**Supplementary Information**
**for "Critical Fluctuations in the Native States of Proteins"**

Qian-Yuan Tang, Yang-Yang Zhang, Jun Wang, Wei Wang, and Dante R. Chialvo

# A. Data and Methods

## 1. Dataset Definition

The dataset used in the present work contains 4988 protein structural ensembles and each ensemble includes no less than 10 different structures (of the same protein) which are determined through solution NMR and are stored in the Protein Data Bank. In general, these structures are recognized as the native ones for the proteins. Note that only globular proteins are considered in the present work, membrane proteins and very short peptide chains (with the number of amino acids $N < 15$) are not included into our dataset due to their non-aqueous molecular environments or the lack of stable native structures.

## 2. Files for Dataset

The file "Datafiles.zip" contains:

- The proteins listed in the file"ProteinDataset_40%.txt" corresponds to those exhibiting a sequence similarity between any two proteins smaller than 40%. Beside the PDB codes of all these proteins, the file include the values of the relevant features ($R_g, s, \chi, \cdots$ as indicated in the file "readme.txt"). For completeness we include similar information for proteins with sequence similarity smaller than 90% in file "ProteinDataset_90%.txt".
- Files containing the $C_\alpha$-atom coordinates of four typical protein structural ensembles adapted from the corresponding PDB files, in the format consistent with the Matlab code included in the Appendices I & II.

The proteins' structures (only the coordinates of the $C_\alpha$ atoms are included), after alignment, can be accessed through the URL:

`https://www.dropbox.com/s/2r3nfs5g152fdxy/Aligned_Structures.7z?dl=0`

## 3. Effects of Sequence Similarity

In the main text, we use a dataset with the sequence similarity between any two proteins is smaller than 40%. This kind of selection for proteins is typical in bioinformatics to ensure statistical independence. Actually, we found that the scale-free feature of proteins does not depend on the criterion of sequence similarity. As an example, two sets of proteins with different sequence similarities (i.e., one set with the largest sequence similarity up to 40% and another set with the largest sequence similarity up to 90%) are investigated (as shown in Fig.S1). For simplicity, only those proteins with almost the same gyration radii (say $R_g \simeq 15$ Å) in both sets are considered in this calculation. As shown in Fig.S1, the correlations $\phi(r)$ are almost identical. These results demonstrate the insensitivity of correlation features to the sequence similarity.

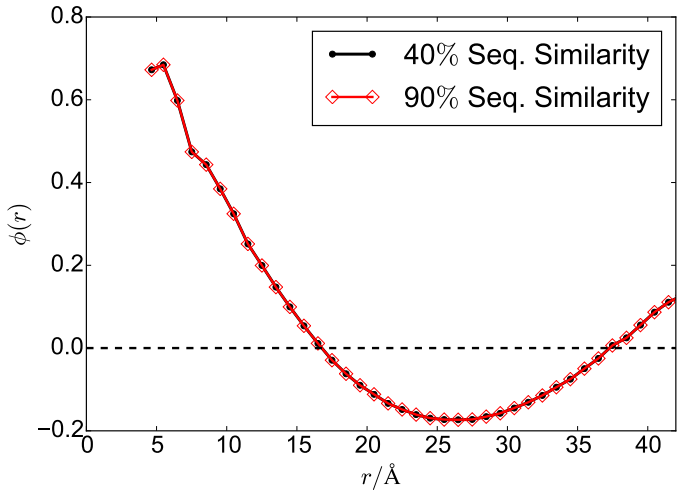



FIG. S1. Correlations $\phi(r)$ for two sets of proteins with maximal sequence similarity being 40% and 90%, respectively. These proteins all have the gyration radii $R_g \sim 15$Å.

## 4. NMR-Determined Structural Ensemble and Native Dynamics

With solution NMR technique, a series of spatial constraints between atoms of a protein in its native state can be obtained. These constraints contain the information from various fluctuating conformations in the native basin. Based on these constraints, a number of molecular models could be derived. Consistently, these models include fluctuating conformations around the native structure, and are related to native dynamics in solution. Consequently, the NMR-determined structural ensemble can be successfully employed to provide realistic models of protein structure to predict chemical shifts [8] which can be accurately measured and can be used to discriminate well the conformations of native and nonnative states of proteins [9]. Moreover, it is observed that models based on NMR-determined ensemble is the most general and accurate to characterize active sites [10]. These successful applications of NMR-determined structural ensembles indicate that these ensembles really reflect the basic information of native dynamics of the proteins.

## 5. Basics of Alignment

In the studies on proteins, it is often necessary to align (or superposition) different structures for comparison. The alignment is realized by removing the degrees of freedom related to translational and rotational motions, typically by minimizing the root-mean-square deviation (RMSD) of specific atoms through translations and rotations of the related structures [11]. Here, the specific atoms are selected based on the model resolution for proteins or the parts that we are interested in. For the alignments of multiple structures, the superposition is typically done by selecting one (arbitrary) structure/conformation as a reference state and then the best alignments of the other structures to this reference structure can be worked out.

## 6. Implementation of Alignment in Our Work

In our work, the calculations related to alignment are done with the Python package "ProDy" [12], with which the first structure in the ensemble is selected as the reference state and the alignments are performed based on the least-squares method and $C_\alpha$-atom representation of the proteins.

## 7. RMSD between Aligned Structures in Ensemble

It is worth noting that the RMSD (root mean squared deviation) between aligned structures in the structural ensembles is typically smaller than 2Å, except the small proteins whose RMSD are a little larger (between 2Å and 3Å). The RMSDs for proteins with various structural characteristics are calculated based different alignment algorithms and model resolutions, as shown in Fig.S1. In more details, there are 12.5 residues in average whose RMSDs are larger than 3 Å. Comparing to the typical length (with $\sim 300$ residues) of proteins, these largely fluctuating residues are only a small fraction of proteins. Additionally, the ratio of disordered residues can be characterized with the contact numbers of residues. It is found that, in our dataset, about 2.97% of all the residues have less than 4 neighboring residues. Here, the cutoff distance for contacts is 7.5Å. These residues may have a larger fluctuation with weaker local interactions. This result is consistent with the RMSD analysis. All in all, these observations ensures that the present dataset is valid to characterize the features of fluctuations of globular proteins.

| | | All $\alpha$ Proteins | All $\beta$ Proteins | $\alpha+\beta$ Proteins | $\alpha/\beta$ Proteins | Small Proteins | Designed Proteins | Multi-domain Proteins* | Multi-chain Protein Assemblies** |
|---|---|---|---|---|---|---|---|---|---|
| **Least Squares Superposition** | $C\alpha$ | 1.385 | 1.818 | 1.274 | 1.201 | 2.932 | 2.117 | 2.034 | 1.304 |
| | $C\alpha + C\beta$ | 1.373 | 1.795 | 1.27 | 1.202 | 2.887 | 2.099 | 2.035 | 1.301 |
| | Backbone | 1.392 | 1.821 | 1.278 | 1.203 | 2.934 | 2.119 | 2.041 | 1.311 |
| | Heavy Atoms | 1.389 | 1.78 | 1.28 | 1.209 | 2.847 | 2.096 | 2.03 | 1.303 |
| | All Atoms | 1.387 | 1.766 | 1.28 | 1.213 | 2.854 | 2.118 | 2.037 | 1.312 |
| **Least Squares Superposition (No Coil)** | $C\alpha$ | 1.274 | 1.653 | 1.168 | 1.201 | 2.775 | 1.997 | 1.849 | 1.251 |
| | $C\alpha + C\beta$ | 1.265 | 1.651 | 1.168 | 1.202 | 2.77 | 1.99 | 1.847 | 1.25 |
| | Backbone | 1.278 | 1.65 | 1.166 | 1.2 | 2.773 | 1.987 | 1.85 | 1.253 |
| | Heavy Atoms | 1.28 | 1.676 | 1.176 | 1.207 | 2.754 | 2.003 | 1.841 | 1.259 |
| | All Atoms | 1.279 | 1.67 | 1.18 | 1.21 | 2.765 | 2.008 | 1.845 | 1.264 |
| **Maximum Likelihood** | $C\alpha$ | 1.24 | 1.45 | 1.116 | 1.183 | 2.734 | 1.89 | 1.7 | 1.189 |
| | $C\alpha + C\beta$ | 1.239 | 1.45 | 1.116 | 1.184 | 2.737 | 1.89 | 1.7 | 1.188 |
| | Backbone | 1.241 | 1.45 | 1.116 | 1.184 | 2.733 | 1.899 | 1.698 | 1.189 |
| | Heavy Atoms | 1.236 | 1.45 | 1.115 | 1.184 | 2.721 | 1.883 | 1.692 | 1.186 |
| | All Atoms | 1.236 | 1.449 | 1.114 | 1.183 | 2.718 | 1.877 | 1.69 | 1.185 |

TABLE S1. the average RMSDs for proteins with various structural characteristics. Different alignment algorithms and model resolutions are considered. The concerned proteins have largely different gyration radii ($6 \leq R_g \leq 20$Å and sizes ($12 < N < 300$).

## 8. Effects of Reference States

For a protein structural ensemble, the alignment of multiple structures may depend on the selected reference structure in principle. However, due to the high similarity between structures in an ensemble, we found that our results are not sensitive to the selection of reference states. As shown in Fig.S2, the correlations $\phi(r)$ using two reference structures are given. All these correlations have the same correlation lengths and collapse together. There are almost no differences for these cases using different reference structures.

