A topological order parameter for describing folding free energy landscapes of proteins

Pham Dang Lan, Maksim Kouza, Andrzej Kloczkowski, and Mai Suan Li

Citation: J. Chem. Phys. **149**, 175101 (2018); doi: 10.1063/1.5050483
View online: https://doi.org/10.1063/1.5050483
View Table of Contents: http://aip.scitation.org/toc/jcp/149/17
Published by the American Institute of Physics

A topological order parameter for describing folding free energy landscapes of proteins

Pham Dang Lan,[1,2,a)] Maksim Kouza,[3,4,a)] Andrzej Kloczkowski,[4] and Mai Suan Li[5,b)]
[1]*Institute for Computational Science and Technology, SBI Building, Quang Trung Software City, Tan Chanh Hiep Ward, District 12, Ho Chi Minh City, Vietnam*
[2]*Faculty of Physics and Engineering Physics, VNUHCM-University of Science, 227, Nguyen Van Cu Street, District 5, Ho Chi Minh City, Vietnam*
[3]*Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland*
[4]*Nationwide Children's Hospital, Battelle Center for Mathematical Medicine, Columbus, Ohio 43215, USA*
[5]*Institute of Physics, Polish Academy of Science, Al. Lotnikow 32/46, 02-668 Warsaw, Poland*

(Received 31 July 2018; accepted 11 October 2018; published online 2 November 2018)

We studied the refolding free energy landscape of 26 proteins using the Go-like model. The distance between the denaturated state and the transition state, X_F, was calculated using the Bell theory and the nonlinear Dudko-Hummer-Szabo theory, and its relation to the geometrical properties of the native state was considered in detail. We showed that none of the structural parameters, such as the contact order, protein length, and radius of cross section, correlate with X_F for all classes of proteins. To overcome this problem, we have introduced the nematic order parameter P_{02}, which describes the ordering of the structured elements of the native state. Due to its topologically global nature, P_{02} is better than other structural parameters in describing the folding free energy landscape. In particular, P_{02} displays a good correlation with X_F extracted from the nonlinear theory for all three classes of proteins. Therefore, this parameter can be used to predict X_F for any protein, if its native structure is known. *Published by AIP Publishing.* https://doi.org/10.1063/1.5050483

I. INTRODUCTION

The construction of a free energy landscape (FEL) of biomolecules is important in understanding their structure-function relationship. However, this problem is very complicated because we have to deal with a multi-dimensional object. One of the possible ways to overcome this difficulty is to map the FEL onto a lower dimensional space using one or more reaction coordinates, while simultaneously not losing important information about the system under investigation.

Recently, single molecule force spectroscopy (SMFP) has become a useful tool for exploring FEL using mechanical force as an additional parameter for probing molecular interactions.[1–4] Depending on the mechanical stability of the biomolecules, various SMFP techniques have been applied. Magnetic and optical tweezers are commonly used to probe the FEL of DNA and RNA, which are highly elastic and malleable,[2] having the rupture force of tens of pN. Atomic force microscopy (AFM) is often used to study the behavior of proteins where a typical force is about 10^2 pN. AFM experiments were carried out mainly for protein unfolding[1–3,5] as well as for protein refolding.[6–11]

In the presence of an external mechanical force f, the free energy landscape changes in such a way that the unfolding barrier is reduced by $\Delta\Delta G^{\ddagger}_{U} = \Delta G^{\ddagger}_{U}(f) - \Delta G^{\ddagger}_{U}(0)$, while the folding barrier levels up by $\Delta\Delta G^{\ddagger}_{F} = \Delta G^{\ddagger}_{F}(f) - \Delta G^{\ddagger}_{F}(0)$ (Fig. 1). Here $\Delta G^{\ddagger}_{U,F}(f)$ and $\Delta G^{\ddagger}_{U,F}(0)$ are the unfolding/folding barriers in the presence and absence of force ($f = 0$), respectively. In weak force or Bell regime,[12] one can approximate $\Delta\Delta G^{\ddagger}_{F} = X_F f$ and $\Delta\Delta G^{\ddagger}_{U} = -X_U f$, where X_F is the distance between the denaturated state (DS) and the transition state (TS), while X_U is the distance between the native state (NS) and TS. Thus, in the linear force regime, parameters X_F and X_U can be obtained from the exponential dependence of refolding[6] and unfolding[2] times on f, respectively. In an experiment with a constant pulling speed v, X_U is usually estimated from the linear dependence of the rupture force on $\ln(v)$. In the non-linear regime, one has to apply a more complicated dependence of the folding/unfolding time on f taking into account the movement of TS under the force.[13]

For protein unfolding under mechanical force, it was shown[2,14] that X_U depends on the contact order[15] (CO) or on the secondary structure in such a way that the larger the β-structure, the smaller the X_U. A similar linear dependence on the β-structure was obtained for the rupture force.[14,16,17]

The question that we ask in this paper is which structural parameter is best to describe the distance between DS and TS, X_F, which characterizes the refolding process. One should note that, in general, X_F is an abstract reaction coordinate. However, Schlierf and Rief[18] proposed that X_F is directly related to the protein end-to-end distance in such a way that it may be extracted from the free energy landscape plotted as a function of the end-to-end distance and their suggestion was confirmed for a number of proteins.[18,19] We have shown that, contrary to the unfolding case,[2] this distance is correlated with the β-structure for β-proteins, but not for α- and mixed α/β-proteins. In addition, none of the existing structural

[a)]P. D. Lan and M. Kouza contributed equally to this work.
[b)]Author to whom correspondence should be addressed: masli@ifpan.edu.pl

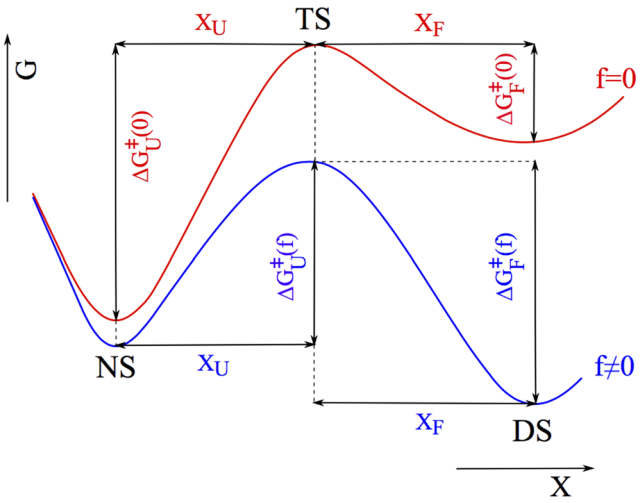

FIG. 1. Schematic plot for one-dimensional FEL in the absence (red) and in the presence (blue) of the external force f.

parameters such as protein length and radius of cross section can be used to predict X_F for all classes of proteins. This prompted us to introduce the "nematic liquid crystal" order parameter P_{02}, which characterizes the ordering of secondary structures in the native state. Using the Go model, we have demonstrated that this order parameter does not improve the correlation much with X_F extracted from the Bell theory, but it shows a reasonable correlation with X_F extracted from Dudko-Hummer-Szabo (DHS) theory.[13] Thus, P_{02} is better than any other parameter in describing the FEL of protein folding.

