Critical Fluctuations in the Native State of Proteins

Qian-Yuan Tang,[1] Yang-Yang Zhang,[1] Jun Wang,[1,*] Wei Wang,[1,†] and Dante R. Chialvo[2,‡]
[1]National Lab of Solid State Microstructure, Collaborative Innovation Center of Advanced Microstructures, and Department of Physics, Nanjing University, Nanjing 210093, China
[2]CEMSC[3], Center for Complex Systems & Brain Sciences, Escuela de Ciencia y Tecnología, Universidad Nacional de San Martín & Consejo Nacional de Investigaciones Científicas y Tecnológicas (CONICET), 25 de Mayo y Francia, San Martín(1650), Buenos Aires, Argentina

(Received 16 February 2016; revised manuscript received 31 December 2016; published 24 February 2017)

Based on protein structural ensembles determined by nuclear magnetic resonance, we study the position fluctuations of residues by calculating distance-dependent correlations and conducting finite-size scaling analysis. The fluctuations exhibit high susceptibility and long-range correlations up to the protein sizes. The scaling relations between the correlations or susceptibility and protein sizes resemble those in other physical and biological systems near their critical points. These results indicate that, at the native states, motions of each residue are felt by every other one in the protein. We also find that proteins with larger susceptibility are more frequently observed in nature. Overall, our results suggest that the protein's native state is critical.

DOI: 10.1103/PhysRevLett.118.088102

Introduction.—Protein molecules are formed by unbranched chains of amino acids (or residues) and have several structural types, i.e., the globular, fibrous, membrane, and intrinsically disordered proteins. Globular proteins, the majority of the proteins in nature, generally fold into globular shapes through diffusive dynamics on a minimally frustrated energy landscape [1]. It is this kind of three-dimensional folded structure, known as a native state, that make proteins capable of performing their biological functions. A protein (hereafter, we mean the globular protein) carries out its functions by switching from one structure to another, even transiently, for instance, when it recognizes and binds with other molecules. To achieve such performance, the structure of the native state of the protein must be susceptible enough to sense the signal and switch to another structure, but it must also be stable enough to warrant functional specificity and structural robustness. Coincidentally, these apparently competing demands are generically exhibited by physical systems near their critical points [2–6], raising the question of whether such competing demands could be mechanistically resolved by certain kinds of critical behavior in proteins.

Critical fluctuations in protein equilibrium dynamics have been emphasized already by a number of results, including the power-law relation between a solvent-accessible surface area and the volume of proteins [7], the fractal-like structure of configuration space [8] and the oscillation spectrum [9], the slowness of relaxation in protein molecules [10,11], the overlap between the low-frequency collective oscillation modes and large-scale conformational changes in allosteric transitions [12–14], critical water fluctuation near hydrated proteins [15], and so on. A few works have already ventured to call the native state an example of self-organized criticality, as in the work by Phillips [16] or in the discussion on pairwise correlations between residues in protein families [3]. Yet, a direct characterization of the critical fluctuations near the native states of proteins based on experimental data is still incomplete.

Here, we study the fluctuations around the native states of a large number of proteins based on their structural ensembles determined by solution nuclear magnetic resonance (NMR). Each structure of the ensemble is conjectured here to be an instantiation of the conformations in the native basin (please see Ref. [17] for some caveats). Note that NMR-based structural ensembles have been used in the characterization of landscape and conformational fluctuations around the native state [18–20] and may encode evolutionary constraints of protein structures [21]. We examine the distance-dependent correlations of position fluctuations of residues and conduct a finite-size scaling analysis, demonstrating that the correlations and susceptibility exhibit features similar to those in other physical systems near their critical points, implying that even weak local perturbations to any residue are felt by every other residue of the entire protein. This may be an additional universal characteristic of natural proteins.