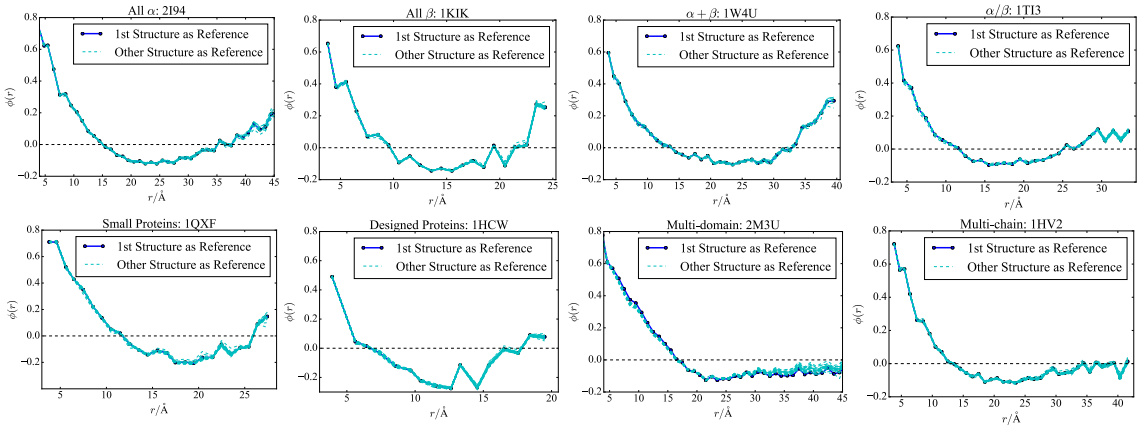



FIG. S2. Correlation $\phi(r)$ for proteins with the first structure (solid) or another arbitrary structure (dashed) in an ensemble as the reference state. The PDB codes of the concerned proteins are 2I94 (all $\alpha$ class), 1KIK (all $\beta$ class), 1W4U ($\alpha+\beta$ class), 1QXF (small proteins), 1HCW (designed protein), 2M3U (multi-domain proteins) and 1HV2 (multi-chain protein assemblies).

## 9. Effects of Alignment Algorithms and Model Resolutions

The scale-free feature is found to be independent of the alignment algorithms or the selections of interested atoms, though the aligned structures may be very different. As examples, we study the correlations $\phi(r)$ and $\phi'(r)$ (referring to Sec.B 11 for definition of $\phi'$) using two kinds of alignment methods (the least-square method and the maximum likelihood methods [13]) and five kinds of resolutions (namely, 1) with $C_\alpha$ atoms, 2) with $C_\alpha$ and $C_\beta$ atoms, 3) with all backbone atoms, 4) with all heavy atoms, and 5) with all atoms) to align our structure ensembles (as Table S2 and S3). Here, a subset of proteins (107 proteins with various structural characteristics and gyration radii from 6 to 20 Å) are used for simplicity. Firstly, the proportional relation between correlation lengths $\xi_\phi$ (or $\xi_{\phi'}$) and gyration radius $R_g$ is found for various alignment methods and model resolutions. The Pearson correlation coefficients for various cases are given in Table S2, and the average Pearson correlation coefficient is $0.865 \pm 0.086$ (or $0.736 \pm 0.099$). Moreover, the proportional relations between the correlation lengths $\xi_\phi$ (or $\xi_{\phi'}$) with different alignment methods and model resolutions are also observed. By taking the case with least-square alignment and the $C_\alpha$-atom resolution as the basic one, comparisons between the basic case and those using other methods and/or model resolutions are calculated. The Pearson correlation coefficients for various cases are given in Table S3, and the average Pearson correlation coefficients are $0.853 \pm 0.135$ (or $0.841 \pm 0.117$). Even, such calculations can be applied for the cases considering only the residues being in $\alpha$ and $\beta$ secondary structures in the native state. The related Pearson coefficients are similar to the above comparisons

(as in Table S2 and S3). All these results demonstrate the insensitivity of scale-free behaviors to the alignment method and/or the selections of interested atoms.

| | | All $\alpha$ Proteins | All $\beta$ Proteins | $\alpha+\beta$ Proteins | $\alpha/\beta$ Proteins | Small Proteins | Peptides | Designed Proteins | Multi-domain Proteins | Multi-chain Protein Assemblies |
|---|---|---|---|---|---|---|---|---|---|---|
| **Least Squares Superposition** | $C\alpha$ | 0.799 (0.815) | 0.854 (0.906) | 0.829 (0.858) | 0.805 (0.961) | 0.633 (0.901) | 0.629 (0.954) | 0.714 (0.854) | 0.692 (0.662) | 0.656 (0.824) |
| | $C\alpha + C\beta$ | 0.779 (0.824) | 0.788 (0.907) | 0.822 (0.84) | 0.796 (0.96) | 0.638 (0.924) | 0.726 (0.956) | 0.767 (0.857) | 0.683 (0.657) | 0.619 (0.82) |
| | Backbone | 0.795 (0.819) | 0.874 (0.916) | 0.653 (0.845) | 0.834 (0.956) | 0.639 (0.908) | 0.582 (0.953) | 0.738 (0.875) | 0.717 (0.652) | 0.625 (0.817) |
| | Heavy Atoms | 0.809 (0.788) | 0.821 (0.893) | 0.72 (0.8) | 0.767 (0.953) | 0.656 (0.911) | 0.725 (0.861) | 0.743 (0.87) | 0.709 (0.641) | 0.672 (0.761) |
| | All Atoms | 0.79 (0.766) | 0.796 (0.897) | 0.721 (0.787) | 0.779 (0.951) | 0.701 (0.909) | 0.732 (0.827) | 0.78 (0.844) | 0.711 (0.655) | 0.672 (0.737) |
| **Least Squares Superposition (No Coil)** | $C\alpha$ | 0.708 (0.883) | 0.751 (0.906) | 0.587 (0.906) | 0.769 (0.947) | 0.68 (0.896) | 0.301 (0.896) | 0.593 (0.765) | 0.668 (0.655) | 0.659 (0.847) |
| | $C\alpha + C\beta$ | 0.703 (0.897) | 0.757 (0.909) | 0.547 (0.898) | 0.772 (0.948) | 0.674 (0.915) | 0.311 (0.9) | 0.632 (0.814) | 0.701 (0.66) | 0.653 (0.841) |
| | Backbone | 0.764 (0.862) | 0.778 (0.903) | 0.584 (0.908) | 0.773 (0.942) | 0.689 (0.908) | 0.343 (0.889) | 0.607 (0.761) | 0.739 (0.655) | 0.664 (0.875) |
| | Heavy Atoms | 0.706 (0.87) | 0.593 (0.941) | 0.663 (0.841) | 0.731 (0.952) | 0.704 (0.896) | 0.392 (0.88) | 0.671 (0.786) | 0.489 (0.624) | 0.672 (0.762) |
| | All Atoms | 0.696 (0.895) | 0.72 (0.929) | 0.583 (0.769) | 0.76 (0.936) | 0.733 (0.899) | 0.385 (0.877) | 0.629 (0.834) | 0.429 (0.824) | 0.674 (0.748) |
| **Maximum Likelihood** | $C\alpha$ | 0.754 (0.907) | 0.919 (0.88) | 0.668 (0.965) | 0.68 (0.906) | 0.948 (0.791) | 0.564 (0.9) | 0.905 (0.961) | 0.754 (0.917) | 0.696 (0.957) |
| | $C\alpha + C\beta$ | 0.755 (0.913) | 0.918 (0.882) | 0.717 (0.966) | 0.721 (0.902) | 0.951 (0.789) | 0.57 (0.733) | 0.902 (0.945) | 0.762 (0.916) | 0.693 (0.959) |
| | Backbone | 0.754 (0.903) | 0.881 (0.881) | 0.655 (0.964) | 0.724 (0.895) | 0.947 (0.787) | 0.624 (0.832) | 0.857 (0.756) | 0.746 (0.921) | 0.696 (0.957) |
| | Heavy Atoms | 0.761 (0.913) | 0.922 (0.891) | 0.609 (0.97) | 0.726 (0.897) | 0.937 (0.8) | 0.606 (0.913) | 0.904 (0.969) | 0.761 (0.924) | 0.758 (0.954) |
| | All Atoms | 0.752 (0.914) | 0.923 (0.896) | 0.625 (0.971) | 0.771 (0.899) | 0.946 (0.779) | 0.534 (0.933) | 0.904 (0.969) | 0.749 (0.924) | 0.757 (0.954) |

TABLE S2. Comparison between the correlation length $\xi_\phi$ (or $\xi_{\phi'}$) with the radius of gyration ($R_g$). Values correspond to the Pearson correlation coefficients between the correlation length $\xi_\phi$ (or $\xi_{\phi'}$) and the gyration radius $R_g$ for the cases with different combinations of alignment algorithms, model resolutions and structural characteristics. Here, three alignment methods (including Least-Square Method, Least-Square Method for the parts without coil structures, and Maximum-Likelihood Method), five kinds of model resolutions (considering 1) $C_\alpha$ atoms, 2) $C_\alpha$ and $C_\beta$ atoms, 3) backbone atoms, 4) heavy atoms, and 5) all atoms), and nine kinds of structural characteristics (including all $\alpha$ class, all $\beta$ class, $\alpha+\beta$ class, $\alpha/\beta$ class, small proteins, peptides, designed proteins, multi-domain proteins, and protein assemblies with multiple chains, respectively) are considered.