II. MATERIAL AND METHODS

A. Coarse-grained model

The computation of folding time, in particular, in the presence of the external force, using all-atom models is beyond the present computational limit. Therefore, coarse-grained models are the only choice. The simplest coarse-grained model is a Go model[20] that favors the native state by distinguishing between native and non-native interactions and is demonstrated appropriate to delineate protein folding process of many proteins.[21] Despite simplicity, Go models provided reasonable results that agree with all-atom simulations and experiments on impact of mechanical force and denaturant on protein folding and unfolding.[19,22–27]

In this paper, we used the Go-like model[21] in which each amino acid is represented by a single bead. The energy of this model is as follows:

$$\begin{aligned} E = & \sum_{bonds} K_r(r_i - r_{0i})^2 + \sum_{angles} K_\theta(\theta_i - \theta_{0i})^2 \\ & + \sum_{dihedral} \left\{ K_\phi^{(1)}[1 - \cos(\phi_i - \phi_{0i})] \right. \\ & \left. + K_\phi^{(3)}[1 - \cos 3(\phi_i - \phi_{0i})] \right\} \\ & + \sum_{i>j-3}^{NC} \varepsilon_H \left[5\left(\frac{r_{0ij}}{r_{ij}}\right)^{12} - 6\left(\frac{r_{0ij}}{r_{ij}}\right)^{10} \right] \\ & + \sum_{i>j-3}^{NNC} \varepsilon_H \left(\frac{c}{r_{ij}}\right)^{12} - \left|\vec{f}.\vec{R}\right|. \end{aligned} \quad (1)$$

Here $\Delta\Phi_i = \Phi_i - \Phi_{0i}$, $r_{i,j}$ is the distance between residues i and j, and θ_i is the bond angle between bonds $(i-1)$ and i. Subscripts "0," "NC," and "NNC" denote the native conformation, native contacts, and non-native contacts, respectively. Beads i and j are in native contact if r_{0ij} is less than a cutoff distance $d_c = 6.5$ Å, where r_{0ij} is the distance between the residues i and j in the native conformation. The chain connectivity is ensured by the first harmonic term in Eq. (1), and the second term refers to the bond angle potential. The potential for dihedral angles is given by the third term in Eq. (1). The interaction energy between residues separated by at least 3 beads is described by 10-12 Lennard-Jones potential. A soft sphere repulsive potential given by the fifth term in Eq. (1) disfavors the formation of non-native contacts. The last term accounts the protein interaction with the force applied to C and N termini along the end-to-end vector R in constant force simulations.

We have chosen $K_r = 100\varepsilon_H/\text{Å}^2$, $K_\theta = 20\varepsilon_H/\text{rad}^2$, $K_\phi^{(1)} = \varepsilon_H$, and $K_\phi^{(3)} = 0.5\varepsilon_H$, where ε_H is the hydrogen bond energy and $c = 4$ Å. Force unit is $[f] = \varepsilon_H/\text{Å} \approx 68$ pN. The Langevin equation was employed to study dynamics of the system and was solved by the Verlet algorithm using time step $\Delta t = 0.005\ \tau_L$, where the characteristic time $\tau_L \approx 3$ ps.[28] More details about this model are available elsewhere.[14,19,29]

In combination with the Langevin dynamics, the Go-like model (1) successfully predicted the correlation between X_U and the secondary structures of NS.[2,14] This model can also reproduce refolding experimental results on X_F for a number of proteins.[19] Thus, we expect that it can be used to study the relationship between X_F and topology of NS.

B. All-atom reconstruction

To obtain all-atom representation from a single bead conformation, the backbone resolution is reconstructed from α-carbon trace by PUNCHRA.[30] Then side chain atoms were added to the backbone structure and were optimized using SCWRL 4.0 package.[31]

C. Contact order

The contact order or relative contact order (RCO) is defined as[15]

$$RCO = \frac{\sum_{ij} \Delta_{ij}|i-j|}{L\sum_{ij} \Delta_{ij}},$$

where $\Delta_{ij} = 1$ if residues i and j are in native contact and $\Delta_{ij} = 0$ otherwise, $|i-j|$ is the distance between residues in the sequence space, L is the number of residues, and the summation is over all possible pairs of residues. The absolute contact order (ACO) is equal to L^*RCO.

D. Radius of cross section

The radius of cross section is defined as S_{ASA}/V_{ASA}, where S_{ASA} is the solvent-accessible surface area, while V_{ASA} is the volume surrounded by this surface.[32] This quantity is proportional to the average radius of the minimal cross section in the center of protein molecule. Because $(S_{ASA}/V_{ASA})^2$ is correlated with protein folding times,[32] we will use this as a structural indicator for describing the free energy landscape.

E. Other structural indicators

According to the quantitative capillarity theory,[33] the folding barrier scales with the protein size like $L^{2/3}$, while the random barrier picture predicted the $L^{1/2}$dependence.[34] These theories have been supported by experimental[35] as well as simulation[36] data. Thus one can expect a correlation between X_F and $L^{2/3}$ and $L^{1/2}$. However, we can show that the correlation levels are nearly the same for these quantities and we will focus only on $L^{2/3}$. In addition, we will consider the relationship between and X_F and L. Because the gyration radius R_g characterizes the compactness of the protein, we will also consider its ability to decipher the refolding free energy landscape.

F. Folding time at T_{min}

The folding time, τ_f, is a median value of folding times, obtained from individual runs. In order to obtain reliable results for each value of force, a pool of 50-100 trajectories has been generated. As in our previous work,[19] all simulations were carried out at temperature T_{min}, at which the folding is fastest in the absence of force.

III. RESULTS AND DISCUSSIONS

A. Correlation between x_F and typical structural parameters

In this section, we consider a possible correlation between X_F and several typical structural parameters, such as contact order (CO), radius of cross section $(V_{ASA}/S_{ASA})^2$, and protein size $(L^{2/3})$.

1. Bell approximation

Protein refolding is very sensitive to the applied force,[4,37–39] and in general, the distance from DS to TS depends on the force intensity. In the Bell approximation, this distance is assumed to be constant in the narrow force regime[37] and then X_F can be extracted from the exponential dependence of the folding time τ_f on the external force,[12]

$$\tau_f = \tau_0 \exp(fX_F/k_BT), \tag{2}$$

where τ_0 is the folding time at zero force.