Fluctuations and correlations: Data and definitions.—Our data set contains 4988 protein structural ensembles from the Protein Data Bank (PDB) [22]. Each ensemble corresponds to a protein and has no less than 10 different structures (see details in Ref. [17]). All proteins have no more than 40% sequence similarity. For simplification, we focus here on the C_α traces of these proteins. As shown in Fig. 1(a), the first structure in each ensemble is selected as the reference structure (in orange), and all the other

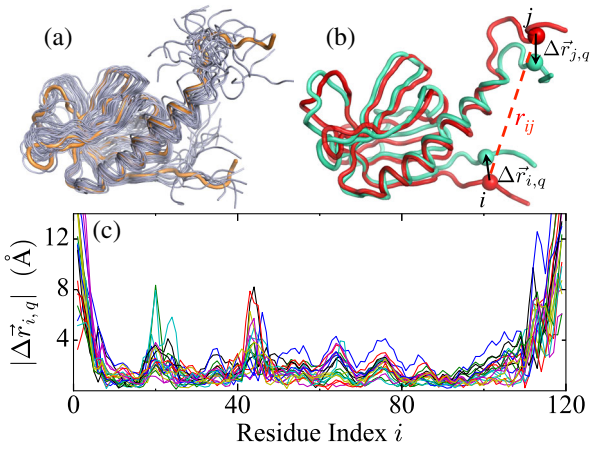

FIG. 1. (a) Example of a structural ensemble (protein PDB code: 1BAK) showing the 20 structures (grey) aligned to the reference structure (orange). (b) The average structure (red) and the qth structure (cyan) for the protein in (a). The displacement $\Delta\vec{r}_{i,q}$ (or $\Delta\vec{r}_{j,q}$) of the ith (or jth) residue from its counterpart in the average structure is marked with black arrows. The distance r_{ij} between residues i and j is marked with red dashes. (c) The magnitude of the residues' position fluctuations $|\Delta\vec{r}_{i,q}|$ for the qth structure ($q = 1, 2, \ldots$) in the ensemble.

structures (in grey) are aligned to it. By minimizing the root-mean-square deviation (RMSD) of the C_α atoms, the degrees of freedom related to the translational and rotational motions are removed. After the alignment, the position fluctuations of every residue (represented by the C_α atom) can be calculated.

Let us introduce the following notation (see Fig. 1): For each ensemble (of a given protein) in our data set, the qth structure (with $q = 1, 2, \ldots, Q$) contains the coordinates of all the residues—say, the ith residue is $\vec{r}_{i,q} = (x_{i,q}, y_{i,q}, z_{i,q})$ (for $i = 1, 2, \ldots, N$). Then, the displacement of the ith residue in the qth structure can be calculated as the deviation from the average structure of the protein,

$$\Delta\vec{r}_{i,q} = \vec{r}_{i,q} - \langle\vec{r}_i\rangle, \tag{1}$$

where $\langle\vec{r}_i\rangle = (1/Q)\sum_{q=1}^{Q}\vec{r}_{i,q}$ is defined as the average structure [see Fig. 1(b)]. Correspondingly, the magnitudes of position fluctuations of every residue can be determined by the norm $|\Delta\vec{r}_{i,q}|$ for $q = 1, 2, \ldots, Q$ structures [see Fig. 1(c)], revealing the well-known variability of residue positions in the structural ensembles [23]. The results are insensitive to reference structures, model resolution, or alignment algorithms (see Ref. [17]). The distance-dependent correlation $G(r)$ of these fluctuations for a given protein can be described as

$$G(r) = \frac{\sum_{i<j}^{N}\mathcal{K}_{ij}\delta(r - r_{ij})}{\sum_{i<j}^{N}\delta(r - r_{ij})}, \tag{2}$$

where $\mathcal{K}_{ij}$ can be calculated based on the fluctuations, and $\delta(r - r_{ij})$ is the Dirac delta function selecting residue pairs at mutual distance r in the reference structure. This definition is similar to that in Refs. [24,25] for other systems and can be extended to the cases with multiple proteins. Thus, the distance-dependent covariance $C(r)$ and orientational correlation $\phi(r)$ can be determined using Eq. (2) with $\mathcal{K}_{ij} = C_{ij} = (1/Q)\sum_q \Delta\vec{r}_{i,q}\cdot\Delta\vec{r}_{j,q}$ and $\mathcal{K}_{ij} = \phi_{ij} = C_{ij}/(C_{ii}C_{jj})^{1/2}$, respectively. Similar calculations can be applied to the magnitude correlations (ϕ', see Ref. [17]).