This kind of insensitivity is also present for different subsets of proteins. For proteins with similar structural characteristics, the Pearson coefficients show the similar features as those for the dataset of all the proteins in this work. Even, for individual proteins, the similar correlation lengths $\xi$ are observed for the cases with different alignment methods and/or the selection of interested atoms (as shown in Fig.S3 and S4). These further demonstrate the robustness of the scale-free behaviors.

| | | All $\alpha$ Proteins | All $\beta$ Proteins | $\alpha+\beta$ Proteins | $\alpha/\beta$ Proteins | Small Proteins | Peptides | Designed Proteins | Multi-domain Proteins | Multi-chain Protein Assemblies |
|---|---|---|---|---|---|---|---|---|---|---|
| **Least Squares Superposition** | $C\alpha$ | 1.0 (1.0) | 1.0 (1.0) | 1.0 (1.0) | 1.0 (1.0) | 1.0 (1.0) | 1.0 (1.0) | 1.0 (1.0) | 1.0 (1.0) | 1.0 (1.0) |
| | $C\alpha + C\beta$ | 0.992 (0.992) | 0.91 (0.996) | 0.997 (0.932) | 1.0 (1.0) | 0.923 (0.994) | 0.651 (0.999) | 0.962 (1.0) | 0.999 (1.0) | 0.989 (0.997) |
| | Backbone | 0.996 (0.999) | 0.99 (0.986) | 0.766 (0.968) | 0.996 (1.0) | 0.942 (0.997) | 0.966 (0.999) | 0.997 (0.979) | 0.996 (0.999) | 0.932 (0.998) |
| | Heavy Atoms | 0.985 (0.988) | 0.887 (0.986) | 0.862 (0.86) | 0.992 (0.989) | 0.832 (0.954) | 0.688 (0.859) | 0.925 (0.991) | 0.982 (0.99) | 0.941 (0.974) |
| | All Atoms | 0.861 (0.973) | 0.856 (0.987) | 0.862 (0.84) | 0.995 (0.99) | 0.76 (0.949) | 0.671 (0.779) | 0.878 (0.992) | 0.982 (0.991) | 0.976 (0.953) |
| **Least Squares Superposition (No Coil)** | $C\alpha$ | 0.834 (0.912) | 0.804 (0.935) | 0.681 (0.86) | 0.998 (0.982) | 0.745 (0.901) | 0.409 (0.887) | 0.745 (0.853) | 0.893 (0.837) | 0.87 (0.931) |
| | $C\alpha + C\beta$ | 0.891 (0.904) | 0.801 (0.935) | 0.657 (0.803) | 0.998 (0.982) | 0.731 (0.902) | 0.224 (0.86) | 0.777 (0.846) | 0.892 (0.832) | 0.87 (0.931) |
| | Backbone | 0.919 (0.911) | 0.817 (0.935) | 0.682 (0.86) | 0.998 (0.983) | 0.742 (0.882) | 0.238 (0.882) | 0.754 (0.857) | 0.833 (0.841) | 0.876 (0.93) |
| | Heavy Atoms | 0.815 (0.914) | 0.651 (0.934) | 0.807 (0.735) | 0.989 (0.978) | 0.767 (0.859) | 0.388 (0.864) | 0.856 (0.942) | 0.647 (0.826) | 0.848 (0.912) |
| | All Atoms | 0.801 (0.897) | 0.686 (0.946) | 0.668 (0.624) | 0.994 (0.966) | 0.777 (0.85) | 0.286 (0.858) | 0.878 (0.938) | 0.726 (0.681) | 0.849 (0.914) |
| **Maximum Likelihood** | $C\alpha$ | 0.92 (0.556) | 0.755 (0.741) | 0.619 (0.8) | 0.955 (0.873) | 0.643 (0.588) | 0.523 (0.934) | 0.754 (0.785) | 0.889 (0.591) | 0.788 (0.759) |
| | $C\alpha + C\beta$ | 0.919 (0.564) | 0.748 (0.747) | 0.651 (0.799) | 0.972 (0.871) | 0.68 (0.59) | 0.607 (0.733) | 0.751 (0.824) | 0.888 (0.584) | 0.787 (0.765) |
| | Backbone | 0.92 (0.551) | 0.744 (0.749) | 0.62 (0.796) | 0.972 (0.861) | 0.642 (0.592) | 0.437 (0.803) | 0.703 (0.883) | 0.889 (0.594) | 0.787 (0.755) |
| | Heavy Atoms | 0.913 (0.57) | 0.761 (0.763) | 0.601 (0.806) | 0.972 (0.865) | 0.637 (0.631) | 0.432 (0.927) | 0.751 (0.842) | 0.889 (0.573) | 0.837 (0.745) |
| | All Atoms | 0.918 (0.57) | 0.756 (0.771) | 0.607 (0.808) | 0.974 (0.868) | 0.631 (0.624) | 0.637 (0.956) | 0.747 (0.845) | 0.891 (0.569) | 0.837 (0.744) |

TABLE S3. A comparison of the correlation lengths with the correlation lengths calculated with $C_\alpha$ least square superposition (calculated with different methods). Values correspond to the Pearson correlation coefficients between correlation lengths $\xi_\phi$ (or $\xi_{\phi'}$) for the basic case and those with different combinations of alignment algorithms, model resolutions, and structural characteristics (see text for details). The alignment methods, model resolutions, and structural characteristics used in this investigation are the same as those for Table.S2

.

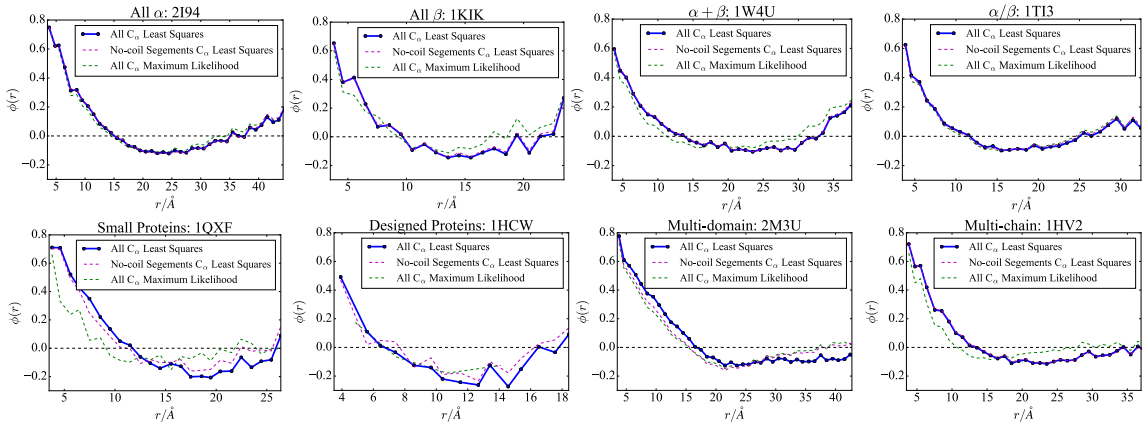



FIG. S3. Correlation functions $\phi(r)$ for proteins with different algorithms and selections of specific atoms. All proteins are represented with $C_\alpha$ atoms. Three cases are considered: 1) least-square method, 2) least-square method and the $\alpha$ and $\beta$ secondary structures selected, and 3) maximum likelihood method, which are shown in figures with different line styles. The concerned proteins with various structural features are the same as those in Fig.S2.

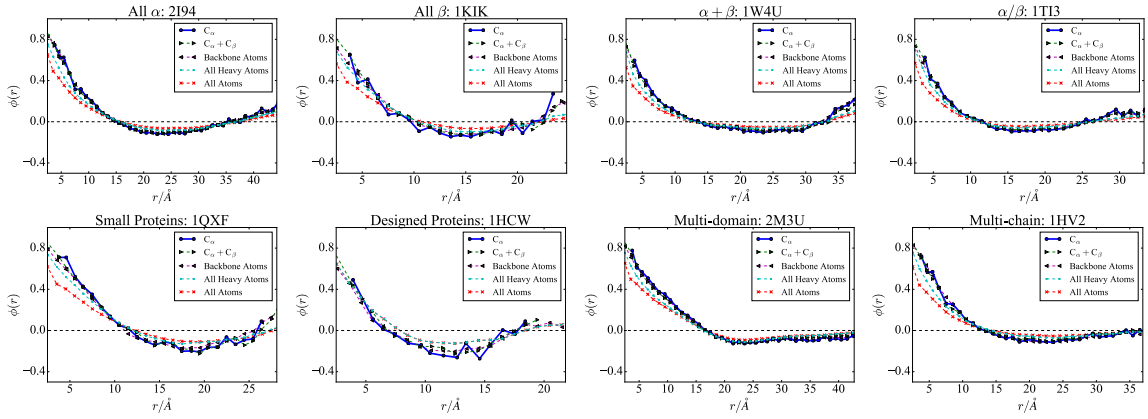



FIG. S4. Correlation functions $\phi(r)$ for proteins with different model resolutions. In these models, the proteins are represented with: 1) $C_\alpha$ atoms, 2) $C_\alpha$ and $C_\beta$ atoms, 3) backbone atoms, 4) all heavy atoms, and 5) all atoms, which are shown in figures with different line styles. The least-square method are used for all the cases. The concerned proteins are the same as those in Fig.S2.

## 10. Effects of Averages of Structures

In our calculation of the fluctuations, we generally assume that every structure is equally weighted as others in the ensemble. Practically, in a structural ensemble determined by NMR, the structures may have different weights. With different weights for structures in the ensemble, the average structure $\{\langle\Delta\vec{r}_i\rangle\}$ may be different. Yet, the introduction of weights for structures would not affect the feature of correlations. As shown in Fig.S5, the correlation functions $\phi(r)$ are calculated for

two cases, one with the equal weights for the structures in the ensemble, and the other with the random weights uniformly distributed in the region [0,1). Based on these results, the scale-free correlations and the corresponding correlation lengths are not sensitive to the changes in weights of the structures in the ensemble.