Recently, we have demonstrated[19] that there is a switch from the weak force regime to the medium force regime. In

TABLE I. Relative contact order (RCO), absolute contact order (CO), radius of cross section $(V_{ASA}/S_{ASA})^2$, protein size $(L^{2/3})$, and order parameter P_{02} of 26 proteins studied in this work. X_{F1} and X_{F2} in low and mean force regimes were calculated in the Bell approximation. The last column refers to X_F extracted from DHS theory.

PDB	RCO	ACO	$(V_{ASA}/S_{ASA})^2$	L	Rg	P_{02}	X_{F1}/X_{F2} (nm) Bell	X_F (nm) (ν = 1/2) DHS
A proteins								
1NTI	0.127	10.95	3.732	86	12.864	0.729	1.12/6.60	0.91
2ABD	0.137	12.70	3.641	86	12.494	0.787	0.28/1.76	0.56
2PDD	0.113	4.86	2.378	43	9.245	0.885	0.40/1.74	0.62
1IMQ	0.118	10.38	3.680	86	11.819	0.835	0.13/1.98	0.37
256B	0.073	7.94	4.844	106	14.162	0.905	0.56/1.85	0.69
1VII	0.112	4.03	1.702	36	8818	0.462	0.76/2.50	1.13
1LMB	0.094	7.50	4.134	80	11.461	0.284	1.09/2.92	0.89
β proteins								
1SRL	0.196	10.95	2.775	56	9.706	0.160	0.85/4.86	0.80
1KSR	0.152	15.22	4.223	100	13.601	0.855	0.61/3.73	0.49
1QJO	0.211	16.87	3.849	80	12.010	0.268	0.69/2.79	0.64
1TEN	0.171	15.43	4.475	89	12.780	0.735	0.20/1.36	0.39
1RSY	0.162	20.41	5.687	126	14.722	0.760	0.36/1.08	0.70
1G1C	0.182	17.77	4.131	98	13.594	0.640	0.28/2.83	0.51
1FNF	0.174	16.95	3.611	94	13.053	0.712	0.19/0.78	0.53
1G6P	0.177	11.69	3.207	66	10.176	0.317	1.07/4.45	0.86
1TIT	0.178	15.84	4.268	89	12.594	0.802	0.12/0.87	0.37
Mixed α/β proteins								
2RN2	0.124	19.27	5.744	155	15.093	0.564	1.42/6.51	1.18
1QYS	0.108	9.94	4.504	92	12.244	0.859	0.16/1.33	0.70
1APS	0.212	20.77	4.772	98	12.577	0.835	0.01/2.12	0.62
1HZ6	0.161	10.01	2.983	62	10.963	0.756	0.29/4.26	0.31
1X0O	0.111	13.25	5.497	119	13.744	0.259	0.49/2.04	1.16
1BRF	0.183	9.68	3.001	53	9.836	0.184	0.45/2.47	0.86
1PGA	0.173	9.69	2.984	56	10.202	0.819	0.23/1.90	0.44
1UBQ	0.151	11.47	3.806	76	11.493	0.283	0.32/1.01	0.71
1ZRP	0.168	8.91	3.235	53	9.732	0.424	0.19/1.29	0.73
1BNR	0.114	12.26	4.827	108	13.476	0.303	0.37/1.72	1.30

the weak force regime, folding pathways are the same as in the absence of force (temperature-driven regime) and the dependence of the folding time on the external force is described by Eq. (2) with the distance X_{F1}. In the medium force regime, which is probed by AFM experiments,[19] the dependence of τ_f on f is also given by Eq. (2) but with the distance X_{F2}, where $X_{F1} < X_{F2}$.

In this work, we also observed two regimes separated by a switch force f_{switch} (Figs. S1–S3 in the supplementary material), but for more proteins than our previous study.[19] Here the proteins are divided into three classes of α-, β-, and mixed α/β-proteins. Using Eq. (2) and the data from Figs. S1–S3, we extracted X_{F1} and X_{F2} for two regimes (Table I). X_F greatly varies among proteins. Some of them such as 1NTI (α-protein) and 2RN2 (mixed α/β-protein) have very large X_F, indicating their high sensitivity to the quenched force and the acquisition of their native state is difficult.

a. Correlation between X_F and structural parameters for β-proteins. It is known that α-rich proteins fold more rapidly than β-rich proteins[15,40] and the folding rate correlates with the parameters specifying NS such as contact order (CO) and its various variants,[15,41,42] the protein length (L),[43] and the radius of the cross section [$(V_{ASA}/S_{ASA})^2$].[32] Studies on mechanical unfolding showed a correlation between X_u and CO.[2,14] An interesting question emerges: does the distance X_F correlate with these structural parameters.

For β-proteins, there is no fit with the relative contact order CO, but ACO has a correlation level $R = 0.67$ and 0.74 for X_{F1} and X_{F2}, respectively (Fig. 2). A similar correlation level is observed for $L^{2/3}$ ($R = 0.64$ for X_{F1} and 0.68 for X_{F2}) and radius of gyration R_g ($R = 0.72$ for X_{F1} and 0.69 for X_{F2}). $(V_{ASA}/S_{ASA})^2$ parameter, which describes the compactness of the protein, also correlates with the distance between DS and NS, but the correlation level is slightly lower than that of ACO and $L^{2/3}$ (R = 0.57 and 0.65 for X_{F1} and X_{F2}, respectively). To be more confident in the correlations between X_F and the structural parameters from linear regression, we calculated the p-values that are shown in the corresponding plots. By adopting the standard criterion that a good fit should have p less than 0.05, ACO and R_g display the best correlation with X_F among the others.

b. No correlation between X_F and structure parameters for α-proteins and α/β-proteins. As evident from Fig. S4, the distance X_F of α-proteins correlates with structural parameters worse than β-proteins, since R remains below 0.27 and p-values are larger than 0.56. This may be due to the fact that proteins of this class are too sensitive to external force, leading to a greater variation in X_F compared to β-proteins. Another possible reason is that our set is not large enough to have good statistics.

