-dimensional descriptions might suggest misleading unfolding pathways, consequence of this projection restriction.

Some recent studies try to address this problem by means of different strategies. It has been reported, for instance, that the mechanical unfolding of the Green Fluorescent Protein (GFP) can occur via two distinct routes which cannot be distinguished by a one-dimensional end-to-end representation.[28–30] Also, new procedures, such as the introduction of mutations, have been suggested in order to obtain more reliable information from single-molecule experiments. Engineered disulfide bonds can create constraints that block an unfolding pathway, allowing their detection.[30–32] In addition, the substitution of key aminoacids in the sequence might destabilize the intermediate states,[33] altering the unfolding mechanism.

In order to explore such aspects, we choose a coarse-grained model protein[34–43] and study it through a force-clamp protocol. The output of the simulations will be analyzed through two different approaches, allowing a comparison between the conclusions yielded by each. First, we build one dimensional free-energy profiles along the end-to-end distance and the fraction of native contacts. Second, we describe the configurational space of the system by using Markov-Model methods[44–47] and obtain the unfolding paths applying transition-path theory.[48–51]

Although recent works cast doubt on a simple low dimensional description of thermal (un)folding processes,[24,52] the one-dimensional approach is usually adopted for mechanical unfolding processes, due to the privileged direction imposed by the force.[15,17] In the case studied here, this fact, together with the simplicity of the protein structure, apparently point to a valid one-dimensional description of the unfolding process. Nevertheless, we find out that one-dimensional profiles lead to deceptive conclusions. In particular, these profiles suggest the existence of a metastable state (the half-stretched configuration, see Fig. 1) as a mechanical intermediate between the native and stretched states. Opposed to this, we find that unfolding occurs through two major routes defined by the existence of two different mechanical intermediates, not identified in the one-dimensional description. Although very stable, the half-stretched configuration plays a marginal role in the unfolding process. This

a)Rafa.T.Rojo@gmail.com

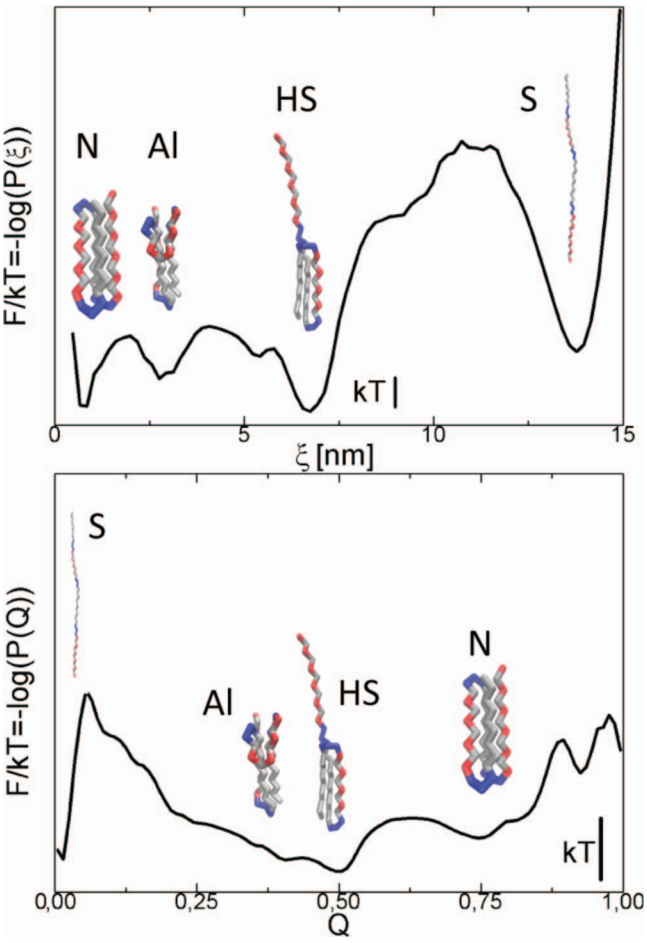

FIG. 1. Potential of mean force as a function of the end-to-end distance ξ and the fraction of native contacts Q.

multi-path picture can never be captured through a one-dimensional description. In addition we are able to systematically define all the individual unfolding pathways calculating their relative weight in the dynamics and yielding a complete and quantitative vision of the protein's landscape that completes the picture described in previous studies on the same system.[38,41,42]

II. MODEL

The BLN model[34,35] is a coarse grained off-lattice protein model in which the residues are represented by "colored" beads, hydrophobic (B), hydrophilic (L), and neutral (N). Due to its rich behavior, despite its simplicity, this model has been widely studied, with several modifications through time.[36–39,41,42] In particular, the 46-residue sequence (BLN-46) $B_9N_3(LB)_4N_3B_9N_3(LB)_5L$ folds into a four-strand β barrel showing nonetheless a frustrated ground state.[39] From now on, we number the β strands, being β_1 the N-terminal all-hydrophobic strand and β_4 the C-terminal.

The potential terms we use account for a stiff nearest-neighbor harmonic potential, a three-body bending interaction, a four-body dihedral interaction and a sequence dependent Lennard-Jones potential:[41,42]

$$V_{BLN} = \frac{1}{2}K\sum_{i=1}^{N-1}(r_{i,i+1}-r_0)^2 + \sum_{i=1}^{N-2}[A\cos\theta_i + B\cos 2\theta_i - V_0] + \sum_{i=1}^{N-3}[C_i(1+\cos\phi_i) + D_i(1+\cos 3\phi_i)] + \sum_{ij}\epsilon_{ij}\left(\frac{1}{r_{ij}^{12}} - \frac{c_{ij}}{r_{ij}^{6}}\right), \quad (1)$$

where r_{ij} is the distance between residues i and j, θ is the bending angle, and ϕ is the dihedral angle. Note that in this model, unlike Gō-models, the Lennard-Jones potential does not depend on the native contacts but includes sequence-dependent parameters. For parameter values see Ref. 41 and Appendix A.

We simulate the system by integrating Langevin equations of motion at constant temperature T and following a force-clamp protocol, where monomer 1 is fixed while a constant force is applied to the last monomer, 46, through a linear spring. Such equations are given by

$$m\ddot{\mathbf{r}}_i = -\gamma\dot{\mathbf{r}}_i - \nabla_i V_{BLN} + \mathbf{F}_i + \boldsymbol{\eta}_i, \quad (2)$$

where m is each residue unitary mass, γ is the friction coefficient, $\mathbf{F}$ is the external force applied in the z direction, and η_i is Gaussian white noise of zero average, holding fluctuation-dissipation theorem $\langle\eta_i\eta_j\rangle = 2T\gamma\delta(t-t')\delta_{ij}$.

This model protein has a well characterized unfolding transition (see Ref. 41 and Appendix C) at T_c and unfolds mechanically at F_U. We work from now on at $T = 0.55T_c$ and $F = 0.8F_U$ in order to maximize the number of configurations visited by the system. Lower forces would not populate the unfolded state while above F_U the unfolding would be irreversible.

III. METHODS

We present here the different methods use to analyze the simulated trajectories in order to understand the mechanical unfolding scenario of our model system.