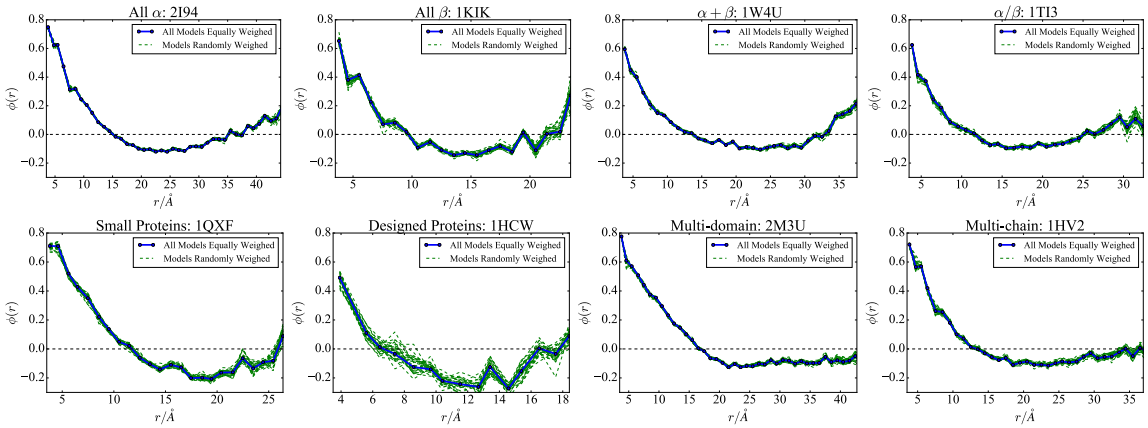



FIG. S5. Correlation function $\phi(r)$ for proteins with different weights of the structures in their ensembles. Two cases are considered: 1) all the structures with equal weights, 2) the weights of structures are generated as random numbers uniformly distributed in the region [0,1). These calculations are carried out with least-square method and $C_\alpha$-atom representation. The proteins used are the same as those in Fig.S2.

# B. Critical Behavior of Proteins

## 11. Correlation Function of Magnitude of Fluctuations

The correlation function $\phi'(r)$ of the magnitude $\Delta r_{i,q}$ (*i.e.* $|\Delta\vec{r}_{i,q}|$) of fluctuations $\Delta\vec{r}_{i,q}$ can be defined based on Eq.(2) in the main text. The related kernel is

$$\mathcal{K}_{ij} = \phi'_{ij} = \frac{\sum_q \left(\Delta r_{i,q} - \langle \Delta r_i \rangle\right) \cdot \left(\Delta r_{j,q} - \langle \Delta r_j \rangle\right)}{\sqrt{\sum_q \left(\Delta r_{i,q} - \langle \Delta r_i \rangle\right)^2 \cdot \sum_q \left(\Delta r_{j,q} - \langle \Delta r_j \rangle\right)^2}}, \tag{1}$$

in which $\langle \Delta r_i \rangle = (1/Q) \cdot \sum_q |\Delta\vec{r}_{i,q}|$. This reflects whether two residues move coherently in any modes. As in Fig.S6(a) and Fig.S6(b), the correlation function $\phi'(r)$ and the corresponding correlation lengths show the similar behaviors as those for orientational correlation $\phi(r)$ shown in the main text. The proportional relation between the correlation length $\xi_{\phi'}$ and gyration radius $R_g$ (Fig.S6(a)) and the scale-free behavior (Fig.S6(c)) are observed, similar to the results in Fig.2(c) and Fig.2(e) in the main text.

## 12. Susceptibility and Distribution of Proteins with Similar Lengths

Based on the finite-size scaling analysis, there is a critical line which can be used to describe the divergence of susceptibility as a function of increasing system size $N$ (as the red line in Fig.3(a)

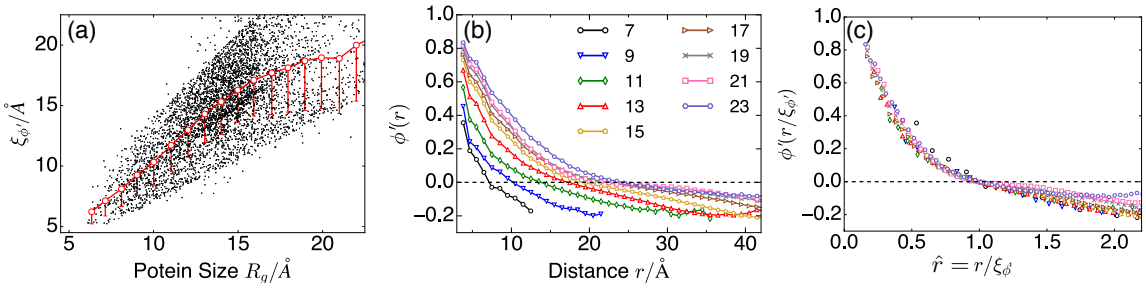



FIG. S6. (a) Scatter plot of correlation length $\xi_{\phi'}$ versus gyration radius $R_g$. The average of correlation lengths for proteins with their radii $R_g \in [R_g - 0.5, R_g + 0.5)$ is plotted as the red line. (b) Correlation function $\phi'(r)$ of proteins with their radii $R_g \in [R_g - 0.5, R_g + 0.5)$. The concerned radii $R_g$ are indicated in Figure. (C) Rescaled correlation function $\phi'(\hat{r}')$ with $\hat{r}' = r/\xi_{\phi'}$.

in the main text and Fig.S7). Interestingly, on the $\chi$-$s$ plane, the proteins are largely distributed around the red line. This result shows the relevance between $s$ and $\chi$ for the proteins, which may suggest an evolutionary constraint and deserve further studies.

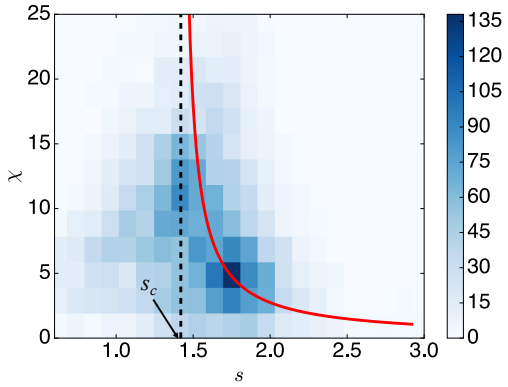



FIG. S7. The number of proteins on the $\chi$-$s$ plane. The region with more proteins is marked with a darker color. The bin sizes for $\chi$ and $s$ are 2 and 0.11, respectively. The red line describes the divergence of susceptibility $\chi$ for longer proteins.

The proteins are largely populated close to the critical shape factor $s_c$ regardless of lengths (as shown in Fig.S8). This means that the proteins with near-critical shape factor (and correspondingly, higher susceptibility) are more frequently observed than those with lower susceptibility. This is consistent with the conclusion obtained from Fig.4(a) in the main text, indicating that natural proteins are basically critical.

## 13. Susceptibility and Distribution of Proteins with Similar $R_g$

Similar to the bird flocks [14], the density of protein systems may vary as a function of the shape factor $s$. Proteins with different lengths may have a similar spatial size (namely, gyration radius $R_g$) and different packing densities. These proteins with a similar radius $R_g$ (say with $R_g \in$

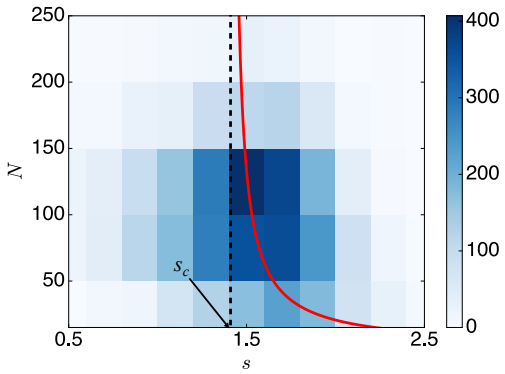



FIG. S8. The number of proteins on the $N$-$s$ plane. The region with more proteins is marked with a darker color. The bin sizes for $N$ and $s$ are 50 and 0.2, respectively. The red line describes the divergence of susceptibility $\chi$ for longer proteins.

$[R_g - 0.5, R_g + 0.5))$ can be regarded as a set of proteins with a similar spatial size. The study of the critical behavior of these subsets of proteins may provide another way to understand critical fluctuations in proteins.

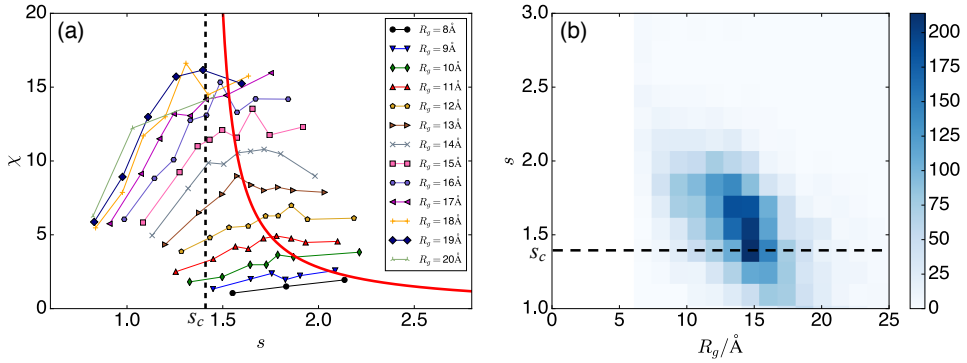



FIG. S9. (a) Susceptibility $\chi$ versus the shape factor $s$ for sets of proteins with their gyration radii $R_g \in [R_g - 0.5, R_g + 0.5)$. The peak height $\chi_m$ and position $s_m$ approximately show a power-law relation $\chi_m \sim (s_m - s_c)^{-\gamma}$ (the red thick line, which is the same as shown in Fig.3(a)). (b) The number of proteins on the $s$-$R_g$ plane. The region with more proteins is marked with a darker color. The bin sizes for $R_g$ and $s$ are 1.3 Å and 0.14, respectively. The critical shape factor $s_c \approx 1.4$ is plotted as the dashed line.

As shown in Fig.S9(a), similar behaviors of the susceptibility is also observed as those in Fig.3(a) of the main text. That is, a series of susceptibility peaks are observed for different $R_g$. These peaks are also around the critical line (the red line) obtained in the main text (Fig.3(a)). Besides, as shown in Fig.S9(b), the proteins are largely distributed around a linear zone on the $s$-$R_g$ plane. Furthermore, the peak positions (i.e. shape factors corresponding to the maximal susceptibility $\chi_m$ in Fig.S9(a)) are approximately the same as the most probable shape factors. This relation is shown in Fig.4(b) in the main text. This abundance of proteins with high susceptibility implies the

possible connection between the structures and critical behavior of the proteins.