Scale invariant correlations.—We find that the correlation function $\phi(r)$ shares common features across proteins. To illustrate this behavior, $\phi(r)$ is computed for two cases: (1) eight proteins representing different structural classes (such as the all α, all β, $\alpha+\beta$, … classes [26]), multidomain proteins, and multichain assemblies, respectively; and (2) seven proteins with similar gyration radii $R_g \simeq 12$ Å. In all cases $\phi(r)$ display a similar behavior; i.e., for each curve, $\phi(r)$ decreases from its maximum at $r \sim 3.8$ Å (the intrinsic distance between two C_α atoms of consecutive residues), crosses zero, reaches a negative minimum, and finally increases or even oscillates again. The correlation length ξ is defined by the distance r at which $\phi(r)$ crosses zero [e.g., the arrow in Fig. 2(a)]. From visual inspection of the results in Fig. 2(b), it is evident that proteins with similar sizes (i.e., similar R_g) have similar ξ. This is further demonstrated by the scattering plot of ξ vs R_g in Fig. 2(c) for all the proteins: We find that the correlation lengths ξ are distributed in a band which is approximately proportional to R_g: $\langle\xi\rangle \sim R_g$ [the open circles in Fig. 2(c)]. The behavior of the correlations is not seen on a null model constructed by randomly perturbing the positions of residues in an arbitrary structure (see Ref. [17]).

To further explore the dependence of correlation $\phi(r)$ on the magnitudes of R_g, we divide our data set into different subsets according to their R_g, i.e., subsets with R_g values inside bins of $[R_g - 0.5$ Å, $R_g + 0.5$ Å$]$. From these subsets we compute the respective $\phi(r)$, which exhibit the same behavior mentioned above [Fig. 2(d)]. For all proteins investigated here, the distance-dependent correlation function is scale free: Rescaling, in each curve, the distance r by its correlation length ξ results in a collapse of all curves of $\phi(\hat{r})$ with $\hat{r} = r/\xi$ [Fig. 2(e)]. We observe similar behavior for the magnitude correlation (see Ref. [17]).

We now explore whether or not the scale-free correlations are related to the residue composition (or interactions) of proteins. We calculate the correlations [$\phi(r)$ and $C(r)$] for various types of residue pairs [27]. Figure 2(f) and its inset show the correlations for four representative residue pairs, i.e., the combinations of hydrophobic (isoleucine), polar (threonine), or charged residues (lysine). We find that all $\phi(r)$ curves are almost coincident and have the same