A. Potential of mean force

The Potential of Mean Force (PMF) is a low dimensional (typically one-dimensional) characterization of the free energy landscape of a system, which relies on the choice of a reaction coordinate X. The PMF is simply $F/k_BT = -\log P(X)$, where $P(X)$ is the probability density of the chosen reaction coordinate X.

We will explore the PMF of the system (Sec. IV A) by using two different reaction coordinates. As the mechanical force imposes a privileged direction, the end-to-end distance $\xi = |\mathbf{r}_N - \mathbf{r}_0|$ appears as a natural choice. This magnitude is indeed widely used in most single molecule force

spectroscopy applications.[6,15,53,54] Additionally, we use the fraction of native contacts Q,[19,20] often reported in computational applications as a good magnitude for describing protein unfolding, based on the importance of topology on protein structure.

B. Principal components analysis

Principal Component Analysis (PCA) is a standard statistical method for reducing the dimensionality of a complex system such as biological molecule.[25–27] PCA performs a linear transformation by diagonalizing the covariance matrix $\mathcal{C}_{ij} = \langle y_i y_j \rangle - \langle y_i \rangle \langle y_j \rangle$, removing thus all internal correlations. The Principal Components (PCs) q_i are calculated as the projection of the trajectory onto each eigenspace. If we order the eigenvalues, the first largest PCs contain most of the fluctuations of the system and can be used as adequate reaction coordinates.

C. Conformational Markov network

In order to characterize the thermodynamical and kinetic properties of our system we build a Markov Model[44,45] by discretizing the state space of our molecule into a set $S = \{1, 2, \ldots, M\}$ of M conformational states defining the Conformational Markov Network of the system.[46,47] For our system, the conformational space is defined as the first three PCs, reducing greatly its dimensionality but keeping its essential features. With these three coordinates we maintain the 75% of the system fluctuations, while the remaining ones account for symmetric thermal fluctuations. Each of the coordinates is discretized into 30 bins of equal volume, thus $M = 27\,000$.

The Conformational Markov Network is built from the dynamical trajectories, by counting the occupation of each of the states π_i and calculating the transition matrix T_{ij} which measures the probability of going from state i to state j within time τ, being τ the time window or lag time used to analyze our trajectories ($\tau = 15$ ps in our case).

The transition matrix $\tilde{T}$ is ergodic and, if the molecule is in equilibrium, the occupation distribution π_i can be recovered as the eigenvector with eigenvalue 1. In such situation, detail balance condition holds, $\pi_i T_{ij} = \pi_j T_{ji}$, and π is the Boltzmann distribution.

D. Basins of attraction network

As the Conformational Markov Network is typically made up of thousands of nodes and links, hardly any relevant physical information can be directly obtained. A clustering or coarse-graining process is usually followed in order to group together nodes with similar physical features leading to an smaller, more meaningful network.

Here we apply the Stochastic Steepest Descent algorithm[47] (see Appendix B 2 for detailed algorithm). The advantage of this algorithm is that the network is systematically split into its basins of attraction, i.e., groups of nodes whose probability flux converges into a single node (minimum). The coarse-graining process does not rely in any arbitrary definition, but on the kinetic properties of the system. Physically, while each node would represent microstates of the system, the basins of attraction represent macrostates.

Onto this network we calculate a new transition matrix T_{ij} and the occupation probability of each basins π_i. Free energy differences from basin i and j are given by $\Delta F_{ij} = -k_B T \log \pi_i/\pi_j$. The mean escape time from basin i is defined as $\langle t_s \rangle = \tau/(1 - T_{ii})$, where τ is the time window used to sample the configurations, while transition times between basins i and j are defined as $\tau_{i \to j} = \tau/T_{ij}$.

E. Transition-path theory

The Markov Network defined above contains all thermodynamic and kinetic information of the system. Nevertheless, we are interested in computing the transition pathways between the set of native conformations to the fully stretched conformation. Transition-path theory provides the necessary tools for doing this.[48–50] We define A as the subset of basins which represent the native conformation while B is the subset of stretched basins. Our question is which is the typical sequence of intermediate I states to go from A to B.

The committor probability q_i^+ is defined as the probability, when starting at state i, to reach set B next rather than A. In our case, this is the unfolding probability. By definition $q_i^+ = 0$ if $i \in A$ and $q_i^+ = 1$ if $i \in B$. Mathematically, the committor probability can be computed by solving the following system of linear equations:

$$-q_i^+ + \sum_{k \in I} T_{ik} q_k^+ = -\sum_{k \in B} T_{ik}. \tag{3}$$

For a molecule in equilibrium, the backward-committor probability q_i^- is simply $q_i^- = 1 - q_i^+$.

The transition matrix T_{ij} contains information from every possible trajectory which appears in the equilibrium ensemble of the molecule. In order to extract the contributions from the unfolding trajectories $A \to B$, we calculate the effective flux f_{ij} defined as the probability flux from $i \to j$ contributing to the $A \to B$ transition:

$$f_{ij} = \pi_i q_i^- T_{ij} q_j^+. \tag{4}$$

If we want to calculate the unfolding flux, removing recrossings which might appear in a $A \to B$ transition, we need to define the net flux as

$$f_{ij}^+ = \max[0, f_{ij} - f_{ji}], \tag{5}$$

f_{ij}^+ defines a network of fluxes that go from A to B. The total unfolding flux F represents the expected number of $A \to B$ transitions per time window τ and is defined as

$$F = \sum_{i \in A} \sum_{j \notin A} \pi_i T_{ij} q_j^+. \tag{6}$$

In order to decompose this flux network onto individual pathways P_i, different approaches can be applied.[50,51] Here we base our strategy on the bottleneck algorithm, where given an individual pathway, the bottleneck (rate limiting step) is identified as the minimal net flux of the path f_i and subtracted

from every remaining net flux f_{ij}^{+}. The process is iterated until the network is fully decomposed into a set of individual pathways P_i.

IV. RESULTS

In order to elucidate the unfolding mechanism under the effect of mechanical force for our model protein, we have performed six long equilibrium simulations. Every simulation starts from the native configuration, is equilibrated for 3 μs and then runs up to 3 ms.

A. One dimensional description: The potential of mean force

Figure 1 shows the PMF calculated along the end-to-end distance ξ and the fraction of native contacts Q of our model protein. The profile for ξ shows four clear minima that can be identified with four different configurations, considering that each of the β strands has a length of $\xi \sim 3$ nm. In the native configuration (N) $\xi \sim 0$ nm, as the extremal β strands are oriented in the same direction. In the aligned configuration (Al) the second strand $(LB)_4$ is bent so that the extremes are aligned in the pulling direction and $\xi \sim 3$ nm. The half-stretched configuration (HS) appears as an stable minima at $\xi \sim 6$ nm, as the fourth $(LB)_5$ strand is unfolded. The fully stretched configuration (S), with $\xi \sim 12$ nm, shows the protein totally unfolded, as an stretched polymer.

These states can also be identified in the Q profile. State S has all contacts broken $Q \sim 0$, while Al and HS maintain around half of the contacts ($Q \sim 0.5$). The N configuration shows a minimum at $Q \sim 0.75$, as thermal fluctuations break on average some of the contacts.