## 14. Criticality of Proteins with Different Folding Behaviors

To investigate the relation between the criticality and folding behaviors of proteins, we studied several examples of proteins, since there is no sufficient information of folding behaviors for all proteins. We take two proteins, namely, Cyclophilin A (PDB code: 2MS4, $R_g = 14.6$ Å) which shows a two-state folding behavior [15], and dihydrofolate reductase (DHFR, PDB code: 2KGK, $R_g = 15.9$ Å) which shows folding with multiple states [16]. As shown in Fig.S10, their correlation lengths are approximately equal to their gyration radii, indicating that both proteins show the scale-free behaviors. This observation suggests that the critical behavior would be shared by proteins with various folding behaviors and reflect some aspects of the intrinsic characteristics of native basin of proteins.

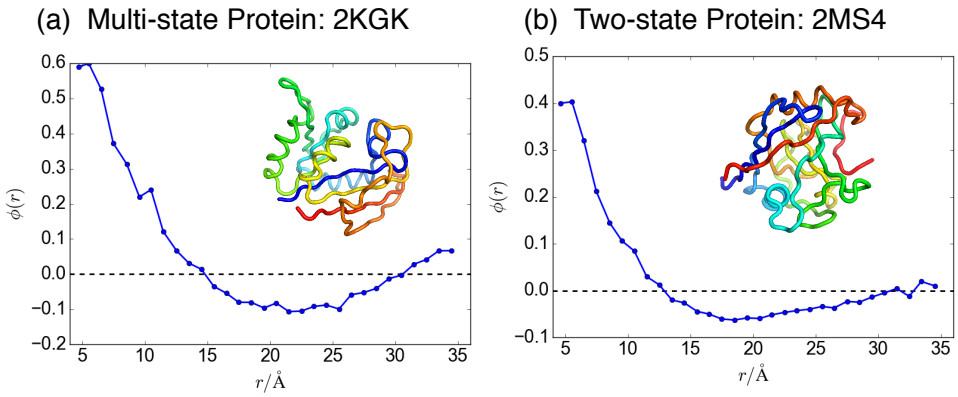



FIG. S10. Correlation function $\phi(r)$ for (a) Multi-state folder (PDB code: 2KGK) and (b) Two-state folder (PDB code: 2MS4). The backbone traces of these two proteins are also shown.

## 15. Criticality of Proteins with Different Structural Characteristics

Based on the structural characteristics of proteins (such as those used in the SCOP classification [17]), our dataset can be classified into seven categories: a) all $\alpha$ proteins, b) all $\beta$ proteins, c) $\alpha + \beta$ proteins, d) $\alpha/\beta$ proteins, e) small proteins f) peptides, g) designed proteins, h) multiple-domain proteins, and i) protein assemblies (composed of multiple chains). For these categories, the relations between correlation length $\xi_\phi$ (red)(or $\xi'_\phi$ (blue)) and gyration radius $R_g$ are calculated (as shown in Fig.S11). The proportionality between the correlation length and the gyration radius is generally kept for various cases. These results further indicate that the scale-free behaviors are kept for proteins with various structural characteristics.

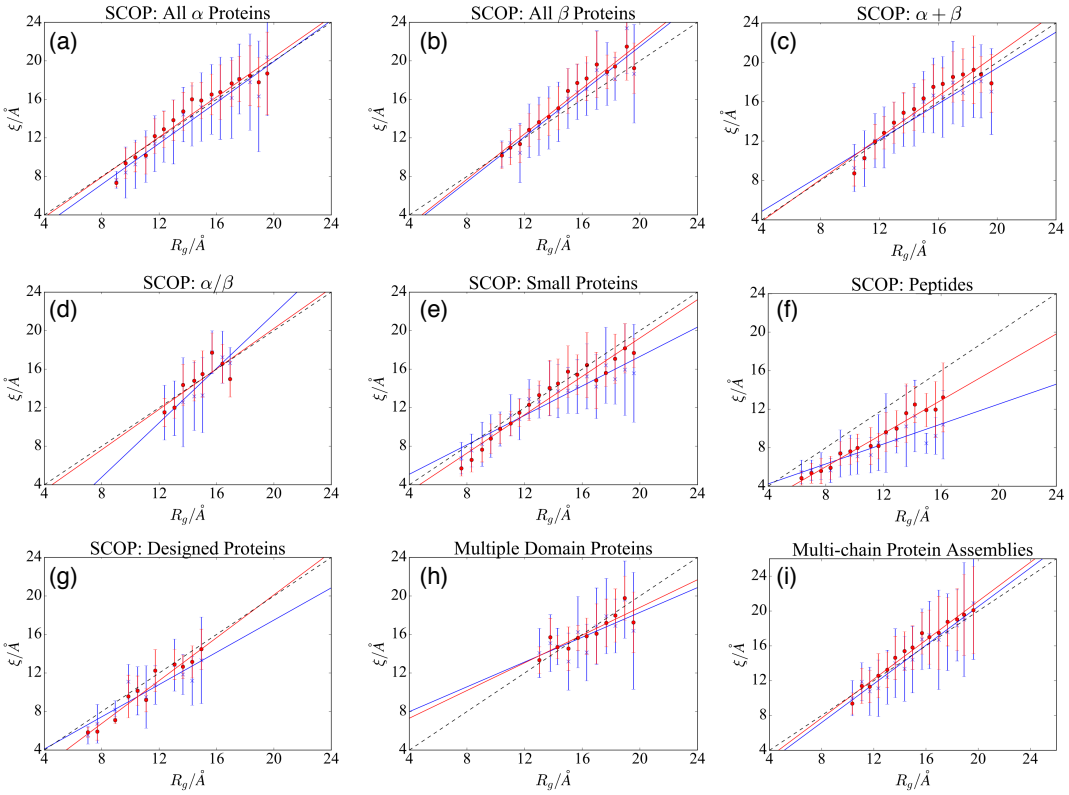



FIG. S11. The relation between correlation length $\xi_\phi$ (or $\xi_{\phi'}$) versus gyration radius $R_g$ for proteins with various structural features, including a) all-$\alpha$ class, b) all-$\beta$ class, c) $\alpha+\beta$ class, d) $\alpha/\beta$ class, e) small proteins, f) peptides, g) designed proteins, h) multi-domain proteins, and i) protein assembly with mupltiple chains. The related data and the fitting are shown in red (blue) color, and the black dashed line is the diagonal.

Furthermore, it could be inferred that the scale-free behaviors are not originated from contact order of proteins [18], since the proteins in different structural categories may generally have distinct contact orders, for example, the contact orders of the all-$\alpha$ proteins are generally smaller than those of the all-$\beta$ proteins.

## 16. Criticality of Proteins with Different Biological Functions

As discussed in the main text, the scale-free correlations for proteins may help the proteins to response efficiently to environmental perturbations. This is universal for various proteins, and may have fundamental contributions to the functions of proteins. As an example, four sets of proteins which correspond to different classes of enzymes are investigated (as shown in Fig.S12). The linear relations between the correlation length and the gyration radius are almost the same for all classes of enzymes. This observation indicates that the scale-free features exist widely for various enzymes.

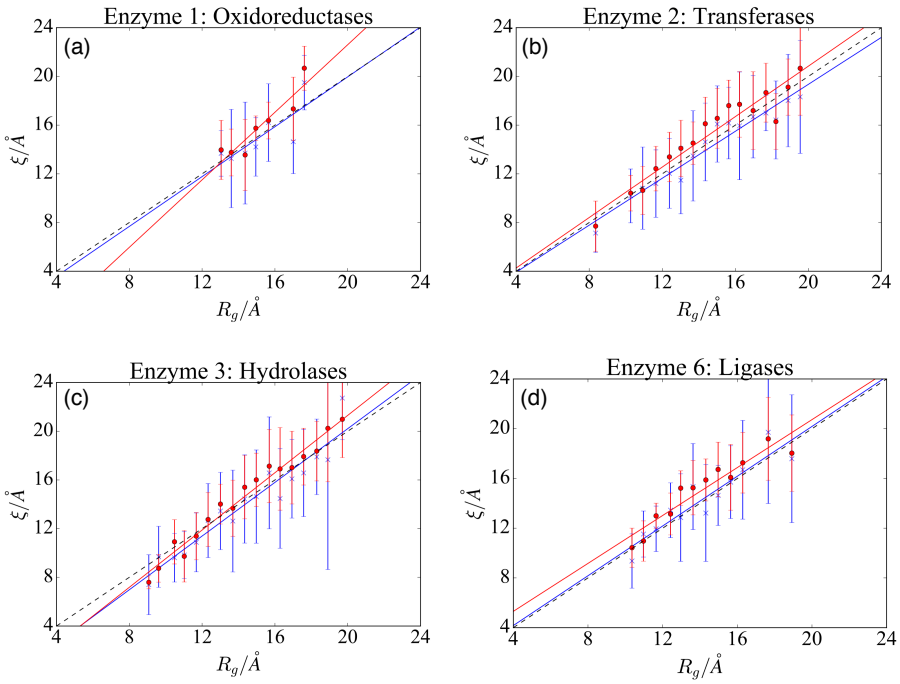



FIG. S12. The relation between correlation lengths $\xi_\phi$ ($\xi_{\phi'}$) versus gyration radius $R_g$ for proteins in different enzyme classes, including a) oxidoreductases, b) transferases, c) hydrolases and d) ligases. The linear fitting for the relation between $\xi_\phi$ ($\xi_{\phi'}$) versus $R_g$ is plotted as the red (blue) line, and the black dashed line is the diagonal.