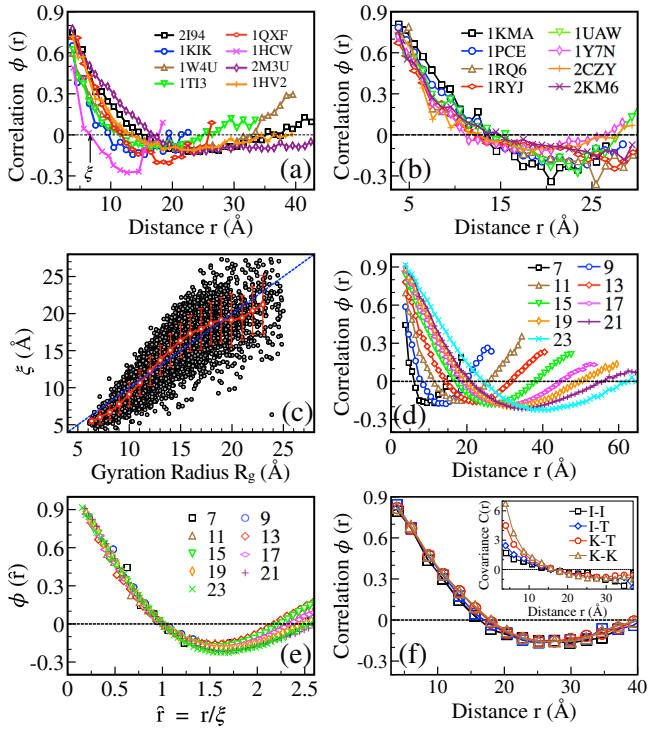

FIG. 2. Correlation $\phi(r)$: (a) $\phi(r)$ for proteins with various structural features: 2I94 (α class), 1KIK (β class), 1W4U ($\alpha+\beta$), 1TI3 (α/β), 1QXF (small), 1HCW (designed), 2M3U (multidomain), and 1HV2 (multichain). The correlation length ξ (for 1HCW) is indicated by an arrow. (b) $\phi(r)$ for proteins with similar $R_g \approx 12$ Å. (c) Scattering plot of ξ and R_g for all 4988 proteins. The red open circles show the average $\langle\xi\rangle$ for proteins with similar R_g. (d) $\phi(r)$ for the subsets of proteins in bins labeled by $R_g = 7, 9, \ldots, 23$ Å (see text). (e) Scaling plot of $\phi(\hat{r})$ with $\hat{r} = r/\xi$ for curves in (d). (f) $\phi(r)$ and covariance $C(r)$ (inset) related to specific types of residue pairs for proteins with $R_g \approx 15$ Å.

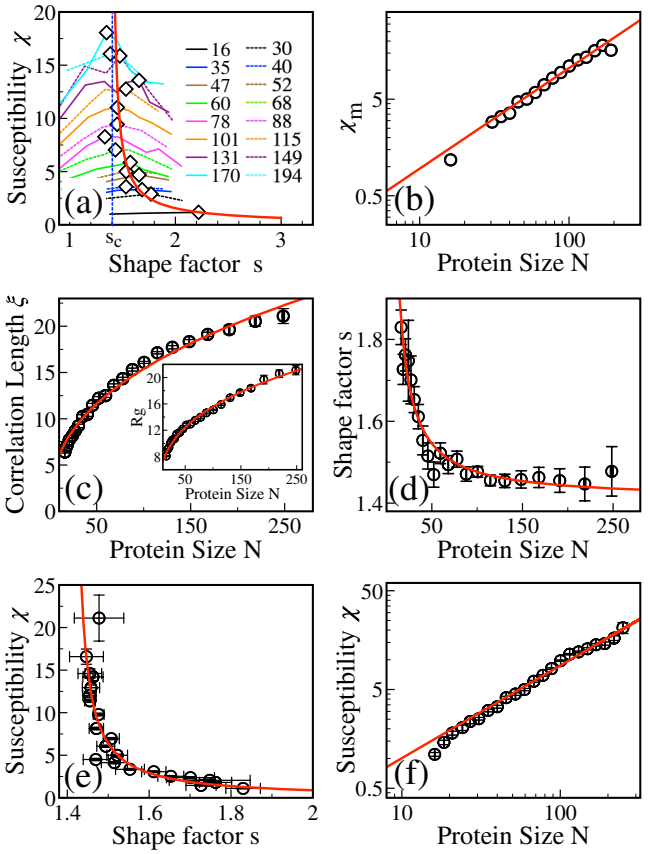

FIG. 3. Finite-size scaling. (a) Susceptibility χ vs shape factor s for proteins with different sizes N. The peak height χ_m and position s_m show a power-law relation $\chi_m \sim (s_m - s_c)^{-\gamma}$ (thick red line). (b) Scaling of peak heights with N as $\chi_m \sim N^{-\alpha\gamma/\mu}$. (c) Correlation length ξ vs protein size N and R_g vs N (inset). (d) Shape factor s vs N. (e) Susceptibility χ vs shape factor s. (f) Susceptibility χ vs protein size N. Error bars are standard errors of the mean.