Remarkably, for this value of the force, the HS configuration correspond to the lowest minimum in both free energy profiles, and thus is the most stable configuration. Its position in the PMF suggests that it also has a relevant role in the stretching pathways, appearing as a clear mechanical intermediate between the native and fully stretched configuration. In addition, it is necessary to jump over a barrier of several k_BT to reach state S while the other states are separated by low barrier. This suggest a fast dynamics between N and HS and longer time scales to visit state S.

B. One-dimensional description: ξ versus time trajectories

In order to complete the one-dimensional vision of the unfolding mechanism, it is useful to look directly on individual trajectories, as it is usually done in single-molecule studies. Figure 2 shows a large snapshot of a simulation representing $\xi(t)$ over 12 μs and includes two complete unfolding transitions. For this value of the force, the protein transits mainly between the N ($\xi \sim 1$ nm) and HS configurations ($\xi \sim 6.5$ nm), with some periods dwelling in the Al configuration ($\xi \sim 3$ nm). The unfolding transition appears as a rare event. The highlighted windows depict two examples of unfolding transitions, which contradict the conclusions derived from the PMF. We can see how both unfolding processes seem

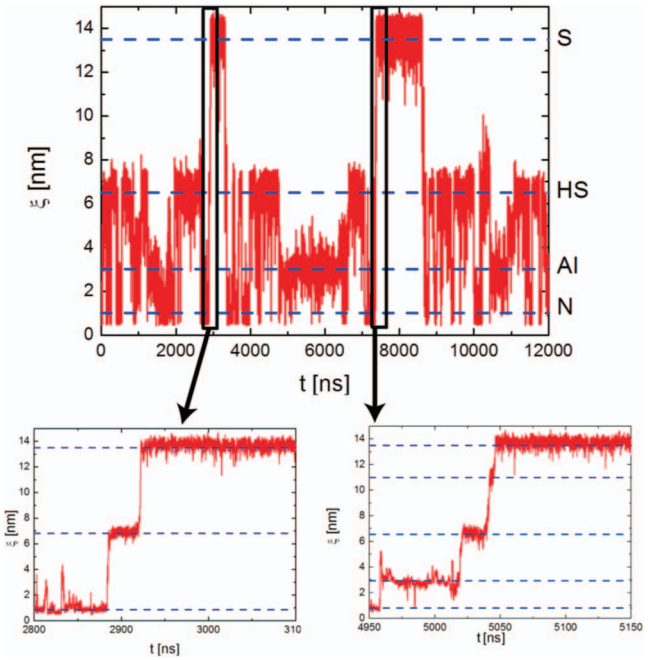

FIG. 2. Time evolution of coordinate ξ over 12 μs. The snapshot includes several transitions between N and HS and two complete unfolding-refolding transitions.

to follow different pathways, so a one-dimensional landscape is incomplete.

The first transition shows a clear intermediate at $\xi \sim 6.5$ nm, while the second one has at least three intermediates of different life times, with at least one with large ξ that cannot be identified from Fig. 1. The exact nature of the intermediates is hard to tell from the one-dimensional trajectories, although the one at 6.5 nm coincides with HS and the one with 3 nm with Al. This point will be clarified later on. Nevertheless, this picture suggests that unfolding occurs through a complex, rough landscape that cannot be simplified through a one-dimensional profile.

C. Two dimensional description: Principal component analysis

Before describing the Markov Model of the system, it is worth to exploit further the information PCA provides. As explained previously, we build the Markov network by discretizing the first three PCs, which define our conformational space, with lower dimensionality, but still capturing the main aspects of the system dynamics.

Figure 3 shows the free-energy landscape along the first two principal components $\Delta F/k_BT = -\log P(q_1, q_2)$. Its basic features agree with the one dimensional landscapes shown in Sec. IV B, as three major wells are found. Nevertheless, we see also clear differences, being the PCs able to capture better the details of the free-energy landscape. Each of these major wells have a rough structure, showing a set of minor wells separated by small energy barriers $\sim 2k_BT$, revealing thus a richer variety of configurations. Moreover, two new low populated wells appear between the folded structures (native and half-stretched) and the fully-stretched configurations. These new

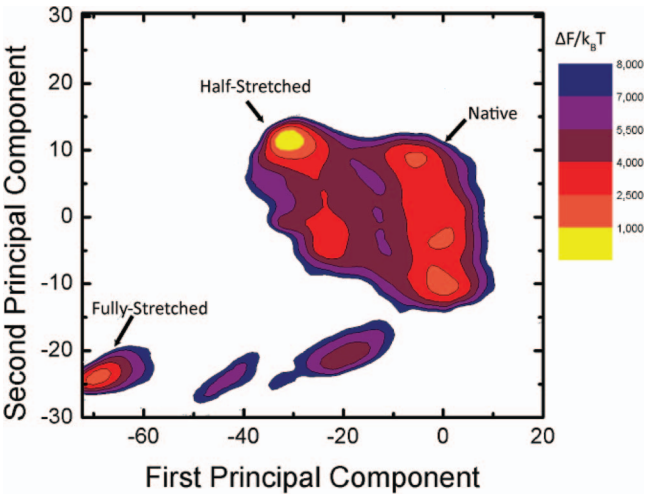

FIG. 3. Free-energy landscape along the first two PCs.

states could suggest the existence of different unfolding pathways, where the half-stretched configuration does not necessarily plays the role of mechanical intermediate.

D. Equilibrium ensemble of the model protein: The basin network

The built microstate network is made up of 1876 nodes related kinetically through 23 995 links. After applying the Stochastic Steepest Descent algorithm,[47] the network is clustered into 30 basins connected through 1290 links. In order to obtain a good description of the system, we keep only those basins which were visited at least 0.001% of the trajectory ($\pi_i > 10^{-5}$), avoiding pathological or extremely rare states. After this refinements, we keep 13 macrostates, connected through 65 edges, including auto-links.

Figure 4 (upper) shows a graphical representation of the basin network, where the size of each bead (node) is proportional to its occupation π_i. The spatial arrangement of the nodes was calculated applying the *Force Atlas* algorithm,[55] where an artificial dynamics is simulated. This dynamics is based in considering each link as a linear spring and including a certain repulsion between nodes, until an equilibrium configuration is obtained. The nodes are colored according to the modularity class they belong to,[56] having five different classes. Lower panel of Fig. 4 shows a representative structure of each basin (macrostate), including the label which identifies them.

Configurations N_1 and N_2 correspond to native-like states and will define the native set A due to its structural similarity and high Q value. The aligned configuration *Al*, already identified in Fig. 1, appears close to N_1 and N_2 in Fig. 4 but does not belong to the native set since it gives very different Q and ξ values. Basin *HS* is the Half-Stretched Configuration, the most stable macrostate under these conditions. State S is the Fully-Stretched Configuration, while the remaining 8 basins are labelled as intermediate states and will be discussed further on.