# C. Other Issues

## 17. Correlation Analysis and Normal-Mode Analysis

Normal mode analysis (NMA) can be used to analyze fluctuations of proteins. In principle, this is in parallel with our correlation analysis. However, these two kinds of analysis have some practical differences,

- NMA based on physical interactions is generally applied based on the unique native conformation, which reflects the local shape of the landscape around native conformations. Differently, the correlation analysis considers conformational transitions between local minima in native basin, and is related to a super-basin. In this sense, correlation analysis is more physical to characterize the fluctuation behaviors of proteins.
- NMA based on elastic network model (ENM) requires the assignment of the equilibrium structure and the the related spring constants, which (especially the spring constants) generally come from some assumptions. Therefore, a specific ENM often is limited to solve some certain problems. Differently, the correlations we analyzed are fully based on experimental data or simulation data, which well reflect the solution dynamics of protein molecules, and avoid artificial biases. Practically, based on our correlation, effective interactions under harmonic

approximation could be estimated. This provides more quantitative information which may help to build a realistic elastic network model.

## 18. Shape Factor and Other Conformation-related Parameters

The shape factor $s$ (used as the control parameter in our work) describes the residue density of proteins with some ability to describe the shape of the protein molecules. Physically, for the systems loosely packed, the correlations between units would be smaller, while for the helical proteins with a highly aspherical shape, the correlation may be limited by their shapes though the density may be larger [19]. That is, proper density and shape are necessary to achieve strong correlation and high susceptibility, as we observed in Fig.3(a) of the main text. In this sense, the quantity $s$ is a compromise of the considerations for density and shape.

Actually, some quantities (asphericity and acylindricity) could describe the shape characteristics more directly. They are defined as asphericity $a_{sph} = L_z^2 - (L_x^2 + L_y^2)/2$ and acylindricity $a_{cyl} = L_y^2 - L_x^2$, where $L_{x,y,z}$ are three principal axes of the concerned molecule satisfying the condition $L_x < L_y < L_z$. Yet, they cannot depict the density issue. They are generally not so good, especially for the cases with small $s$, as show in Fig.S13.

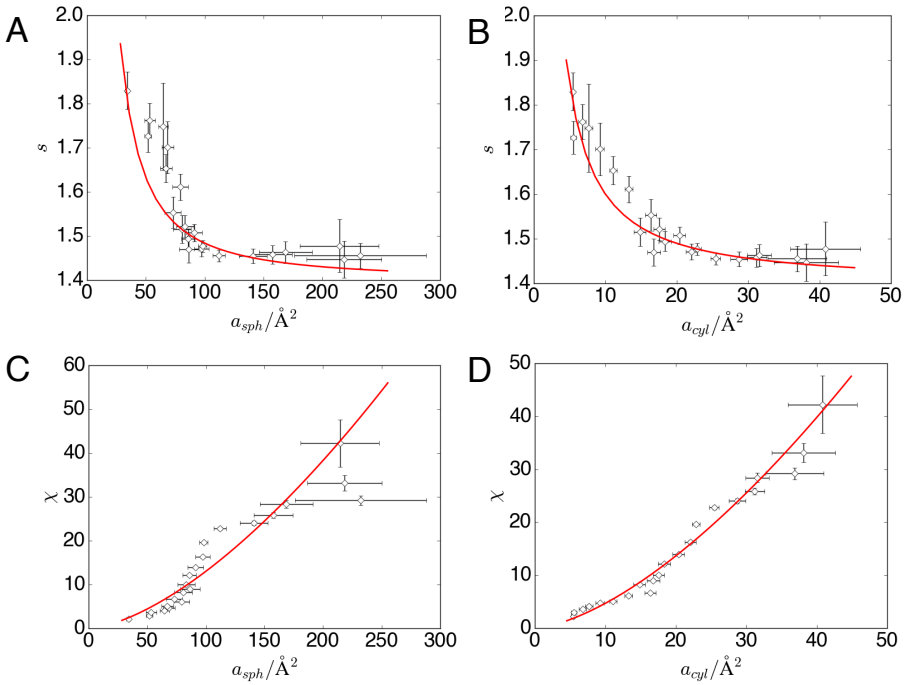



FIG. S13. The relations between (a) shape factor ($s$) versus asphericity ($a_{sph}$), (b) shape factor ($s$) versus acylindricity ($a_{cyl}$), (c) susceptibility ($chi$) versus asphericity ($a_{sph}$), and (d) susceptibility ($\chi$) versus acylindricity $a_{cyl}$. The red lines are the power-law fits.

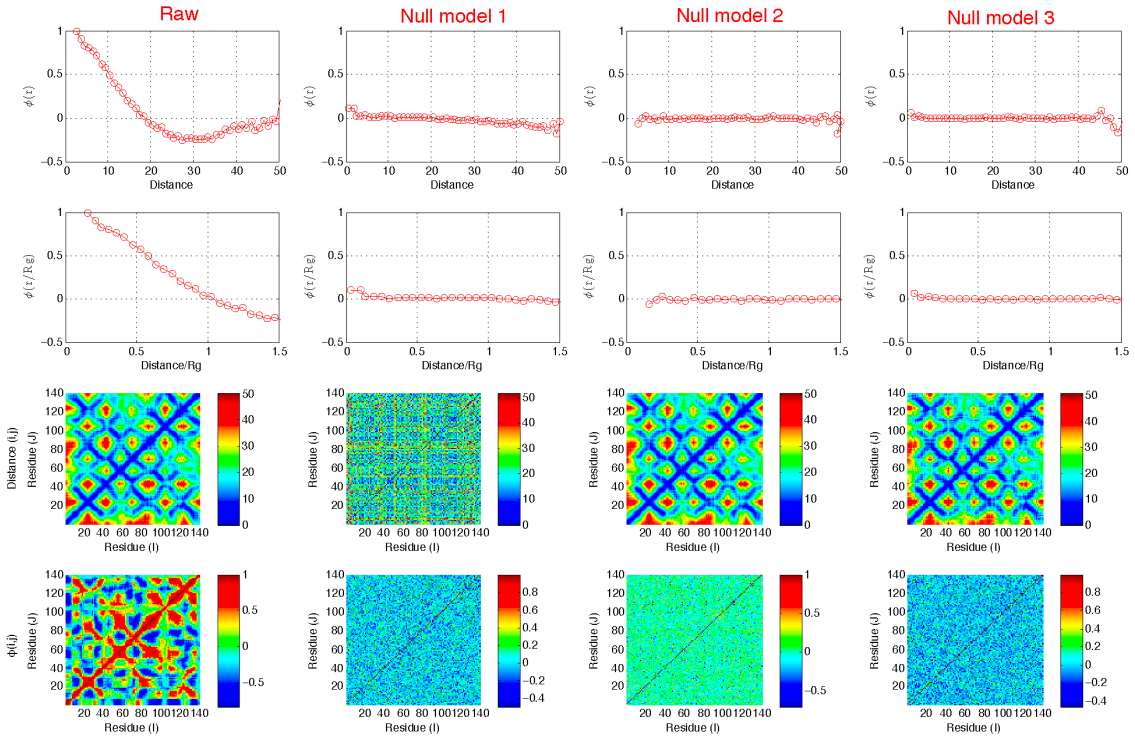



FIG. S14. Examples of null models in which the scale-free correlations breakdown due to different random permutations. Column one corresponds to the results for the raw data of the protein (PDB code"2d8k") and columns two to four to different null models. Top two rows show the correlations $\phi(r)$ and $\phi(r/R_g)$, respectively. The color graphs of the bottom two rows show the matrix of the residues mutual distances as well as the matrix of the mutual $\phi(i,j)$ correlations.

## 19. Null models

Fig. S14 shows three simple null models following the same calculations of distance-dependent correlations used in the main text. Each of the four columns shows from top to bottom $\phi(r)$, $\phi(r/R_g)$, the mutual distance (distance(i,j)) between residues and the correlation matrix $\phi(i,j)$. The first column corresponds to the results obtained from the raw structure of the protein (PDB code"2d8k"). The second column illustrates the results from the same protein in which the residues x, y and z coordinates were randomly permuted. As expected, all correlations vanish. In the third column, for the same protein, the spatial coordinates are preserved (as shown by the distance matrix) but the deviations from the mean ( i.e., $\Delta\vec{r}_i$ in Eq 1 of main text) are randomly permuted. This result also in the breakdown of the correlations. Finally the fourth column corresponds to the results of selecting (arbitrarily) one structure of the same ensemble and adding to each residue' coordinate a relatively small random Gaussian deviation (SD=1), implying no correlations across members (i.e., models) of the ensemble.

## 20. The Validity of NMR Ensemble Analysis.

*(1) The similarity between protein dynamics and flocking dynamics.*

Because the work on flocks might seem foreign to that in proteins, here we discuss similarities and differences between amino-acid distance vectors obtained by structure alignment and velocity vectors for a flock of birds. According with Cavagna's results[3, 14] birds in a flock interact with their neighbors via their velocity fluctuations. In this sense, the evolution of the flock is guided by an evolving energy landscape in the space spanned by birds' velocities. For the proteins, the dynamics (or fluctuations) of proteins are governed by the energy landscape in the space spanned by the positions of residues. It is then, from this viewpoint, reasonable to map the fluctuations of proteins (namely the variance of positions) to the velocity fluctuations of birds in a flock.

With that clarification in mind the fluctuations can be considered as fluctuations on a rugged landscape. Within the picture of minimally frustrated funnel, the native basin can be described with the harmonic approximation as

$$E = \frac{\gamma}{2}\Delta\vec{r}^{T} \cdot G \cdot \Delta\vec{r},$$

where $\Delta\vec{r}$ describes the displacement vector of residues away from equilibrium, and $G$ characterizes the effective elasticity. The validity of such kind of approximation is supported by the maximal-entropy analysis of the NMR structural ensemble in [1]. Clearly, this kind of effective interaction shares the same formula as that for bird flocks by replacing the fluctuation velocity of birds with the fluctuation displacement of residues, and it was the mindset which triggered our exploration of criticality in proteins native states. Under this harmonic approximation, the transition velocities related to eigen modes have proportional relations with the corresponding eigen vectors, and this is consistent with a velocity/distance mapping.