correlation lengths, and all $C(r)$ curves exhibit a similar trend, indicating that the scale-free behavior is independent of the residue composition. Even so, some additional information concerning residue-residue interactions can be revealed by comparing the $C(r)$ of different types of residue pairs. For example, the covariance $C(r)$ of isoleucine-isoleucine pairs (I-I) at short distances is generally smaller than that of other types of residue pairs, which means smaller fluctuations and stronger interactions. This is attributed to the fact that isoleucines are strongly hydrophobic and usually deeply buried inside proteins. Physically, the interaction strength can be estimated by $\epsilon_{ij} \sim C_{ij}/(C_{ii}C_{jj})$; thus, $\sim 1/C_{ij}$ [i.e., $1/C(r)$] when the correlations $\phi(r)$ are almost the same for all residue pairs [see Fig. 2(f)]. Consistently, a large covariance (so weak interactions) for lysine-lysine pairs (K-K) at short distances is also observed. These results reflect some detailed interaction features of residue pairs, and they are in line with previous work [20]. These aspects may help us to refine the force field for coarse-grained models and to elucidate the evolutionary constraint in proteins [28], and they deserve further study.

Additional hints from finite-size scaling.—The types of correlations described above resemble the collective behaviors observed in a variety of biological and physical systems [2,3,24,25,29] in which the correlations are amplified around the vicinity of the critical points. However, most often, the system sizes are very small with respect to their thermodynamic limit, such that the value of the control parameter at which the susceptibility peaks depends on system sizes. In turn, Attanasi *et al.* used this finite-size limitation to probe the features of criticality near a critical point [25]. We adopted Attanasi's strategy for our data, first defining a dimensionless shape factor $s = Na^3/(L_1L_2L_3)$ as the pseudocontrol parameter for a protein, in which $a = 3.8$ Å is the residue size, and L_1, L_2, and L_3 are lengths of the principle axes of the protein with $L_1 \le L_2 \le L_3$. Such a control parameter can also be understood as

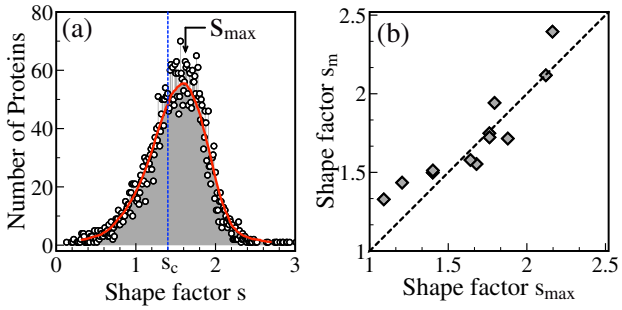

FIG. 4. (a) Shape factor distribution for all proteins. (b) The most frequent shape factor $s_{\max}$ vs the position of the maximal susceptibility s_m for subsets of proteins with similar R_g.

"packing density" since $L_1L_2L_3$ is proportional to the volume of an ellipsoid. For densely packed proteins, s is relatively large, and conversely, s decreases for loosely packed proteins and those with loops. Then, the susceptibility χ of a protein is defined as in Ref. [25],

$$\chi = \frac{s}{N}\sum_{i<j}^{N}\phi_{ij}\theta(\xi - r_{ij}), \tag{3}$$

describing the total correlation in a unit volume within the correlation length. As in Fig. 3(a), for sets of proteins with different sizes N, the average susceptibility χ exhibits a series of maximum χ_m at the corresponding s_m (open diamonds). As N increases, the peak χ_m becomes sharper and the peak position s_m approaches a critical value $s_c \approx 1.4$, following the scaling relations $\chi_m \sim (s_m - s_c)^{-\gamma}$ and $\chi_m \sim N^{\alpha\gamma/\nu}$ [red lines in Figs. 3(a) and 3(b)].