Table I shows information about each of the identified macrostates. π_i is the occupation of basin i, $\langle t_s \rangle$ the mean

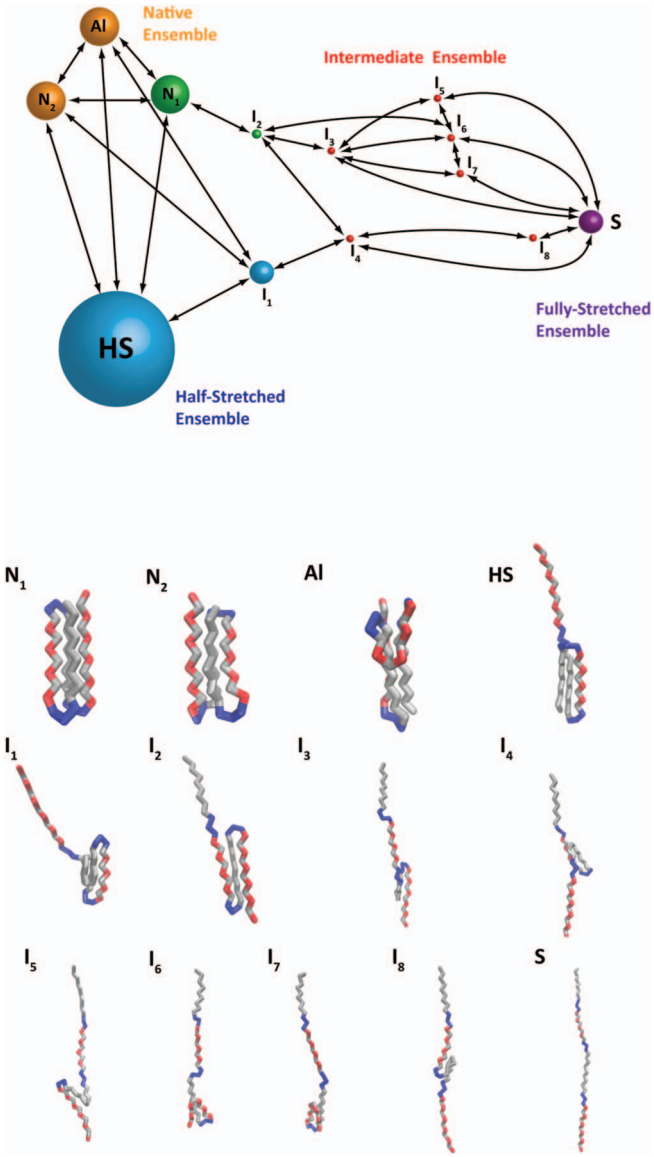

FIG. 4. Basins of attraction Markov Network (upper). We represent the 13 basins with $\pi > 10^{-5}$ where the size of the bead is proportional to π_i. The bidirectional arrows connecting nodes represent allowed transitions (the magnitude of T_{ij} is not shown). Each basin is labelled according to the configuration they encode. Representative structure associated to each basin (lower).

TABLE I. Description of the basins of attraction.

#	π_i	$\langle t_s \rangle$ (ps)	$\langle Q \rangle$	$\langle \xi \rangle$ (nm)	f_{NN}	q_i^+
N_1	0.15	559	0.75	0.8	0.13	0.0
N_2	0.14	495	0.73	0.9	0.30	0.0
Al	0.14	272	0.40	2.6	0.60	1.4×10^{-4}
HS	0.44	2982	0.46	6.5	0.18	9.2×10^{-4}
I_1	0.07	362	0.25	4.8	0.66	1.2×10^{-3}
I_2	0.01	2586	0.35	6.8	0.40	0.12
I_3	6.67×10^{-5}	120	0.12	9.0	0.23	0.29
I_4	1.3×10^{-4}	198	0.11	10.1	0.54	0.34
I_5	1.9×10^{-5}	64	0.10	9.6	0.60	0.51
I_6	3.9×10^{-4}	163	0.14	8.55	0.30	0.53
I_7	3.3×10^{-4}	176	0.13	9.35	0.50	0.58
I_8	2.5×10^{-5}	56	0.09	10.5	0.70	0.71
S	0.06	75000	0.01	13.7	0.00	1

escape time (defined above), $\langle Q \rangle$ the mean fraction of native contacts and $\langle \xi \rangle$ the mean end-to-end distance, both calculated from the marginal distributions of such magnitudes on each basin. It is remarkable that in many cases such distributions are not unimodal, so the actual meaning of the average must be taken with care. q_i^+ are the committor probabilities from the native (N_1 and N_2) to the stretched (S) configuration this is: the unfolding probability of basin i. We show also an additional magnitude f_{NN}, the fraction of *non-native* contacts. The model we use allows non-native interactions which can stabilize configurations which would not form in Gō-like models.

It is important to stress the difference between the two native basins N_1 and N_2, as they have very different connectivity features in the network, belonging to different modularity classes. Configuration N_1 is closer to the native structure, given the arrangement of the neutral turns, while N_2 shows larger fluctuations, leading to a loss of some native contacts, and the formation of non-native ones. Interestingly, N_1 is more connected to the Intermediate States than N_2, which shows fast transition times to *HS*, $\tau_{N_2 \to HS} = 557$ ps, while $\tau_{N_1 \to HS} = 13.5 \times 10^6$ ps. In fact, they are both scarcely connected, $\tau_{N_2 \to N_1} = 14 \times 10^3$ ps and $\tau_{N_1 \to N_2} = 15 \times 10^3$ ps, reason why they belong to a different modularity class. In this regard, in spite its structural similarity which overlap both states in the PMF description, their actual role in the configurational space is quite different.

In this sense, the first contradictions with the conclusions yielded by the PMF description appear here. While both descriptions agree coarsely in the main features of the equilibrium ensemble of the system, revealing three major states (native, half-stretched, and fully-stretched), the role of such states and the presence of other relevant configurations is hidden in the one-dimensional projection. N_1 and N_2 states are integrated into the same high Q or low ξ minimum, will the intermediate low-populated states which connect to the stretched state are impossible to be identified in the one-dimensional representation.

E. The unfolding pathways: Transition path theory

In order to decipher the actual unfolding mechanism of our model protein under the effect of a mechanical force, we apply transition-path theory to the basin network, as explained in Sec. III.

We define the native set A as basins N_1 and N_2, while the stretched set B is just made up of basin S. According to this definitions, we calculate the committor probabilities, shown in Table I. Figure 5 shows the net flux network, being the thickness of the arrows proportional to the net flux f_{ij}^+. The total unfolding flux is $F = 2.9 \times 10^{-7}$ ps^{-1}, meaning that we observe an unfolding transition every 3.5 μs, approximately.

We decompose the net flux network by identifying first the strongest pathway, remove it from the network and repeat the process until there is no path from set A to set B. Due to the size of our network, this process can be done manually, although computational applications can be used.[50,51] We identify a total of 9 different paths leading from A to B. After decomposing the network into these 9 paths, unconnected regions still remain due to the presence of *trap states*[49] that carry around 20% of the flux. Figure 6 shows the 6 more relevant paths, which carry 89% of the *unfolding* flux.