Correlation is better in the case of α/β-proteins (Fig. S5), but the fit remains poor, with the exception of $L^{2/3}$ and L for which we obtained $R = 0.65$, $p = 0.044$ and $R = 0.68$, $p = 0.029$, respectively.

c. Why X_F of β-proteins correlates with structure parameters? In order to qualitatively understand why X_F correlates with the structural parameters of β-proteins but not with other

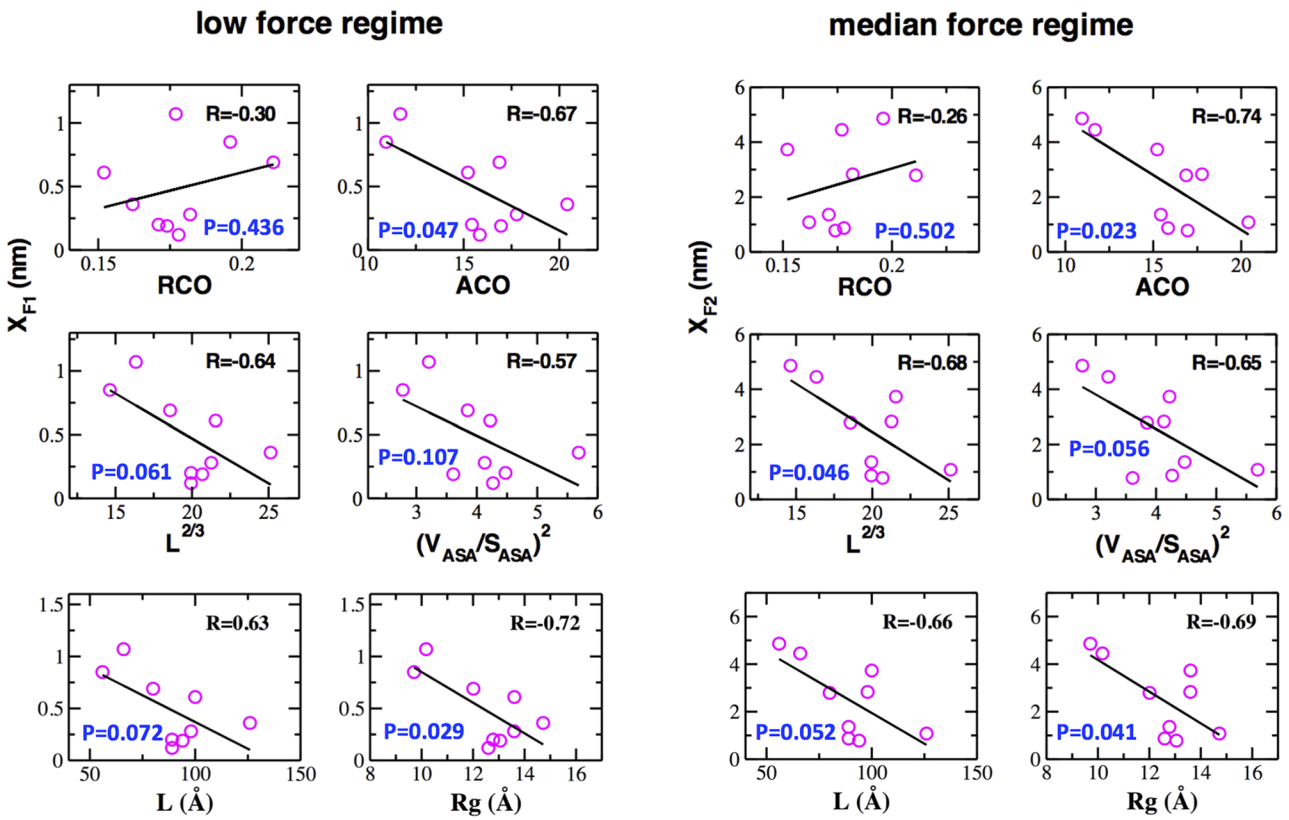

FIG. 2. Correlation between X_F and relative contact order (RCO) and absolute contact order (ACO); $L^{2/3}$ and radius of cross section $(V_{ASA}/S_{ASA})^2$; and L and radius of gyration R_g for β-proteins in the low (left panel) and medium (right panel) force regime.

classes of proteins, we present FEL as a function of the end-to-end distance and the fraction of native contacts for 1LMB (α-protein) and 1G1C (β-protein) (Fig. 3). For 1G1C, in the both low ($f = 3.5$ pN) and medium force ($f = 8.16$ pN) regimes, the most representative structures contain tertiary contacts implying that the refolding is not so sensitive to external force. As a result, X_F values are not very diverse among proteins leading to its correlation with structural parameters, as can be seen in our simulation.

The situation becomes different for the -protein 1LMB, where secondary structures were formed in the representative structures in the medium force regime ($f = 6.8$ pN), but tertiary contacts do not yet occur making it difficult to acquire the native structure. Thus, the effect of external force on the refolding of α-proteins is strong and leads to a large variation of X_F or, ultimately, to a poor fit between X_F and structural parameters.

2. Nonlinear Dudko-Hummer-Szabo (DHS) theory

Although Bell theory is widely used in experiments to obtain information about FEL, X_F is determined only in a narrow force range and estimation of τ_0 from Eq. (2) might be inaccurate.[37] This is partly because in the linear theory the movement of TS under the action of an external force is neglected. On the other hand, as seen above, in the Bell approximation, the correlation between X_F and structural parameters is poor, in particular, for α-proteins and α/β-proteins. Therefore, we attempted to go beyond this approximation using DHS theory, which took into account the displacement of TS under an external force.[13] Then X_F can be defined over a broader range of forces by the following equation:

$$\tau(f) = \tau_0 \left(1 + \frac{\nu f X_F}{\Delta G^{\ddagger}}\right)^{1-1/\nu} e^{\frac{\Delta G^{\ddagger}}{k_B T}\left[1-\left(1+\nu f X_F/\Delta G^{\ddagger}\right)\right]^{1/\nu}}, \quad (3)$$

where $\Delta G^{\ddagger}$ is the height of the barrier between DS and TS, and the parameter ν is related to the shape of the barrier: $\nu = 2/3$ for a linear-cubic barrier, $\nu = 1/2$ for a cusp-like barrier, and $\nu = 1$ returns the Bell theory [Eq. (2)]. This theory was successfully used to explore FEL of proteins and RNA hairpin.[4,29,44]

Since X_F extracted from Eq. (3) is almost the same for $\nu = 1/2$ and 2/3, we present the results for $\nu = 1/2$. Figure 4 illustrates the application of the DHS theory to protein 1KSR. Obviously the fit works for a wider range than the weak force regime, but it fails to cover the entire medium regime. In general, the region, where this fit works, depends on proteins. For some proteins, it is valid for the weak force regime, but for other proteins it covers both weak and medium force regimes (Figs. S1–S3). X_F predicted by the DHS theory for all studied proteins is given in Table I.

a. The DHS theory works best for α/β proteins. Contrary to the Bell theory, for β-proteins X_F, predicted by the DHS theory does not correlate with any structural parameter (Fig. 5). The best fit was obtained for R_g, but the corresponding correlation level is rather low ($R = 0.58$, $p = 0.103$). For α-proteins, the DHS approximation works better than the Bell theory (compare Fig. 5 and Fig. S4), but the correlation remains poor because the best correlation, obtained for ACO, is only 0.5.