*(2) The validity of the conjecture of structural ensembles.*

Despite the conjecture that structural ensembles based on NMR reflects some key characteristics of native basin of proteins, it might be reasonable to doubt that the fluctuations discussed in the manuscript represent actual transitions executed by folded proteins, because of the way the proteins data is obtained (i.e., NMR).

However, the harmonic approximation for the native basin of proteins (as discussed above) suggests that the characteristics of fluctuations (transitions) in the native basin (especially the low-frequency components) can be detected from the NMR structure ensembles, as suggested in the literature [2]. Though the roughness of energy landscape may introduce heterogeneity of transition rates, the minimal-frustration principle may guarantee the validity of the above picture. Furthermore, the argument is tilted favoring the picture since each of the computations and scaling relations computed in the present work agrees with the conjectured scenario. Indeed, detailed descriptions for the nature of the transitions requires future analysis based on molecular dynamics simulations and other experiments such as SAXS and neutron scattering and deserve further investigations.

# APPENDIX I:

Matlab code to compute from the examples files the normalized covariation on $C_\alpha$ $< x, y, z >$ coordinates discussed in the paper:

```
%% XYZ
% Computing the distance dependent correlation of native states as
% the normalized covariation on < x,y,z > C_alpha coordinates
%Notabene: Coding is purposely non-optimal, only for demonstration
% Typical data files used in the paper
% "Critical fluctuations in proteins native states".
% The data files correspond to 4 proteins of different sizes
% (Rg raging from ~ 7 to ~ 16), the original data and the aligned
% structures as described in the paper are provided.
% The filenames starting with "aligned_" are aligned structures,
% and the filenames starting with "original_"   contain the coordinates
% of the original PDB files (only Carbon alphas are listed in these file).
%% PDB-Code  N    Rg
% 1l2y  20  7.001172073
% 1j7m  60  10.0015842
% 2m5e  100  13.00164718
% 2d8k  141  16.00002502
%% In each file, for each line, the data field are:
% (1) The number of the model
% (2) The amino acid type
% (3) The chain ("A" or "B", some proteins may be made up of two chains)
% (4) The residue number (It may not start from 1, a few cases the structures
% is taken from a larger protein complex)
% (5) x Carbon alphas coordinates
% (6) y Carbon alphas coordinates
% (7) z Carbon alphas coordinates
% For the first model, the coordinate in both files are the same, since is taken
% as a reference for the alignment as explained in the paper.
clear all ;
path1='CA_AtomCoordinates/';
file{1}='original_1l2y.txt';
file{2}='aligned_1l2y.txt';
file{3}='original_1j7m.txt';
file{4}='aligned_1j7m.txt';
file{5}='original_2m5e.txt';
file{6}='aligned_2m5e.txt';
file{7}='original_2d8k.txt';
file{8}='aligned_2d8k.txt';
%% Number of residues
nf(1)=20; nf(2)=20;nf(3)=60;nf(4)=60;nf(5)=100;nf(6)=100;nf(7)=141;nf(8)=141;
%% Radius of Gyration
rg(1)=7; rg(2)=7; rg(3)=10; rg(4)=10; rg(5)=13; rg(6)=13;rg(7)=16;rg(8)=16;
%% Create figure
close all
figure1 = figure('rend','painters','pos',[20 20 500 700]);
```

```
panel(1)=axes('position',[0.1 0.8 0.35 0.175]);
panel(2)=axes('position',[0.55 0.8 0.35 0.175]);
panel(3)=axes('position',[0.1 0.55 0.35 0.175]);
panel(4)=axes('position',[0.55 0.55 0.35 0.175]);
panel(5)=axes('position',[0.1 0.3 0.35 0.175]);
panel(6)=axes('position',[0.55 0.3 0.35 0.175]);
panel(7)=axes('position',[0.1 0.05 0.8 0.175]);

%% Main  loop ( to look at the 4 proteins)
for u=1:1:8

clear x y z data xyz Dr Dx Dy Dz xd yd yz
filename=[path1 file{u}]; a=importdata(filename);
textdata=a.textdata; data=a.data; Nn=data(:,1);
        N=nf(u); % number N of residues
        Q=ceil(length(Nn)/N); % number of models in the file
%% Rename /reformat the data into models
p=0;
for q=1:Q % number of models q=1 is taken as the reference
for i=1:N  % index of each residue in each model
    p=p+1;
x(i,q)=data(p,2);y(i,q)=data(p,3); z(i,q)=data(p,4);
end
end
% Amplitude  displacement from the average
%Compute the mean (over all models) of the fluctuation in
%  each axis at each residue j ;
for i=1:N
   xmean(i)=mean(x(i,:));
   ymean(i)=mean(y(i,:));
   zmean(i)=mean(z(i,:));
end

%% subtract the mean of the fluctuations to each residue i ;
for q=1:Q
   Dx(:,q)=x(:,q)- xmean';
   Dy(:,q)=y(:,q)- ymean';
   Dz(:,q)=z(:,q)- zmean';
end
%% Compute the correlation between vectors
for i=1:N
    for j=1:N
        clear A B
        qt=1;
  for q=1:Q  % place  in A and B all xyz's at residues i j for all Q models
  A(qt)=  Dx(i,q);
  A(qt+1)=Dy(i,q);
  A(qt+2)=Dz(i,q);
```

```
  B(qt)=Dx(j,q);
  B(qt+1)=Dy(j,q);
  B(qt+2)=Dz(j,q);
  qt=qt+3;
  end

        r(i,j)=dot(A,B)/ sqrt( dot(A,A) * dot(B,B) );
% Distance between i and j residues (Here we use model 1
% or, alternative the mean)
         xd=x(i,1)- x(j,1);
         yd=y(i,1)- y(j,1);
         zd=z(i,1)- z(j,1);
%          xd=xmean(i)- xmean(j);
%          yd=ymean(i)- ymean(j);
%          zd=zmean(i)- zmean(j);
         dista(i,j)=sqrt( ( xd*xd ) +(yd*yd)+ (zd*zd)); % Distance(i,j)
    end
end

%% From here to the end it is just graph related
rlin=reshape(r,1,N*N);
dlin=reshape(dista,1,N*N);
%% Bin the data (using distances) and average r inside each bin
h=0; paso=1;
for ejex=0:paso:max(dlin)
    adentro=find(dlin > ejex & dlin <= ejex+paso);
    if length(adentro > 0)
       h=h+1;
       erre(h)=nanmean(rlin(adentro));
       eje(h)=nanmean(dlin(adentro));
    end
end
candidates=find(erre < 0);
y0=eje(candidates(1)-1); y1=eje(candidates(1));
x0=erre(candidates(1)-1);
x1=erre(candidates(1));
delta=(y1-y0)/(x0-x1);
%% Extract the zero crossing
CL(u)=y0+(x0/delta);  % Correlation length derived from < C(r) >
%% %%%%%%%%Graphs%%%%%%%%%%%%%%%%
%%%%%%%%----------------------%%%%%%%%%%%%
if mod(u,2) == 0
                axes(panel(3))
title('Aligned')
    plot(eje,erre,'-or')
    xlabel('Distance'); ylabel('$\phi$(r)', 'interpreter', 'latex')
    grid on; hold on
                axes(panel(4))
```

```
title('Aligned')
    plot(eje/rg(u),erre,'-or')
    axis([0 1.5 -.5 1])
    grid on; hold on
    drawnow
    xlabel('Distance/Rg');ylabel('$\phi$(r/Rg)', 'interpreter', 'latex')
end

if mod(u,2) == 1
                 axes(panel(1))
title('Non-Aligned')
    plot(eje,erre,'-ob')
    hold on; grid on
    xlabel('Distance'); ylabel('$\phi$(r)', 'interpreter', 'latex')
                  axes(panel(2))
title('Non-Aligned')
    plot(eje/rg(u),erre,'-ob')
    axis([0 1.5 -.5 1])
    hold on; grid on;drawnow
    xlabel('Distance/Rg'); ylabel('$\phi$(r/Rg)', 'interpreter', 'latex')
end
end
%% Correlation and Distances matrix (for protein aligned 2d8k )
                axes(panel(5))
pcolor(dista); title('Distances(i,j), 2d8k')
xlabel('Residue #, I'); ylabel('Residue #, J')
colorbar; shading flat
                axes(panel(6))
pcolor(r); title('$\phi$(i,j)', 'interpreter', 'latex')
xlabel('Residue #, I'); ylabel('Residue #, J')
colorbar; shading flat
%% Example of X coordinates displacements (for protein 2d8k)
                axes(panel(7))
for q=1:Q
    plot(Dx(:,q)+q,'-')
    hold on
end
xlabel('Residue #, i'); ylabel('X axis Disp.(i,k) (Shifted)')
axis([0 N 0 23]); title('Protein 2d8k (aligned)')
```