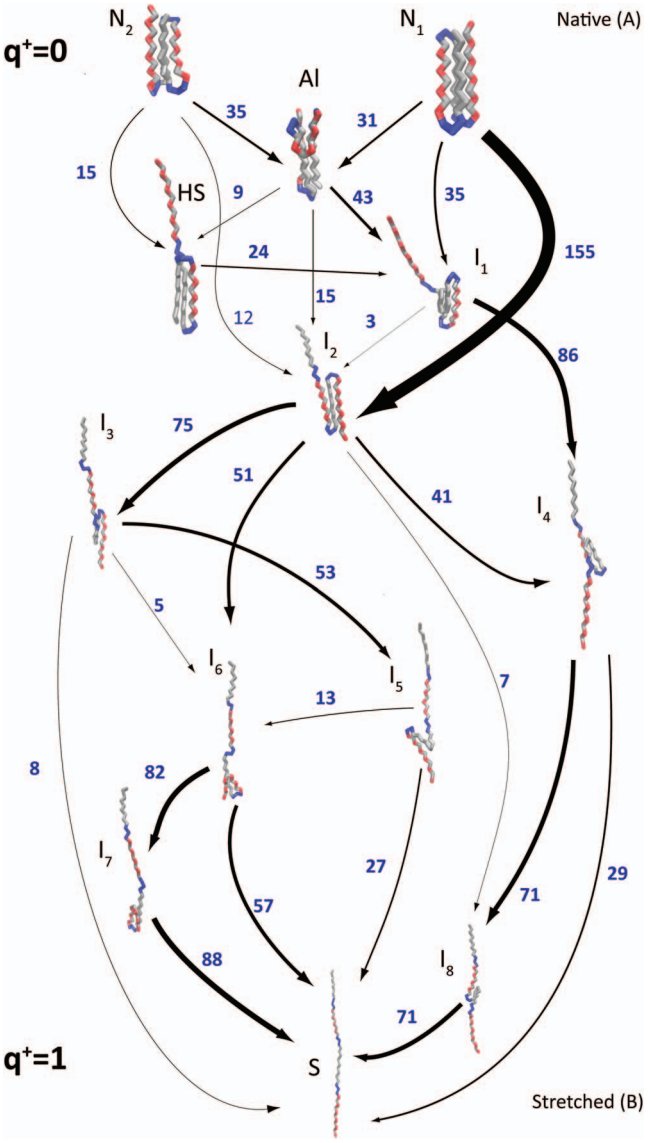

FIG. 5. Folding flux for the model protein. The network depicts the 85% most relevant unfolding pathways for the 46-*mer BLN* model protein Each of the 13 configuration identified with the Stochastic Steepest Descent algorithm are shown here, together with the label which identifies them. The configurations are arranged vertically according to their committor probability (not in scale). The arrows connecting configurations represent the unfolding net flux f_{ij}^+, with their thickness is proportional to the magnitude of the flux. The numbers next to the arrows give the flux magnitude in 10^{-9} ps^{-1}.

From the 9 pathways, 7 start from conformation N_1 while just 2 from N_2. This is a remarkable fact, being N_1 closer to the native structure than N_2, as discussed in Sec. IV D. In addition, states I_1 and I_2 appear as the actual intermediates for the unfolding mechanism: $A \to B$ is forbidden in case these two states are removed from the net flux network. Out of the 9 pathways, 6 of them pass through state I_2 and 3 through state I_1.

The construction of the Markov Model from the PCs and the use of transition-path Theory help us to unveil the actual

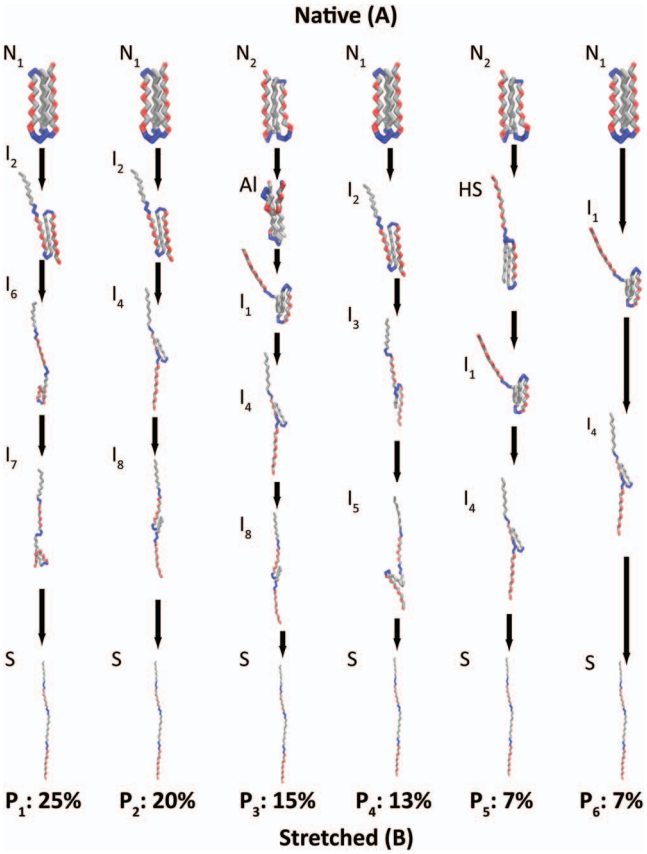

FIG. 6. Model protein unfolding pathways. The six pathways carrying most of the total flux (up to 89%) are explicitly shown.

unfolding mechanism and its driving process. First of all, being *HS* a notably relevant metastable state ($P_e = 0.44$), its role in the unfolding mechanism is completely marginal, as it just appears in path P_5, with a low weight (7%). Configuration *HS* is mainly involved in the fast transitions to the native ensemble, which actually occupy the largest fraction of the trajectory, as seen in Fig. 2. In *HS* strand β_4 is detached, aligning the terminal strands in the direction of the force while keeping the hydrophobic core formed by β_1 and β_3 interaction.

Configuration I_2 and *HS* have similar ξ values (see Table I), as β_1 strand is unfolded. Due to this, both states overlap in the one dimensional description, although I_2 low population ($P_e = 0.01$) makes its contribution negligible, and thus hard to be directly identified. Nevertheless, its role is remarkably different as I_2 plays a central role in the unfolding process, since the loss of the hydrophobic core destabilizes this structure, driving the unfolding mechanism.

The other major folding route includes I_1 as the intermediate. This configuration might look similar to *HS* as it also has β_4 strand unfolded. However, the core adopts a compact, globular structure that is sustained by a large number of non-native interactions (nearly the 70% of the contacts) between the hydrophobic residues in strands β_1 and β_4. This state is also likely lost in the one-dimensional profile. This configuration drives then the unfolding through states I_4 and I_8, which are also stabilized through a large proportion of non-native contacts (see Table I).

The possibility of forming non-native contacts is responsible also of structure *Al*, with relevant stability ($P_e = 0.14$), and a 60% of non-native interactions. This structure plays a certain role in the unfolding pathways (as it allows to reach I_1 from native state N_2), but also participates in the fast dynamics between *HS* and the native set, as can be seen in Fig. 2.