For α/β proteins, the DHS theory works much better than the Bell theory with two regimes (Fig. 5 and Fig. S5). With the exception of ACO ($R = 0.32$, p = 0.37) and RCO

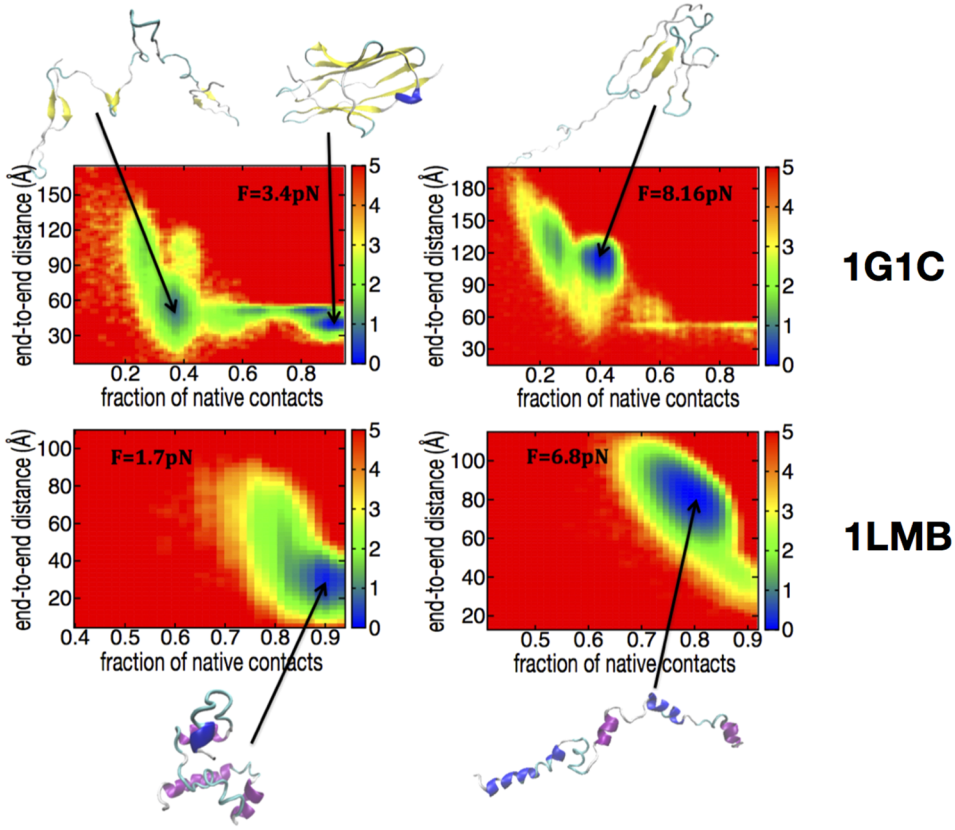

FIG. 3. Two-dimensional free energy landscapes of β-protein 1G1C and α-protein 1LMB. The end-to-end distance and the fraction of native contacts were used as the reaction coordinates.

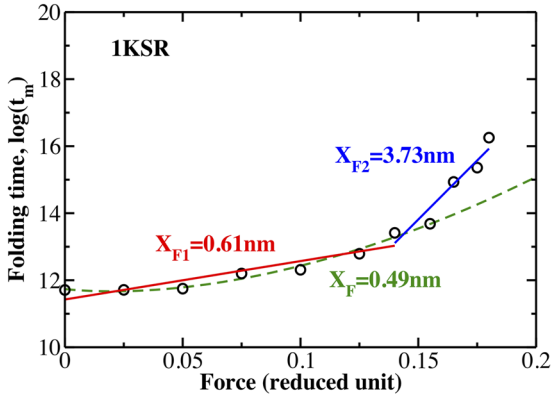

FIG. 4. Application of the linear (two force regimes) and nonlinear DHS theory to protein 1KSR. Using Eq. (3) and ν = 1/2, we obtaind X_F = 0.49 nm (green curve). For ν = 2/3, we obtained the same results.

($R = 0.6$, $p = 0.066$), which have low correlation, we obtained $R = 0.69$, 0.73, 0.71, and 0.69 for $L^{2/3}$, $(V_{ASA}/S_{ASA})^2$, L, and R_g, respectively.

B. Order parameter P_{02}

1. Definition of P_{02}

Because none of the structural parameters are capable to characterize X_F for all classes of proteins, we introduce a "nematic liquid crystal" order parameter P_{02} as follows:

$$P_{02} = \frac{1}{2N}\sum_{i=1}^{N}\frac{3}{2}\left(\vec{u_i}.\vec{d}\right)^2 - \frac{1}{2}. \quad (4)$$

Here $\vec{u_i}$ (i = 1, . . ., N) are the unit vectors connecting the two Cα-ends of secondary structures including the helix and beta (Fig. 6), N is the number of secondary structure elements, and $\vec{d}$ is the unit eigenvector, which defines the collective direction of the vectors $\vec{u_i}$. The STRIDE software was employed to obtain secondary structures of native conformation retrieved from the Protein Data Bank (PDB).[45] The vectors $\vec{u_i}$ were calculated by the *cafit_orientation* module[46] plugged in the Pymol[47] software. Note that parameter P_{02} was introduced[48] to describe the fibril structure of protein. In our case, this parameter measures the order of structured segments. P_{02} of all studied proteins is shown in Table I. Proteins with small P_{02} are structurally less ordered than proteins with large P_{02}. Thus, it is expected that proteins, which have small P_{02}, will be sensitive to external force.

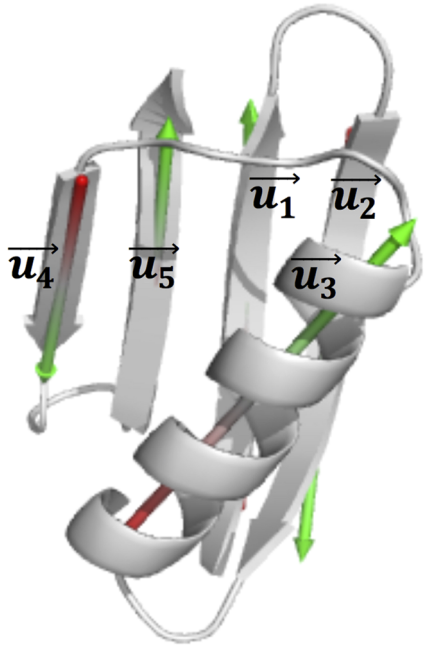

FIG. 6. A cartoon representation of 1PGA protein that has 5 structured segments, identified by STRIDE. Vectors connecting two Cα-ends of secondary structure elements are $\vec{u_1}$ (residues 2 → 8), $\vec{u_2}$ (residues 13 → 19), $\vec{u_3}$ (residues 23 → 37), $\vec{u_4}$ (residues 42 → 46), and $\vec{u_5}$ (residues 51 → 55). For 1PGA, P_{02} = 0.82.