# APPENDIX II:

Matlab code to compute from the examples files, the norm of the modulus of the covariation on $C_\alpha < x, y, z >$ coordinates discussed in the paper:

```
%  Modulus of the covariation on < x,y,z >
% Computing the distance dependent correlation of Native states using
%  the modulus of the covariation on < x,y,z > C_alpha coordinates
%% Notabene: Coding is purposedly non-optimal, only for demonstration
```

```
%Typical data files used in the paper
% "Critical fluctuations in proteins native states".
% The data files correspond to 4 proteins of different sizes
% (Rg raging from ~ 7 to ~ 16), the original data and the aligned
% structures as described in the paper are provided.
% The filenames starting with "aligned_" are aligned structures,
% and the filenames starting with "original_"   contain the
% coordinates of the original  PDB files
%(only Carbon alphas are listed in these file).
%% PDB-Code  N    Rg
% 1l2y  20  7.001172073
% 1j7m  60  10.0015842
% 2m5e  100  13.00164718
% 2d8k  141  16.00002502
%% In each file, for each line, the data field are:
% (1) The number of the model
% (2) The amino acid type
% (3) The chain ("A" or "B", some proteins may be made up of two chains)
% (4) The residue number
% (5) x Carbon alphas coordinates
% (6) y Carbon alphas coordinates
% (7) z Carbon alphas coordinates
% For the first model, the coordinate in both files are the same, since is taken
% as a reference for the alignment as explained in the paper.
clear ; close all
path1='CA_AtomCoordinates/';
file{1}='original_1l2y.txt';
file{2}='aligned_1l2y.txt';
file{3}='original_1j7m.txt';
file{4}='aligned_1j7m.txt';
file{5}='original_2m5e.txt';
file{6}='aligned_2m5e.txt';
file{7}='original_2d8k.txt';
file{8}='aligned_2d8k.txt';
%% Number of residues
nf(1)=20; nf(2)=20;nf(3)=60;nf(4)=60;nf(5)=100;nf(6)=100;nf(7)=141;nf(8)=141;
%% Radius of Gyration
rg(1)=7; rg(2)=7; rg(3)=10; rg(4)=10; rg(5)=13; rg(6)=13;rg(7)=16;rg(8)=16;
%% Create figure
close all
figure1 = figure('rend','painters','pos',[20 20 500 700]);
panel(1)=axes('position',[0.1 0.8 0.35 0.175]);
panel(2)=axes('position',[0.55 0.8 0.35 0.175]);
panel(3)=axes('position',[0.1 0.55 0.35 0.175]);
panel(4)=axes('position',[0.55 0.55 0.35 0.175]);
panel(5)=axes('position',[0.1 0.3 0.35 0.175]);
panel(6)=axes('position',[0.55 0.3 0.35 0.175]);
panel(7)=axes('position',[0.1 0.05 0.8 0.175]);
```

```
%% Main  loop ( to look at the 4 proteins)
% u=odd are original files (blue plots), u=even are re-aligned (red plots)
for u=1:1:8

clear x y z data xyz Dr Dx Dy Dz rlin dlin erre eje tempx bins valu
filename=[path1 file{u}]; a=importdata(filename);
textdata=a.textdata; data=a.data; Nn=data(:,1);
        N=nf(u); % number of residues
       Q=ceil(length(Nn)/N); % number of models in the file
%% Rename /reformat the data into models
p=0;
for q=1:Q % number of models q=1 might be taken as the reference
for i=1:N  % index of each residue in each model
    p=p+1;
x(i,q)=data(p,2);y(i,q)=data(p,3); z(i,q)=data(p,4);
end
end
% Amplitude  displacement from the average at each residue
% Compute the mean (over all q models) of the fluctuations in
% each axis at EACH residue i ;
for i=1:N
   xmean(i)=mean(x(i,:));
   ymean(i)=mean(y(i,:));
   zmean(i)=mean(z(i,:));
end
%% subtract the mean of the fluctuations above to each residue i ;
for q=1:Q
   Dx(:,q)=x(:,q)- xmean';
   Dy(:,q)=y(:,q)- ymean';
   Dz(:,q)=z(:,q)- zmean';
end
%% Compute the magnitude (i.e., norm) of the fluctuations
for q=1:Q
Dr(:,q)=sqrt((Dx(:,q).* Dx(:,q)) + (Dy(:,q) .* Dy(:,q)) + (Dz(:,q) .* Dz(:,q)));
end

%% at this point all what is left is the "amplitude of the fluctuations

for q=1:Q
Dr(:,q)= Dr(:,q)- mean(Dr(:,q));
end
%% Compute the correlations and the distances
for i=1:N
    for j=1:N
        clear A B
 A=Dr(i,:);
 B=Dr(j,:);
```

```
        A=A -mean(A);
        B=B -mean(B);
          r1=dot(A,B) / sqrt( dot(A,A) * dot(B,B) );
         r(i,j)=r1;
%% Distance between i and j (Here I use prot. 1 or the mean
        xd=x(i,1)- x(j,1);
        yd=y(i,1)- y(j,1);
        zd=z(i,1)- z(j,1);
%          xd=xmean(i)- xmean(j);
%          yd=ymean(i)- ymean(j);
%          zd=zmean(i)- zmean(j);
        dista(i,j)=sqrt( ( xd*xd ) +(yd*yd)+ (zd*zd)); % Distance(i,j)
    end
end

%% From here to the end it is just averaging and plotting routine
rlin=reshape(r,1,N*N);
dlin=reshape(dista,1,N*N);
%% Bin the data (using distances) and average "r" inside each bin
h=0; paso=1;
for ejex=0:paso:max(dlin)
    inside=find(dlin > ejex & dlin <= ejex+paso);
    if length(inside > 0)
       h=h+1;
       erre(h)=nanmean(rlin(inside));  % The correlation
       eje(h)=nanmean(dlin(inside));   % the distance
    end
end
candidates=find(erre < 0);
y0=eje(candidates(1)-1); y1=eje(candidates(1));
x0=erre(candidates(1)-1); x1=erre(candidates(1));
delta=(y1-y0)/(x0-x1);
%% Extract the appr. zero crossing.
CL(u)=y0+(x0/delta);  % Correlation length derived from < C(r) >
%% Graphs
%%%%%%%%%%%%%%%%%%%%%%%%%%%%
if mod(u,2) == 0
                axes(panel(3))
title('Aligned')
    plot(eje,erre,'-or')
    xlabel('Distance'); ylabel('$\phi$(r)', 'interpreter', 'latex')
    grid on; hold on
                axes(panel(4))
title('Aligned')
    plot(eje/rg(u),erre,'-or')
    axis([0 1.5 -.5 1])
    grid on; hold on
```

```
    drawnow
    xlabel('Distance/Rg');ylabel('$\phi$(r/Rg)', 'interpreter', 'latex')
end
if mod(u,2) == 1
                axes(panel(1))
title('Non-Aligned')
    plot(eje,erre,'-ob')
    hold on; grid on
    xlabel('Distance'); ylabel('$\phi$(r)', 'interpreter', 'latex')
                 axes(panel(2))
title('Non-Aligned')
    plot(eje/rg(u),erre,'-ob')
    axis([0 1.5 -.5 1])
    hold on; grid on;drawnow
    xlabel('Distance/Rg'); ylabel('$\phi$(r/Rg)', 'interpreter', 'latex')
end
end
%% Correlation matrix and Distances matrix (just for protein 2d8k )
                axes(panel(5))
pcolor(dista); title('Distances(i,j), 2d8k')
xlabel('Residue #, I'); ylabel('Residue #, J')
colorbar; shading flat
                axes(panel(6))
pcolor(r); title('$\phi$(i,j)', 'interpreter', 'latex')
xlabel('Residue #, I'); ylabel('Residue #, J')
colorbar; shading flat
%% Example of X coordinates displacements (for protein aligned 2d8k)
                axes(panel(7))
for q=1:Q
    plot(Dr(:,q)+q,'-')
    hold on
end
xlabel('Residue #, i'); ylabel('Magnitude Disp.(i,k) (Shifted)')
axis([0 N 0 23]); title('Protein 2d8k (aligned)')
```


[1] T.R. Lezon, and I. Bahar. *PLoS Comput. Biol.* **6**, e1000816 (2010).
[2] N. Trbovic, *et al. Proteins.* **71**(2): 684–694 (2008).
[3] A. Attanasi, *et al.*, Phys. Rev. Lett. **113**, 238102 (2014).
[4] J.N. Onuchic, Z. Luthey-Schultenm, and P.G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997).
[5] J.R. Banavar and A. Maritan, Annu. Rev. Biophys. Biomol. Struct. **36**, 261 (2007).
[6] A. E. Garcia and G. Hummer. Proteins **36**, 175 (1999).
[7] F. Rao and A. Caflisch. J. Mol. Biol. **342**, 299 (2004).
[8] J. Lehtivarjo, K. Tuppurainen, T. Hassinen, R. Laatikainen, and M. Perakyla, *J. Biomol. NMR* **52**, 257 (2012).
[9] A. Cavalli, X. Salvatella, C. M. Dobson, and *et al. Proc. Natl. Acad. Sci. U.S.A.* **104**, 9615 (2007).
[10] K.L. Damm, and H.A. Carlson., *J. Am. Chem. Soc.* **129**, 8225 (2007).

[11] E.A. Coutsias C. Seok, and K.A. Dill, *J. Comput. Chem.* **25**, 1849 (2004).
[12] A. Bakan, L. M. Meireles, and I. Bahar, *Bioinformatics* **27**, 1575 (2011).
[13] D. L. Theobald and D. S. Wuttke. *Bioinformatics* **22**, 2171 (2006).
[14] 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).
[15] A.Y. Istomin, D.J. Jacobs, and D.R. Livesay, *Protein Sci.* **16**, 2564 (2007).
[16] R.M. Ionescu, V.F. Smith, J.C. O'Neil, and *et al.*, *Biochemistry* **39**, 9540 (2000).
[17] A. G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia, *J. Mol. Biol.* **247**, 536 (1995); L. Lo Conte, S.E. Brenner, T.J.P. Hubbard, C. Chothia, A. Murzin, *Nucl. Acid Res.* **30**, 264 (2002); A. Andreeva, D. Howroth, S.E. Brenner, T.J.P. Hubbard, C. Chothia, A.G. Murzin, *Nucl. Acid Res.* **32**, D226 (2004); A.Andreeva, D. Howorth, J.-M. Chandonia, S.E. Brenner, T.J.P. Hubbard, C. Chothia, A.G. Murzin, *Nucl. Acid Res.* **36**, D419 (2008).
[18] K.W. Plaxco, K.T. Simons, D. Baker, **J. Mol. Biol. 277**, 985 (1998); D.N. Ivankov, *et al.*, *Protein Sci.* **12**, 2057 (2003).
[19] A. Maritan, C. Micheletti, A. Trovato, *et al.*, *Nature* **406**, 287 (2000).