V. CONCLUSIONS AND DISCUSSION

In this paper we have presented the detailed analysis of the unfolding process of a model protein under the presence of a mechanical pulling force. This scenario mimics force clamp single molecule experiments, where proteins or nucleic acids are subject to a constant external force that drives their unfolding. Due to the limitation of available observables, these experiments are often analyzed by reconstructing their free-energy landscape along the pulling direction through different existing techniques.[6,15–18,53,54] This approach is often followed in many computational studies by using different reaction coordinates.[19–24]

In this sense we wanted to reproduce a similar protocol and explore the conclusions yielded by a one-dimensional analysis and a multidimensional Markov model approach. The simplicity of our model protein, and the fact that the force sets a privileged direction invites to a one-dimensional characterization. Nonetheless, we have seen how both approaches lead to contradictory conclusions. The PMF description shows the existence of three major states, the native, the stretched or denatured and a metastable half-stretched configuration, which seems to play the role of mechanical intermediate in the unfolding process. A close look to a one-dimensional trajectory casts doubt on this conclusion, revealing a rough landscape, where different unfolding pathways seem to be possible. Nevertheless, this simple approach is not able by itself to provide a detailed vision of the unfolding mechanism, as the one provided by the method applied here.

A more detailed multidimensional study changes dramatically the unfolding picture. Being the most populate one, *HS* state plays a marginal role in the unfolding pathway, with just 7% of the unfolding flux passing through it. The true mechanical intermediates are states I_1 and I_2, building the two major unfolding routes, both related to the loss of the hydrophobic core that destabilizes the structure and drives the unfolding process. Due to their low population, both states are lost in the projection onto a single coordinate. The two one-dimensional pathways shown in Fig. 2 are now clear, as the state at $\xi \approx 6.5$nm would actually correspond to I_2, and the multi-pathway scenario is systematically revealed with all the intermediates. Interestingly, the configuration space of our system stresses the importance of non-native interactions, as states like *Al* or I_1 have low values of Q, while a large fraction of non-native contacts, so they would not have been identified with a Gō version of the model.

In this regard, due to the existence of multiple pathways, independently of the chosen reaction coordinate, a one-dimensional picture would *never* be enough to characterize the unfolding pathway of this system. Thus, our work

differs from those which put attention on the proper choice of the reaction coordinate.[17,18] The necessity of multidimensional descriptions indeed has been warned in the last years to understand thermal unfolding, where the protein transits from a low-entropy state (native) to a high-entropy one (denatured).[24,52] Also, recent works on mechanically pulled proteins, warn about the possibility of various unfolding pathways or the existence of multiple mechanical intermediates, which can be worked out by a combination of experiments and computer simulations, and the use of engineered proteins to force the unfolding route through a modified free-energy landscape.[30–32] Nevertheless, the one-dimensional picture, is still vastly assumed in mechanical unfolding processes, both in experimental and computational applications.

Regarding our analysis Markov Model protocol, we stress two major differences when compared to most works of this community. First, it is important to note that we are actually using the PCs as reaction coordinates in order to reduce the system dimensionality. Nevertheless, these coordinates has been proven to capture successfully the most relevant dynamical events of complex systems such as biomolecules. In our case, three coordinates are enough, as the remaining ones account merely for gaussian thermal fluctuations. Second, we stress on the importance of the coarse-graining mechanism applied to the original Conformational Markov Network,[47] which is able to systematically cluster the network based only on the kinetic properties of the system.

Although extremely simple molecular assays such as DNA or RNA hairpins could fit into a single reaction coordinate description,[6] increasing slightly the complexity of the molecule leads to a dramatical rise in the complexity of the actual free energy landscape in the system, requiring more detailed studies. In this sense, molecules such as multiple nucleic-acid hairpins,[57] protein-ligand complexes[58] or any mechanically pulled protein,[59] appear as potential systems where a one-dimensional description takes the risk of leading to a clear misunderstanding of the actual complexity of their conformational space and the dynamical processes to which they are subject.

ACKNOWLEDGMENTS

The authors acknowledge support from the Spanish MINECO, Project No. FIS2011-25167 cofinanced by FEDER funds, and Gobierno de Aragón (FENOL group). R. T-R. is supported by Spanish government fellowship FPU-2012-2608.

APPENDIX A: MODEL PARAMETERS AND SIMULATION PROTOCOL

We simulate our system using the following adimensional parameters in Eq. (1),:

- $\boldsymbol{V_1}$: $K = 50$, $r_0 = 1$.
- $\boldsymbol{V_2}$: $A = 5.118$, $B = 5.308$, $V_0 = -5.295$.
- $\boldsymbol{V_3}$: $C_i = 0$ and $D_i = 0.2$ if two or more aminoacids are neutral, and $C_i = D_i = 1.2$ otherwise.
- $\boldsymbol{V_4}$: there are three different cases, according to the character of the aminoacids.
  1. $c_{ij} = 0$ and $\epsilon_{ij} = 4$ if i or j are neutral.
  2. $c_{ij} = 1$ and $\epsilon_{ij} = 4$ if i and j are hydrophobic.
  3. $c_{ij} = -1$ and $\epsilon = 8/3$ in the remaining cases.

All simulations were carried out using self-built code, integrating the overdamped Langevin equations described above with an stochastic second order Runge-Kutta algorithm.[60]

Physical units can be easily recovered in the following way. Length unit is defined by the $C_\alpha - C_\alpha$ distance $r_0 = 0.38$ nm. Energy units are defined as the energy of a hydrogen bond $\epsilon_H \approx 1.7 k_B T$, being force units $\tilde{F} \approx 17.3$ pN. Mass unit is that of an average aminoacid $m_a \approx 3 \times 10^{-22}$ kg. In this sense our time units $\tilde{t} = \sqrt{m_a r_0^2/\epsilon_H} \approx 3$ ps, and the damping is that of water $\gamma \approx 10 \frac{m_a}{\tilde{t}}$.

Six trajectories at $F = 0.8 F_U$ were simulated (with $F_U \approx 20$ pN), were monomer 1 was kept fixed while force was exerted to monomer N through a linear spring. Each simulation covered a total time of 3 ms, with a previous thermalization process of 3 μs. The integration step is $dt = 0.005\tilde{t}$ and the time window to sample the trajectories $\tau = 5\tilde{t}$.

APPENDIX B: ANALYSIS METHODS

1. Conformational Markov network

The Conformational Markov Network (CMN)[46,47] appears as a useful coarse-grained representation of large stochastic trajectories. This picture is obtained by discretizing the conformational space explored by the system and considering the dynamical jumps between the discretized configurations along the simulation. In this sense, the nodes of the complex network are defined by the discretized states, while the links account for the observed transitions between them. The arising network is thus a weighted and directed graph.