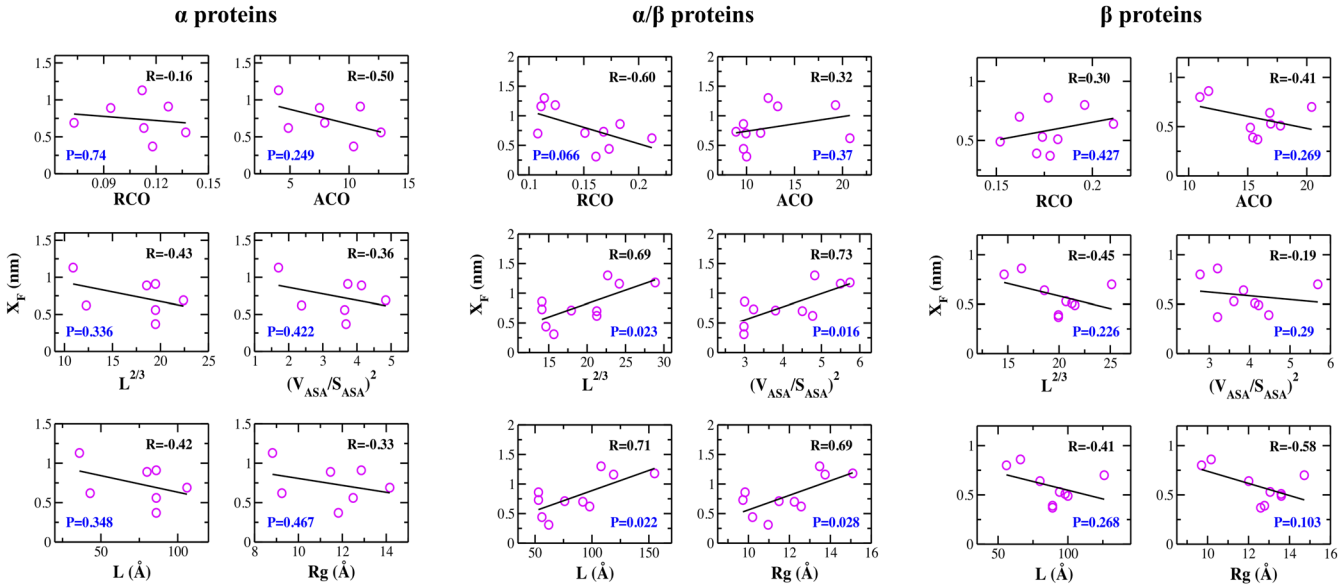

FIG. 5. Correlation between X_F, extracted from DHS theory, and relative contact order (RCO), absolute contact order (ACO), $L^{2/3}$, radius of cross section $(V_{ASA}/S_{ASA})^2$, L, and radius of gyration R_g for three classes of proteins.

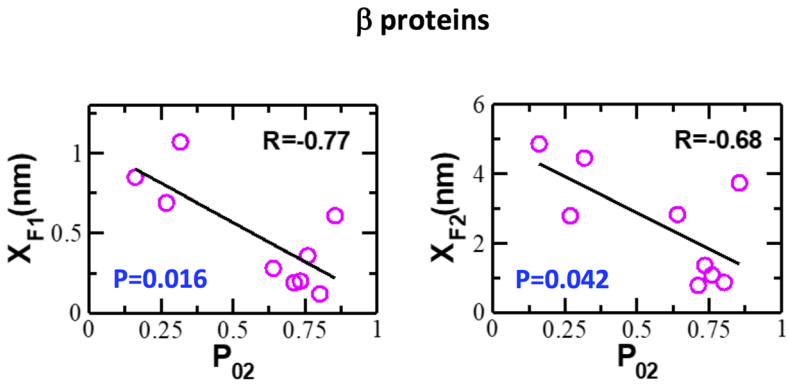

FIG. 7. Correlation between P_{02} and X_{F1} and X_{F2} for β-proteins.

2. Bell theory

a. P_{02} parameter correlates with X_F of β-proteins. In the low force regime, parameter P_{02} of β-proteins correlates with X_{F1} (R = 0.77 and p = 0.016, Fig. 7) better than other structural parameters (Fig. 2). However, in the medium force regime, the correlation X_{F2} with P_{02} (R = 0.68 and p = 0.042, Fig. 7) is compatible with ACO and Rg (Fig. 2). For both fits, p-value is less than 0.05 providing additional support for this correlation. The fact that P_{02} fits well with the Bell theory is not surprising because, except RCO, other structural parameters also correlate with it due to the high stability of β-proteins.

b. P_{02} parameter displays a good correlation with X_F for combined α- and β-proteins. As can be seen from the FEL of 1LMB (α-protein), the formation of the secondary structures depends little on the force intensity (Fig. 3). The primary effect of force on α-proteins is to prevent their packaging into the native structure. So we obtained a marginally correlation of P_{02} in the low force regime (R = 0.66, Fig. S6). However, in the medium force regime, when the force action becomes strong, P_{02} ceases to correlate with X_{F2} (R = −0.19) (Fig. S6). This is partly due to the fact that protein 1NTI has too high X_{F2} (6.6 nm, Table I).

For the studied proteins, almost no correlation has been observed for mixed α/β-proteins as the correlation level R is below 0.24 and $p \gg 0.05$ (Fig. S7). However, by removing 1NTI and combining the data, obtained for α-proteins and β-proteins, we obtained R = 0.76, p = 0.0009 and R = 0.67, p = 0.0058 for X_{F1} and X_{F2}, respectively (Fig. 8).

3. Correlation between P_{02} and X_F extracted from DHS theory

Figure 9 shows X_F obtained in the nonlinear theory [Eq. (3)], as a function of P_{02} for each class of proteins and for all proteins combined together. As expected, the best fit was obtained for the β-proteins (R = 0.76). Even for brittle α-proteins, we have a good correlation with R = 0.69. For mixed α/β-proteins, a slightly weaker match was obtained (R = 0.62), but this result shows that the DHS theory works better than the Bell approximation, where correlation was not observed for this class of proteins (Fig. S7). More importantly, there is a rather good correlation of P_{02} with X_F of all three classes of proteins (R_{all} = 0.6 and p = 0.001, Fig. 9). This result is important because all of typical structural parameters fail to have correlation even with X_F extracted from the DHS theory (Fig. S8 in the supplementary material). Thus, our main result is that the new parameter P_{02} is better than other structural parameters in describing protein refolding under the quenched force.