If the behavior of the susceptibility corresponds to critical dynamics, the following relations are expected to hold: $\xi \sim N^{\alpha}$, $s - s_c \sim N^{-\alpha/\nu}$, and $\chi \sim N^{\alpha\gamma/\nu}$. Despite some variations, the fittings of the results [red lines in Figs. 3(c)–3(f)] closely follow the expected scaling functions. From these fittings, the following exponents are determined. First, based on $\xi \sim N^{\alpha}$, we get $\alpha = 0.40$ [Fig. 3(c)], which is similar to the result $\alpha = 0.32$ based on $R_g \sim N^{\alpha}$ [inset of Fig. 3(c)]. This indicates that proteins are tightly packed, and it is consistent with the critical shape factor s_c [30]. Second, for $s - s_c \sim N^{-\alpha/\nu}$, we have $1/\nu \approx 2.87$ (or $\nu \approx 0.35$) [Fig. 3(d)]. Third, γ can be determined from the relations between χ and s (or N). Figure 3(e) shows the relation $\chi \sim (s - s_c)^{-\gamma}$ with $\gamma \approx 1.05$, and Fig. 3(f) depicts the relation $\chi \sim N^{\alpha\gamma/\nu}$ with $\gamma \approx 1.03$; these relations are comparable to the fitting result ($\gamma \approx 1.01$) in Fig. 3(a). By taking integers, the approximated exponents are $\alpha = 1/3, \nu = 1/3, \gamma = 1$. Thus, these scale-invariant features obey similar scaling relations as in other critical systems, demonstrating that the fluctuations of the proteins' native states are critical.

The results in Fig. 4(a) show some preference for proteins to be near the critical state since the most frequent shape factor ($s_{\max} = 1.5$) is very close to the critical shape factor $s_c \approx 1.4$ [i.e., Fig. 3(a)], where the susceptibility is diverging in the thermodynamic limit. The analysis for proteins with similar R_g is also consistent; as Fig. 4(b) shows, the most probable shape factors $s_{\max}$ are found to be roughly the same as s_m with the maximal susceptibility (see Ref. [17]), indicating that proteins with high susceptibility are most probable in nature. This suggest that there would be some evolutionary preference for certain shapes, which may be relevant to the previous observations on evolution constraints in structure and dynamics [21,28,31,32].

In summary, for a protein's native state determined by NMR, the correlation functions of the structural fluctuations exhibit universal features obeying distinctive scaling behaviors of critical dynamics. Such criticality comes together with short-range interactions and global kinetic modes, and it is in line with the picture of minimal frustration. Moreover, the uncovered relative abundance of proteins with high susceptibility opens a novel window to investigate the underlying relation between structure and function, shedding light on the dynamics of protein folding, misfolding, and aggregation as well as design.

This work is supported by National Natural Science Foundation of China, (NSFC) (Grants No. 11334004 and No. 11174133), 973 program of China (Grant No. 2013CB834100), and by CONICET of Argentina. Q. Y. T. and D. R. C. acknowledge the hospitality of the Max Planck Institute for the Physics of Complex Systems at Dresden (Germany).

*wangj@nju.edu.cn
†wangwei@nju.edu.cn
‡dchialvo@conicet.gov.ar

[1] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997); J. R. Banavar and A. Maritan, Annu. Rev. Biophys. Biomol. Struct. **36**, 261 (2007); A. E. Garcia and G. Hummer, Proteins **36**, 175 (1999); F. Rao and A. Caflisch, J. Mol. Biol. **342**, 299 (2004).
[2] P. Bak, *How nature works: the science of self-organized criticality* (Copernicus, New York, 1996).
[3] T. Mora and W. Bialek, J. Stat. Phys. **144**, 268 (2011).
[4] A. R. Honerkamp-Smith, S. L. Veatch, and S. L. Keller, Biochim. Biophys. Acta **1788**, 53 (2009).
[5] D. R. Chialvo, Nat. Phys. **6**, 744 (2010).
[6] H. Chaté and M. Muñoz, Physics **7**, 120 (2014).
[7] M. A. Moret and G. F. Zebende, Phys. Rev. E **75**, 011920 (2007); M. A. Moret, Physica (Amsterdam) **390A**, 3055 (2011).
[8] T. Neusius, I. Daidone, I. M. Sokolov, and J. C. Smith, Phys. Rev. Lett. **100**, 188103 (2008).
[9] S. Reuveni, R. Granek, and J. Klafter, Phys. Rev. Lett. **100**, 208101 (2008).
[10] H. P. Lu, L. Xun, and X. S. Xie, Science **282**, 1877 (1998).
[11] X. Hu, L. Hong, M. D. Smith, T. Neusius, X. Cheng, and J. C. Smith, Nat. Phys. **12**, 171 (2016).