In our case, the conformational space is defined by the three first principal components, in order to reduce the number of degrees of freedom, keeping indeed the essential features of our system. We divide each of the principal component into 30 cells of equal volume. Our discretized conformational space is thus made up of 30^3 possible states, which may be or not occupied within the stochastic trajectory. We assign each node a weight π_i accounting for the fraction of trajectory that the system has visited within the trajectory. The normalization condition $\sum_i \pi_i = 1$ holds. Second, the value T_{ij} is assigned to each directional link accounting for the dynamical jumps from node j to i. Self-loops can exist, and thus $T_{ii} \neq 0$. Finally, the normalization condition $\sum_i T_{ij} = 1$ is forced. According to this, the CMN is totally defined by the occupancy vector $\Pi = P_i$ and the transition matrix $\tilde{T} = \{T_{ij}\}$. The matrix $\tilde{T}$ is the transition probability of the Markov chain defined by

$$\Pi(t + \Delta t) = \tilde{T}\,\Pi(t), \tag{B1}$$

where $\Pi(t)$ it the probability distribution at time t. If the trajectory is long enough, $\tilde{T}$ is ergodic and time invariant, vector Π coincides with the stationary distribution associated with the Markov chain $\Pi = \tilde{T}\,\Pi$. Moreover, the detailed balance

condition must hold

$$T_{ji}\pi_i = T_{ij}\pi_j. \tag{B2}$$

2. Stochastic steepest descent

Once we have translated the molecular dynamics trajectories onto a CMN, we apply the stochastic steepest descent (SSD) algorithm[47] in order to split it into its basins of attraction in an efficient way, obtaining in turn useful thermostatistical information about the system. The SSD algorithm is inspired in the deterministic steepest descent algorithm used to find minima in a multidimensional surface. We define the assisting vector $\mathbf{U} = \{u_i\}$, where i labels the nodes. The steps of the SSD algorithm are as follows:

1. We start with $\mathbf{U} = \mathbf{0}$.
2. Select randomly a node l with $u_l = 0$ and write an auxiliary list of nodes adding l as first entry.
3. Select within the neighbors of l the node m that follows the maximum probability flux, this is $T_{ml} = \max\{T_{jl,\forall j \neq l}\}$. Check which of the following conditions is fulfilled:
   (a) If $T_{ml} > T_{lm}$ and $u_m = 0$, add m to the list and go back to 3, using m instead of l.
   (b) If $T_{ml} > T_{lm}$ and $u_m \neq 0$ write the labels of all the nodes in the list as $u_j = u_m$. Go back to step 3.
   (c) If $T_{ml} \leq T_{lm}$ remove link $l \rightarrow m$ from the graph. Return to point 3.

This process ends when every node in the CMN has been labelled, this is $u_i \neq 0 \forall i$. Then, the whole conformational space has been characterized and every node is connected with its local minima in the FEL. All nodes with the same label belong to the same basin in this FEL and therefore we can associate them with the same conformational state.

Given the basin partition, a new CMN network can be built, taken the basins themselves as new nodes. The occupation probabilities will now be defined as $\pi_\alpha = \sum_{i\in\alpha}\pi_i$, while the new transition matrix $\tilde{T}$ is built, with elements $T_{\beta\alpha} = \sum_{i\in\alpha}\sum_{j\in\beta}T_{ji}\pi_i/\sum_{i\in\alpha}\pi_i$. From these definitions, transition times can be easily calculated as $t_{\alpha\beta} = \tau/T_{\beta\alpha}$, being τ the time window used for the network construction. The relative free energy of basin α with respect to basin β is simply $\Delta F_\alpha = -k_B T \log(\pi_\alpha/\pi_\beta)$.

APPENDIX C: THERMAL AND MECHANICAL CHARACTERIZATION

We start by characterizing the protein from a thermal and mechanical point of view, in order to know the suitable range of force and temperature to work with. Although more detailed characterizations have been made in previous works[42] we focus on the thermodynamical transition at T_c, reflected on a peak in the heat capacity, as it can be seen in Fig. 7. The heat capacity is calculated as $C_p = (k_B T)^{-2}[\langle E^2\rangle - \langle E\rangle^2]$, with E the total internal energy. We work at $T = 0.55T_c$, below the transition, but with allowing enough fluctuation for the system to explore its configurational space.

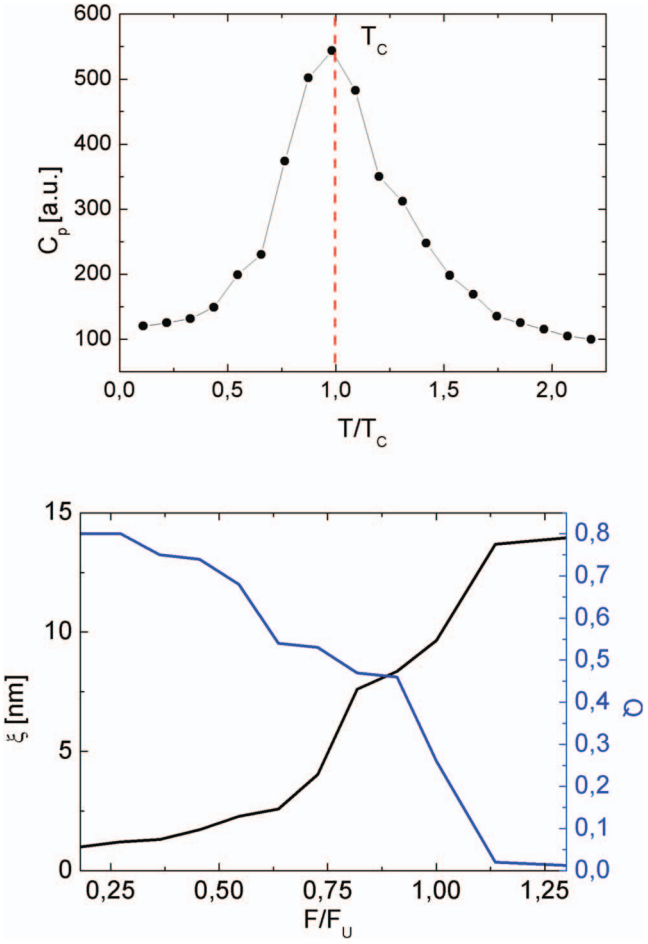

FIG. 7. Thermal and mechanical characterization of the model protein. At T_c it exhibits a thermodynamical unfolding transition, reflected in a peak on the heat capacity (in arbitrary units). Force also induces an unfolding transition at F_U, leading to the fully stretched conformation.

When applying force to the protein, it exhibits also a transition at F_U, where the protein unfolds mechanically to the fully stretched configuration. At this force, the end-to-end distance ξ increases abruptly, while the fraction of native contacts Q drops to 0. Around $F = 0.75F_U$ a first change of behavior can be seen, due to the population of the half-stretched configuration, which leads to a drop to $Q \sim 0.5$ and $\xi \sim 7$ nm.

[1]J. N. Onuchic and P. G. Wolynes, Curr. Opin. Struct. Biol. **14**, 70 (2004).
[2]T. R. Sosnick and D. Barrick, Curr. Opin. Struct. Biol. **21**, 12 (2011).
[3]C. D. Snow, H. Nguyen, V. S. Pande, and M. Grubele, Nature **420**, 102 (2002).
[4]B. Onoa, S. Dumont, J. Liphardt, S. B. Smith, I. Tinoco, Jr., and C. Bustamante, Science **299**, 1892 (2003).
[5]K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, Science **334**, 517 (2011).
[6]J. Liphardt, B. Onoa, S. B. Smith, I. J. Tinoco, and C. Bustamante, Science **292**, 733 (2001).
[7]F. Ritort, J. Phys.: Condens. Matter **18**, R531 (2006).
[8]W. J. Greenleaf, M. T. Woodside, and S. M. Block, Annu. Rev. Biophys. Biomol. Struct. **36**, 171 (2007).
[9]R. Liu, S. Garcia-Manyes, A. Sarkar, C. L. Badilla, and J. M. Fernández, Biophys. J. **96**, 3810 (2009).
[10]M. Carrión-Vázquez, A. F. Oberhauser, S. B. Fowler, P. E. Marszalek, S. E. Broedel, J. Clarke, and J. M. Fernández, Proc. Natl. Acad Sci. U.S.A. **96**, 3694 (1999).