We stress that the correlation of 0.6 is not as high as that found by Plaxco and Baker for the relationship between the relative contact order and the experimental folding rate in the absence of force ($R \approx 0.8$).[15] In our opinion, the reason for this is that the protein folding is sensitive to the external force, making the variation of folding times in a much wider range compared to the case of folding without quenched force. Consequently, X_F is sensitive to secondary structures, which results in a lower correlation level.

4. Comparison with experiments

To our best knowledge, X_F was experimentally measured for 7 proteins (Table II). There is a low correlation between

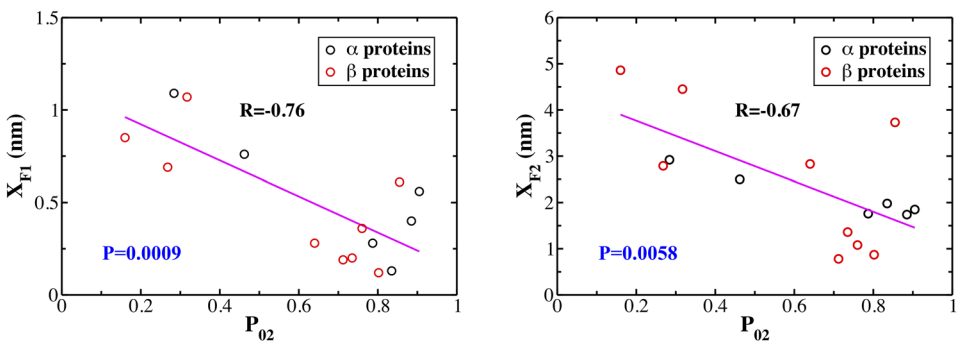

FIG. 8. Correlation between P_{02} and X_{F1} and X_{F2} for all α- and β-proteins. 1NTI protein was excluded from the fitting.

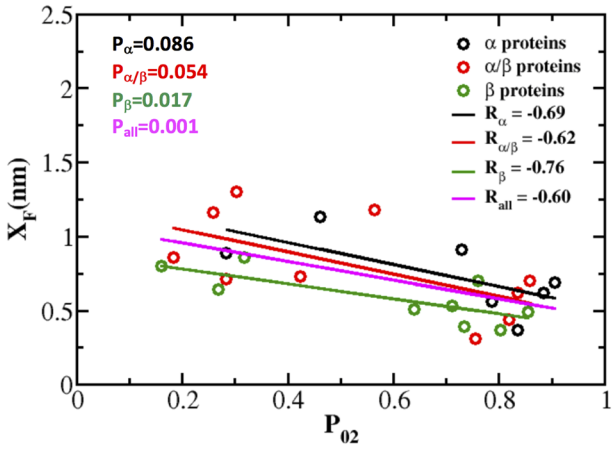

FIG. 9. Correlation between X_F estimated by using DHS formula [Eq. (3)] and P_{02} for each class of proteins and for all proteins.

the experimental data on X_F and parameter P_{02} (Fig. 10) as $R = -0.43$ and $p = 0.337$. This is probably because the data set is small, but more importantly because in the experiment X_F was estimated using the Bell theory, i.e., from the linear dependence of the rupture force on the logarithm of the pulling speed. From this point of view, the experimental data agree with our result about the poor relationship between P_{02} and X_F estimated within the linear theory. It would be interesting to extract X_F from fitting experimental data, which are collected in both weak and medium force regimes, with the DHS Eq. (3)

TABLE II. Experimental values of X_F for seven proteins.

No.	Protein	Class	P_{02}	Experimental X_F (nm)	Reference
1	1UBQ	α/β	0.283	0.8	6
2	1SRL	β	0.16	5.25	8
3	1NTI	α	0.720	6.7	9
4	1CFC	α	0.613	8.0	10
5	1PGA	α/β	0.819	2.1	11
6	1BZ6	α	0.147	14.9	7
7	1TIT	β	0.802	2.3	5

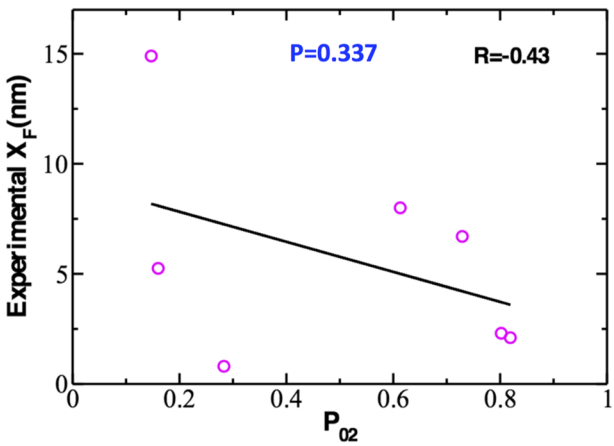

FIG. 10. Correlation between P_{02} and experimental X_F for 7 proteins listed in Table II.

and verify our main prediction that they are correlated with P_{02}.

IV. CONCLUSION

We have calculated X_F for 26 Go proteins using the Bell and DHS theory. The relationship between this distance and various structural parameters has been analyzed in detail. It was shown that none of the existing parameters (RCO, ACO, radius of cross section, R_g, and protein length) is correlated with X_F obtained for all classes of proteins. To overcome this problem, we have introduced parameter P_{02} which characterizes the ordering of the β- and α-structures of NS. Due to this topologically global nature, for all proteins, the new parameter correlates with X_F extracted from DHS theory in such a way that the smaller the P_{02}, the larger the X_F. Thus, contrary to the distance between NS and TS, X_U, which is determined by the contact order,[2,14] the relationship between X_F and the geometry of NS is more complicated.

Our simulations show that P_{02} is an useful parameter for exploring FEL of protein folding. It can be used to predict X_F of proteins, based solely on information about their NS. It would be important to reaffirm our result on the relationship between P_{02} and X_F for a larger data set not only by simulation but also by experiment.

SUPPLEMENTARY MATERIAL

See supplementary material for dependence of folding time on quenched force for three classes of proteins with linear and DHS fitting (Figs. S1–S3); correlation between X_F extracted by Bell formula and relative contact order (RCO), absolute contact order (ACO), $L^{2/3}$, L, radius of cross section $(V_{ASA}/S_{ASA})^2$, and radius of gyration (R_g) for α-proteins (Fig. S4) and α/β-proteins (Fig. S5); correlation between X_F, extracted by Bell formula, and P_{02} for α-proteins and α/β-proteins (Figs. S6 and S7); and correlation between X_F, extracted by using the DHS equation, and relative contact order (RCO), absolute contact order (ACO), $L^{2/3}$, L, radius of cross section $(V_{ASA}/S_{ASA})^2$, and radius of gyration (Rg) for all 26 proteins.