[12] I. Bahar, A. R. Atilgan, M. C. Demirel, and B. Erman, Phys. Rev. Lett. **80**, 2733 (1998).
[13] I. Bahar, T. R. Lezon, L.-W. Yang, and E. Eyal, Annu. Rev. Biophys. **39**, 23 (2010).
[14] L. Yang, G. Song, and R. L. Jernigan, Biophys. J. **93**, 920 (2007).
[15] A. J. Patel, P. Varilly, S. N. Jamadagni, M. F. Hagan, D. Chandler, and S. Garde, J. Phys. Chem. B **116**, 2498 (2012).
[16] J. C. Phillips, Physica (Amsterdam) **415A**, 440 (2014); Proc. Natl. Acad. Sci. U.S.A. **106**, 3107 (2009); **106**, 3113 (2009).
[17] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.118.088102 for a detailed description of the data set and methods.
[18] R. B. Best, K. Lindorff-Larsen, M. A. DePristo, and M. Vendruscolo, Proc. Natl. Acad. Sci. U.S.A. **103**, 10901 (2006).
[19] A. G. Palmer, III., Chem. Rev. **104**, 3623 (2004).
[20] T. R. Lezon and I. Bahar, PLoS Comput. Biol. **6**, e1000816 (2010).
[21] H. Lammert, J. K. Noel, E. Haglund, A. Schug, and J. N. Onuchic, J. Chem. Phys. **143**, 243141 (2015).
[22] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, J. Mol. Biol. **112**, 535 (1977).
[23] K. Lindorff-Larsen, R. B. Best, M. A. DePristo, C. M. Dobson, and M. Vendruscolo, Nature (London) **433**, 128 (2005).
[24] A. Cavagna, A. Cimarelli, I. Giardina, G. Parisi, R. Santagati, F. Stefanini, and M. Viale, Proc. Natl. Acad. Sci. U.S.A. **107**, 11865 (2010).
[25] A. Attanasi *et al.*, Phys. Rev. Lett. **113**, 238102 (2014).
[26] A. G. Murzin *et al.*, J. Mol. Biol. **247**, 536 (1995).
[27] For a $\mathcal{X}$-$\mathcal{Y}$ pair, with $\mathcal{X}$ and $\mathcal{Y}$ being certain kinds of residues, correlations are calculated by replacing $\sum_{ij}$ in Eq. (1) with $\sum_{ij}\delta(T_i,\mathcal{X})\delta(T_j,\mathcal{Y})$. Here T_i represents the kind of residue i in the pth protein, and $\delta(A,B)=1$ if A and B are the same, and 0 otherwise.
[28] Y. Liu and I. Bahar, Mol. Biol. Evol. **29**, 2253 (2012).
[29] K. Christensen and N. R. Moloney, *Complexity and Criticality* (Imperial College Press, London, 2005).
[30] Based on fitting for the inset of Fig. 3(c), $R_g \approx wN^{1/3}$ with $w = 3.4$ Å. For a spherical molecule, $s \approx Na^3/R_g^3 = (a/w)^3 \approx 1.4$ (namely, s_c).
[31] G. D. Friedland, N.-A. Lakomek, C. Griesinger, J. Meiler, T. Kortemme, and R. Nussinov, PLoS Comput. Biol. **5**, e1000393 (2009).
[32] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Proc. Natl. Acad. Sci. U.S.A. **106**, 67 (2009).