[11]H. Li, A. F. Oberhauser, S. B. Fowler, J. Clarke, and J. M. Fernández, Proc. Natl. Acad. Sci. U.S.A. **97**, 6527 (2000).
[12]R. B. Best, S. B. Best, J. L. Toca-Herrera, and J. Clarke, Proc. Natl. Acad. Sci. U.S.A. **99**, 12143 (2002).
[13]B. Schuler, E. A. Lipman, and W. A. Eaton, Nature **419**, 743 (2002).
[14]B. Schuler and W. A. Eaton, Curr. Opin. Struct. Biol. **18**, 16 (2008).
[15]O. K. Dudko, G. Hummer, and A. Szabo, Phys. Rev. Lett. **96**, 108101 (2006).
[16]O. K. Dudko, G. Hummer, and A. Szabo, Proc. Natl. Acad. Sci. U.S.A. **105**, 15755 (2008).
[17]O. K. Dudko, T. G. W. Graham, and R. B. Best, Phys. Rev. Lett. **107**, 208301 (2011).
[18]M. T. Woodside and S. M. Block, Annu. Rev. Biophys. **43**, 19 (2014).
[19]P. G. Wolynes, Q. Rev. Biophys. **38**, 405 (2005).
[20]P. G. Wolynes, J. N. Onuchi, and D. Thirumalai, Science **267**, 1619 (1995).
[21]R. B. Best, G. Hummer, and W. A. Eaton, Proc. Natl. Acad. Sci. U.S.A **110**, 17874 (2013).
[22]E. R. Henry, R. B. Best, and W. A. Eaton, Proc. Natl. Acad. Sci. U.S.A. **110**, 17880 (2013).
[23]S. Piana, K. Lindorff-Larsen, and D. E. Shaw, Proc. Natl. Acad. Sci. U.S.A. **110**, 5915 (2013).
[24]A. Altis, P. H. Nguyen, R. Hegger, and G. Stock, J. Chem. Phys. **126**, 244111 (2007).
[25]A. E. García, Phys. Rev. Lett. **68**, 2696 (1992).
[26]A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, Proteins **17**, 412 (1993).
[27]G. G. Maisuradze, A. Liwo, and H. A. Scheraga, Phys. Rev. Lett. **102**, 238102 (2009).
[28]H. Dietz and M. Rief, Proc. Natl. Acad. Sci. U.S.A. **101**, 16192 (2004).
[29]C. Hyeon, R. I. Dima, and D. Thirumalai, Structure **14**, 1633 (2006).
[30]M. Mickler, R. I. Dima, H. Dietz, C. Hyeon, D. Thirumalai, and M. Rief, Proc. Natl. Acad. Sci. U.S.A. **104**, 20268 (2007).
[31]M. Bertz and M. Rief, J. Mol. Biol. **378**, 447 (2008).
[32]M. Bertz, H. Chen, M. J. Feige, T. M. Franzmann, J. Buchner, and M. Rief, J. Mol. Biol. **400**, 1046 (2010).
[33]P. E. Marszalek, H. Lu, H. Li, M. Carrion-Vazquez, A. F. Oberhauser, K. Schulten, and J. M. Fernandez, Nature **402**, 100 (1999).
[34]J. D. Honeycutt and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **87**, 3526 (1990).
[35]J. D. Honeycutt and D. Thirumalai, Biopolymers **32**, 695 (1992).
[36]S. Brown, N. J. Fawzi, and T. Head-Gordon, Proc. Natl. Acad. Sci. U.S.A. **100**, 10712 (2003).
[37]S. Brown and T. Head-Gordon, Protein Sci. **13**, 958 (2004).
[38]D. J. Lacks, Biophys. J. **88**, 3494 (2005).
[39]D. J. Wales and P. E. J. Dewsbury, J. Chem. Phys. **121**, 10284 (2004).
[40]M. A. Miller and D. J. Wales, J. Chem. Phys. **111**, 6610 (1999).
[41]A. Imparato, S. Luccioli, and A. Torcini, Phys. Rev. Lett. **99**, 168101 (2007).
[42]S. Luccioli, A. Imparato, S. Mitternacht, A. Irbäck, and A. Torcini, Phys. Rev. E **81**, 010902(R) (2010).
[43]D. J. Wales and T. Head-Gordon, J. Phys. Chem. B **116**, 8394–8411 (2012).
[44]*An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*, Advances in Experimental Medicine and Biology, edited by G. R. Bowman, V. S. Pande, and F. Noé (Springer, 2014).
[45]S. J. Klippenstein, V. S. Pande, and D. G. Truhlar, J. Am. Chem. Soc. **136**, 528 (2014).
[46]F. Rao, and A. Catfisch, J. Mol. Biol. **342**, 299 (2004).
[47]D. Prada-Gracia, J. Gómez-Gardeñes, P. Echenique, and F. Falo, PLoS Comput. Biol. **5**, e1000415 (2009).
[48]E. Vanden-Eijnden, J. Stat. Phys. **123**, 503 (2006).
[49]F. Noe, C. Schutte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, Proc. Natl. Acad. Sci. U.S.A. **106**, 19011–19016 (2009).
[50]P. Metzner, C. Shutte, and E. Vanden-Eijnden, Multiscale Model. Simul. **7**, 1192 (2009).
[51]R. Banerjee and R. I. Cukier, J. Phys. Chem. B **118**, 2883 (2014).
[52]S. V. Krivov and M. Karplus, Proc. Natl. Acad. Sci. U.S.A. **101**, 14766 (2004).
[53]G. Hummer and A. Szabo, Proc. Natl. Acad. Sci. U.S.A. **98**, 3658 (2001).
[54]M. Li, A. M. Gavovich, and A. I. Voitenko, J. Chem. Phys. **129**, 105102 (2008).
[55]M. Bastian, S. Heymann, and M. Jacom, in ICWSM'09: Proceedings of the International AAAI Conference on Weblogs and Social Media, San Jose, California (AAAI, 2009).
[56]V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, J. Stat. Mech: Theory Exp. **10**, P1000 (2008).
[57]A. Alemany, A. Mossa, I. Junier, and F. Ritort, Nat. Phys. **8**, 688 (2012).
[58]Y. Suzuki and O. K. Dudko, Phys. Rev. Lett. **110**, 158105 (2013).
[59]J. Alegre-Cebollada *et al.*, Cell **156**, 1235 (2014).
[60]H. S. Greenside and E. Helfand, Bell Syst. Tech. J. **60**, 1927 (1981).