ACKNOWLEDGMENTS

This work was supported by the Department of Science and Technology at Ho Chi Minh city, Vietnam (Grant No. 32/2017/HD-KHCNTT), and the Polish NCN Grant No. 2015/19/B/ST4/02721, Poland. M.K. acknowledges the Polish Ministry of Science and Higher Education for financial support through "Mobilnosc Plus" Program No. 1287/MOB/IV/2015/0.

[1] T. Hoffmann, K. M. Tych, M. L. Hughes, D. J. Brockwell, and L. Dougan, Phys. Chem. Chem. Phys. **15**, 15767 (2013).
[2] S. Kumar and M. S. Li, Phys. Rep. **486**, 1 (2010).
[3] M. L. Hughes and L. Dougan, Rep. Prog. Phys **79**, 076601 (2016).
[4] Y. Chen, S. E. Radford, and D. J. Brockwell, Curr. Opin. Struct. Biol. **30**, 89 (2015).
[5] M. Carrion-Vazquez, A. F. Oberhauser, S. B. Fowler, P. E. Marszalek, S. E. Broedel, J. Clarke, and J. M. Fernandez, Proc. Natl. Acad. Sci. U. S. A. **96**, 3694 (1999).
[6] J. M. Fernandez and H. Li, Science **303**, 1674 (2004).

[7]P. J. Elms, J. D. Chodera, C. Bustamante, and S. Marqusee, Proc. Natl. Acad. Sci. U. S. A. **109**, 3796 (2012).
[8]B. Jagannathan, P. J. Elms, C. Bustamante, and S. Marqusee, Proc. Natl. Acad. Sci. U. S. A. **109**, 17820 (2012).
[9]P. O. Heidarsson, I. Valpapuram, C. Camilloni, A. Imparato, G. Tiana, F. M. Poulsen, B. B. Kragelund, and C. Cecconi, J. Am. Chem. Soc. **134**, 17068 (2012).
[10]J. P. Junker, F. Ziegler, and M. Rief, Science **323**, 633 (2009).
[11]Y. Cao and H. Li, Nat. Mater. **6**, 109 (2007).
[12]G. I. Bell, Science **200**, 618 (1978).
[13]O. K. Dudko, G. Hummer, and A. Szabo, Phys. Rev. Lett. **96**, 108101 (2006).
[14]M. S. Li, Biophys. J. **93**, 2644 (2007).
[15]K. W. Plaxco, K. T. Simons, and D. Baker, J. Mol. Biol. **277**, 985 (1998).
[16]A. Gabovich and M. S. Li, J. Chem. Phys. **131**, 024121 (2009).
[17]H. Dietz and M. Rief, Phys. Rev. Lett. **100**(4), 098101 (2008).
[18]M. Schlierf and M. Rief, Biophys. J. **90**, L33 (2006).
[19]M. Kouza, P. D. Lan, A. M. Gabovich, A. Kolinski, and M. S. Li, J. Chem. Phys. **146**, 135101 (2017).
[20]N. Go, Ann. Rev. Biophys. Bioeng. **12**, 183 (1983).
[21]C. Clementi, H. Nymeyer, and J. N. Onuchic, J. Mol. Biol. **298**, 937 (2000).
[22]M. Kouza, C.-K. Hu, and M. S. Li, J. Chem. Phys. **128**, 045103 (2008).
[23]J. I. Sułkowska and M. Cieplak, Biophys. J. **95**, 3174 (2008).
[24]P. C. Whitford, J. K. Noel, S. Gosavi, A. Schug, K. Y. Sanbonmatsu, and J. N. Onuchic, Proteins: Struct., Funct., Bioinf. **75**, 430 (2009).
[25]M. Cieplak, T. X. Hoang, and M. O. Robbins, Proteins: Struct., Funct., Bioinf. **49**, 114 (2002).
[26]D. Klimov and D. Thirumalai, Proc. Natl. Acad. Sci. U. S. A. **97**, 2544 (2000).
[27]E. P. O'Brien, G. Ziv, G. Haran, B. R. Brooks, and D. Thirumalai, Proc. Natl. Acad. Sci. U. S. A. **105**, 13403 (2008).
[28]T. Veitshans, D. Klimov, and D. Thirumalai, Folding Des. **2**, 1 (1997).
[29]M. S. Li, M. Kouza, and C.-K. Hu, Biophys. J. **92**, 547 (2007).
[30]P. Rotkiewicz and J. Skolnick, J. Comput. Chem. **29**, 1460 (2008).
[31]G. G. Krivov, M. V. Shapovalov, and R. L. Dunbrack, Proteins: Struct., Funct., Bioinf. **77**, 778 (2009).
[32]D. N. Ivankov, N. S. Bogatyreva, M. Y. Lobanov, and O. V. Galzitskaya, PLoS One **4**, e6476 (2009).
[33]P. G. Wolynes, Proc. Natl. Acad. Sci. U. S. A. **94**, 6170 (1997).
[34]D. Thirumalai, J. Phys. I **5**, 1457 (1995).
[35]M. S. Li, D. K. Klimov, and D. Thirumalai, Polymer **45**, 573 (2004).
[36]M. Kouza, M. S. Li, E. P. O'Brien, C. K. Hu, and D. Thirumalai, J. Phys. Chem. A **110**, 671 (2006).
[37]M. T. Woodside and S. M. Block, Ann. Rev. Biophys. **43**, 19 (2014).
[38]M. Yao, B. T. Goult, B. Klapholz, X. Hu, C. P. Toseland, Y. Guo, P. Cong, M. P. Sheetz, and J. Yan, Nat. Commun. **7**, 11966 (2016).
[39]X. Zhao, X. Zeng, C. Lu, and J. Yan, Nanotechnology **28**, 414002 (2017).
[40]S. E. Jackson, Folding Des. **3**, R81 (1998).
[41]M. M. Gromiha and S. Selvaraj, J. Mol. Biol. **310**, 27 (2001).
[42]H. Zhou and Y. Zhou, Biophys. J. **82**, 458 (2002).
[43]A. V. Finkelstein, N. S. Bogatyreva, and S. O. Garbuzynskiy, FEBS Lett. **587**, 1884 (2013).
[44]O. K. Dudko, T. G. Graham, and R. B. Best, Phys. Rev. Lett. **107**, 208301 (2011).
[45]H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, Nucleic Acids Res. **28**, 235 (2000).
[46]See https://github.com/Pymol-Scripts/Pymol-script-repo/blob/master/anglebetweenhelices.py for the method for calculation of vector $\vec{u_i}$ using a python script.
[47]W. L. DeLano, http://pymol.org (2002).
[48]P. H. Nguyen, M. S. Li, G. Stock, J. E. Straub, and D. Thirumalai, Proc. Natl. Acad. Sci. U. S. A. **104**, 111 (2007).